

Ss. Cyril and Methodius University in Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING



2015

Proceedings of the 12th International Conference for Informatics and Information Technology

Held at Hotel Molika, Bitola, Macedonia
24-26th April, 2015

Editors:
Mile Jovanov
Biljana Risteska Stojkoska

ISBN 978-608-4699-05-7

Conference on Informatics and Information Technology 2015
Web-site: http://ciit.finki.ukim.mk
Email: ciit@finki.ukim.mk

Publisher:
Faculty of Computer Science and Engineering, Skopje, Macedonia
Ss. Cyril and Methodius University – Skopje, Macedonia
Address: Rugjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, Macedonia
Web-site: http://www.finki.ukim.mk/
Email: contact@finki.ukim.mk

Proceedings Editors:
Mile Jovanov
Biljana Risteska Stojkoska

Technical editing: Ilinka Ivanoska
Cover page: Vangel Ajanovski

Total print run:150
Printed in Skopje, Macedonia, 2015

ISBN: **978-608-4699-05-7**

CIP - каталогизација на публикација
Народна и универзитетска библиотека „Св.Климент Охридски“, Скопје

004.7:621.39(062)
004(062)

PROCEEDINGS of the 12th Conference on Informatics and Information Technology (12; 2015; Bitola) Proceedings of the 11th Conference on Informatics and Information Technology: CIIT 2015, April, 24-26, Molika, Bitola, Macedonia / editors Mile Jovanov and Biljana Risteska Stojkoska. - Skopje : Faculty of Computer Science and Engineering, 2015. - 351 стр. : граф. прикази ; 30 см

Библиографија кон трудовите

ISBN 978-608-4699-05-7
1. Jovanov, Mile [уредник] 2. Risteska Stojkoska, Biljana [уредник]

Preface

This volume contains the papers presented at CIIT 2015: the 12th International Conference on Informatics and Information Technologies held on April 22-25, 2015 in Bitola, Macedonia. The conference was organized by the Faculty of Computer Science and Engineering (FCSE), within the Ss. Cyril and Methodius University in Skopje, Republic of Macedonia, and the Computer Society of Macedonia.

The CIIT conference has grown and continues to evolve. In the twelfth edition, the key conference mission remained to provide an opportunity for MSc and PhD students to present their work to a wider research community, to discuss their progress and methodology with fellow PhD colleagues – mature researchers, as well as to gain external feedback to their work. Building on the success of the past eleven conferences, this year conference attracted a large number of submissions resulting in presentations of 79 short and full papers. The program also included the following 4 satellite events: International Olympiad of Informatics (IOI) workshop on Creating an International Informatics Curriculum for Primary and High School Education, MARnet Campus Best Practice Workshop, Workshop on FCSE Cisco Academy – 10 Years of Active Participation, and High-school student conference, in collaboration with the IOI initiative, MARnet, FCSE CISCO academy and the Macedonian academy of art and science, respectively.

Five distinguished key note lecturers gave plenary sessions covering the different areas of the conference. Prof. Veljko Milutinovic from the University of Belgrade gave a talk on Data Flow Super Computing, Dr. Danko Ilik from INRIA gave a lecture about the advances in Proof Theory, Prof. Dr. Petraq Papajorgji from the Canadian Institute of Technology presented a talk on modeling complex information systems. Dr. Ali Hurson and Dr. Sahra Sedigh Sarvestani, both from Missouri University of Science and Technology, gave talks on "Energy-Efficient Algorithms for Data Retrieval from Indexed Parallel Broadcast Channels" and "Towards quantitative modeling of reliability for sustainable critical infrastructure: challenges and advances", respectively.

This year, the audience of the conference expanded beyond the boundaries of South East Europe, confirming the need for doctoral students and young researchers to come together and share their ideas. For the first time since the very beginning of CIIT, we had participants from all over Europe, as well as from the USA, Brazil, and Iran..

Traditionally, the conference also included the work of the best undergraduate students, selected on the basis of their submitted projects, prepared during the previous year. The IT company SEAVUS gave a generous support for the student session by providing the transport and accommodation for the students. IEEE Computer Chapter Macedonian Section supported the student session, with the organization of a best paper competition, providing awards to the first 3 selected papers.

All in all, this year the CIIT conference has outgrown the role of being an excellent opportunity for students to present innovative ideas and practice their

presentation and communication skills in front of a broad audience, to a more premier role, i.e. to bring researchers together for establishing collaborative links between disciplines, for testing the ground for innovative ideas and for engaging the wider academic community.

November 2, 2015
Skopje

Mile Jovanov
Biljana Risteska Stojkoska
Alexandru I. Nicolin
Marcin Paprzycki

Organization

Conference chairs

Biljana Risteska Stojkoska	Assistant professor - Faculty of Computer Science and Engineering, Skopje
Mile Jovanov	Assistant professor - Faculty of Computer Science and Engineering, Skopje
Alexandru I. Nicolin	Horia Hulubei National Institute of Physics and Nuclear Engineering, Romania
Marcin Paprzycki	Systems Research Institute of the Polish Academy of Sciences, Poland

Organizing Committee

Ilinka Ivanoska	Teaching/research assistant - Faculty of Computer Science and Engineering, Skopje
Magdalena Kostoska	Assistant professor - Faculty of Computer Science and Engineering, Skopje
Emil Stankov	Teaching/research assistant - Faculty of Computer Science and Engineering, Skopje
Vangel Ajanovski	Assistant professor - Faculty of Computer Science and Engineering, Skopje

Program Committee

Adrijan Bozhinovski	School of Computer Science and Information Technology, University American College Skopje, Macedonia
Aleksandra Mileva	Faculty of computer science, Goce Delcev University Shtip, Macedonia
Ali R. Hurson	Missouri University of Science and Technology, USA
Ana Madevska-Bogdanova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Andreja Naumoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Andrijana P. Avramova	Department of Photonics Engineering, Technical University of Denmark, Lyngby, Denmark
Antun Balaz	Institute of Physics, Belgrade, Serbia
Bojan Marinkovic	Mathematical Institute of the Serbian Academy of Sciences and Arts, Serbia

Bojan Pepikj	Max-Planck-Institut fr Informatik, Germany
Boro Jakimovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Daniela Remenska	Vrije Universiteit Amsterdam, Netherlands
Danko Ilic	INRIA, France
Dejan Gjorgjevikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Dejan Spasov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Emil Slusanschi	Politehnica University of Bucharest, Romania
Emilia Todorova	Faculty of Mathematics and Informatics, St. Cyril and St. Methoius University - Veliko Turnovo, Bulgaria
Goce Armenski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Goce Ristanoski	Information and Communications Technology Research Centre of Excellence (NICTA), Australia
Goran Velinov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Haris Gavranovic	International University of Sarajevo, Bosnia and Hercegovina
Ionut Traian Vasile	Horia Hulubei National Institute for Physics and Nuclear Engineering, Romania
Ivana Nieti Kosovi	Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Croatia
Ivaylo Donchev	Faculty of Mathematics and Informatics, St. Cyril and St. Methoius University - Veliko Turnovo, Bulgaria
Katerina Zdravkova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Kire Trivodaliev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Laurentiu Serban	Horia Hulubei National Institute of Physics and Nuclear Engineering, Romania
Marija Slavkovik	University of Bergen, Norway
Mihaela Carina Raportary	Horia Hulubei National Institute for Physics and Nuclear Engineering, Romania
Mihai Ciubancan	Horia Hulubei National Institute for Physics and Nuclear Engineering, Romania
Mihnea Dulea	Horia Hulubei National Institute for Physics and Nuclear Engineering, Romania
Nestor Avramov	Schneider Electric, Smart Home Business, Copenhagen, Denmark
Nevena Ackovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia

Petraq Papajorgji	Canadian Institute of Technology, Albania
Riste Gligorov	VU University Amsterdam, Netherlands
Sahra Sedign Sarvestani	Missouri University of Science and Technology, USA
Sashko Ristov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Sasho Gramatikov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Smile Markovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Smilka Janeska Sarkanjac	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Stevce Stefanoski	Carnegie Institution for Science, USA
Stojan Trajanovski	Delft University of Technology, Netherlands
Stoyan Kapralov	Technical University - Gabrovo, Bulgaria
Valentina Kukenska	Department of Computer System and Technologies, Technical University - Gabrovo, Bulgaria
Verica Bakeva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Vesna Dimitrova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University Skopje, Macedonia
Vesna Prckovska	Neuroimmunology group, IDIBAPS, Spain
Zharko Aleksovski	Philips Research Eindhoven, Netherlands

Table of Contents

eLearning (eWorld)

A new brand of math competition for university students 1
Stoyan Kapralov and Mladen Manev

Binary Search Trees and Modern C++ 5
Emilia Todorova and Ivaylo Donchev

Training in Music Informatics Master and Bachelor Programs 9
Emilia Todorova

E-learning module for digital design courses 13
Valentina Kukenska, Paolina Radeva and Petar Minev

Interactive modules for computer graphics learning 17
Paolina Radeva, Valentina Kukenska and Petar Minev

Internet safety for children in education 21
Venera Krasniqi and Marija Mihova

A New Tool for Calculation of a New Source Code Metric 25
Emil Stankov, Mile Jovanov, Kjiro Gjorgjiev and Ana Madevska Bogdanova

Computer Architecture, GRID Computing and Parallel Processing

Effective polynomial and non-polynomial Schrödinger equations for the 0K dynamics of Bose-Einstein condensates 31
Mihaela Carina Raportaru and Alexandru Nicolin

Code Generation for Parallel Differential Equation Solvers 35
Dániel Berényi, Péter Lévai, Sándor Varró and Vladimir V. Skokov

Possibilities for multicore processor and GPU parallelization of neural network training in Matlab 40
Aleksandra Dedinec, Sonja Filiposka and Dimitar Trajanov

Computational Physics on the Magurele Physics Platform: Some challenges ahead 45
Alexandru Nicolin and Virgil Baran

Performance comparison throw running job in Hadoop by defining the number of map and reduce 49
Zirije Hasani

Virtuoso, System for Saving Semantic Data 54
Zirije Hasani

Concept of utomated Data Transfer Process by Using Linked Servers in Microsoft SQL Server 2012 58
Jelena Jovkovska, Adrijan Bozinovski and Veno Pacovski

Cloud Computing

Cloud application for water resources based on open source software 63
Blagoj Delipetrev

Business Analysis for Hosting an e-Ordering System in the Cloud 67
Filip Dimitrievski, Sashko Ristov, Marjan Gushev and Goce Armenski

Overview of Cloud Computing Security Standards and Challenges 73
Elena Markoska, Sasko Ristov, Katerina Zdravkova and Marjan Gushev

Translating BPMN to WS-BPEL 77
Elena Markoska, Sasko Ristov and Marjan Gushev

Cloud Computing – Challenges and Demands of the Business Sector 83
Kocho Hrisafov, Adrijan Božinovski, Ivana Stojanovska and Biljana Stojcevska

In-memory vs. traditional databases - advantages and challenges 87
Petar Tagasi, Veno Pachovski and Adrijan Bozinovski

Simplifying parallel implementation of algorithms on Hadoop with Pig Latin 92
Eftim Zdravevski, Petre Lameski, Andrea Kulakov, Sonja Filiposka and Dimitar Trajanov

Artificial Intelligence, Robotics, Bioinformatics (Part 1)

Automatic Injection Of Web Semantics To Static And Dynamic Web Sites 98
Gorjan Jovanovski and Dimitar Jovanov

Statistical Approach to Meta-optimization of Classification Models Applied on SVM Models 102
Kliment Mahoski and Zhaneta Popeska

Multiple GO Tools Analysis in Discovering Enriched CRC Signature 106
Monika Simjanoska and Ana Madevska Bogdanova

Plant images classification based on the angles between the leaf shape-contour points 110
Petre Lameski, Eftim Zdravevski, Andrea Kulakov and Dejan Gjorgjevikj

Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education 114
Vlatko Nikolovski, Riste Stojanov, Igor Mishkovski, Ivan Chorbev and Gjorgji Madjarov

Finding Similar Answers Using Some Text Mining Techniques 119
Bojan Ilijoski and Zaneta Popeska

eWorld & Multimedia

Public health facilities spatial distribution in Macedonia 123
Stefce Aksoski, Dejan Burjakoski, Nikola Koteli, Andreja Naumoski and Kosta Mitreski

Case-based retrieval aided with image data 127
Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski and Suzana Loskovska

Semantic Web integration with SPARQL autocomplete 131
Aleksandar Andreevski, Riste Stojanov, Milos Jovanovik and Dimitar Trajanov

Making impact with data in a University Setting – the case of SEE-University 135
Blerta Abazi Chaushi, Agron Chaushi and Zamir Dika

New online applications and facilities at IFIN-HH / ELI-NP 141
Laurentiu Serban and Alexandru Nicolin

Visual Novel - Potentials for Education and Socializing 144
Katerina Bashova, Veno Pachovski and Adrijan Božinovski

Artificial Intelligence, Robotics, Bioinformatics (Part 2)

Detecting protein binding sites by using the fuzzy nearest neighbours clasifier 150
Georgina Mirceva and Andrea Kulakov

Autonomious robot for temperature monitoring in the datacenter 154
Vladislav Bidikov and Nevena Ackovska

Homogeneity Optimization in Protein Interaction Networks 156
Kire Trivodaliev, Ilinka Ivanoska, Slobodan Kalajdziski and Biljana Risteska Stojkoska

The challenges of indoor movement analysis 162
Ivana Nizetic Kosovic

Object tracking using a mobile camera as a sensor on a hexapod robot .. 168
Dragan Nakevski, Kosta Cincaroski, Ivan Ilioski, Nevena Ackovska and Vesna Kirandziska

Complexity of software navigation and implementation of small UAVs ... 172
Igor Sekuloski and Nevena Ackovska

Wireless and Mobile Computing & Computer Networks

Noise-pollution mapping, measurement and analysis system using volunteered geographic information 178
Natalija Najdova and Ivan Chorbev

SDN Controller: Where to Drop the Pin 182
Mile Piperkoski and Igor Mishkovski

Real-World ARP Attacks and Packet Sniffing, Detection and Prevention on Windows and Android Devices 186
Blagoj Nenovski and Pece Mitrevski

Wireless Charging – Concept, Theory and Applications 192
Nenad Trajkovski, Biljana Stojchevska and Ivana Stojanovska

A Survey on data partitioning techniques for spatial and attribute data .. 197
Goran Nikolov, Goran Velinov and Margita Kon-Popovska

Better IT Administration with the help of Network Data Acquisition 203
Arbnor Grubi, Biljana Stojcevska and Ivana Stojanovska

Towards Integration of Wireless Sensor Networks into Smart Home 207
Biljana Risteska Stojkoska

Theoretical Foundations of Informatics

Primitive polynomials as a tool in generation of (t,s)- sequences 212
Vesna Dimitrievska Ristovska and Vassil Grozdanov

Comparative Study of Various Approaches for Solving the Torsional Schrodinger Equation 216
Bojana Koteska, Ljupco Pejov and Anastas Mishev

Symbolic calculations of the static and dynamical properties of Bose-Einstein condensates 221
Mihaela Carina Raportaru and Alexandru Nicolin

Implementation of well known metrics and their practical use in PHP ... 225
Elena Jovanovska and Nevena Ackovska

Applying soft computing methods to discriminant analysis 231
Milan Djordjevic

Uper Bound on q-ary Codes 235
Dejan Spasov

Security and Cryptography & Signal Processing

pplication of Machine Learning Techniques Towards Speech Separation .. 238
Tamara Dimitrova

Semantic Web Access Control Aspects 243
Slobodanka Trajkova, Riste Stojanov and Dimitar Trajanov

Exploitation and distribution of setuid and setgid binaries on Linux systems 247
Eva Tanaskoska

Federating Any on-premise established identity management system with Office 365 cloud services 253
Vasko Sazdovski and Boro Jakimovski

Plenary Session

Learning models of abiotic influence on the biodiversity indices in Lake Prespa 257
Andreja Naumoski

Text Classification of Macedonian News 262
Dimitar Venov and Igor Trajkovski

Chalenges, Benefits and Expectations from Implementing an E-health System in Macedonia 268
Ristana Mitrovska, Veno Pachovski and Adrijan Božinovski

Publishing Skopje Air Quality Data as Linked Data 273
Kostadin Mishev, Nikola Koteli, Angjel Kjosevski, Nikola Kalemdzhievski, Milos Jovanovik, Dimitar Trajanov and Kosta Mitreski

SEAVUS Student Session 1

Improving learning in mathematics trough 3d digital game play 278
Jovan Shalamanoski and Biljana Stojkoska

Semantic Analysis and Data Visualization of Songs from Different Languages 282
Goran Petkovski, David Zafirovski, Blagoja Evkovski, Usamedin Nuhiju and Dimitrij Mijoski

Experimental comparison of few classification algorithms and discussion . 286
Milan Petrov, Dimitrij Mijoski, Duko Serafimovski and Filip Pancevski

High-Level Synthesis, an Introduction and Overview of Present-day Tools 290
Dimitrij Mijoski

Performance comparison of AES candidates 294
Eva Tanaskoska, Ivana Bozhinova, Sandra Delovska and Andrej Gashteovski

Parallel Sorting of Arrays with Open MP and the Quick-Sort Algorithm . 298
Goran Petkovski and Darko Tosev

Benchmarking And Comparison Between Node.js Real Time Engines 302
Irina Arsovska, Monika Jandrevska, Vlado Velichkovski and Kristijan Trajkovski

Twitter sentiment analysis: generating automatically labeled sets 306
Ana Tanevska and Stevica Bozhinoski

PageRank Approach to Ranking National Football Teams 310
Verica Lazova and Lasko Basnarkov

SEAVUS Student Session 2

Information retrieval over Macedonian test collection using word forms and transliteration 314
Ivana Bozhinova, Stevica Bozhinoski and Jasmina Armenska

Isolating steps in periodic movement 318
Dimitar Spasovski, Viktor Janevski, Aleksandar Kuzmanoski, Mile Jovanov and Marija Mihova

Generating Boolean-Unlinear quasigroups of order 8 and their cryptographic properties 322
Bojan Petrovski, Tome Stojkovski, Atanas Todorovski and Marija Mihova

An interactive tool for drawing graphs and running graph algorithms 326
Petar Grchevski, Biljana Zaeva and Elena Golubovska

Educational game with Question-Answer method 330
Angela Arsova, Martin Stojanovski, Sara Miftar, Vladimir Jovanovski, Mile Jovanov, Bojan Ilijoski and Mario Ampov

On the Way to Future's High Energy Particle Physics Transport Code . . . 334
Gábor Bíró, Gergely Gábor Barnaföldi and Endre Futó

Microwave imaging: current trends and future perspectives 339
Danail Tavchioski

Defining Virtual Networks using SDN 344
Bojan Petrovski and Ivona Najdenkoska

Author index

..... 347

A New Brand of Math Competition for University Students*

Stoyan Kapralov
IEEE Education Society Member
Dept. of Mathematics
Technical University
Gabrovo, Bulgaria
s.kapralov@tugab.bg

Mladen Manev
Dept. of Applied Informatics
Technical University
Gabrovo, Bulgaria
ml.manev@gmail.com

***Abstract*—Math education, and more generally the STEM (Science, Technology, Engineering, and Mathematics) education are of great importance to society. At the same time it is apparent that contemporary math education is in a deep crisis and that it requires major shift from teaching manual calculation techniques towards computer-based solution techniques.**

CompMath is a competition among university students in the field of Computer Mathematics. It serves as an excellent teaching tool for using computer algebra systems in solving mathematical problems.

The third Edition of the competition – CompMath-2014 was held in November 2014. The next competition – CompMath-2015 will be held on November 2015 in Burgas Free University, Bulgaria. Interested colleagues and students from abroad are kindly invited to take part in the competition.

Keywords*—*STEM education; math education; math problem solving; computer algebra systems; student competition

I. INTRODUCTION

CompMath – www.compmath.eu – is a mathematical contest for university students enrolled on a bachelor or master degree course, which is organized once per year. The competition aims to increase students' interest in mathematics and computer algebra systems, as well as creating conditions for sharing experience among academic staff.

The CompMath competition is unique all over the world and it is a major educational innovation. The purpose of the contest is in full accordance with IEEE Educational Activities [1].

Students shall participate in the Olympiad individually. The contestants are divided into groups according to their subject area:

- Group A – Mathematics, Informatics and Computer Science;
- Group B – Engineering and Natural Science.

The objective of the contest is to solve 30 problems with the help of a computer algebra system like Mathematica, Maple, Maxima, Derive, MATLAB. Each problem is worth 2 points. The time given is 4 full hours.

The research is supported in part by Grant C1407/2014 of Technical University of Gabrovo, Bulgaria

The syllabus includes the following topics:

- Polynomials
- Matrices and Determinants
- Systems of linear equations
- Vector algebra and its applications in geometry
- Analytic geometry
- Sequences and series
- Differential calculus and its applications
- Integral calculus and its applications
- Differential equations
- Linear operators (for group A only)
- Congruences and systems of congruences (for group A only).

In preparation of the competition problems, the authors try to satisfy the following conditions:

- Computer algebra systems (CAS) should be helpful for briefly solving the tasks;
- The tasks can be solved with either CAS with approximately the same difficulty;
- The test should consist of approximately equal numbers easy, medium and difficult tasks.

The first Edition of the Student Olympiad in Computer Mathematics was held in 2012 in the Technical University of Gabrovo.

The host of the CompMath-2013 was the University of Ruse.

The third Edition of the competition – CompMath-2014 was hosted by the Sofia University. Students from 9 universities took part in the event.

In Section II we present contests' problems and students' results in CompMath-2014.

II. THE COMPMATH-2014 COMPETITION

A. *The Competition in Group A*

Problems:

1. Calculate the value of
$$\left(\sqrt{2+\sqrt{3}}+\left(2-2\cos\frac{11\pi}{6}\right)^{\frac{1}{2}}\right)\left(\sqrt{2+2\cos\frac{\pi}{6}}-\sqrt{2-\sqrt{3}}\right)^{-1}.$$

2. If $F(x,y)=\dfrac{\sin(x-y)+\cos(x+2y)}{x+2y}$, calculate $F\left(2F(x,y),F(y,x)\right)$ for $x=4{,}5$ and $y=3{,}8$.

3. Check if the number 2760727302517 is simple.

4. Factor the polynomial $1-x-7x^2-10x^3-7x^4-x^5+x^6$ into irreducible polynomials with real coefficients.

5. Simplify the expression
$$2(a+b)^{-1}(ab)^{\frac{1}{2}}\left(1+\frac{1}{4}\left(\sqrt{\frac{a}{b}}-\sqrt{\frac{b}{a}}\right)^2\right)^{\frac{1}{2}}$$ if a and b are real numbers.

6. The quadratic equation $x^2-4x+1=0$ has roots x_1 and x_2. Calculate the value of $\left(\sqrt{x_1}+\sqrt{x_2}\right)^{-14}$.

7. Prove the identity $\displaystyle\sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{1}{2k+1}=\frac{2^{2n}(n!)^2}{(2n+1)!}$ for $n\in N$.

8. Find the integers n, for which n^2+2 is a divisor of $2014n+2$.

9. Find the greatest natural number $n\le 1000$ that can be represented in the form $n=\dfrac{a^2+b^2}{ab-1}$, where a and b are natural numbers.

10. Calculate f_{2014}, given the sequence $\{f_n\}_{n=1}^{\infty}$ where $f_1=2$, $f_2=1$, $f_{3n}=3f_n$, $f_{3n+1}=3f_n+2$, $f_{3n+2}=3f_n+1$.

11. Calculate $f(A)$ given the function $f(x)=x^{2014}-x^{1989}$ and the matrix $A=\begin{pmatrix}-1 & -2 & 2\\ 4 & 5 & -4\\ 3 & 3 & -2\end{pmatrix}$.

12. Find an irreducible fraction that is a solution of the equation $x=\sqrt{2015-x}\sqrt{2014-x}+\sqrt{2014-x}\sqrt{2013-x}+\sqrt{2013-x}\sqrt{2015-x}$.

13. Solve the equation $\operatorname{arctg}x-\dfrac{1}{2}\operatorname{arctg}\dfrac{2x}{1-x^2}=\dfrac{\pi}{2}$.

14. Find a solution of the system $\begin{cases}24x^3-10x^2y-3xy^2+y^3=0\\ x^2+5x=y^2\end{cases}$ such that x has the greatest value.

15. Find the nearest to 10 root of the equation $e^{-x^2}-\cos x=0$.

16. Find the values of the real parameter m, for which the equation $(1-m)x^3-3mx^2-3mx+4-m=0$ has three real roots.

17. Find the number of the real solutions of the $\begin{cases}5x_{k+1}=x_k^5+3,\ k=1,\ 2,\ \ldots,\ 2013\\ 5x_1=x_{2014}^5+3\end{cases}$.

18. Calculate $\lim\limits_{n\to\infty}\left(\sqrt[n+1]{(n+1)!}-\sqrt[n]{n!}\right)$.

19. Calculate the integral $\displaystyle\int_0^{2014}\frac{\ln x\,dx}{\sqrt{2014x-x^2}}$.

20. Find the length of the curve $x^{\frac{2}{3}}+y^{\frac{2}{3}}=1$ and the area of the figure, inside that curve.

21. Let the triangle ABC be a right isosceles triangle with hypotenuse AB equal to 4. The point D lies on a circle with centre C and radius 1. Find the smallest possible perimeter of the triangle ABD.

22. The lines CA and CB pass through the point $C(3,-1)$ and are tangent to the ellipse $\Gamma: 2x^2+3y^2+x-y-5=0$ ($A\in\Gamma$, $B\in\Gamma$). Find the area of the triangle ABC.

23. Draw that part of the sphere $x^2+y^2+z^2=1$ which is inside the cylinder $x^2+y^2=x$.

24. Find the volume of the body, defined by the inequality $\left(x^2+y^2+z^2+8\right)^2\le 36\left(x^2+y^2\right)$.

25. Find a natural number N for which the equations $x^2-2014x+N=0$ and $x^2-2014x-N=0$ have integer roots.

26. The sequence $\{p_n\}_{n=1}^{\infty}$ of natural numbers is such that the sequence $\left\{\dfrac{p_n}{n}\right\}_{n=1}^{\infty}$ is strictly decreasing. Find the least possible value of p_{681} if $p_{2014}=10000$.

27. Find the number of the squares with area smaller that 201400, which vertices have integer coordinates satisfying the equality $x^4+x^3y^3=y^4+xy$.

28. Find the 2014th digit of the number 2014^{2014}.

29. The natural number n and the positive numbers $x_1,\ x_2,\ \ldots,\ x_n$ are such that $\displaystyle\sum_{k=1}^{n}x_k=2014$. Find the greatest value of $\displaystyle\prod_{k=1}^{n}x_k$.

30. Let $\dfrac{1}{1-x-14x^2+x^3}=\displaystyle\sum_{n=0}^{\infty}a_nx^n$. Calculate the limit $\lim\limits_{n\to\infty}\dfrac{a_{n+1}}{a_n}$.

Results:
The results of the 45 students which took part in group A are presented in Fig. 1. The post exam relative difficulty of the problems is shown in Fig. 2.

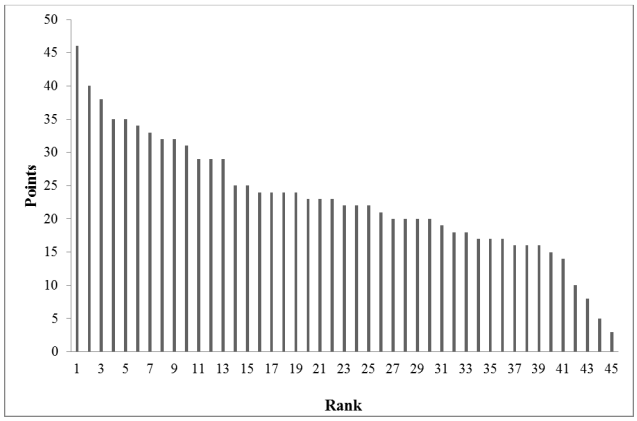


Fig. 1. Results in group A.

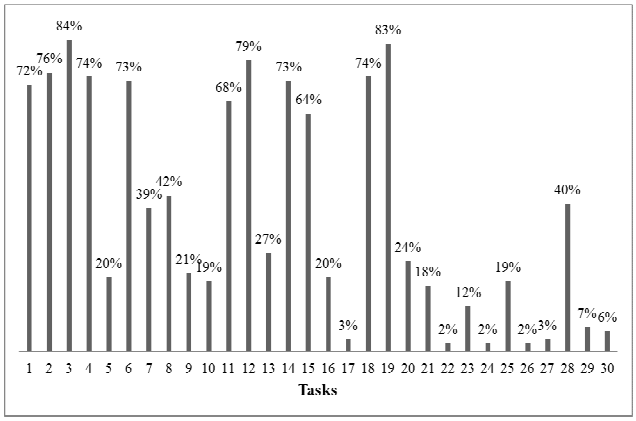


Fig. 2. Relative difficulty of the problens for group A.

B. The Competition in Group B

Problems:

1. Find a polynomial of forth degree whit leading coefficient 1, which zeros are less than zeros of $x^4+4x^3-12x^2-29x+56$ with 3.

2. Find a polynomial of least possible degree such that for $x=1,\ 2,\ 3,\ 4$ is equal respectively to $y=3,\ 1,\ 7,\ 27$.

3. Find real a and b, if the number $2-i$ is a root of the equation $2x^4+x^3-x^2+ax+b=0$.

4. What are the values of a for which the numbers x, y and z derived from the system $\begin{vmatrix} 3x-y+z=5 \\ ax+y-z=0 \\ x+2y+az=17 \end{vmatrix}$ form an arithmetic progression in that order?

5. What are the values of the parameter a such that the system $\begin{vmatrix} ax+3y+z=0 \\ x+(a+1)y-z=0 \\ (2a-1)x+2y+4z=0 \end{vmatrix}$ has non-zero solution?

6. Prove that if a, b, c and d are different integers, then $(a-c)^2+(b-d)^2$ is a divisor of the determinant $\begin{vmatrix} a & b & c & d \\ b & c & d & a \\ c & d & a & b \\ d & a & b & c \end{vmatrix}$.

7. The point C lies on the line $a: x+y=2014$. If $A(2,0)$, $B(1,4)$ and the area of the triangle ABC is 2015, then find the coordinates of the point C.

8. Find the area of the figure, defined by the curve $y=2-4x^2+4x^3-x^4$, abscise axis and the lines $x=x_1$ and $x=x_2$, if y has local maximum in the points x_1 and x_2.

9. Find all four-digit numbers, that are equal to the sum of the forth degrees of their digits.

10. Calculate the limit

$$\lim_{n\to\infty}\frac{1}{n}\left(\frac{1}{2+\cos\frac{\pi}{n}}+\frac{1}{2+\cos\frac{2\pi}{n}}+\cdots+\frac{1}{2+\cos\frac{n\pi}{n}}\right).$$

11. Prove the identity $\int_0^{\sin^2 x}\arcsin\sqrt{t}\,dt+\int_0^{\cos^2 x}\arccos\sqrt{t}\,dt=\frac{\pi}{4}$.

12. Compare the value of the expression $\operatorname{arctg}\frac{1}{2}+\operatorname{arctg}\frac{1}{5}+\operatorname{arctg}\frac{1}{8}$ with the number $\frac{\pi}{4}$.

13. Solve the equation $\sin x=\log_{10} x$.

14. Calculate the sum $1+\frac{x}{1!}+\frac{x^2}{2!}+\cdots+\frac{x^n}{n!}+\cdots$ and compare the value of the integral $\int_0^{2014} e^{-x}\left(1+\frac{x}{1!}+\cdots+\frac{x^{2014}}{2014!}\right)dx$ with 2014.

15. Find the values of the real parameter m, for which the equation $(1-m)x^3-3mx^2-3mx+4-m=0$ has three real roots.

16. On the hyperbola $xy=-1$ the points A_n and B_n are chosen such that their abscises are respectively $\frac{n}{n+1}$ and $\frac{n+1}{n}$. Let M_n be the centre of the circle passing through the points A_n, B_n and the point $C(1,-1)$. Find the limit of the sequence of points $\{M_n\}$ for $n\to\infty$.

17. Find the intersection point of the planes $\alpha: 7x-5y+2z-41=0$, $\beta: 4x+3y-11z+49=0$ and $\gamma: 2x+3y+4z-20=0$.

18. Find the greatest value of the function $y=2\text{tg}x-\text{tg}^2x$ in the interval $\left(0,\frac{\pi}{2}\right)$.

19. The local extreme value of the function $y=\frac{ax+b}{(x-20)(x-14)}$ is equal to 1 and is reached for $x=2014$. Find a and b and determine the kind of the extreme.

20. Solve the equation $\text{arctg}\,x-\frac{1}{2}\text{arctg}\frac{2x}{1-x^2}=\frac{\pi}{2}$.

21. Calculate the limit

$$\lim_{n\to\infty}\left(1+\frac{1}{n^2}\right)\left(1+\frac{2}{n^2}\right)\cdots\left(1+\frac{n-1}{n^2}\right)\left(1+\frac{n}{n^2}\right).$$

22. Find a differential equation which family of solutions curves is $y=\frac{x^2-C^2}{2C}$.

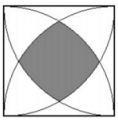
23. Tangent lines to the parabola $y=-x^2+5x-6$ pass through the point $M(2,4)$. Find the area of the "curved" triangle, formed by the two tangent lines and the parabola.

24. Two circles are given. The first has a radius 3 and a centre $M(0,5)$ and the second – a radius 2 and centre $N(10,7)$. Find a point A on the first circle, a point B on the second circle and a point C on the abscise axis, such that the sum $AB+BC+AC$ is minimal. Plot the circles and the solution on one drawing.

25. Find the length of the curve $x^{\frac{2}{3}}+y^{\frac{2}{3}}=1$ and the area of the figure defined by that curve.

26. Which is the 2014th digit of the number 2014^{2014} ?

27. The vertices of a square are used as centres of four circles (but only one fourth of each one is drawn). Find the area of the shaded part of the figure.



28. A light beam starting from point $A(-3,0)$ is directed towards point $B(0,2)$ and is reflected by the ellipse $\frac{x^2}{25}+\frac{y^2}{16}=1$ first at point C, and then at point D. Find the coordinates of the point D.

29. Find the volume of the body, defined by the planes $y=1$ and $z=0$, parabolic cylinder $y=x^2$ and the paraboloid $z=x^2+y^2$.

30. Prove that the inflexion points on the graph of the function $y=\frac{x+1}{x^2+1}$ lie on a straight line.

Results:

The results of the students in group B and the relative difficulty of the problems are presented in Fig. 3 and Fig. 4.

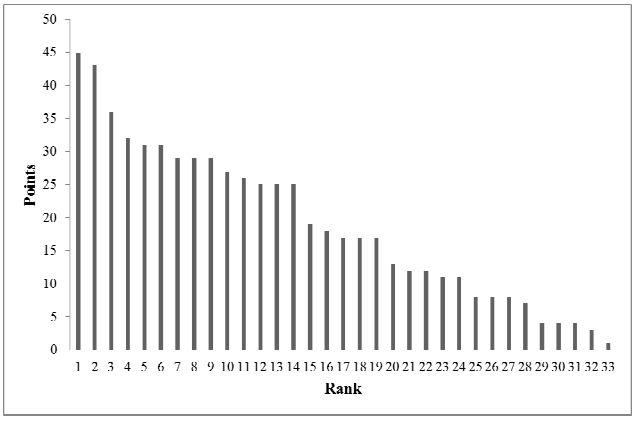


Fig. 3. Results in group B.

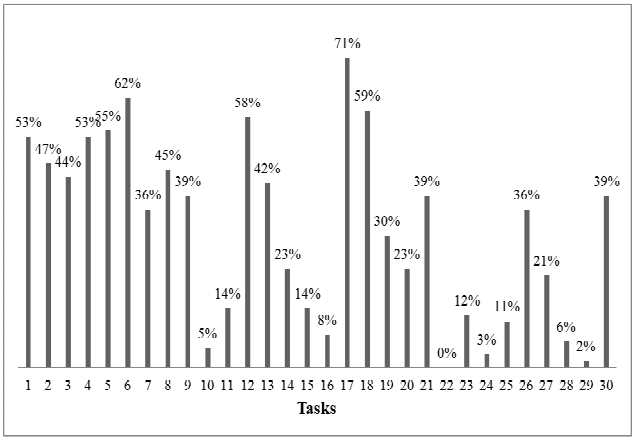


Fig. 4. Relative difficulty of the ptoblems in group B.

REFERENCES

[1] EA Insight, The Newsletter of IEEE Educational Activities, vol. 2, issue 1, 2015.

Binary Search Trees and Modern C++

Ivaylo Donchev, Emilia Todorova

Department of Information Technologies, Faculty of Mathematics and Informatics
St Cyril and St Methodius University of Veliko Turnovo , Veliko Turnovo, Bulgaria
e-mail: i.donchev@abv.bg, e.todorova@uni-vt.bg

***Abstract*— Study of binary trees has prominent place in the training course of DSA (Data Structures and Algorithms). Their implementation in C++ however is traditionally difficult for students. To a large extent these difficulties are due not so much to the complexity of algorithms as to the complexity of the language in terms of memory management by raw pointers. The programmer must consider too many details to ensure a reliable, efficient and secure implementation. Evolution of C++ regarded to automated resource management, as well as our experience in implementation of linear lists by means of C++ 11/14 made us try to implement of binary search trees (BST) via smart pointers as well. In the present paper we share our experience in this direction. Some conclusions about the pedagogical aspects and effectiveness of the new classes, compared to the traditional library containers and implementation with built-in pointers, are made.**

Keywords— abstract data structures; binary search trees; C++; smart pointers; teaching and learning

I. INTRODUCTION

The search for surveys of similar approaches in the investigation of implementation of dynamic data structures in the course of DSA did not give satisfactory results, which explains the short list of references. Comparisons of the given implementations are carried out in respect of the available ones in the standard libraries.

From the C language we know that pointers are important but are a source of trouble. One reason to use pointers is to have reference semantics outside the usual boundaries of scope [1]. However, it can be quite difficult to ensure that the life of the pointer and the life of the object to which it points will coincide, especially in cases where multiple pointers point to the same object.

We must have in mind too many details if we want to accomplish dynamic implementation of ADS (Abstract Data Structures) and often the time for this exceeds the time remaining to comment the structures and operations on them. Moreover, there are rare cases where we have a working implementation of a structure with carefully designed interface and methods written according to the best methodologies, but we identify gaps in the management of memory only when the fall in non-trivial situations such as copying large structures, transfer of items from one structure to another, or destruction of a large recursive structure. For each class representing ADS the programmer must also provide characteristic operations as well as correctly working copy and move semantics, exception handling, construction and destruction. This requires both time and expertise in programming at a lower level. The teacher will have to choose between emphasizing on language-specific features and quality of implementation or to compromise with them and to spend more time on algorithms and data structures. In an attempt to escape from this compromise, we decided to change the content of our CS2 course in DSA, and include the study of smart pointers for resource management and with their help to simplify implementations of ADS, and avoid explicit memory management which is widely recognized as error-prone [2].

In our work the emphasis was on the implementation of linear structures and binary trees. This paper discusses only part of this work dedicated to binary search trees.

Our initial hypothesis was that a correct and effective implementation of BST is possible, which could relieve our work in two directions:

- operations with whole structures (trees): not having to implement copy and move semantics methods;
- shorter and easier to understand implementation of operations with elements of BST – include (insert element), search, delete.

Section II is a brief overview of language features for managing dynamic memory. In section III an implementation of BST is presented and compared to those based on build-in pointers. Section IV discusses effectiveness of the implemented structures and algorithms compared to the similar realization of the library container `std::set`. In section V some conclusions are made and recommendations are given for smart pointers usage in the DSA course.

II. DEVELOPMENT OF LANGUAGE FEATURES FOR DYNAMIC MEMORY MANAGEMENT

Before introducing of `new` and `delete` for work with dynamic memory, inherited from the C language functions `malloc, calloc, realloc` and `free` are used, which are still available in C++ by including the header file `<cstdlib>.`

Memory blocks allocated by these functions are not necessarily compatible with those returned by `new`, so each must be handled with its own set of functions or operations. The problems with using these functions are related to unnecessary type conversions and error-prone size calculations (with `sizeof`).

Introduction of `new` and `delete` operators simplifies the syntax, but does not solve all problems. Especially in applications that manipulate complicated linked data structures, it may be difficult to identify the last use of an object. Mistakes lead to either duplicate de-allocations and possible security holes, or memory leaks [2].

All the potential problems with locally defined naked pointers include:

- leaked objects;
- premature deletion;
- double deletion.

One way to circumvent these problems is to simply use a local variable, instead of a pointer, but if we insist to use pointer semantics, the usual approach to overcome such problems is the use of "smart pointers". Their "intelligence" is expressed in that they "know" whether they are the last reference to the object and use this knowledge to destroy the object only when its "ultimate owner" is to be destroyed.

We can consider that a "smart pointer" is RAII (Resource Acquisition Is Initialization) modeled class that manages dynamically allocated memory. It provides the same interfaces that ordinary pointers do (`*`, `->`). During its construction it acquires ownership of a dynamic object in memory and deallocates that memory when goes out of scope. In this way, the programmer does not need to care himself for the management of dynamic memory.

For the first time the standard C++98 introduces a single type of smart pointer – `auto_ptr` which provides specific and focused transfer-of-ownership semantics. `auto_ptr` is most charitably characterized as a valiant attempt to create a `unique_ptr` before C++ had move semantics. `auto_ptr` is now deprecated, and should not be used in new code. It works well in trivial situations – template `auto_ptr` holds a pointer to an object obtained via `new` and deletes that object when it itself is destroyed (such as when leaving block scope). Here `auto_ptr` is "smart" enough, but it appears that the problems entailed outweigh the benefit from it:

- copying and assignment among smart pointers transfers ownership of the manipulated object as well.
- `auto_ptr` can not be used for an array of objects.
- because `auto_ptr` does not provide shared-ownership semantics, it can not even be used with Standard Library containers like `vector`, `list`, `map`.

Practice shows that to overcome (or at least limit) problems as described above it is not sufficient to use only one smart pointer class. Smart pointers can be smart in some aspects and carry out various priorities, as they have to pay the price for such intelligence [1], p. 76. Note that even now, with several types of smart pointers their misuse is possible and programming of wrong behavior.

In the standard [3] instead of `auto_ptr` several different types of smart pointers are introduced (also called Resource Management Pointers) [4], modeling different aspects of resource management. The idea is not new – it formally originates from [5] and was originally implemented in the Boost library and only in 2011 became a part of the Standard Library. The basic, top-level and general-purpose smart pointers are `unique_ptr` and `shared_ptr`. They are defined in the header the file `<memory>`.

The class `unique_ptr` is designed for pointers that implement the idea of exclusive (strict) ownership, what was intended `auto_ptr` to do. It ensures that at any given time only one smart pointer may point to the object. As a result, an object gets destroyed automatically when its `unique_ptr` gets destroyed. However, transfer of ownership is permitted. This class is particularly useful for avoiding leak of resources such as missed `delete` calls for dynamic objects or when exception occurs while an object is being created. It has much the same interface as an ordinary pointer. Operator `*` dereferences the object to which it points, whereas operator `->` provides access to a member if the object is an instance of a class or a structure. Unlike ordinary pointers, smart pointer arithmetic is not possible, but specialists consider this an advantage, because it is known that pointer arithmetic is a source of trouble. `unique_ptr` uses include passing free-store allocated objects in and out of functions (rely on move semantics to make return simple and efficient).

Copying or assignment between unique pointers is impossible if we use the ordinary copy semantics. However, we can use the move semantics. In that case, the constructor or assignment operator transfers the ownership to another unique pointer.

A point of interest is the situation when `unique_ptr` is passed as a parameter of a function by rvalue reference, created by `std::move()`. In this case the parameter of the called function acquires ownership of `unique_ptr`. If then this function does not pass ownership again, the object will be destroyed at its completion.

Using a unique pointer, as a member of a class may also be useful for avoiding leak of resources. By using `unique_ptr`, instead of built-in pointer there is no need of a destructor because the object will be destroyed while destroying the member concerned. In addition `unique_ptr` prevents leak of resources in case of exceptions which occur during initialization of objects.

`The shared_ptr` class implements the concept of shared ownership. Many smart pointers can point to the same object, and the object and its associated resources are released when the last reference is destroyed. The last owner is responsible for the destroying. To perform this task in more complex scenarios auxiliary classes `weak_ptr`, `bad_weak_ptr`, `enable_shared_from_this` are provided.

The class `shared_ptr` is similar to a pointer with counter of the number of sharings (reference counter), which destroys the pointed object when this counter becomes zero. Shared pointer can be used as an ordinary pointer – to assign, copy and compare, to have access to the pointed object via the operations `*` and `->`. We have a full range of copy and move constructions and assignments. Comparison operations are applied to stored pointers (usually the address of the owned object or `nullptr` if none). `shared_ptr` does not provide index operation.

By using shared pointers we can avoid the problems with dangling pointers, which arise while pointers are stored in containers.

A problem with reference-counted smart pointers is that if there is a ring of objects that have smart pointers to each other, they keep each other "alive" – they will not get deleted even if

no other objects are pointing to them from "outside" the ring. Such a situation often occurs in implementations of recursive data structures. C++11 includes a solution: "weak" smart pointers: these only "observe" an object but do not influence its lifetime. A ring of objects can point to each other with `weak_ptrs`, which point to the managed object but do not keep it in existence. Like raw pointers, the weak pointers do not keep the pointed-to object "alive". The cycle problem is solved. However, unlike raw pointers, the weak pointers "know" whether the pointed-to object is still there or not and can be interrogated about it, making them much more useful than a simple raw pointer would be.

In practice often happens a situation when we hesitate which version of a smart pointer to use – `unique_ptr` or `shared_ptr`. The advice is to prefer `unique_ptr` by default, and we can always later move-convert to `shared_ptr` if needed. In our case, however, we had from the very beginning to start with `shared_ptr`, because being recursive by definition, binary trees that we tried to implement with smart pointers cannot do without shared ownership.

III. Implementation of Binary Search Trees

Most attention in our course is given to binary search trees, so here we will focus only on its implementation.

In our traditional implementation with raw pointers beside the special member functions we add methods to insert, search and remove elements, and various deep-first (inorder, preorder, postorder) and breath-first traversals. We include a number of additional functions the implementation of which is a question of interest, for example, calculating the height of the tree and, if there is enough time, balancing. For the implementation of these operations we prefer recursive algorithms because they are shorter and more intuitive. Most difficulties we meet with the deletion, which is normal – the algorithm is most complex.

Since we count on the reliability, in the course we try to follow the methodology for verification of object-oriented programs as proposed in [6].

In order to simplify the technical part and to focus on algorithms, implementing the operations on trees from 2013-2014, we went to implementation with smart pointers. Our initial expectation was that it was possible to avoid all methods of copy and move semantics, destructors for nodes and whole trees.

The interface of smart pointers implementations with which we started was the following:

```
template <typename T>
class Tree {
  struct Node {
         T key;
         shared_ptr<Node> left;
         shared_ptr<Node> right;
         Node():key(), left(), right(){}
         Node(T x):key(x),left(), right(){}
  };
  shared_ptr<Node> root;
//...
public:
        Tree():root(){}
        ~Tree();
        Tree(Tree&&) = default;
        Tree& operator =(Tree&&) = default;
        Tree(const Tree&);
        Tree& operator =(const Tree&);
        bool push(T);
        bool remove(T);
        void inorder();
        shared_ptr<Node> find(T x) {
               return find(x, root);
        }
        void breath_first();
};
```

Because of recursive algorithms that we use for each operation we had to write two functions – one private, with additional parameter the node from which to start. So public method is very short and just calls the corresponding private method that implements the algorithm. For example the public method for deleting:

```
template <typename T>
bool Tree<T>::remove(T x){
         return remove(x, root);
}
```

calls the private method `remove(T, shared_ptr<Node>&)` where the second parameter is the root of the tree:

```
template <typename T>
bool Tree<T>::remove(T x, shared_ptr<Node>& p) {
   if(p && x < p->key)
        return remove(x, p->left);
   else if(p && x > p->key)
        return remove(x, p->right);
        else if(p && p->key == x) {
           if(!p->left)
                p = p->right;
           else if(!p->right) p = p->left;
                else {
        shared_ptr<Node> q = p->left;
        while(q->right) q=q->right;
        p->key = q->key;
        remove(q->key, p->left);
        }
        return true; }
        return false;}
```

TABLE 1. TEST RESULTS

Operations	Binary Search Tree Implementations		
	Row Pointers	*Smart Pointers*	*std::set*
Add element	438	453	156
Search	31	32	15
Remove	47	46	32

Note: time in milliseconds

We note that the code for this method is 37% shorter than the code for the corresponding raw pointers implementation

(due mainly to the fact that there is no need to call `delete`). In addition readability of code was improved. For inserting a node there is no difference between the amounts of code – both methods have 16 rows.

For educational purposes all operations with a single tree ran normally, but when we tested a larger tree, we got a "stack overflow" error during the automatic destruction of the tree at the end of the program. With a standard size of 1 MB stack error occurs even for the destruction of a tree of 29,000 integers. Because of the recursive links a situation occurs where one node keeps "alive" the whole structure. This on one hand requires a large stack, and on the other – can lead to significant delays in the demolition of the structure. So we decided to add a destructor, instead of increasing the stack size from the settings of the linker.

As for the implementation of special member functions, default-tion of move constructor and move assignment operator worked and we did not have to implement move semantics, but copy semantics required to write the appropriate methods, because you need to copy the entire tree structure, so as to obtain a true copy of the tree, not just tree, which contains the same elements.

Comparing the overall implementation of trees with raw pointers, we can conclude that smart pointers give us short and easy to understand code without apparent loss of efficiency (see table).

IV. Performance Evaluation

In order to evaluate the efficiency of smart pointers implementation we carried out an experiment in which we compare the times for typical operations with binary trees, implemented with and without smart pointers.

We compare the 3 conversions: our traditional row pointer implementation, our new smart pointer and library implementation `std::set` (Table 1). Note that `std::set` is typically implemented in libraries as a red-black tree. This may adversely affect the generation time of the tree (for coloring and balance), but improves the search speed.

The same data is used in the experiment: 100’000 randomly generated unique strings of length of 20 stored in a text file. They are used to construct the trees. The first operation "Add element" reads all strings from the file and stores them in the relevant tree. For each tree the text file is opened and read again. For unbalanced versions we obtained tree with height of 38.

In testing for search and remove elements we use another file, which recorded 10,000 strings that are found in the tree. The algorithm makes search and remove for exactly these elements.

The results show that there is practically no differences in performance between implementation of operations with build-in and smart pointers, which is a good argument to continue to study smart pointers in the course DSA. Some surprise for us is almost is the time for `std::set` in operations creating structure (adding operation), which is three times better. Apparently extra time for coloring and balancing the tree is offset by the lower height of the red-black tree – `std::set` for these input data theoretically the tree can get a height of 12, and as mentioned before, the tree in our implementations has height 38. For the same reasons, the search time in our implementations is 2 times worse, and time for removing elements – 1.5 times worse the library implementation.

V. Conclusion

Our initial hypothesis regarding the implementation of BSTs with smart pointers was proven partially. We could not do entirely without implementation of methods of copy and move semantics, but their code turned out to be short, clear and easily understandable for students. Moreover, move semantics in our case can be provided by defaulted move constructors and assignment operators. We consider the second part of the hypothesis, namely the shorter and clearer implementation of the basic operations with data structures for fully achieved. In addition, smart pointer versions do not require user-defined exception handling.

Since we do not have enough empirical data, we can not prove the advantage of this way of teaching DSA yet, but even without helding a strictly formal pedagogical experiment, we can confirm that the results of students tests, homeworks and exams are comparable to those demonstrated by their colleagues trained in previous years under the old program.

The implementation of ADS with smart pointers is more clear and concise, but requires spending time to study in additionally templates and essential elements of the STL, though not in detail. This could be facilitated by reorganizing CS1 course Programming Fundamentals, where to underlie learning C++11/14 and STL. Note that for our implementations it is not needed even to know the full interface for work with smart pointers. In most situations the interface of build-in pointers is sufficient plus function `make_shared` and possibly member function `reset`.

References

[1] Josuttis, N. M. (2012). *The C++ Standard Library: A Tutorial and Reference*. Addison-Wesley Professional; 2nd edition (April 9, 2012).

[2] Boehm, H. & Spertus, M. (2009). *Garbage Collection in the Next C++ Standard*. Proceedings of the 2009 international symposium on Memory management, pp. 30-38. ACM New York. doi>10.1145/1542431.1542437

[3] ISO/IEC. (2011). *International Standard ISO/IEC 14882:2011(E) Information technology – Programming languages – C++* (3rd ed.)

[4] Stroustrup, Bj. (2013). *The C++ Programming Language, 4th Edition*. Addison-Wesley Professional; 4th edition (May 19, 2013)

[5] Dimov, P., Dawes, B. & Colvin, G. (2003). *A Proposal to Add General Purpose Smart Pointers to the Library Technical Report*. C++ Standards Committee Papers. Document number: N1450=03-0033 http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2003/n1450.html

[6] Todorova, M., Kanev, K. (2012). *Educational framework for verification of object-oriented programs,* in Proceedings of the 2012 Joint International Conference on Human-Centered Computer Environments, ACM, New York, pp. 23-27

Training in Music Informatics Master and Bachelor Programs

Emilia Todorova
Department of Information Technologies, Faculty of Mathematics and Informatics,
St Cyril and St Methodius University of Veliko Turnovo
Veliko Turnovo, Bulgaria
e-mail: emilia_todorova@yahoo.co.uk

***Abstract*— Music Informatics (MI) is a relatively new program in some European and American universities. It studies music using tools from a wide range of disciplines, including Computer Science, Cognitive Sciences, Mathematics, and Library Science. The field is propelled by the many new areas of application which study requires viewing music as data, whether in audio, graphical, performance, or a symbolic. form. Music Informatics is an expanding area that combines skills from many diverse disciplines. This paper examines the curricula of the pioneers in teaching MI.**

Keywords—bachelor program; curriculum; master program; music informatics;

I. INTRODUCTION

The term "Music Informatics" (Music Informatics, MI) combines numerous interdisciplinary research. Some of them are not new, as applications of mathematics in music theory. In many universities there are research groups and laboratories working in this field. What is new is that MI goes beyond the research groups: many departments of Music taught discipline "Music Informatics" and in the last few years appear universities offering training to students in Master [3], and even bachelor [1, 2, 4] degree in "Music Informatics." This paper examines the content of the curricula of the pioneers in this endeavor, in an attempt to answer the question could we develop and sustain such a program in our university.

II. MUSIC INFORMATICS IN BACHELOR DEGREE

Most universities offering MI are in the UK: University of Sussex, University of Westminster, and a similar program "Computer Science with Music Technology" in City University London. Britain has traditionally strong research in the field of education. The education in informatics and computer science [5] is not an exception.

A matter of great interest are the qualifications and skills of the graduate students, as well as the perspectives for their future development and , as opposed to training in the reviewed master's degree, here students do not have and cannot have previous training. University of Sussex even offers training for acquiring two different undergraduate degrees in MI: Bachelor of Arts (BA - Bachelor of Arts) and Bachelor of Science (BSc - Bachelor of Science).

Leading departments in training are Music Department and the Department of Informatics and students in both versions of the degree have access to all available courses, computer and music labs. Here, however, students have a BA and hours for learning musical instrument. For the Bachelors of Arts the emphasis is placed on disciplines in the field of music, and for those in the sciences - on computer science. Unfortunately, the University of Westminster does not give details of the curricula in "Music Informatics."

Appendix 1 presents Music Informatics Bachelor of Arts Course curriculum, developed by University of Sussex.

III. MASTER COURSE IN MUSIC INFORMATICS

A. Range of Courses

Leading in this respect is Indiana University, Bloomington, USA. It was the first in the country, and so far the only, offering training in degree "doctor" in MI and "master" in MI. This specialty is studying various musical means from the perspective of a wide range of fields: computer science, cognitive science, mathematics, library and information activities. The emergence of such specialty is the result of the development of many new fields of application in which music is represented as data, no matter what is the form - audio, graphical, symbolic form or another. Some of the guidelines of the development and application of knowledge and skills in MI include:

- Restoration of musical information. The essence is to search a database for a musical pattern;
- Constructing of musical recommendations - to search for music by genre, style, etc.;
- Study and synthesis of musical expressions;
- Convertion of audio to score (similar to convert speech into text format);

Table 1. Training areas according to UNESCO/IFIP

UNESCO/IFIP	ACM
Computer Engineering	Computer Engineering
Computing	Computer Science
Computer Science	
Software Engineering	Software Engineering
Artificial Intelligence	
Information Systems	Information Systems
Computer Information Systems	
Management Information Systems	Information Theory
Information Technologies (IT)	Information Technologies (IT)
Information and Communication Technologies	

- Music analysis - harmonic analysis, studying the structure of phrases, spelling of notes - all this leads to algorithmic research;
- Creation of musical accompaniment systems;
- Play score - the developed techniques explore the correlation between symbolic representation of music and real audio playback;
- Optical music recognition (OMP);
- The disassembling of a music source. This is a very difficult, sometimes impossible task. Separation of the source to parts need very detailed and meaningful information;
- Music for computer games;
- Transformation of MIDI (Musical Instrument Original Interface) format to another symbolic representation of scores. For years, the MIDI file format is the most common for symbolic representation of scores. With the advent of standardized and qualitative tools (eg MusicXML) become necessary algorithms for converting MIDI formats.

B. Requirements for the Applicants for MC in MI

Applicants for Master degree in MI should cover to a very large extent the requirements for the qualification of "Bachelor" in Informatics. UNESCO (United Nations Educational, Scientific and Cultural Organization), IFIP (the International Foundation of Information Processing) and ACM (Association for Computing Machinery) identified the areas that need to be included in the curriculum for obtaining the degree of "Bachelor" in Informatics [6].

Proposals for UNESCO curricula were developed first in 200-0 in cooperation with IFIP. Table 1 shows the areas in which it is necessary to carry out training for acquisition of the specified qualification.

Table 2. Music Informatics Master Course Curriculum

COURSES	CREDITS	TYPE
CORE (4) - Information Management -Music Information Processing: Symbolic - Music Information Processing: Audio - Introduction to Music Informatics	3 3 3 3	CORE
COGNITIVE SCIENCES (3)	3	ELECTIVE
COMPUTER SCIENCE (19)	2	ELECTIVE
EDUCATION (3)	1-3	ELECTIVE
MUISIC (8)	3	ELECTIVE
SPEECH AND HEARING (1)	3	ELECTIVE

Similar specialty develops City University London: Computer Science with Music Technology [7].

Table 2 represents Music Informatics Master Course curriculum, developed by Indiana University, Bloomington. The students must have 12 credits from core courses, and 18 credits from elective courses.

Cognitive sciences module of electives include Mathematics and Logic of Cognitive Science, Methods in Cognitive Science and Models in Cognitive Science. Computer science elective courses are numerous, among them various disciplines in programming, (C++ Programming, Object-Oriented Programming in C++, Programming language Principles, etc.), Data Structures, Algorithms Design and Analysis, Artificial Intelligence, Network and Distributed Computing, Network technologies and Administration and more.

The module of educational electives includes - Instructional Design and Development and Computer - Mediated Learning.

Music is the other big module of electives, and contains eight courses, among them Electronic Text Processing and Distribution, Developing Computer-Based Music Instruction**,** Music Information representation, search, and retrieval, Music Information Processing.

Even the very name of the program implies the focus on computer science. The training is implemented in three modules. The first two are mandatory for all students of

computer science, and the third module is responsible for specialization: it offers elective courses, as for the specialization certain disciplines are recommended.

A detailed review of the curricula of these universities shows that the University of Sussex, especially for applicants for Bachelor of Arts in MI has requirements for ability to work with a musical instrument and also some basic knowledge of music theory, harmony, composition and so on.

IV. Music Informatics Groups and Laboratories

More and more universities maintain research centers and laboratories for specialization both in research and in training of students in music informatics.

The Music Informatics group at New York University is concerned with research aimed at extending the understanding and usefulness of digital music data. Their research interests include the automatic modeling of temporal dynamics in music, the creation of interactive visualizations of music and the content-based analysis of digital music collections [8].

The Music Informatics group at Georgia Tech designs algorithms and software to enable new ways of producing, accessing and enjoying digital music. The specialists are involved in working on technologies to improve the instrument learning experience by giving detailed and instant feedback on the student's performance. More specifically, we adapt automated music analysis and machine learning approaches to enable the tutoring system to evaluate and grade the recording [9].

The Music Informatics and Cognition Research Group at Aalborg University, Denmark, uses computational and mathematical methods to advance understanding of musical structure and the cognitive processes that underlie the ability to perceive and process musical structure. Design of algorithms and development of computer models that derive structural descriptions from musical surfaces are of importance. The research falls within the fields of music informatics, computational music analysis, mathematical music theory, music information retrieval and computational music cognition. Recently the group focuses on the following topics:

- geometric pattern discovery in music;
- compression-based music analysis;
- using wavelets for analysis, segmentation and classification of symbolic music data;
- automatic analysis and generation of 12-tone music.

The Music Informatics Group is a specialized research group within the Department of Computer Science and Artificial Intelligence at the University of the Basque Country, San Sebastián. Music Informatics is the study of computational models of music analysis, music generation, and music information retrieval. Members of the research group work on statistical modeling, computational musicology, music knowledge representation, combinatorial optimization, and pattern discovery [10].

The Music Informatics Research Group at the University of Edinburg, formerly known as the Artificial Intelligence and Music Research Group, is part of the Division of Informatics at the University of Edinburgh, although it has strong links with other research groups and institutes within and out Edinburgh. The group's research interest is the application of Artificial Intelligence techniques and methods to modeling human musical behavior and communication, and so to support music analysis, performance, education and composition research [11].

V. Conclusion

We considered the prospect of developing of Music Informatics program at our university. At first glance, all the prerequisites are present: Department of Music with renowned teachers and traditions; Faculty of mathematics and Informatics with two mathematical departments, a department of information Technologies, a Department of Computer Systems and Technologies; and a Department of Library and Information activities. The University has the necessary qualified teaching staff in different specialties. We have no answer to the question whether teachers are willing to go beyond the boundaries and to work in an interdisciplinary field. Next: funds are needed to build the laboratories and studios. Third, but actually perhaps first in importance, there is the question whether society needs specialists in Music Informatics: it is noteworthy that even in developed societies, where the entertainment industry is very strong and profitable, there are very few students in these programs. On these latter issues our research could not answer, it is a matter of sociological studies.

APPENDIX 1: Music Informatics Bachelor Course Curriculum

Term/Year	Status	Module	Credits
Autumn /1st year	2 of these options	Approaches to Composition and Performance	12
		Music and Society	12
		Popular Music Cultures	12
	Core	Introduction to Multimedia	12
	Core	Introduction to Programming	12
Spring and Summer /1 year	Core	Further Programming	18
	Core	Introduction to Electronic Music	18
	Core	Introduction to Music Studies: Harmony , Analysis and Context	36
Autumn/2 year	Core	Creative Music Technologies 1	15
	Core	Natural Language Engineering	15
	Option	3D Modelling and Rendering	15
		Compilers and Computer Architecture	15
		Databases	15
		Ensemble Performance	15
		Music, Stage and Screen 1: from opera to film	15
		Program Analysis	15
Spring/2 year	Core	Computer Music	15
	Core	Creative Music Technologies 2	15
	Core	Machine Learning	15
	Option	Music, Stage and Screen 2	15
		Solo Performance	15
Autumn/3 year	Core	Advanced Computer Music	15
	Option	Advanced Composition and Arrangement	30
		Dissertation: Historical and Contextual	30
		E-Business and E-Commerce Systems	15
		Film Music and Audiovisual Project	30
		Human-Computer Interaction	15
		Intelligence in Animals and Machines	15
		Knowledge & Reasoning	15
		The Musical	30
		Web Computing	15
Spring/3 year	Core	Generative Creativity	15
	Option	Advanced Performance	30
		Dissertation: Case Studies	30
		Music, Media and Culture	30
		Project 2 Spring	30
		Studio Project	30

REFERENCES

[1] City University London Home Page. www.soi.city.ac.uk/organisation/doc/research/mi/

[2] University of Sussex Home Page. www.cogs.susx.ac.uk/courses/music-**informatics**/

[3] Music Informatics, Indiana University. http://www.soic.indiana.edu/graduate/degrees/informatics/music-informatics/

[4] University of Westminster Home Page. www.wmin.ac.uk/mad/page

[5] Informatics In The Uk: Current Perspective. http://www.ics.heacademy.ac.uk/italics/vol6iss3/white_and_irons.pdf

[6] Connecting Music and Computer Science through a Music Informatics Programof Study and Counterpoint Software:websrv.ewu.edu/groups/academicaffairs/strategicplanning/SPGrantMiddleton_06.pdf

[7] Computer Science with Music Technology: http://www.soi.city.ac.uk/ugdetails/mi.html

[8] Music Informatics Group, New York University. http://steinhardt.nyu.edu/marl/research/music_informatics

[9] Music Informatics group, Georgia Tech. http://www.gtcmt.gatech.edu/research/music-informatics

[10] Music Informatics Group, University of the Basque Country, San Sebastián. http://www.ehu.eus/cs-ikerbasque/conklin/mig/

[11] Music Informatics Research Group, University of Edinburg. http://www.dai.ed.ac.uk/groups/aimusic/

E-LEARNING MODULE FOR DIGITAL DESIGN COURSES

Valentina Kukenska, Paolina Radeva, Petar Minev
Faculty of Electrical Engineering and Electronics
Technical University of Gabrovo
Gabrovo, Bulgaria
e-mail: {vally, polly, pminev}@tugab.bg

***Abstract*— This report describes the approach and the implementation of an e-learning module for training in the design of digital hardware for students of CST at the Technical University of Gabrovo.**

The aim of the electronic training module is to prepare students in several disciplines related to computer-aided design of digital circuits. Thus students always have electronically-based learning materials on hand and more time remains in the practical exercises for the development of more complex projects and the clarification of the issues accompanying this process.

Keywords*— *digital design, education, e-learning module

I. INTRODUCTION

The introduction of e-learning in Bulgaria's higher education system raises the question about its incorporation into the training courses for students in electronic engineering. It could not fully replace classroom learning, as hands-on training needs the use of hardware resources (printed circuit boards, stands, testers, etc.). The same goes for students in Computer Systems and Technologies regarding their knowledge of hardware and digital hardware design [9].

Electronic engineering courses could be used as auxiliary education. They could help reduce the duration of training in correspondent subjects. Students can prepare themselves for the practical sessions, and the latter can take place in laboratories.

The development of e-learning modules means one should define which parts of the subject matter could be transformed into e-learning content to be used for self-preparatory work and auxiliary to training. Once the subject matter is distributed into one or more disciplines (in one field), it should undergo further fragmentation into smaller sections called e-learning units (modules).

II. BODY OF PRESENTATION

In general, the subjects taught in the Computer Systems and Technologies department could be divided in two categories: hardware ones and software ones.

A special place among them is reserved for the subjects in the field of computer-aided design, which is in the core of this paper. They also fall into two sections - hardware and software. The software part features tools for automation of the design process, as well as hardware description languages VHDL and Verilog. The hardware part is the final result, the digital unit or device obtained in the process. The subject matter dealing with the aforementioned software devices and hardware description languages could be presented in an electronic form (e-form). What could not be included in this e-form is the physical implementation of the hardware project, which requires the proper training printed circuit board, like the one shown in Fig. 1.

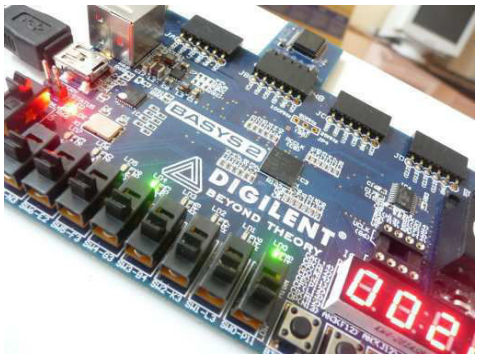


Fig. 1. Basys 2 training printed circuit board for computer-aided hardware design.

Ready made e-learning modules presented in this paper allow students to do some self-preparatory work for their practical sessions. The physical implementation of a digital unit or a device requires the use of a personal printed circuit board or the use of in-house board for the laboratory [2].

Students can prepare their tasks by following all stages of the design process on their own. The final stage or the physical implementation, i.e. launching the readied project into the printed circuit board and verifying its operational capabilities, could be done at the Automated Design laboratory of the Computer Systems and Technologies department.

A. Contents of the E-Learning Module for Computer-Aided Digital Hardware Design

The hereby presented e-learning module for computer-aided digital hardware design contains the following lessons:

- Stages in the design of digital circuits. Levels of presentation of digital circuits

- Designing digital circuits with programmable logic devices
- Specification languages. Hardware description language VHDL
- Development environment for computer-aided digital circuits design
- Designing of combinatory circuits
- Designing of sequential circuits
- Designing of finite-state machine

The first three lessons offer theoretical presentation of the methodic and the tools for computer-aided designing of digital circuits [1], [7].

The following four lessons are more practical. They present the employed computer-aided design system, as well as the use of the hardware description language in the design of the combinatory, registry and the managing part of the digital circuits [3], [4], [5], [6], [8].

Each lesson is accompanied by a test to assess the level of preparation and self-preparation of each student.

All e-lessons can be found on the Moodle e-learning platform.

B. Summarized Structure of the E-lesson in the E-learning Module for Computer-Aided Digital Circuit Design

The e-lessons are in the form of presentations and have the summarized structure shown in Fig. 2.

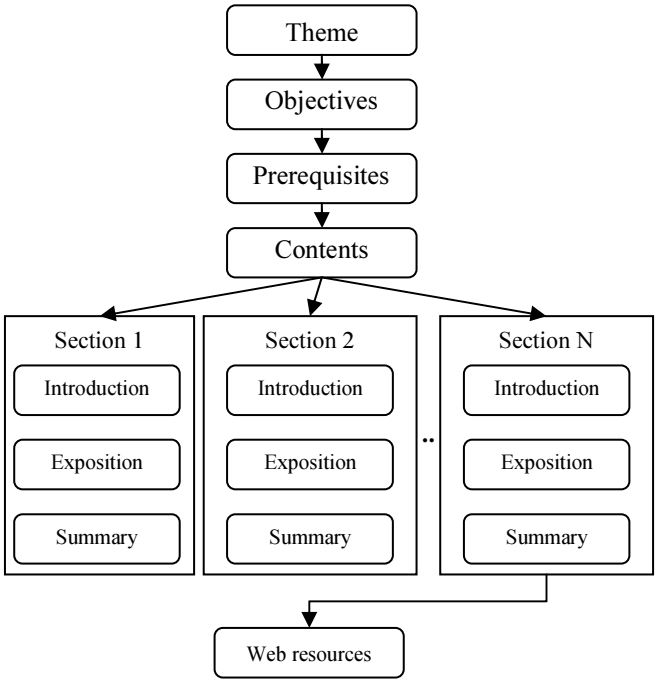


Fig. 2. Summarized structure of the e-lesson in the e-learning module for computer-aided digital circuit design

This way of presenting the subject matter, i.e. the subsequent accumulation of knowledge, helps students get information in small portions and make step-by-step progress.

Further reading on the topic could be found in the online resources listed in the latest section called Web resources.

Figures 3, 4, 5, 6, 7, 8 and 9 show fragments from several of our e-lessons.



Fig. 3. Fragment from the e-lesson "Development environment for computer-aided design of digital circuits"

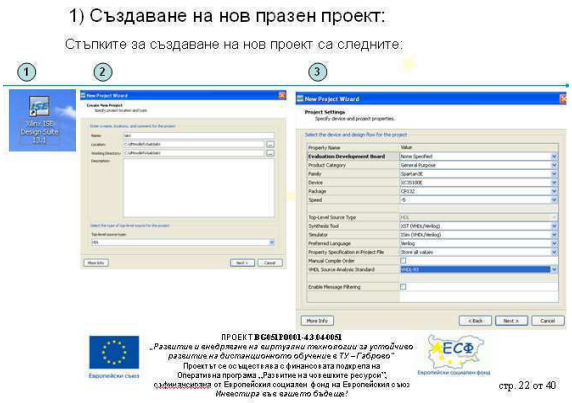


Fig. 4. Another fragment from the e-lesson "Development environment for computer-aided design of digital circuits"

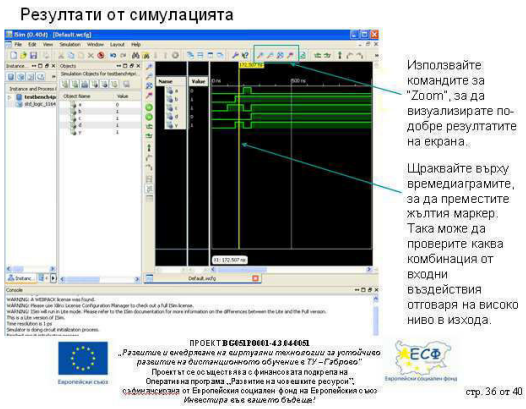


Fig. 5. Another fragment from the e-lesson "Development environment for computer-aided design of digital circuits"



Fig. 6. Fragment from the e-lesson "Design of combinatory circuits"



Fig. 7. Another fragment from the e-lesson "Design of combinatory circuits"



Fig. 8. Fragment from the e-lesson "Stages in the design of digital circuits. Levels of presentation of digital circuits"



Fig. 9. Web resources in the e-lesson "Specification languages. Hardware description language VHDL"

C. *E-lesson Format*

The hereby presented e-lessons are transformed into a SCORM 1.2 compatible format with the help of the iSpring Free programme (Fig. 10). It is in this format that they are uploaded in the university's e-learning environment http://dmoodle.tugab.bg in the section "Modules from the school year Faculty EE Activity 9.3."



Fig. 10. Printscreen of the iSpring Free tool for transforming presentations into SCORM 1.2 compatible formats.

D. *Tests Accompanying the E-lessons*

Tests accompanying each e-lesson have been created with the help of tools available in the Moodle environment. Test questions are contained in a pool, arranged in different categories depending on the lesson theme. For each of category there are fourteen closed questions. Questions and answers are randomly distributed every time a student wants to try the test (Fig. 11). A test is considered passed, when the correct answers are more than 60%.

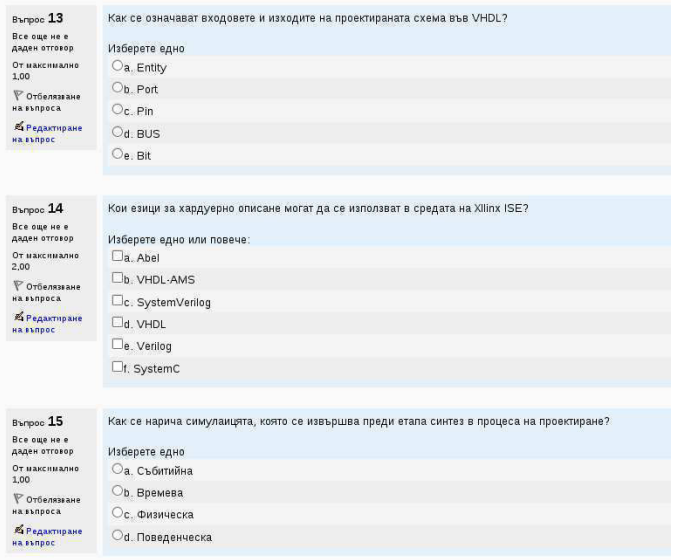


Fig. 11. Fragment from a test for the e-lesson "Development environment for computer-aided design of digital circuits" in the Moodle environment

III. CONCLUSION

The e-learning modules presented in this paper have been developed in line with the training plan of the Computer Systems and Technologies department of the Technical University of Gabrovo. They can be used for self-study of the subject matter, for helping the practical training in computer-aided designing of digital hardware, and for preparing of theses.

The hereby presented e-learning modules are opened for further expansion and upgrade with new themes, such, for example, designing digital hardware on registry level, as well as designing on higher levels of abstraction, application of the hardware description language Verilog, etc.

REFERENCES

[1] Chu P., "FPGA Prototyping by VHDL Examples: Xilinx Spartan-3 Version", Wiley, 2008.

[2] Haskell R., D. Hanna, "Introduction to Digital Design Using Digilent FPGA Boards - VHDL/Active-HDL Edition", 3rd Ed., 2012.

[3] Hebib E., "Design of hardware control module for traffic light", Proc. of Student Scientific Conference 2011, Technical University of Gabrvo, Bulgaria, October 21, 2011, pp. 83-87.

[4] Kukenska V., I. Simeonov, "Modeling and optimization of sequence logic circuits", The 11 International Scientific and Applied Science Conference, ELECTRONICS ET'2002, Sozopol, Bulgaria, September 25-27, 2002, book 4, pp 14-19.

[5] Kukenska V., I. Varbov, P. Minev, "Modeling and analysis of interface module for keyboard", Proc. of the International Scientific Conference UNITECH'12, Gabrovo, November 16 – 17, 2012, vol. I, pp 419 - 422.

[6] Kukenska V., I. Simeonov, "Design of combinational circuits", Proc. of the International Scientific Conference UNITECH'02, Gabrovo, November 21-22, 2002, pp. 227–229.

[7] Kukenska V., L. Tzekov, "Computer models synthesis of digital systems of high-level language", Second Scientific Conference on Electrical Engineering, Electronics and Communications "Sliven-2003", Sliven, October 3-4, 2003.

[8] Minev P., V. Kukenska, "Modelling and synthesis of finite state machines with VHDL", Proc. of the International Scientific Conference UNITECH '07, Gabrovo, November 23–24, 2007, vol. I, pp 361–365.

[9] Minev P., V. Kukenska, "VHDL training for students of the Computer Systems and Technologies (CST) department", Proc. of the Scientific Conference, Stara Zagora, March 17 – 18, 2007, pp 160 – 164..

INTERACTIVE MODULES FOR COMPUTER GRAPHICS LEARNING

Paolina Radeva, Valentina Kukenska, Petar Minev
Faculty of Electrical Engineering and Electronics
Technical University of Gabrovo
Gabrovo, Bulgaria
e-mail: { polly, vally, pminev }@tugab.bg

Abstract — **An interactive computer based system, designed by the authors, is presented. The main educational purpose of the system is the learning of computer graphics basis. An accent is put on the creation and use of interactive educational tools for applying the fundamentals in the process of learning computer graphics equipments.**

Keywords — computer graphics, education, computer-based modules

I. INTRODUCTION

The introduction of e-learning in education system in Bulgaria raises the questions about its incorporation into students' courses in engineering. E-learning could be used as auxiliary education.

The e-learning materials are structured into parts, which are separated in two modules. First module consists of thirteen lessons. Second module consists of ten lessons. Students can prepare themselves for practice and they can solve practical problems.

Using of e-learning modules reduces students' preparation time for design and graphic projects realization.

II. EXPOSITION

In the curriculum of "Computer Systems and Technologies" at the Technical University - Gabrovo are set courses "Computer Graphics" and "Computer graphics systems." They learn the basics of computer graphics and its applications in engineering tasks.

Computer Graphics Module includes thirteen lessons and Computer Graphic Systems Module includes ten lessons. Each lesson has theme, objectives and contents. The contents are structured in sections. Each section has introduction, exposition and summary. This way presenting the subject matter helps students to get information in small parts. Students can prepare their tasks by following all stages of the computer graphic design.

All lessons are accompanied by tests to assess the level of students' preparation. Test control shows self-preparation of each student.

E-learning modules contain information of using coordinate systems and coordinate projections, coordinate transformations, curves and surfaces, forms and sizes, places and orientation, shapes and objects, image file formats. Students can study about fundamentals of computer graphics. They can use tools and environments for computer image processing. They can train themselves using filters, layers and etc.

The students can use modules to train themselves in modeling of two-dimensional and three-dimensional objects Modules presents tools and environments methods, tools and environments for polygons and curves modeling. There are illustrated primitives and their components, extrusions, add divisions, smoothing, revolving, connecting surfaces and etc. The lessons in the modules presented using of reference images image planes.

E-learning modules contain information of light and color characteristics. Students can study and train themselves in using materials, colors, reflection, transparency and lighting. Render choice, render settings, rendering if a single image and animation rendering are shown.

The lessons presented animation types and rigging. The students study about skin and object skeleton creation and train using frames, key frames, frame rate, motion paths and camera settings.

All lessons have the structure shown in fig.1. Every lesson of modules consists of three types of files: theory, presentation and examples.

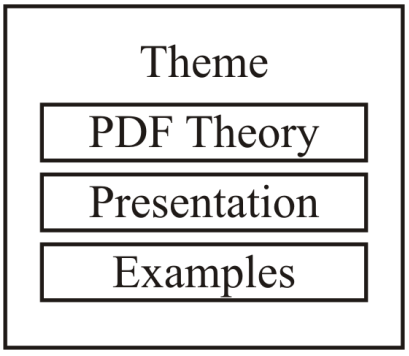


Fig. 1. Lessons' Structure.

The hereby e-learning modules for computer graphics contain the following thirteen lessons:

- Computer graphic fundamentals
- Computer graphic equipments
- Graphic file formats
- Coordinate systems
- Graphic primitives
- Coordinate transformations
- 2D graphics
- 3D graphics
- Curves and surfaces
- Color, gradient and pattern filling
- Layers
- Filters
- Animation

The hereby e-learning modules for computer graphic systems contain the following ten lessons:

- Computer graphic systems fundamentals
- Polygonal modeling
- NURBS modeling
- Extrude and revolve
- Color models and palette
- Materials and texture
- Light sources
- Rendering
- Rigging
- Computer animation

The theory of lessons is presented as Portable Document Format (PDF) files (fig.2). The text is formed by using the word processing program Microsoft Office Word. The text materials are converted into PDF – Microsoft Office Word supports conversion as standard function which simplifies process.

The presentations were made by using the product Microsoft Office PowerPoint (fig.3). Conversion to SCORM 1.2 compatible format can be done by iSpring Free. Then they can by uploaded in the university's e-learning environment.

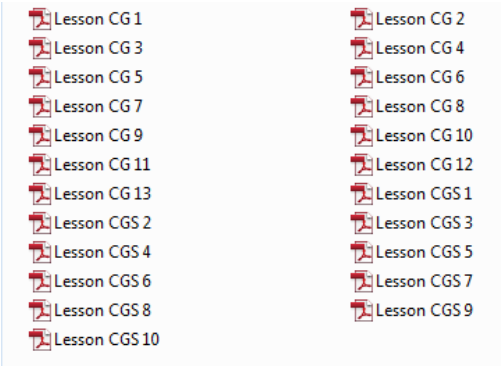


Fig. 2. Computer Graphics and Computer Graphic Systems PDF Files.



Fig. 3. Computer Graphics and Computer Graphic Systems Presentation Files.

The practice parts of e-learning modules are organized in video files. They show tools, action order and parameter for students tasks needs. The number of frame per minute, resolution, cursor position and movements, used key and keyboard shortcuts are important. All video lessons are created by free screen recorders. The final result is video file in .avi or .mp4 file format.

The examples in modules are video files with computer animation, visual effects, morphing, riggings, particles and cloth simulations (fig.4). There are many examples with static images in .jpg and .png file formats.

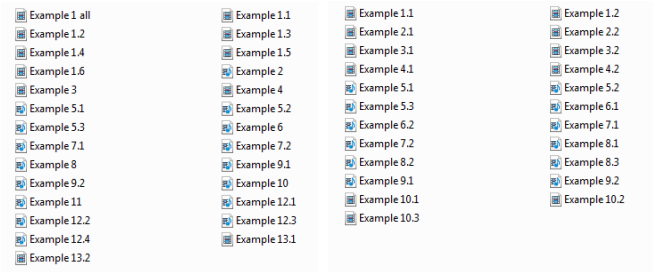


Fig. 4. Computer Graphics and Computer Graphic Systems Examples.

In paper is presented local program environment which provides quick access to different lessons in the modules.

The structure of computer graphic systems module is shown in fig. 5.

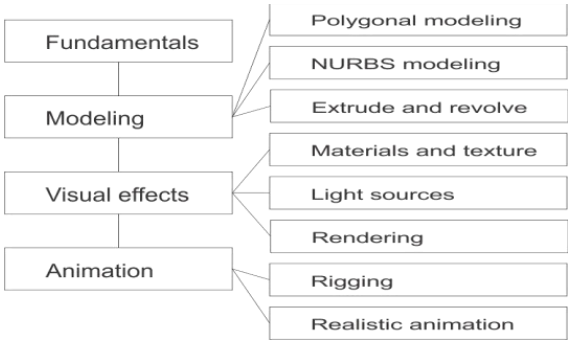


Fig. 5. Structure of Computer Graphic Systems Module.

The screenshots of modules are shown in fig. 6 and fig.7. The screenshot of theory part is shown in fig.8 and the screenshot of presentation part is shown in fig.9.

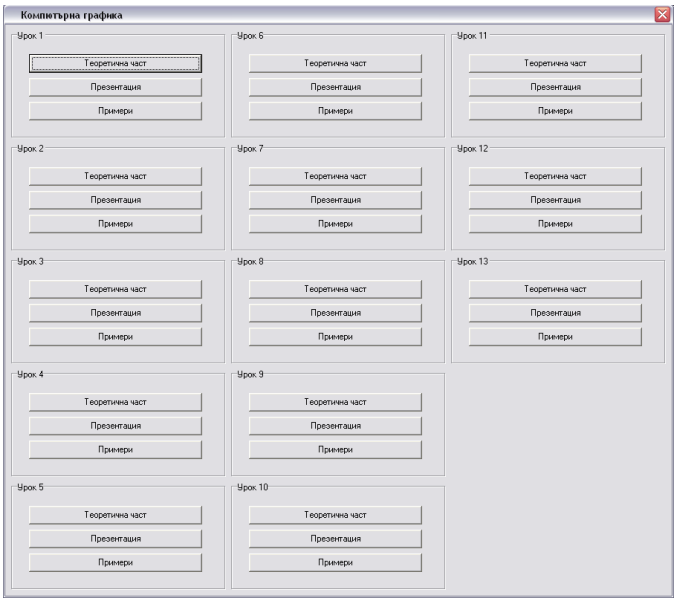


Fig. 6. omputer Graphics Module.

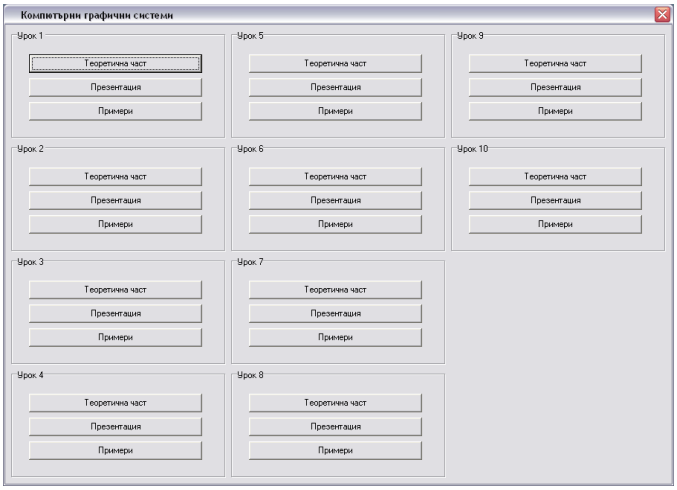


Fig. 7. Computer Graphic Systems Module.

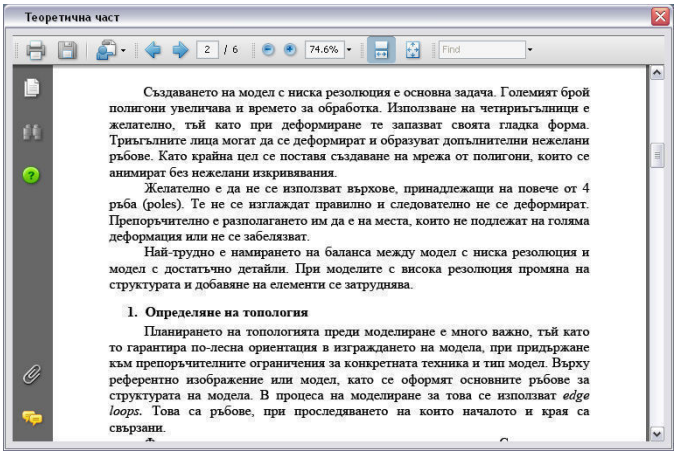


Fig. 8. Part of Portable Document Format File.

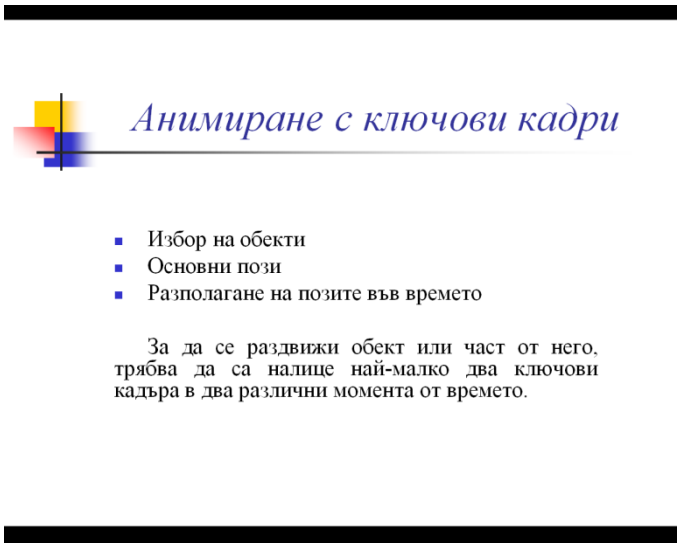


Fig. 9. Part of presentation.

Video and image examples are shown in fig.10, fig.11 and fig.12.

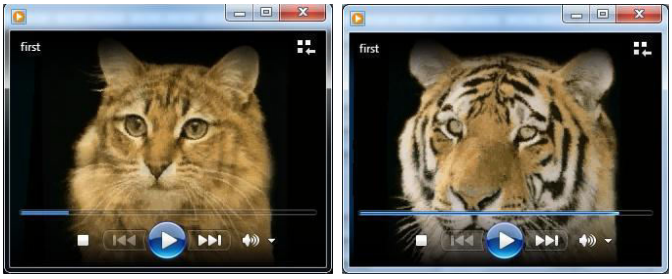


Fig. 10. Video Example – morphing.



Fig. 11. Video Examples – visial effect.



Fig. 12. Image Examples.

All lessons show and explain different ways and tools, parameters and short keys that help students to study and then to create their own projects.

In fig.13 is shown extrusion of faces. In fig.14 is shown creation of skeleton and in fig.15 is shown result – animated robot.

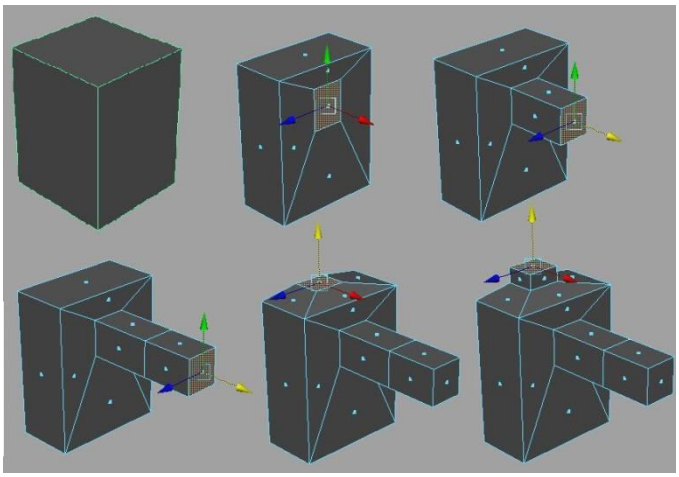


Fig. 13. Modelling by extrusion.

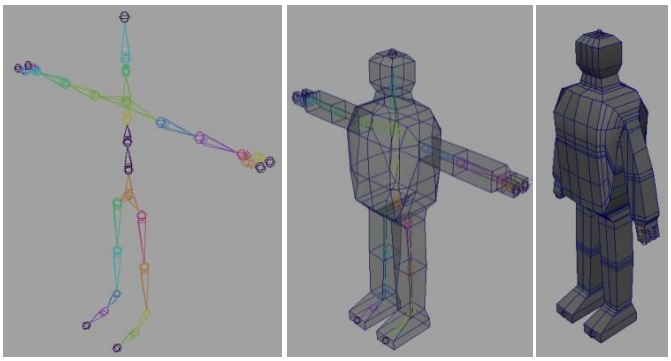


Fig. 14. Rigging.

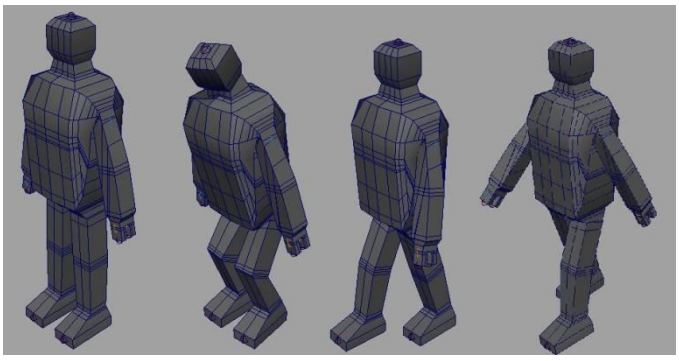


Fig. 15. Animation.

In fig. 16 is shown result of students' model created after watching of lesson about modeling.

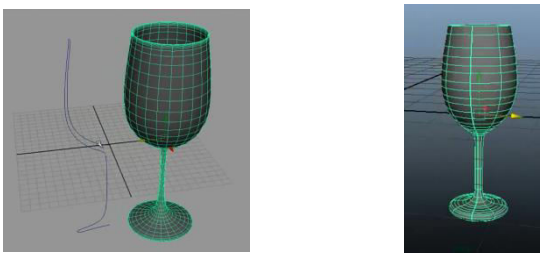


Fig. 16. Part of lesson and result of students' modelling by revolve.

After education students can create their own projects like architectural objects, game heroes and animations. A student's model of human hand is shown in fig.17.

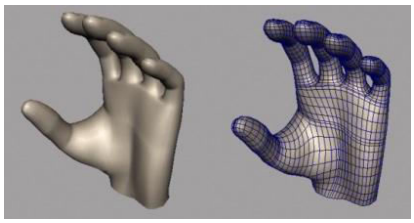


Fig. 17. Student's 3D model of human hand.

Modules increased students' interest and their exam passing success. At last year more than 75% of students passed their exam from the first time and 9% of students reached maximum points.

Some of students present their projects in the student scientific conferences in Technical university of Gabrovo. Two of them take a prize for third place.

III. CONCLUSION

The e-learning modules presented in this paper have been developed in line with the training plan of the Computer System and Technologies Department of the Technical University of Gabrovo, Bulgaria. They can be used for students' self-study. The modules help the practical training in computer modeling and computer animation. The modules are opened for expansion and upgrade.

References

[1] Feyzulah F., Development of modules for three dimensional modeling, Student Scientific Conference 2013, Technical University of Gabrovo, Bulgaria, October 25, 2013, pp 346-321.

[2] Radeva P., Basic characteristics and problems of image processing and image recognition, Proceedings of the International Scientific Conference UNITECH'01, Gabrovo, Bulgaria, November 22 – 23, 2001, vol. I, pp 376 – 380.

[3] Radeva P., G. Kondev, Modules for modeling of three dimensional objects and visual effects, Proceedings of the International Scientific Conference UNITECH'13, Gabrovo, Bulgaria, November 22 – 23, 2013, vol. II, pp 349 – 352.

[4] Radeva P., D. Atanasov, Image Processing Program Environment, Proceedings of papers TELSIKS 2003, Serbia and Montenegro, Niš, October 1-3, 2003, Volume 2, 542 – 545.

[5] Radeva P., I. Hristov, Modeling and realtime visualizing of three dimensional objects in the Internet, Proceedings of the International Scientific Conference UNITECH'01, Gabrovo, Bulgaria, November 22 – 23, 2001, vol. I, pp 381 – 385.

Internet Safety for Children in School

Venera Krasniqi, Marija Mihova -Faculty of Computer Science and Engineering

University "Ss Cyril and Methodius Skopje"

Abstract— **in this paper we will elaborate and analyze the internet safety of the students in the high school, how much the students uses the internet and what are the risks of using it. Through this paper we will give the results gathered from the survey we conducted with 50 students in high school "Center of Competence", Ferizaj -Kosovo. We know that the internet world is often full of examples of how the children and young people in high schools are at risk and under the constant threat from some of the dangers by the Internet. As in any environment, online or offline, real or virtual world, always will be danger.**
Understanding the environment in real world will help to manage the perceived risk of these dangers. The same stands for online world. As such, the biggest issues that educational professionals in High schools in general face regarding Internet Safety and responsible use is that many do not understand the impact that online environment has had on the behavior and expectations of young people. The overall objective of this paper is to measure the High School children's online behavior and access their knowledge regarding internet safety issues by first mapping their online activities and analyzing the risks associated with these activities.

Index Terms— **Internet, Safety, Education, Children**

I. INTRODUCTION

Most children today begin their online experience at very young age. As an increasing number of children use the internet, using a variety of electronic devices to surf the web, they are also exposed to different type of risks [5]. If used properly, internet access can help children learn more about the world and serve as useful tool for their development. On the other hand, if is not properly or left unmonitored by adults, can lead to serious physical and psychological dangers, thus begin a challenge to Kosovo's Society.

Many existing approaches to promoting Internet safety make use of the risk loaded environment which can inflame parents and educational professionals to adopt an approach with excessive filtering and restrictive access [1]. Whilst not wishing to minimize in any way the possibility for damage, this paper presents research carried out at the "Center of Competence" High School, which considers education and permission as complementary aspects to a toolkit of resources and strategies in promoting Internet Safety. The more comprehensive method has focused on the empowerment of young people to promote Internet Safety. We know that the internet world is often full of examples of how the children and young people in high schools are at risk and under the constant threat from some of the dangers by the Internet. As in any environment, online or offline, real or virtual world, always will be danger [3]. The biggest issues that educational professionals in High schools in general face, regarding Internet Safety and responsible use is that many do not understand the impact that online environment has had on the behavior and expectations of young people. Millions of young people can greatly enhance the education and life experience through the use of the Internet and other communication technologies [4]. Through these technologies they open up new opportunities for young people, improvement their creativity and significantly contribute to achieving the aims of raising e-skills and digital competence. However, the technological developments and their use by children and young people pose risks and present challenges such as : giving out personal information, seeing pornography, violent or hateful content online, being bullied (i.e. cyberbullying), receiving unwanted sexual comments and meeting an online contact offline [4]. According to the Convection on the Rights of the Child, children have human rights such as "The right to develop to the fullest" and "The right to protection from harmful influences". We believe that through this paper we'll inform Kosovo policy- makers about the dangers of non-implementing these core human rights of children. In order to stay safe, children and young people need to be informed about both online opportunities and risks and how to deal with these risks – they need to be empowered to use the internet I n a safe and responsible manner [4].

The overall objective of this paper is to measure children's online behavior and assess their knowledge regarding internet safety issues. This paper tackles issues such as how children specific high school in Ferizaj use the Internet, their frequent online activities, and various risks and threats such as watching inappropriate content. The most valuable asset of Kosovo are children as they present the future of our country. Furthermore, they are the most vulnerable members of the society and their protection against crime must be a priority for the policy makers. Unfortunately the same improvements in technology that allow children to grip more knowledge and develop new communication tools are also leaving them vulnerable to the exploitation an harm by various online activities [2]. The paper aims at informing the children, parents, teachers, and policy

This work was supported in part by the University of "St. Cyril and Methodius" –Skopje (FINKI).
Author Venera Krasniqi, PhD student at University of "St. Cyril and Methodius"-Skopje, FINKI. E-mail: VeeneraKrasniqi@gmail.com.

Marija Mihova Ph.D. , Associate professor at University of "St. Cyril and Methodius" –Skopje (FINKI). Email: marija.mihova@finki.ukim.mk

makers in Kosovo about the complexities and risks from online activities.

II. METHODOLOGY

The methodology used to gather data in this paper is a single-country comparative report which seeks to firstly describe children's online behavior and then explore the risks associated with internet usage. Taking into consideration the lack of data in Kosovo, we have developed a survey which was sent to a specific high school in Ferizaj " Center of Competence" who is the first high school financed by Luxemburg. The sample size consist of 50 students in the beginning then we'll expand in other high schools in Ferizaj including a greater number of students.

Alongside receiving permission to conduct the survey in the schools from the Departments of Education in Ferizaj and Management of high school "Center of Competence" we have asked every child to have the permission of their parents. The survey contained 20 questions, including 3 questions about sexual content online.

Aiming to have the best possible answers in the survey, each questions was first read form the interviewer after which the children had the opportunity to ask (loud or individually) for more clarification, if there is any unclearness.

III. EMPIRICAL RESULTS

In order to define how children use the internet as well as levels of internet usage we conducted a survey in which we divided the questions into three categories:

Internet usage – the survey included several question about children's internet usage which refers to all online activities from any of the devices used at any level and how they access. The questions in this category are:

Q1: How often do you (children) use the internet?
Q2: How long do you (children) use the internet daily?
Q3: From which devices do you (children) use the internet?
Q4: Where do you use the internet?

Online activities – we have asked high school students about which activities they take up, in order to understand the opportunities they have and provide a background to identify and discover online risks.

Q5: How safe do you (children) feel while using the internet?
Q6: Do you know to identify illegal content in the internet?
Q7: Which are your (children's) preferred online activities?
Q8: What is the number of your friends in social networks?
Q9: What type of personal information do you put in your own profiles in Internet?
Q10: Does your personal data are misused?
Q11: Does someone posted disturbing content in your Wall?
Q12: Does someone stole your e-mail?
Q13: Do you play Online games?
Q14: What kind of games do you (children) play?

Risks – Some of the risks addressed in the survey and also in this paper are watching inappropriate content online (pornographic, sexual content, aggressive videos), meeting with new people, missuses of personal information, and attack by viruses.

Q15: Have you seen images with pornographic/sexual content in the past year?
Q16: Have you been bothered /disturbed from seeing sexual images online?
Q17: From which devices you're used to watch disturbing pornographic content?
Q18: From which place of internet access you watch disturbing pornographic content?
Q19: Do you contacted someone online, or went to meet face-to-face with stranger?
Q20: Does your computer is attacked from any virus?

TABLE I. FREQUENCY DISTRIBUTION FOR THE CHRACTERISTICS OF THE SAMPLE

Gender			**Residence**		
	Freq.	%		Freq.	%
F	33	66.0	F	23	46.0
M	17	34.0	Q	27	54.0
Department			**Age**		
Audiometry	19	38.0	15	6	12.0
Pharmacy	6	12.0	16	34	68.0
Optics	25	50.0	17	10	20.0

Most of our data are categorical. The sample description is given in Table I, where we give the frequency tables for, gender, residence (F means village and Q means City) department and age.

The results we gained analyzing the frequency distribution are:

- Respondents were asked to show the frequency of internet usage from which the survey shown that most of the high school students use the internet daily 50 or (100%) for one to two hours 17 or (34 %) out of 50 respondents.
- Internet usage differs depending on the age of the user. From the results that we gained from the frequency tables, we can conclude that children in High School " Center of Competence"- Ferizaj, Kosovo mainly use the internet for communication with each other (e.g. Facebook) 22 or (44%) and entertainment purposes (e.g. Games) 18 or (26%) out of 50 respondents. It is very important to note that using the internet for educational and academic searches contains less risks for children then using the internet for educational and academic searches contains less risks for children then using it for networking. The more children use the internet for communication (Facebook, Twitter, and Instagram) or entertainment (listening to music, watching YouTube videos) the higher their exposure to online risks such as cyber bullying, meeting new people and abuse of personal data.

- Some websites contain illegal materials, while others may be legal, but are meant only for adults or have unregulated advice which can be harmful for children. Generally, children may access inappropriate pornographic content accidentally or intentionally from curiosity. For ethical purposes, we have been more careful and tried to be more sensitive about questions with pornographic content in the way that children not to be asked explicitly, for to do so might induce new ideas to children who might not be aware of such phenomena. From the frequency tables we can conclude that 25 respondents or (50 %) have answered that they have seen pornographic/sexual content in the past year whereas the other 25 respondents or (50 %) answered that they have not seen pornographic/sexual content in the past year.

A. Descriptive Statistics

The goal of this research was to collect data, first in one high school then proceeding with other high schools and see whether the children of age:15-17 years old are safe while searching or surfing in the internet. In order to have better results we compiled very carefully three question about the pornographic/ sexual content to 50 students that we thought nowadays present a rising risk for children. Therefore, in this paper we are concentrate on analyzing questions Q15 and Q16 according to residence, age and gender.

In order to see is there difference in the answers of this questions between the students from rural and urban environments we made chi square tests. The cross tables are given in Table II, from where it is obvious that the around the half of students from F and Q answered that **have seen and no have seen images with pornographic/sexual content in the past year**. Also around the half of students from F (village) and Q (City) answered that **have been disturbed from seeing sexual images online.** Consequently, the Pearson Chi square test accept that the answers of this questions do not depend from residence of the students, for Q15 with probability $p = 0.395$, and for Q16 with probability $p = 0.586$. Moreover, the continuity correction test accept the same for Q15 with probability 0.57, and for Q16 with probability 0.794.

TABLE II. **CROSSTAB RESIDENCE*Q15 AND RESIDENCE*Q16**

Count

		Q15			Q16		
		Yes	No	Total	Yes	No	Total
Residence	F	10	13	11	11	12	23
	Q	15	12	15	15	12	27
Total		25	25	50	26	24	50

In order to see is there difference in the answers of this questions between the students from 15, 16 and 17 years old again we made chi square tests. The cross tables are given in Table III, from where it is obvious that the around the half of students in each age answered that **have seen images with pornographic/sexual content in the past year** and that **have been and bothered /disturbed from seeing sexual images online.** Consequently, the Pearson Chi square test accept that the answers of this questions do not depend from the age of the students, for Q15 with probability $p = 0.587$, and for Q16 with probability $p = 0.745$.

TABLE III. **CROSSTAB AGE*Q15 AND RESIDENCE*Q16**

Count

		Q15			Q16		
		Yes	No	Total	Yes	No	Total
Gender	F	15	18	33	14	19	33
	M	10	7	17	12	5	17
Total		25	25	25	50	26	24

In order to see is there difference in the answers of this questions between the students from of different gender, again we made chi square tests. The crosstabs are given in Table IV, The results here are quite different. An interesting notation is that males more frequently deal with pornographic contents then females. The Pearson Chi square test accept that there is no correlation between genders for the question Q15 with probability $p = 0.059$, which is small probability. For Q16 The Pearson Chi square test accept that there is not correlation between genders with probability $p = 0.37$, so for this question we may conclude that boys and girls are near equal been **bothered /disturbed from seeing sexual images online.**

IV. CONLCUSION

Ways of going online are changing rapidly. For many Kosovar children (especially Ferizaj city), internet use is now thoroughly embedded in children daily lives and everyday routines. The overall objective of this paper is to measure the High School children's online behavior and access their knowledge regarding internet safety issues by first mapping their online activities and analyzing the risks associated with these activities. We can conclude that the results from survey that we conducted for 100 students in High School "Center of Competence" suggests that children's in high schools are very active users of the internet and generally use it for entertainment and communication purposes.

Also, from our survey results we can conclude that the internet usage by children in Ferizaj, Kosovo is highly personalized and internet access and usage is very difficult to be monitored by educational professionals and parents as well. The fact that children watch pornographic content in friend's home, internet cafes, their own, show that their teacher nor their parents are able to put restrictions in terms of what we pages children can access. In addition to this, the use of very personal devices such as phones or personal PCs and laptops furthermore proves that children in High School "Center of Competence " can make the difference between inappropriate and appropriate content, as they use more personal devices.

Therefore, this paper outlines rising issues of Internet safety for children and in other hand will inform policy makers in Kosovo that they need to be conscious of if they wish to be

ambassadors for the Internet Safety for children in high schools of Ferizaj, Kosovo.

V. REFERENCES

[1]http://www.communications.gov.au/__data/assets/pdf_file/0016/204064/Discussion_Paper_-_Enhancing_Online_Safety_for_Children.pdf
[2] http://www.socialresearchmethods.net/kb/statdesc.php

[3].http://www.bloxx.com/assets/downloads/US/bloxx_whitepaper_Internet%20Safety_us.pdf

[4].http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/121EN.pdf

[5].http://futureofchildren.org/futureofchildren/publications/docs/18_01_FullJournal.pdf

A New Tool for Calculation of a New Source Code Metric

Emil Stankov, Mile Jovanov, Kjiro Gjorgjiev and Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
{emil.stankov, mile.jovanov, ana.madevska.bogdanova}@finki.ukim.mk, kire9dk@gmail.com

***Abstract*—Teaching programming is an activity that becomes more and more popular. Assessment of the students that attend introductory courses in programming can partly be done through presentation of simple source code fragments to them. Students should be able to provide the answer to the question: "What is the output of the given code?" When preparing the code segments, teachers should be aware of the complexity ('weight') of the code, and should also try to provide same or similar complexity tasks for all students. Nowadays, when there is a lot of research on the issue of automatic question production, the necessity of having a way to automatically measure the weight of some code is indisputable.**

In this paper we present a new source code metric that helps defining the weight of the code and a new tool that employs it.

Keywords—source code analysis; source code weight; teaching programming

I. INTRODUCTION

Teaching programming is an activity that is becoming more and more popular. Undoubtedly, this is due to the popularity of computer science nowadays, and moreover of programming, as its essential part. One of the most important challenges that teaching programming brings, particularly in courses attended by large numbers of students, is assessment.

Assessment of students' work, as a significant and almost inevitable part of the process of education on high school and university level, can be implemented in different ways, depending on the type of knowledge (theoretical, practical, or both) that is expected to be gained by the students. When it comes to programming courses, especially introductory ones, assessment can partly be done by presenting simple source code fragments to students and determining their basic level of knowledge and understanding of the programming language in which the respective fragments have been written. One clear means to do this is by asking them questions of the form: "What is the output of the given code?" Although these types of questions can hardly assess capabilities such as problem solving, algorithmic thinking or deep logical reasoning, they can (at least) give a good insight into the understanding of the basic programming constructs of the underlying programming language, as well as the comprehension of some basic programming concepts in general. According to the Bloom's taxonomy of educational objectives [1], comprehension is a stage that precedes application, so it is a good practice to examine the students' ability for comprehension, before asking them to apply their knowledge.

In order to achieve objective and fair assessment on a particular course exam, all students taking the exam must be asked questions of same or very similar complexity, i.e. questions that require the same level of knowledge to provide a correct answer. In the context of teaching programming and the types of questions mentioned previously, this means that teachers should be aware of the program codes complexity when preparing these questions, and should always try to provide questions containing same or similar complexity code fragments for all the students. As noted previously, this becomes particularly challenging when working with large groups of students, since teachers must maintain the consistency in creating the required large number of "same complexity" questions for the students' tests.

Programming is a compulsory subject in every computer science educational curriculum, and thus, usually lots of computer science students enroll in these courses. This means that vast majority of programming teachers have to deal with the problem of consistency of question complexity mentioned above. A possible solution to this problem is to provide a means of automatic production of questions. In the past ten years there has been a significant research on the issue of automatic production of questions of good quality for educative assessment needs.

Our wider research is concentrated on the automatic production of questions containing a source code or a chunk of a source code. In order to achieve the desired complexity consistency in the process of automatic production of questions for programming courses, we must have a way to automatically measure the complexity ('weight') of source codes.

In this paper we present a source code metric acceptable for the previously mentioned purpose (calculation of the complexity of a given code), and a new tool that employs the metric to produce new source codes.

The paper proceeds as follows. In Section II, some common software metrics typically used to measure the complexity of a given source code and their drawbacks are described. In Section III we present our new proposed metric. Section IV

presents the tool, and Section V gives the results of a case study of the tool. The conclusion and remarks on the future work are given in Section VI.

II. Related Work

In this section we will describe some common software metrics typically used to measure the complexity of a given source code. The examples will include the Halstead Complexity [2] and the Cyclomatic Complexity (McCabe's Complexity) [3]. We will discuss the advantages of using each of these metrics, as well as their respective drawbacks, and we will consider the possibilities for their application in the domain of teaching programming.

A. The Halstead Complexity

The Halstead complexity metrics [2] are among the oldest measures of source code complexity. They were introduced in 1977 by Maurice Halstead, as a principal attempt to quantitatively estimate the effort of the programmer when writing a particular program's source code. The goal of Halstead's research at that time was to identify measurable properties of software, as well as to establish the relations between them.

Halstead interprets the source code of a given program as a sequence of tokens, and classifies each of the tokens as an operator (traditional operator such as '+', '*', or '>', statement separator such as ';', or a keyword such as 'if', 'return' or 'continue') or an operand (literal expression, constant or variable). The four basic metrics defined by Halstead are the following:

- n_1 – number of unique (distinct) operators
- n_2 – number of unique (distinct) operands
- N_1 – total number of operators
- N_2 – total number of operands

All the basic metrics are calculated by counting the frequencies of each of the operator and operand tokens in the program's source code. The other Halstead metrics are derived from them as explained below:

- Program vocabulary (n) – it is defined as the sum of the number of distinct operators and the number of distinct operands, i.e. $n = n_1 + n_2$.
- Program length (N) – it is defined as the sum of the total number of operators and the total number of operands, i.e. $N = N_1 + N_2$.
- Calculated program length ($\hat{N}$) – this metric provides a way of measuring the relationship between the program length N and the program vocabulary n. It is given by $\hat{N} = n_1 \cdot \log_2(n_1) + n_2 \cdot \log_2(n_2)$.
- Program volume (V) – this metric describes the size of the implementation of a given algorithm, expressed in mathematical bits. It can be calculated as the program length times the logarithm (base 2) of the size of the program vocabulary, i.e. $V = N \cdot \log_2(n)$.
- Difficulty (D) – this metric is also known as error proneness. According to Halstead, the level of difficulty of a program (or its error proneness) is proportional to the number of unique operators, as well as to the ratio between the total number of operands and the number of unique operands, i.e. $D = \frac{n_1}{2} \cdot \frac{N_2}{n_2}$. This means that if we use the same operand(s) many times in our program, it will be more prone to errors. The metric also suggests that sources of program difficulty are repetition of operands and introduction of new operators in the program.
- Program level (L) – it is defined as the inverse of the difficulty level of the program, i.e. $L = \frac{1}{D}$. This means that a high level program is less prone to errors than a low level program.
- Effort (E) – this metric refers to the effort required to implement or to understand a program. Halstead suggests that the effort is proportional to the level of difficulty and the volume of the program: $E = V \cdot D$.
- Time (T) – it refers to the actual coding time, i.e. the time required to implement or to understand a program, expressed in seconds. As expected, this time is proportional to the effort required to write the program. Halstead has experimentally found that a good approximation for the time can be obtained by dividing the effort by 18 ($T = \frac{E}{18}$), but further experiments may be conducted to calibrate the measure.

Some of the advantages of the Halstead's complexity metrics are the facts that they are simple to calculate, applicable to any programming language, and that they do not require in-depth analysis of the program's structure. These metrics measure the overall quality of programs and can predict the maintenance effort. Many studies support the use of the Halstead's metrics in prediction of the programming effort, as well as of the number of programming errors.

B. The McCabe's Complexity

The McCabe's complexity [3] (also known as cyclomatic complexity or program complexity) is one of the most widely accepted software metrics, and undoubtedly – the most widely used static software metric. It was developed in 1976 by Thomas J. McCabe. This metric measures the number of linearly independent execution paths through a program's source code. For example, if the source code under consideration does not contain decision points (such as if

statements or for/while loops), its complexity will be 1, since there exists only a single path through this code. On the other hand, if the source code contains a single decision point, then there will be two paths through the code: one path where the condition corresponding to the decision point evaluates as logically true and the other one where the condition evaluates as logically false.

Formally, the cyclomatic complexity (M) of a structured program is defined by $M = E - N + 2 \cdot P$, where E is the number of edges, N is the number of vertices and P is the number of connected components in the control flow graph of the program. For a single program, P always equals 1, so the formula becomes $M = E - N + 2$.

McCabe proved that the cyclomatic complexity of any structured program with a single entrance point and a single exit point is equal to the number of decision points plus one. However, we must note that this applies only to decision points on the lowest level (machine-level instructions). When writing programs in high-level languages, programmers often use compound conditions and these decision points should be counted in terms of the predicate variables involved in the compound condition (for example, 'if condition1 and condition2 then …' should be counted as two decision points, since it is equivalent to 'if condition1 then if condition2 then …' at machine level). For programs with more than one exit point, the cyclomatic complexity can be calculated as $d - e + 2$, where d is the number of decision points and e is the number of exit points.

One of the most important advantages of the cyclomatic complexity metric is that it can be used to guide the process of dynamic testing of the functionality of the programs (using test cases). Because the cyclomatic number describes the control flow complexity, it is clear that programs with high cyclomatic number need more test cases than programs with low cyclomatic number. Just like the Halstead's complexity metrics, cyclomatic complexity is also easy applicable to any programming language, but it can be computed earlier in the life cycle of the program than the Halstead's metrics (the program doesn't have to be completed to be able to calculate its complexity value).

Both presented metrics, as well as the others that can be found in the literature, rely on the complete source code, and not on the 'actually visited source code' based on the known values of the program variables. Because this is an important issue, we decided to propose a new metric. It takes into consideration the fact that sometimes, even with a very complex code, the student may simply calculate the output of the source code, if it doesn't depend on large portions of the code. The metric is presented in the following section.

III. Our Source Code Complexity Metric

In this section we define a new metric that can be used to measure the complexity of a program's source code, written in the C++ programming language. The same metric can be extended for usage with source codes written in any programming language.

In our approach, we assume that all of the branch statements of the C++ language (if, while, do-while and for) and the most commonly used C++ operators (the arithmetic operators: +, –, *, / and %; the relational operators: <, >, <=, >=, != and ==; the logical operators: !, && and ||; and the remaining binary operators, such as the assignment operators =, +=, etc.) are assigned a specific weight value. Each of these weight values represents the effort required (from a human) to perform the corresponding operation or execute the corresponding branch statement manually. If the weights of all the operators and branch statements are known, we define the complexity C of a given C++ source code using the following equation:

$$C = \sum_{i=1}^{n} w_i \cdot e_i \tag{1}$$

where n is the number of lines in the source code, w_i is the weight assigned to line i, and e_i is the number of executions of line i in a single execution of the source code. The weight assigned to a line is the sum of the weights of all the operators and branch statements present in that line.

This metric should, more precisely, calculate the complexity of the source code from perspective of the student's effort to calculate the output of the code. In the next section, we will present our tool that employs this metric.

IV. Our Tool for Calculation of the Proposed Source Code Complexity Metric

For the purpose of calculating our proposed complexity metric for a given source code discussed in the previous section, and to provide a means of automatic production of similar complexity source codes on the basis of the initial code, we have created an appropriate software tool. The tool represents a Java web application that can be accessed using a web browser. Currently, it works only with programs written in the C++ programming language, and it supports Windows and Linux platforms.

The following technologies and libraries were used in the development process of the tool:

- Maven – a software project management and comprehension tool based on the Project Object Model (POM). It can manage a project's life cycle from a centralized XML file [4];
- Eclipse CDT (C/C++ Development Tooling) API [5];
- FreeMarker – a "template engine"; a generic tool to generate text output (anything from HTML to auto generated source code) based on templates [6];
- Gcov – a tool used in conjunction with GCC to test coverage of programs [7];
- Java Server Faces – Java based framework which implements Model View Controller (MVC) design pattern in a stateful manner. The usage of this

framework allowed well refining of the application layers as well as tracking of beans state [8];

- Spring – used for inversion of control via dependency injection and bean life cycle manipulation [9].

Our tool represents an extension of a tool that uses an initial code to generate a user-specified number of codes by altering literal values and/or operators in it. With the help of the new tool, the newly produced codes can have complexities that do not exceed the complexity of the initial code plus a threshold value. We will refer to the new codes generated by the tool in this way from an initial code as code variations of the initial code.

The combined tool consists of two main parts: 1) Uploading and editing of an initial source code, and generation of code variations; 2) Configuration of the weights associated to each of the operators/statements.

In the first part of the tool, accessed via its home page, there is a wizard that guides the user through the process of generation of code variations. A view of the first step of this process is shown in Fig. 1. In this step, the user can input the initial code – the source code of the program for which he/she wants to generate variations. As can be seen from Fig. 1, in the upper left corner there is a text box in which the user can enter the number of code variations to be generated, and in the upper right corner there is a check box which enables the user to select whether he/she wants modification of the operators in the different code variations. If this check box is not selected, the operators will remain unchanged in all of the generated code variations and will be exactly the same as those in the initial code.

To proceed to the second step of the process, the user has to click on the button labeled 'Next' in the bottom right corner of the page. The wizard will allow this only if no errors have been made in the first step, and otherwise it will show an appropriate error message. Possible error messages at this point are: 'missing value', 'improper value' or 'compiler error'. Codes that cannot be compiled and executed are rejected.

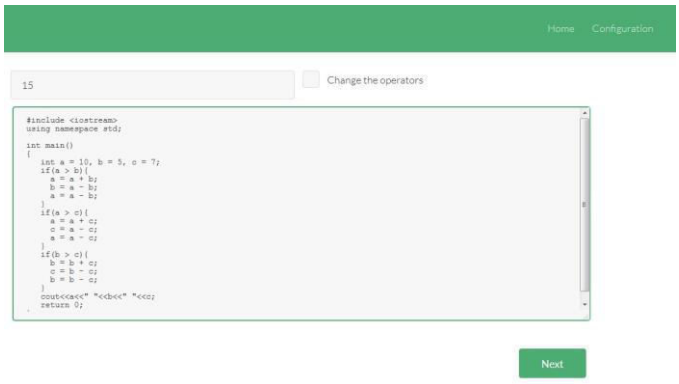


Fig. 1. A view of the first step of the process of generation of code variations.

The second step of the process of generation of code variations enables configuration of the domain for the values of the locations of interest in the code (Fig. 2). Locations of interest in a given code are the positions in the code where literal constants (numerical or non-numerical) are present. For numerical locations of interest, the user can configure the range of values from which each of the locations can be filled. Furthermore, the user can explicitly specify a set of values that should be excluded from this range, i.e. which must not appear in the particular location of interest in any of the generated code variations. If the values entered are floating point numbers, then the generated values will contain at most a single digit after the decimal point. For non-numerical locations of interest, the user can configure the character set from which the locations will be filled. The allowed character set can be configured to include (or exclude) digits, lowercase letters, uppercase letters and special characters, by selecting (or leaving unselected) the corresponding check box. Both numerical and non-numerical locations of interest can be left unchanged (by selecting the check box 'Do not change value'), which means that they will keep their default values – the values present at the same locations in the initial code.

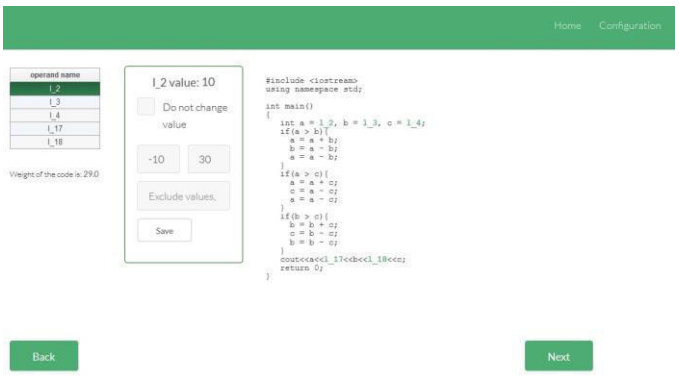


Fig. 2. A view of the second step of the process of generation of code variations.

As shown in Fig. 2, at this moment the user can see the calculated complexity (weight) for the initial code under consideration. The calculated complexity value represents a referent value for the process of generation of code variations. This means that all the code variations that will be generated at the end of the process will have a complexity not greater than the sum of the initial code's complexity value and a predefined threshold. Generated codes with complexities greater than this sum are discarded.

Clicking the 'Next' button in the second step of the generation process starts the actual generation of the code variations. The time required to complete the generation depends on the number of codes that will be generated. After the completion of the codes generation, the wizard brings the user to the third and final step, where he/she can see the results (Fig. 3). The page shows the number of generated codes, and presents the first of them, together with its output and the calculated complexity value. The user can then browse through the other generated codes (by clicking on the left/right arrows above the area where the codes are shown) to see their respective outputs and complexity values.

The second part of the tool provides an interface for configuring the values of the weights that are used in the process of calculation of the complexity of a given code. It can be accessed by clicking on the Configuration hyperlink from

the green menu in the upper right corner at any step of the generation process. The interface provides a tabular view of all these weights (Fig. 4), where each row corresponds to a single weight assigned to a particular operator or statement. Each weight has a default value, as can be seen from Fig. 4. The weight values can be easily changed by entering new values in the appropriate fields in the second column of the table. Here, the user can also modify the threshold value by supplying an appropriate value in the last row of the table. In this way, he/she can control the allowed deviation of the complexities of the generated code variations from the initial code's complexity.

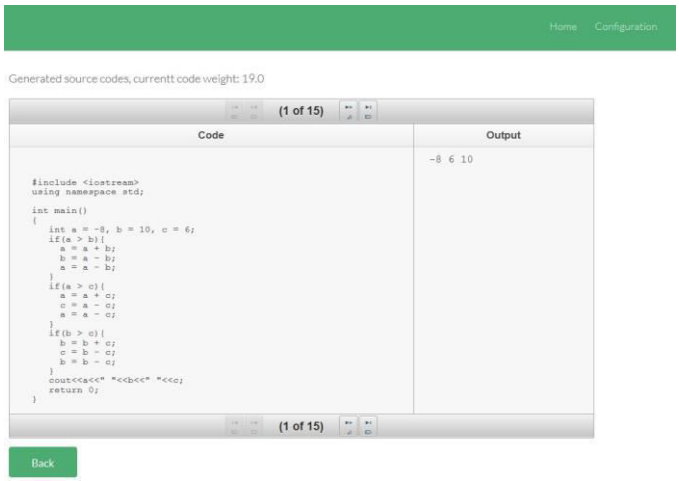


Fig. 3. A view of the third and final step of the process of generation of code variations.

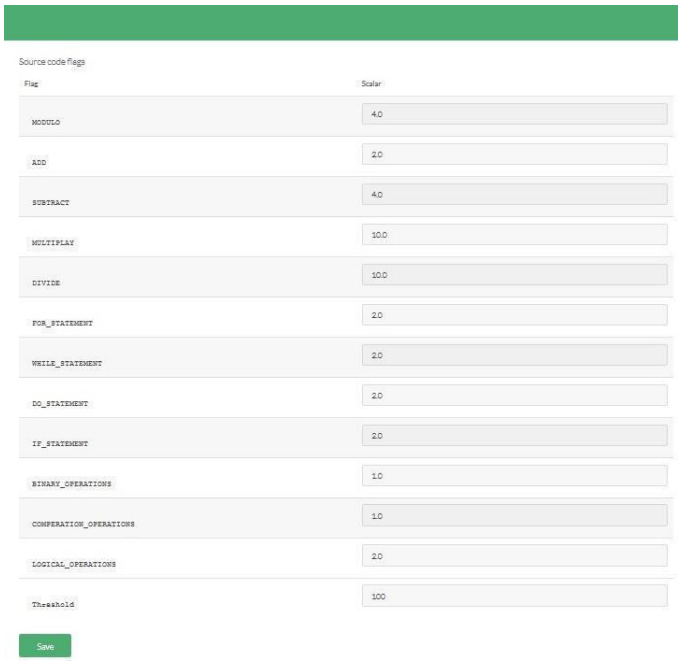


Fig. 4. Configuration of the weight values and the threshold value to be used in the calculation of the complexity of each code.

V. A Case Study

In this section we will describe a case study of an application of our tool on a program's source code in order to calculate its complexity and generate code variations with complexities that have a desired deviation. The example C++ source code is shown in Fig. 5.

```
#include <iostream>
using namespace std;

int main()
{
   int a = 10, b = 5, c = 7;
   if(a > b){
     a = a + b;
     b = a - b;
     a = a - b;
   }
   if(a > c){
     a = a + c;
     c = a - c;
     a = a - c;
   }
   if(b > c){
     b = b + c;
     c = b - c;
     b = b - c;
   }
   cout<<a<<" "<<b<<" "<<c;
   return 0;
}
```

Fig. 5. An example source code for the case study.

Given this initial code, let's suppose that we need to generate 15 different code variations, but without changing any operators in the code. Specifying these input parameters to the tool brings us to the second step of the generation process, as described in the previous section. Here we can specify domains for the locations of interest in the code. We have three numerical and two non-numerical locations of interest in this case, which correspond to the numerical (10, 5 and 7) and the non-numerical literals (the two strings in the 'cout' statement, each of them containing a single white-space character) present in the code. We don't want to modify the non-numerical literals, so we select the corresponding 'Do not change value' check box for each of them. For the numerical literals, we specify the following ranges of integer values as domains:

- [–10, 30] for the value of the location of interest corresponding to the variable *a*,
- [–20, 20] for the value of the location of interest corresponding to the variable *b*, and
- [–30, 10] for the value of the location of interest corresponding to the variable *c*.

We will leave the predefined weight values (as shown in Fig. 4) for the operators and statements unchanged. The calculated complexity of our initial source code with these settings is 29. We will set the threshold value to 20, so that we

don't get code variations with complexities that exceed the value 49.

The results obtained in the final step of the generation process are shown in Table I. As we can see, the desired number of 15 codes have been generated which means that there hasn't been a rejection of a generated code (due to a too large complexity). The minimum complexity value of a generated code is 9, and the maximum complexity is 29, so only codes with complexities that are less or equal to the initial code's complexity have been generated.

TABLE I. STATISTICS FOR THE GENERATED CODE VARIATIONS

Observed statistic	Value
Minimum complexity value	9
Maximum complexity value	29
Average complexity value	24.33
Average deviation from initial code's complexity	6.22
Number of generated codes	15
Number of codes with same complexity as the initial code	10
Number of codes with larger complexity than the initial code	0
Number of codes with smaller complexity than the initial code	5

The results show that even when the initial code is fairly simple (as the one observed in this case study) and the threshold value is set to be relatively small, we may obtain code variations with complexities that may significantly differ from the initial code's complexity. This proves that it is essential to have a tool for checking the complexities of the generated code variations.

VI. CONCLUSION AND FUTURE WORK

In this paper we described the need for maintenance of consistency of question complexity that appears in the assessment process in the domain of teaching programming, which is especially difficult when working with large course classes. Teachers have to produce questions with same or similar complexity for the students' tests, in order to achieve objective and fair assessment of the students' knowledge. A possible and already widely employed solution to this problem is to use automatic production of questions. However, in order to achieve complexity consistency in the process of automatic production of questions that contain source code fragments, which are commonly used in programming course exams, we must have a way to automatically measure the complexity ('weight') of source codes.

Further in the paper we described the software metrics that are most commonly used to measure the complexity of a given program's source code. We explained their advantages and drawbacks, with emphasis on the reason why they're not suitable to be used for the problem under consideration. Next, we proposed a new metric that considers the source code complexity from a perspective of the student's effort required to manually calculate the output of the program (if the input is known), and thus, is well suited for the problem. The metric measures the complexity using user-specified weight values assigned to each of the operators and branch statements in the code. We also described our new tool that employs this metric to calculate the complexity of an initial source code, and generate a desired number of new source codes (code variations) with same or close enough complexity (using a user-defined threshold value to control the complexity deviation).

For our future work, we plan to conduct an extensive research in order to determine weight values that will accurately represent the students' effort required to manually perform the operations and execute the statements in a given source code.

ACKNOWLEDGMENT

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] B.S. Bloom, M.D. Engelhart, E.J. Furst, W.H. Hill, and D.R. Krathwohl, Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain. New York: David McKay Company, 1956.

[2] M.H. Halstead, Elements of Software Science. New York: Elsevier North-Holland, 1977.

[3] T.J. McCabe, "A complexity measure", IEEE Transactions on Software Engineering, vol. 2, no. 4, pp. 308-320, December 1976.

[4] Apache Maven [Online]. Available: https://maven.apache.org/

[5] D. Piatov, A. Janes, A. Sillitti, and G. Succi, "Using the Eclipse C/C++ Development Tooling as a robust, fully functional, actively maintained, open source C++ parser," IFIP Advances in Information and Communication Technology, pp. 399, 2012.

[6] FreeMarker [Online]. Available: http://freemarker.org/index.html

[7] "Introduction to Gcov," GCC online documentation [Online]. Available: https://gcc.gnu.org/onlinedocs/gcc/Gcov.html#Gcov

[8] E. Burns et al., "JSR 344: JavaServerTM Faces 2.2", Java Community Process [Online]. Available: https://www.jcp.org/en/jsr/detail?id=344

[9] Spring framework [Online]. Available: http://projects.spring.io/spring-framework

Effective polynomial and non-polynomial Schrödinger equations for the 0 K dynamics of Bose-Einstein condensates

Alexandru Nicolin and Mihaela Carina Raportaru
Department of Computational Physics and Information Technologies
Horia Hulubei National Institute for Physics and Nuclear Engineering (IFIN-HH)
Magurele, Romania
alexandru.nicolin@nipne.ro, mraportaru@nipne.ro

***Abstract* — In this article we review a series of previous works on the effective equations typically used to describe the 0 K dynamics of Bose-Einstein condensates and distinguish between two classes of equations: those with polynomial nonlinearities and those with non-polynomial nonlinearities. We present the main features of these two classes of equations and emphasize the importance of the symbolic computing platforms used to determine these effective equations.**

Keywords — polynomial and non-polynomial equations, Gross-Pitaevskii, Bose-Einstein condensates

I. INTRODUCTION

The first Bose-Einstein condensation of atomic species which took place simultaneously in 1995 in the groups of Eric Cornell and Carl Wieman at the University of Colorado in Boulder, Wolfgang Ketterle at the Massachusetts Institute of Technology, and that of Randall G. Hulet at Rice University in Houston was a landmarking experimental achievement that opened up an almost unprecedented series of both theoretical and experimental investigations (see the textbook [1] for a survey of the most active directions of research). In fact, atomic Bose-Einstein condensates have turned into the testbed of choice for many of the research groups interested in nonlinear phenomena in quantum systems [2]. While the theoretical predication of the Bose-Einstein condensation precede the experimental observation by seven decades, the long list of engineering developments in the trapping and subsequent cooling of atoms allow one to realize experimentally atomic condensates which are both robust in terms of stability to parasitic excitations and very tunable with respect to the properties of the condensate (e.g., the two-body scattering length) and those of the experimental setup (e.g., the values of the frequencies of the underlying magnetic trap). The early investigations were focused on (single-species and vector) condensates with binary short-range interactions for which there were numerous results concerning the properties of solitons and vortices, and then gradually shifted towards the effects of three-body short-range interactions, condensates with competing long-range dipolar interactions, the role of thermal and quantum fluctuations, fermionic ultra-cold gases, bosonic-fermionic hybrid systems, the BEC-BCS crossover in fermion gases and the condensation of quasi-particles in solid-state settings (exciton-polaritons, magnons, and cavity photons), quantum phase transitions and quantum turbulence and many others [2]. For a review of some of the research fields developed over the past two decades since the first experimental realization of Bose-Einstein condensation we refer the interested readers to the Special Issue of Romanian Reports in Physics which celebrates this outstanding scientific achievement, in particular the introductory review which also serves as a Resource Letter of the main research directions [3].

In addition to their experimental maneuverability atomic condensates have also been appealing due to existence of an accurate theoretical formalism that permitted both the theoretical reinforcement of the early experimental results as well as the prediction of numerous. The main equation that describes the O K dynamics of bosonic condensates is the Gross-Pitaevskii equation (GPE) that can account for both short- and long-range interactions, namely:

$$i\hbar\frac{\partial}{\partial t}\psi(\boldsymbol{r},t)=\left[-\frac{\hbar^2}{2m}\nabla^2+V(\boldsymbol{r})+UN|\psi(\boldsymbol{r},t)|^2\right]\psi(\boldsymbol{r},t) \quad (1)$$

For condensates with short-range interactions the GPE takes the form of a standard partial differential equation with a polynomial nonlinearity for which there are numerous numerical treatments. The most popular ones are the Fortran [4,5] and C packages [6] that can describe the stationary and non-stationary solution of the GPE for one-, two- and three-dimensional condensates loaded in axially-symmetric traps using real- and imaginary-time propagation based on a split-step Crank-Nicolson method. The C codes are particularly appealing due to the OpenMP parallelization that allows a substantial computational speed-up even on a standard multi-

core desktop computer. Other methods of choice include the pseudo-spectral and finite-difference methods. Long-range interactions, however, a more problematic as the equation that describes dipolar condensate contains both differential and integral terms, namely the dipolar Gross-Pitaevskii equation

$$i\frac{\psi(\boldsymbol{r},t)}{\partial t}=\left[-\frac{1}{2}\nabla^2+V(\boldsymbol{r})+4\pi aN|\psi(\boldsymbol{r},t)|^2\right.$$
$$\left.+N\int U_{dd}(\boldsymbol{r}-\boldsymbol{r}')\,|\psi(\boldsymbol{r}',t)|^2 d^3\boldsymbol{r}'\right]\psi(\boldsymbol{r},t) \qquad (2)$$

where the dipolar interaction is given by $U_{dd}(\boldsymbol{R})=3a_{dd}(1-3\cos^2\theta)/R^3$ with $\boldsymbol{R}=\boldsymbol{r}-\boldsymbol{r}'$. Here $V(\boldsymbol{r})$ is the confinig axially symmetric harmonic potential, $\psi(\boldsymbol{r},t)$ the wave function at time t, with the normalization condition $\int|\psi(\boldsymbol{r},t)|^2d\boldsymbol{r}=1$, a the atomic scattering length, θ the angle between $\boldsymbol{R}$ and the polarization direction z. The constant $a_{dd}=\mu_0\bar{\mu}^2m/(12\pi\hbar^2)$ is a length characterizing the strength of dipolar interaction. We note here that there is no standardized numerical treatment available. Outside of full numerical treatments there are also numerous variational and hydrodynamic approaches that give a good description of the static and dynamical properties of condensates.

One particularly efficient method used to describe the dynamics of elongated and oblate fully three-dimensional condensates relies on reducing the effective dimensionality of the condensate by variational means such that the resulting equation carries a smaller computational load. In a nutshell, for highly elongated (or oblate) condensate with short-range interactions a suitable *Ansatz* for the radial (or longitudinal) component of the wave function can be used to integrate out the radial (or longitudinal) dynamics in the Lagrangian description of the condensate. The Gaussian and the q-Gaussian functions are two of the most common candidates that offer a good trade-off between the analytical tractability of the *Ansatz* and the accuracy of the resulting equation of reduced dimensionality. The most popular equations are those introduced by Salasnich and co-workers in Ref. [7], namely

$$i\hbar\frac{\partial}{\partial t}f=\left[-\frac{\hbar}{2m}\frac{\partial^2}{\partial z^2}+V+\frac{gN}{2\pi a_\perp^2}\frac{|f|^2}{\sqrt{1+2a_sN|f|^2}}\right]$$
$$+\frac{\hbar\omega_\perp}{2}\left(\frac{1}{\sqrt{1+2a_sN|f|^2}}+\sqrt{1+2a_sN|f|^2}\right)f \qquad (3)$$

for a cigar-shaped condensate and

$$i\hbar\frac{\partial}{\partial t}\psi=\left[-\frac{\hbar}{2m}\nabla_\perp^2+W+gN\frac{\eta^{-1}}{(2\pi)^{1/2}}|\psi|^2\right.$$
$$\left.+\left(\frac{\hbar^2}{2m}\eta^{-2}+\frac{m\omega_z^2}{2}\eta^2\right)\right]\psi \qquad (4)$$
$$\frac{\hbar^2}{2m}\eta^{-3}-\frac{1}{2}m\omega_z^2\eta+gN\frac{\eta^{-2}}{2(2\pi)^{1/2}}\left||\psi|\right|^2=0 \qquad (5)$$

for a pancake-shaped condensate, commonly referred to as NPSE (an abbreviation for non-polynomial Schrödinger equations), which have proven useful in describing the dynamics of solitons and vortices, the emergence of parametrically-excited density waves, etc. For a q-Gaussian *Ansatz* are obtains the following equation:

$$i\hbar\frac{\partial f(z,t)}{\partial t}=\left\{-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial z^2}\right.$$
$$\left.+2\hbar\omega_\perp\left[\sqrt{a_s|f(z,t)|^2N}-\frac{2^{1/3}}{3}(a_s|f(z,t)|^2N)^{1/6}\right]\right\}f(z,t) \qquad (6)$$

which was first introduced in Ref. [8]. A similar equation (though obtained through a different reasoning) was introduced in Ref. [9]. We mention on the technical side that, unlike the equations determined for Gaussian *Ansatze*, equations such as the previous one cannot be determined using only paper-and-pencil calculations and one has to employ a computer algebra system. MATHEMATICA seems to be one of the most popular one due to its very large library of functions (with emphasis on special functions) and wide set of analytic operators. Please note that the method of obtaining these equations is quite classical and that the main difficulty comes from the analytic manipulation of the Lagrangian which requires a specialized program such as Maple, MATLAB (in particular the Symbolic Math Toolbox), Octave, Python (in particular the SymPy library), etc. (see Ref. [10]). On the side of purely numerical computing we stress that all such equations allow the reduction of the dimensionality at the cost of a non-polynomial nonlinearity that usually increases the computational load. The best solution to go around this computational slow-down is to integrate the radial (or the longitudinal) component of the wave function directly on the GPE such that the resulting equation (of reduced dimensionality) preserves the structure of the nonlinear term. This task is straightforward but for the final equation to be accurate one requires, however, precise knowledge of the structure of the radial (of longitudinal) component of the wave function. As we will show in this paper, the best candidate for the job is the q-Gaussian function that captures accurately the static properties of the condensate and is maneuverable enough to allow the straightforward integration of the radial (or longitudinal) component of the wave function.

II. SCHRÖDINGER POLYNOMIAL EQUATIONS

The effective equations or reduced dimensionality detailed in the first section of this article come at the cost of a significantly increased computing time. This is due to their non-polynomial structure that increases the computing cost of evaluating the nonlinear term by almost one order of magnitude, if one is to make a direct comparison to the standard cubic nonlinearity. Even more, some of the effective low-dimensional equations which stem from variational recipes based on q-Gaussian *Ansatze*, in particular the two-dimensional equations which are not shown above, usually involve special functions which add an extra load to the computing time. Naturally, the numerical solution of NPSE is less time-consuming than that of the fully three-dimensional GPE, if one is to restrict the analysis at the number of operations and the complexity of the numerical method, but for detailed numerical investigations their efficiency is insufficient to address some of the current research problems such as the dynamics of vortices in pancake-shaped condensates [11].

One solution to this problem – applicable to effectively two-dimensional condensates – is to integrate directly the transverse component of the wave function in the GPE and solve numerically the resulting partial-differential equation. For this one needs a relatively simple analytical approximation for the transverse component of the wave function which is amenable to symbolic integration. The Gaussians are the ideal choice in terms tractability of the symbolic integration, but they fail to describe quantitatively high-density condensates. The previously mentioned q-Gaussian functions can be used to construct *Ansatze* which are exact solutions to the GPE in both the low- and the high-density regime, but the variational parameters come only in numerical form from the solution of a set of nonlinear algebraic equations. The ground state of a pancake-like condensate is well described by

$$\psi(\rho,r,t)=A\left(q_\rho,q_z,\omega_p,\omega_z\right)\left(1-\frac{\rho^2\left(1-q_\rho\right)}{2\omega_\rho^2}\right)^{\frac{1}{1-q_\rho}}\times\left(1-\frac{z^2(1-q_z)}{2\omega_z^2}\right)^{\frac{1}{1-q_z}} \quad (7)$$

where A is chosen such that the wave function is normalized to unity and the variational parameters stem from the numerical solution of

$$\begin{aligned}
&5\sqrt{2}gmN(q_r-2)^2(q_r-2)(q_r+1)(25q_z-126)\\
&\quad+3528\pi B\omega_z(m^2(q_r+1)\omega_r^4\Omega_r^4\\
&\quad-(q_r-3)(q_r-2)\hbar^2)=0\\
&5\sqrt{2}gmN(q_r-3)^2(1+q_z)(3q_z-7)(25q_z-126)\omega_z\\
&\quad+1764\pi B\omega_r^2(8m^2(1+q_z)\omega_z^4\Omega_z^2\\
&\quad-(q_z-5)(3q_z-7)\hbar^2)=0\\
&5\sqrt{2}gmN(q_r-7)(q_r-3)(2+q_r-q_r^2)^2(25q_r-126)\\
&\quad+3528\pi B^2\omega_z(m^2(1+q_r)^2\omega_r^4\Omega_r^2\\
&\quad-4(q_r-2)^2\hbar^2)=0\\
&125\sqrt{2}gmN(q_r-3)^2(7+(4-3q_z)q_z)^2\omega_z\\
&\quad+5292\pi B\omega_r^2(4m^2(1+q_z)^2\omega_z^4\Omega_z^4\\
&\quad-(7-3q_z)^2\hbar^2)=0
\end{aligned}$$

where $B=q_r-5$. Following some lengthy algebra (detailed in Ref. [12]) we have that the effectively two-dimensional equation which describes the condensate is given by

$$\begin{aligned}
i\frac{\sqrt{3}}{4}\pi\sqrt{\omega_z}\frac{\partial f}{\partial t}=&-\frac{\sqrt{3}\pi\sqrt{\omega_z}\hbar^2}{8m}\left(\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}\right)\\
&+\frac{\sqrt{3}m\pi\sqrt{\omega_z}\Omega_r^2}{8}(x^2+y^2)f\\
&+\frac{\sqrt{3}m\pi\omega_z^{5/2}}{32}\Omega_z^2 f+\frac{9\sqrt{3}gN\pi}{64\sqrt{\omega_z}}f^3 \quad (8)
\end{aligned}$$

which is the main result of the paper. Our equation does not account for the interplay between the radial and the transversal modes of the condensate, but describes accurately the transversal dynamics through its rescaled effective nonlinearity.

III. NUMERICAL RESULTS

For a realistic ^{87}Rb condensate the aforementioned NPSE equations describe very well (both qualitatively and quantitatively) a wide range of nonlinear phenomena, such as the emergence of pattern-forming dynamical instabilities and have been used with great success to describe the appearance of Faraday waves in cigar-shaped condensates. Following Refs. [13, 14], we depict in Fig. 1 the emergence of such a parametrically excited density waves by monitoring in time the radially integrated density profile of the condensate. Complementary, while in Figs. 2, 3 and 4 we show the actual density profile of the condensate at three different times, such that we can see the emergence of the density wave. We have no standard numerical implementation for the polynomial equations presented in the previous section but the preliminary results suggest that they can describe efficiently the dynamics of effectively longitudinal (transverse) nonlinear waveforms through the rescaled effective nonlinearity in the case of cigar-shaped (pancake-shaped) condensates.

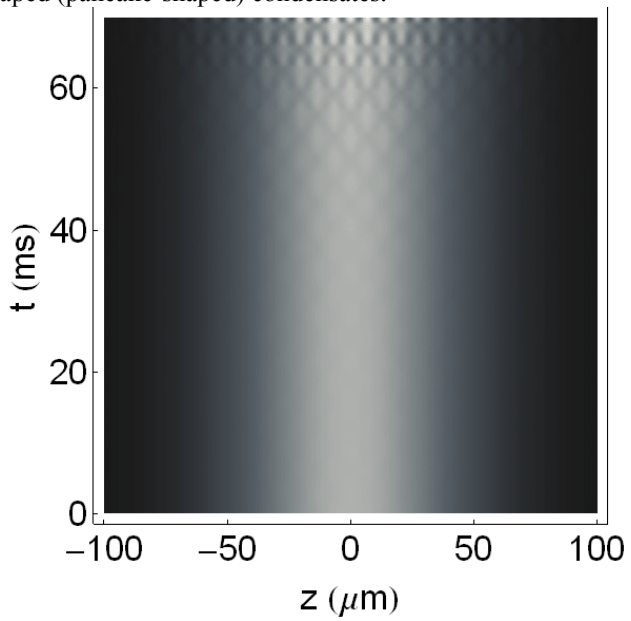


Fig. 1 The emergence of Faraday density waves in a cigar-shaped Bose-Einstein condensate. The plot depicts the time-evolution of the of radially-integrated density profile of the condensate. (Reproduced after Ref. [13].)

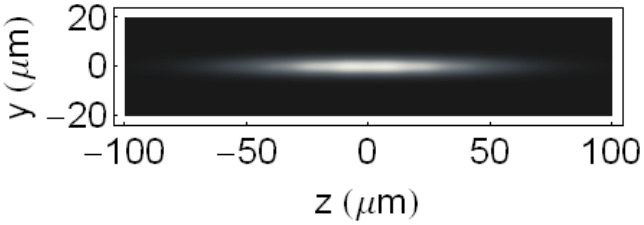


Fig. 2 The density profile of a cigar-shaped Bose-Einstein condensate before the onset of a density wave. (Reproduced after Ref. [14].)

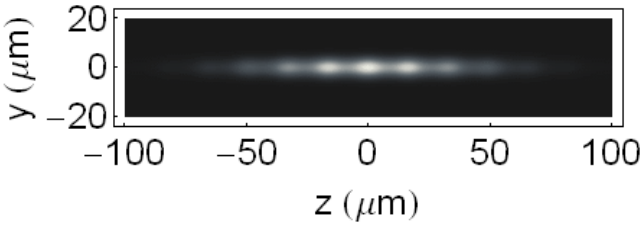


Fig. 3 The density profile of a cigar-shaped Bose-Einstein condensate at the onset of a density wave. (Reproduced after Ref. [14].)

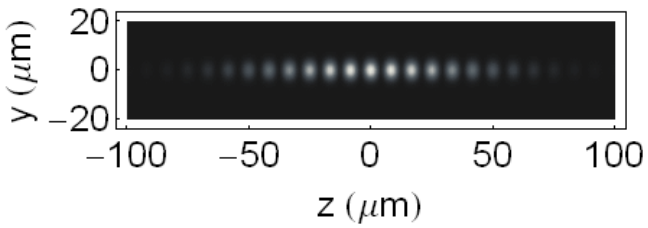


Fig. 4 The density profile of a cigar-shaped Bose-Einstein condensate after the onset of a density wave. (Reproduced after Ref. [14].)

IV. CONCLUSIONS

We have reviewed a series of results on the effective equations typically used to describe the 0 K dynamics of Bose-Einstein condensates to separate between the properties of the two classes of equations: those with polynomial nonlinearities and those with non-polynomial nonlinearities. Most of the existing investigations have relied on NPSEs but the simplicity of the polynomial equations suggests that they will become quite popular in the future.

ACKNOWLEDGMENT

For this work A.I.N. was supported by a grant of the Romanian Ministry of Education, CNCS-UEFISCDI, project PN-II-RU-PD-2012-3-0154.

REFERENCES

[1] C.J. Pethick and H. Smith, Bose-Einstein condensation in dilute gases, Cambridge: Cambridge University Press, 2008.

[2] P.G.Kevrekidis, D. J. Frantzeskakis, and R. Carretero-González, Eds., Emergent nonlinear phenomena in Bose-Einstein condensates, Berlin: Springer-Verlag, 2008.

[3] V.S. Bagnato *et al.*, "Bose-Einstein condensation: Twenty years after," Rom. Rep. Phys., vol 67, 5-50, 2015.

[4] P. Muruganandam and S.K. Adhikari, "Bose-Einstein condensation dynamics in three dimensions by the pseudospectral and finite-differece methods," J. Phys. B: At. Mol. Opt. Phys., vol 36, 2501, 2003.

[5] P. Muruganandam and S.K. Adhikari, "Fortran programs for the time-dependent Gross-Pitaevskii equation in a fully anisotropic trap," Comp. Phys. Comm., vol. 180, pp. 1888-1912, 2009.

[6] D. Vudragović, I. Vidanović, A. Balaž, P. Muruganandam and S.K. Adhikari, "C programs for solving the time-dependent Gross-Pitaevskii equation in a fully anisotropic trap," Comp. Phys. Comm., vol 183, 2021-2025, 2012.

[7] L. Salasnich, A. Parola, and L. Reatto, "Effective wave equations for the dynamics of cigar-shaped and disk-shaped Bose condensates," Phys. Rev. A, vol 65, 043614, 2002.

[8] A.I. Nicolin and M.C. Raportaru, "Faraday waves in high-density cigar-shaped Bose-Einstein condensates," Physica A, vol. 389, 4663, 2010.

[9] A. Muñoz Mateo and V. Delgado, "Effective mean-field equations for cigar-shaped and disk-shaped Bose-Einstein condensates," Phys. Rev. A, vol. 77, 013617, 2008.

[10] M.C. Raportaru and A.I. Nicolin, "Nonlinear dynamics of Bose-Einstein condensates by means of symbolic computations," ICT Innovations 2012 Web Proceedings ISSN 1857-7288, 9-15.

[11] R. Navarro et al., "Dynamics of a few cocirculating vortices in Bose-Einstein condensates," Phys. Rev. Lett., vol. 110, 225301, 2013.

[12] A.I. Nicolin, M.C. Raportaru, "Effective low-dimensional polynomial equations for Bose-Einstein condensates," Rom. Rep. Phys., vol. 67, 143-147, 2015

[13] A.I. Nicolin, "Resonant wave formation in Bose-Einstein condensates," Phys. Rev. E, vol. 84, 056202, 2011

[14] A.I. Nicolin, "Faraday waves in Bose-Einstein condensates subject to anisotropic transverse confinement," Rom. Rep. Phys., vol 63, 1329, 2011.

Code Generation for Parallel Differential Equation Solvers

Dániel Berényi, Péter Lévai
Institute for Particle and Nuclear Physics
Wigner RCP
Budapest, Hungary
berenyi.daniel@wigner.mta.hu

Sándor Varró
Institute for Solid State Physics and Optics
Wigner RCP
Budapest, Hungary

Vladimir V. Skokov
Department of Physics
Western Michigan University
Kalamazoo, MI, USA

Abstract— **Differential equations are fundamental concepts in physics and as such they are central subjects in optimizing and parallelizing high-performance scientific simulations. With the rapid changing of hardware and software environments it is challenging to keep the performance and hardware utilization of simulations at maximum without investing disproportional work in developing the necessary simulation codes. We are investigating the possibility of generating differential equation solver codes right from the mathematical definitions of the equations. Such generated codes can take advantage of the massive parallelism of GPUs with minimal input from the user while retaining as much performance as possible w.r.t handcrafted and optimized codes. We report initial results on the implementation of our proposed differential equation system solver based on the expression of code generation and higher order abstraction via Embedded Domain Specific Languages.**

Keywords— differential equation solvers, strong field pair production, GPUs, EDSLs, dynamic code generation

I. INTRODUCTION

One of the main challenges of programmers in the XXI century is to properly utilize the ever widening spectrum of computing hardware while keeping the development time at minimum. Available hardware based on traditional CPUs ranges from supercomputers through grids and clusters down to PCs. However an increasing proportion of high-performance computing is done via graphical processors (GPUs) because of their relative cheapness and advantage in certain use cases. The newest change on the market is the proliferation of embedded, portable ARM based devices that are usually cheap but provide surprising computing capacity. While these devices are not expected to drive out CPUs and GPUs from traditional scientific simulations but they are expected to infiltrate into experimental devices and data acquisition systems and may finally manifest to the developers when the prediction and measurement codes should be interfaced together.

This work was supported by the Hungarian OTKA grants No.: 106119 and No.: 104260

To access the power of GPUs, or parallelism in general, one has to choose an Application Programming Interface (API, such as CUDA or OpenCL) or a library built upon these APIs. The various constraints and compromises of these tools drastically contribute to the already high development time and cost of parallel and efficient simulations. High-Performance Computing (HPC) experts and programming language theorists already aware of these issues and propose to use code-generation [1] to give high-level access, for example in the form of Domain Specific Languages (DSLs) to end-users while maintaining performance. An active area of research is what parts of the simulation problem should be formulated at a higher-level via DSLs and what parts of the codes should be generated in which target language. We bring some recent directions from the literature. DSLs are used in [2] to specify stencils of operations that appear in the discretization of Partial Differential Equations (PDEs) and image processing and use code generation to target GPUs from C++ / OpenCL. Python is used to generate C code from text templates or syntax trees at run-time to target CUDA and OpenCL in [3]. Expressing codes or parts of it in DSLs is still not user-friendly enough. Researchers would like to give mathematical models of computations or simulations as an input. Such a formulation is investigated in [4] for PDEs arising in physics. There, equations are expressed in a separate text based language and then code-generation is used to target GPUs via C++ and CUDA. Very similar is the direction of the FENICS project that uses the Unified Form Language to formulate PDEs and Finite Element Method (FEM) solvers [5]. A similar but even greater project is Cactus/Kranc that uses Mathematica to generate C/Fortran or intermediate cactus files that combined with the Cactus framework's mathematical DSL can utilize GPU parallelism and use code generation for FEM solution of PDEs [6, 7]. Code-generation and automatic tuning is investigated in [8] in the Haskell language.

Ordinary Differential Equations are massively encountered in biophysics and neuroscience. Grouping and parallel solution of ODEs is investigated in [9] via code-generation in CUDA.

Declarative / markup specification of explicit solvers is utilized in [10] for the same purpose. Finally the Brian project generates parallel codes based from equational models embedded in Python [11].

It is easy to see the advantages of mathematical formulation for end-users. Separate languages for mathematical formulations have the disadvantage that a complete (possibly new) compiler is necessary for processing and it is harder to programmatically modify them. Formulating them in a plain text as in [4] is error prone and using an untyped language like Python is only marginally better [11].

The solvers used in the literature are usually pre-implemented [9] and hand tuned for specific cases or parametrized and automatically generated for a chosen numerical scheme [6, 7, 10]. Modifying certain parts of the solvers to meet specific use cases without losing the advantages provided by the automation is difficult. Also, when a new API / language becomes available these specialized solutions has to be ported to the new platform.

To avoid mixing unnecessarily many languages, and providing a type checked environment for formulation, manipulation and programming we propose a code-generation scheme via two embedded domain specific languages (EDSLs) in C++, one for the mathematical formulation and one for the solver program formulation, that makes higher-order / functional abstractions available in form independent of the target language. We argue that using such a strongly typed language as a host the type system and the compiler can be utilized to catch errors early, reject invalid constructs and decrease debugging times while making code reuse and generic (meta-)programming possible. The language independent formulation makes porting to new targets easier.

In this paper we describe the outline of our framework under development and discuss some of its internal details. The second section shows some typical use cases that we aim to solve. The third section discusses our internal representation via abstract syntax trees and shows how to build a complete pipeline from the mathematical representation and solver templates to the final source codes. The fourth section discusses some current results, further developments and applications.

II. Use cases and Design considerations

The authors' physics research is motivated by strong field physics where the quantum process of pair production is investigated in different physical theories. In the simplest case one can imagine extremely strong counter propagating laser fields which, when collide, create electron-positron pairs from vacuum as predicted by the theory of Quantum ElectroDynamics (QED). One mathematical formulation of the physical process is known as the Dirac-Heisenberg-Wigner (DHW) formalism that leads to a complicated partial differential equation system in 7 dimensions (time+3 spatial+3 momentum) and 16 variables [12-13]. Furthermore there is an analogue process in the strong interaction modelled by Quantum Chromodynamics (QCD) where quark-antiquark pairs are produced after energetic heavy-ion collisions (for example at the Large Hadron Collider at CERN). The same DHW construction leads to a similar, but even more complicated equation system in QCD [14-16]. The common point of these simulations is that one may want to investigate a wide range of different field configurations (models of the laser field, or the details of colliding nuclei) which require only a small, but functional/vectorial change in the program codes. One needs different derivatives of the input field function and, based on its vectorial direction, some components of the equations may become trivially zero matrices. Making these modifications and simplifications by hand each time is cumbersome and error prone.

Discussions with physicists in high energy physics, general relativity, solid state physics, environmental sciences, chaotic dynamics etc., lead to the insight that many other physically relevant problems lead to a similar situation where a simulation should be parametrized over functional dependencies or vector/algebraic quantities whose change affects the program code in far reaching but potentially automatable ways. Thus together with the Wigner GPU Lab we started to investigate the generic way to solve these problems with the additional design goal of making use of parallelism provided by GPUs.

By making use of the many use cases brought together by researchers in different areas the following key points were noticed:

1. The fundamental equations should be used as the input from the user, as this information has far reaching consequences on the numerical schemes but is hard to reconstruct from any other derived representation,
2. The equations should be expressed in a type checked language to avoid malformed inputs at the very first stage, but should be comprehensible to the end user,
3. The numerical solvers should be developed separately in a fully generic way so that they can be efficiently coded, and they can be specialized when all the necessary information becomes available,
4. To minimize errors as much information should be extracted from the equations as possible and must be used to decide with minimal user input which parts should be calculated on the CPU and on GPU.

Points 1 and 2 need some kind of language support. As we do not want to develop yet another language we chose to build an embedded domain specific language for mathematical expressions.

For points 3 and 4 there are two possible solutions:

A. Target an existing language that supports enough generic programming features such that the solver templates can be efficiently expressed and at the same time can be interfaced with existing parallel APIs, or

B. Create a language independent, fully controllable intermediate representation of the numeric solvers that can be exported to a less expressive target language.

Both methods have advantages and disadvantages. The first method requires less effort as the language's compiler can take over large amount of work from the programmer both in

coding and design time (the language features are already well designed and available). However the programmer is tied to a certain language and is limited by expressiveness and the internal constraints of the compiler. The second method requires more effort to create, but the implementer is completely free to add new expression constructs and manipulations and is in complete control of the implementation. Also the (programming) language independence makes it easier to port to many different target languages.

In this paper we outline the method B and discuss some details of the implementation of a language independent program representation and manipulation in C++, with the GPU part written in OpenCL. We chose C++ because it has a good low-level performance and API support, while at the same time (especially since the C++11 and C++14 updates) it has a sufficiently expressive and high-level type system. With this choice we can cover all points in one language. Other languages lack some desirable properties (C/Fortran is not expressive enough, rust is in beta stage and is less covered by APIs, Haskell is too unusual to end users) while a mixture of a high- and a low-level language (e.g.: C/C++ and Haskell) would raise interfacing issues. The choice of OpenCL is based on its portability and its independent run-time compilation so it can be better integrated into our workflow, as will be seen later.

III. Abstract Syntax Trees

As the mathematical representation of the equations and the specialization of the numerical solver program codes have to share information it is desirable to bring them as close as possible. The observation that both mathematical expressions and programming constructs can be represented by trees is widely known in the programming language theory and compiler designer communities. Compilers generate so-called *syntax trees* from the parsed source code that is a natural representation of the language, hence the name *abstract syntax trees* (AST). Since we don't want to write parsers and force the users to learn a new syntax we chose to embed the mathematical expression as a domain specific language with expression templates. The so-called Curiously Recurring Template pattern [17] is a very similar method that is used to build template trees to represent expressions. This would be the way if we chose to implement all the AST building and manipulation at template level, but, as discussed in section II, it would bring too much constraint into the model. The other way is that we build an expression tree in the memory (i.e. at run-time) and we use the template level only to ensure the syntactic validity of the expressions combined from the built in primitives. These primitives are the following: identifiers, declarations, definitions, lambda expressions, structs, statement blocks, and expressions combinators. This way the building blocks are simple, but they stop compilation if the user tries to call simple variables, mix declarations and statements etc. This is the basic syntactic check of the EDSL.

When the expressions are evaluated at runtime, we have a tree of all the mathematical expressions that the user gave as an input. At this point we can make algebraic manipulations, symbolic differentiation, series expansions etc. We especially focus on pseudo spectral collocation methods [18] in the form:

$$f(x) \approx \sum a_i \, \Phi_i(x) \tag{1}$$

Using symbolic derivation rules we can express all the derivatives of the functions as derivatives of the basis functions Φ_i and discretize over the proper collocation set $\{x_j\}$ to arrive at dense matrix constructs. It is also possible to generate the finite difference stencils for a given order from the symbolic expressions. After these steps the equations are ready to be merged with the solver templates.

As mentioned earlier, the solver templates are coded separately in an intermediate representation (see II./B), but they use the same AST based representation to ease expression manipulation and merging with the equations. At this point the solver codes are represented in a C-like EDSL. But to provide the generic formulation discussed in section II. there are two notable extensions: higher order functions and parallel functions.

Higher order functions are functions that return or take as an argument other functions. They are native citizens in more advanced functional programming languages as they make abstraction and generic programming extremely easy and comprehensible. Unfortunately in C like languages functions are not first class citizens and their manipulation and passing around is much more intricate. If one wants to functionally parametrize another function in the most efficient way, in C the best result is by passing function pointers, in C++ one can play with templates and lambda functions (or use std::function which is even slower). Furthermore when one considers GPU programming these techniques are not available, since there is no portable language that supports function pointers, and lambda like abstractions are very limited, so at the end only traditionally declared functions can be used. To work around this problem while providing higher order abstractions to the solver language, the ASTs are processed such that functional arguments are inlined (defunctionalisation) and new function declarations and definitions are generated automatically. This is transparent to the solver developer.

Parallel functions provide the basic, potentially parallel evaluable building block as known from libraries and APIs in GPU programming and other libraries. These functions depend on a special argument that describes the layout (extents) of the threads that can be executed theoretically in parallel. In our representation these functions have a special declaration and they need a special first argument: a *domain* object that at this point represents a 1D integral range. This domain object can be used as an index in the function body. When the parallel function is processed in the AST, the body is meant to be an OpenCL kernel, and all the appearances of the domain object in the body are replaced by the actual index variable of the thread, so in OpenCL these domain variables will translate to *get_global_id* function calls.

Of course the main point is that *parallel functions* can be *higher order functions*. Good examples are the parallel numerical integrators, where the function or r.h.s. of a differential equation is a functional input that comes from the user.

```
int main(){
  MathExpr t(L"t", 1, 1);
  MathExpr x(L"x", 1, 1);
  MathExpr f(L"f", 1, 1);
  MathExpr D(L"D", 1, 1);
  MathExpr v(L"v", 1, 1);
  MathExpr t0(L"t0", 1, 1);
  MathExpr pi(L"pi", 1, 1);

  SymbolicDE DE;
  DE.DimensionsSymbols() << t << x;
  DE.UnknownSymbols() << f;
  DE.Equations() <<
  diff(t, f) - D*diff(x, diff(x, f)) + v*diff(x, f);
  DE.Constants() <<
  (t0, 0.0) << equate(D, 0.75) << equate(v, 1.2);

  DE.BoundaryConditions() <<
  f(t0, x) - exp(-sq(x+2.)/(4*D*t0))/sqrt(4*pi*D*t0);

  DE.SpectralBases() <<
  SpectralExpansion(L"Chebyshev", 42, 1.0, 5.0) <<
  SpectralExpansion(L"RationalChebyshev",
                    42, 0.0., 1.5);

  DE.ProcessAsFullSpectral();

  arr<double> sample; sample << 4.0 << 0.0;
  DE.SampleSolutionToFile2(L"result.txt",
  1.0, 5.0, 0.25, 0, -12.0, 12.0, 0.3, 1, sample);
}
```

Fig. 1. Listing of a complete C++ function from our first project to express the Fokker-Planck equation and parametrize the numerical scheme. The syntax is expected to be even simpler in further revisions. The output is shown on Fig 2.

When the mathematical AST is merged with the solver templates the final code tree emerges and the above detailed defunctionalisation and parallelisation transformations are applied. Finally the AST is traversed and the target source code is emitted. In our case this means the host side C++ and the client side OpenCL kernel code.

IV. Current status and preliminary results

The framework detailed above is still in development and as a proof-of-concept investigated in two pilot projects:

The first project focuses on the mathematic / symbolic part, and the resulting expression trees are used to feed hand coded numerical solvers. Within this project simple ordinary and partial differential equations were tested successfully up to 3 components and 1 and 1+1 dimensions, respectively. The tested equations include: simple harmonic oscillator, Fokker-Planck equation, wave equation and different quantum Vlasov equations. These equations can be formulated symbolically in the mathematical EDSL in 30-40 lines including symbolic initial conditions and sampling of the result function to a text file. A complete user program showing the solution of the Fokker-Planck equation is shown on Fig. 1. and the output is shown on Fig. 2.

The second project investigates the solver template representation and code-generation part. The higher order and

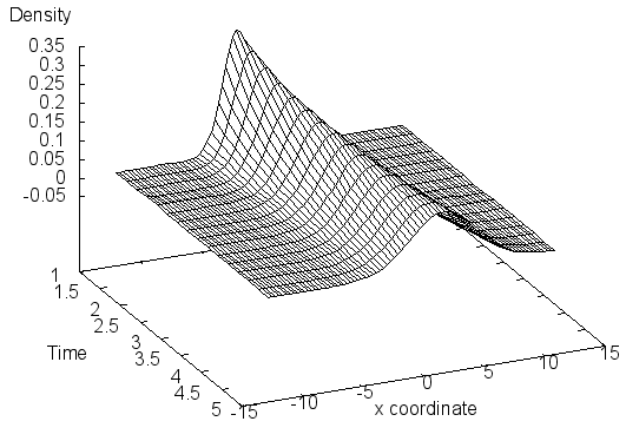


Fig. 2. Output of the program shown on Fig. 1.

parallel functionalities are already operational and have been used to implement numerical integrators (quadratures) and finite differential time steppers for ordinary differential equations. The EDSL used to express these solvers results approximately in the same line-of-code amount as a native version and the syntax is constantly being improved to provide an easy to comprehend meta-code. In the current implementation once the code in EDSL is expressed it is forwarded for AST manipulation, code-generation and compilation. The result is a dynamic library that is loaded back into the program, and an interface is given back to the user in the form of a functor object. This object can be called with buffers given by the user and these are bound to the very same symbolic variables that were used in expressing the EDSL code. In the current state of the project there is an 8th order adaptive Runge-Kutta solver [19] implemented that can be rapidly specialized to evolve ordinary differential equations simultaneously from a large set of initial values and thus can be used to perform GPU accelerated parameter scans. At this moment we can achieve a 30x speedup w.r.t the one thread CPU solution. Such scans are especially useful in pair production studies where numerical integration over the phase space initial values is necessary besides the time evolved solution for each parameter to obtain the total number of pairs, but also exploring fractal shapes, drawn out by an ensemble of trajectories, in the study of chaotic dynamical systems [20] can be accelerated easily. In fact our framework is already being utilized in a collaboration with Gábor Drótos (Eötvös University) to study chaotic saddles emerging in the restricted three body problem.

The main directions of development at this point are the following:

1. Improve the type checking features of the EDSLs in the mathematical part as well as the solver template part, while keeping the syntactic noise low,
2. Improve the AST transformations to include cost estimation, automatic decision of parallelisation and reordering based on data dependencies,

3. Merge the two projects so that the full featured mathematical representation can be fed into the solver templates with minimal user intervention.

V. Summary

In this article we reported on the development of a unified framework that makes the programming of high-performance parallel simulations easier for non-computing expert scientists to achieve. The primary goal is to bridge the gap between the fundamental mathematical models and the programming implementation. Tools like this framework are extremely needed to bring the most out of researchers in their respective fields without having them to delve into the complicated programming of parallel hardware. Our implementation targeted C++ because of its relatively generic type system and high-performance and OpenCL due to its portability. We discussed the implementation of two Embedded Domain Specific Languages to express the mathematical equations and relations of the scientific models and to represent numerical solvers in a natural way. While the development is not finished yet, preliminary applications of the framework show that the ideas can be applied in real life use cases and even at this stage it can significantly decrease development time for scientific end-users.

While the ongoing developments in computing APIs make easier to program the parallel devices in more higher level languages, the gap between mathematics and implementation is still going to be a bottleneck of scientific development, which keeps our addressed goals relevant.

Acknowledgment

This research is supported by experts of the Wigner GPU Lab [21].

References

[1] B. Aktemur, Y. Kameyama, O. Kiselyov, C-c. Shan, Shonan challenge for generative programming", Proceedings of the ACM SIGPLAN 2013 workshop on Partial evaluation and program manipulation, 2013, http://dx.doi.org/10.1145/2426890.2426917.

[2] J. Holewinski, L.-N. Pouchet, and P. Sadayappan, "High-performance Code Generation for Stencil Computations on GPU Architectures", Proceedings of the 26th ACM International Conference on Supercomputing ICS '12, 2012, pp. 311-320, http://doi.acm.org/10.1145/2304576.2304619.

[3] Andreas Klöckner et. al., "PyCUDA and PyOpenCL: A scripting-based approach to GPU run-time code generation", Parallel Computing, vol. 38, no. 3, 2012, pp. 157-174, http://dx.doi.org/10.1016/j.parco.2011.09.001.

[4] K. A. Hawick and D. P. Playne, "Automatically Generating Efficient Simulation Codes on GPUs from Partial Differential Equations", Computational Science Technical Note CSTN-087, http://complexity.massey.ac.nz/cstn/087/cstn-087.pdf.

[5] G. R. Markall, D. A. Ham, P. H.J. Kelly, "Towards generating optimised finite element solvers for GPUs from high-level specifications", Procedia Computer Science, vol. 1, no. 1, 2010, pp. 1815-1823, http://dx.doi.org/10.1016/j.procs.2010.04.203.

[6] M. Blazewicz et. al., "From physics model to results: An optimizing framework for cross-architecture code generation", Scientific Programming, vol. 21, no. 1, 2013, pp. 1-16, http://dx.doi.org/10.3233/SPR-130360.

[7] S. Husa, I. Hinder, C. Lechner, "Kranc a Mathematica application to generate numerical codes for tensorial evolution equations", Computer Physics Communications., vol. 174, no. 12, 2006, pp. 983-1004. http://dx.doi.org/10.1016/j.cpc.2006.02.002.

[8] Takayuki Muranushi, "Paraiso: An Automated Tuning Framework for Explicit Solvers of Partial Differential Equations", Computational Science & Discovery, vol. 5, no. 1, 2012, pp. 015003. [arXiv:1204.4779 [astro-ph.IM]].

[9] A. Hagiescu et. al, "GPU Code Generation for ODE-based Applications with Phased Shared-data Access Patterns", ACM Trans. Archit. Code Optim., vol. 10, no. 4, 2013, pp. 55:1--55:19, http://doi.acm.org/10.1145/2541228.2555311.

[10] F. R. Punzalan et. al., "A CellML simulation compiler and code generator using ODE solving schemes", Source Code for Biology and Medicine, 2012, vol. 7, no. 11, http://dx.doi.org/10.1186/1751-0473-7-11.

[11] M. Stimberg, D. F. M. Goodman, V. Benichoux and R. Brette, "Equation-oriented specification of neural models for simulations", Frontiers in Neuroinformatics, vol. 8, no. 6, 2014, http://dx.doi.org/10.3389/fninf.2014.00006.

[12] I. Bialynicki-Birula, P. Gornicki and J. Rafelski, "Phase space structure of the. Dirac vacuum", Phys. Rev. D44, 1991, pp. 1825-1835.

[13] F. Hebenstreit, R. Alkofer and H. Gies, "Schwinger pair production in space- and time-dependent electric fields: Relating the Wigner formalism to quantum kinetic theory", Phys. Rev. D82, 2010, pp. 105026.

[14] S. Ochs and U. W. Heinz, "Wigner functions in covariant and single-time formulations", Annals Phys. 266, 1998, pp. 351-416. [hep- th/9806118].

[15] A.V. Prozorkevich, S.A. Smolyansky and S.V. Ilyin, "Progress in Nonequilibrium Green's Functions II", Eds. M. Bonitz and D. Semkat, World Scientific, 2003, p. 401.

[16] V. V. Skokov and P. Levai, "Transverse and longitudinal momentum spectra of fermions produced in strong SU(2) fields", Phys. Rev. D 78, 2008, pp. 054004. [arXiv:0710.0229 [hep-ph]].

[17] J. O. Coplien, "Curiously Recurring Template Patterns", *C++ Report*, vol. 7, no. 2, 1995, pp. 24-27.

[18] J. P. Boyd, "Chebyshev and Fourier Spectral Methods" Second edition, Dover, New York, 2001.

[19] Numerical Recipes Webnote No. 20. http://www.nr.com/webnotes/nr3web20.pdf.

[20] E. Ott, "Chaos in dynamical systems", 2nd edition. Cambridge University Press, Cambridge, 2002.

[21] gpu.wigner.mta.hu

Possibilities for multicore processor and GPU parallelization of neural network training in Matlab

Aleksandra Dedinec, Sonja Filiposka, Dimitar Trajanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
aleksandra.kanevche@finki.ukim.mk, sonja.filiposka@finki.ukim.mk, dimitar.trajanov@finki.ukim.mk

Abstract— **Training of neural networks for large data sets is a computationally intensive and time consuming task. Different approaches are suggested in the literature to reduce the effort, whereby most of them use some kind of parallelization techniques. Nowadays, the computer systems are hybrid, which means that besides the multicore processor which is designed to run general programming tasks very well, they also contain graphical processing unit (GPU) which is becoming more powerful and more generalized. This allows, both - the multicore processor and the GPU to be applied to general purpose parallel computing tasks with great efficiency. The purpose of this paper is to explore the possibilities of achieving optimal performance of this hybrid system through workload balancing, when using Matlab. A case study of electricity load forecasting neural networks is used. As a result, in this paper, we present how certain parameters that the neural network training depends on, affect the optimal allocation of work, minimizing the execution time. Based on our findings, we propose further guidelines for efficient parallelization of the neural network training using Matlab on multicore processor and GPU architectures.**

Keywords—parallel processing, Matlab, neural networks, GPU, multicore processor.

I. INTRODUCTION

Today, most computer systems are hybrid, which means that they are equipped with both multicore processors and GPUs that have high computational power. By using the hybrid computing approach the processing units are collaboratively combined in order to perform a particular task, while maximizing the utilization of the computational power. CPU cores and GPUs are quite different in terms of their architecture, level of parallelism and clock rates. Typically, there are fewer CPU cores with higher clock rates. GPU cores are usually slower and require data transfer between main memory and the video memory, but their number is higher.

Proper allocation of work is very important if we want to achieve maximum utilization of the available processing units. For this purpose, given a certain task it is important to investigate whether and in what cases (different workload distribution) the different processing elements of a computer system will be used in the most efficient manner.

The problem of speeding up neural network (NN) simulations via parallelization is an important research topic that is being investigated. In [1] and [2] parallel implementation of a feed forward neural network using GPU is considered. The method for obtaining speedup using parallelization during the learning phase of a neural network is presented in [3]. Parallel implementation of a NN for financial forecasting applications is given in [4]. In [5] acceleration the NN processing for short term demand forecasting using GPU is presented. The hybrid approach with CPU and GPU computing is also reviewed in the literature. For example in [6] it is applied to topographic reconstruction, while in [7] irregular wavefront propagation algorithm is presented.

In this paper, we aim to explore how the neural network training can be parallelized using workload distribution on a hybrid system. As a widely spread engineering software, we have used the possibilities for parallel processing offered by Matlab. Particularly, we have investigated the opportunities for taking full advantage of the processing power available on a computer system by using Matlab, without making major changes in the program code and without low-level programming. A neural network for hourly electricity load forecasting is analyzed as a case study. When analyzing neural networks, the question whether the GPU or the distribution of workload between the processing units can give better results than the multicore processor is complex and depends on many parameters. Therefore, our goal is to provide guidelines on how these parameters affect the optimal workload balance when using Matlab.

The paper is organized as follows. The second section gives description of a neural networks as well as their possibilities for parallelization. In the third section the neural network parallelization on a hybrid system using MATLAB is described. In the last section the results and proper discussion is presented.

II. NEURAL NETWORKS

Artificial neural networks are widely used for load forecasting because they employ a process of learning, by which the relationship between the input and the output variables is defined. [8]. They are able to focus on the variables that have high influence on the output and to ignore the information that has little significance to the output [9].

A. *Description of the Neural Network*

As an example case study for our investigation we use artificial neural networks as a tool for short-term (hour ahead) electricity load forecasting. The model is applied to the data for the Macedonian hourly electricity consumption [10-11].

The neural network used in this paper has multi-layer feed-forward perceptron (MLP) structure, where the network is represented as a directed acyclic graph, whose structure has three layers - an input layer, one hidden layer and an output layer. The input layer gathers the model's inputs vector x while the output layer yields the model's output vector y [12]. The neurons in the hidden layer are activated by the tangent hyperbolic function.

The non-linear mapping between the input x and the output y is given with the following equation:

$$y = \sum_{j=0}^{h}\left[w_j f\left(\sum_{i=0}^{d} w_{ji} x_i \right)\right] \quad (1)$$

where d represents the input variables, h represents the hidden layer neurons and the variable y is a single linear output. The parameters w_j and w_{ji} represent the weights and biases that connect the layers [12].

The structure of the neural networks is presented on Fig. 1. As it is presented, there are eight input variables: hour of day, day of week, holiday flag, previous day's average load, load for the same hour of the previous day and the load for the same hour-day combination of the previous week, air temperature and cheap tariff variable. There is only one output variable representing the forecasted load in the given hour.

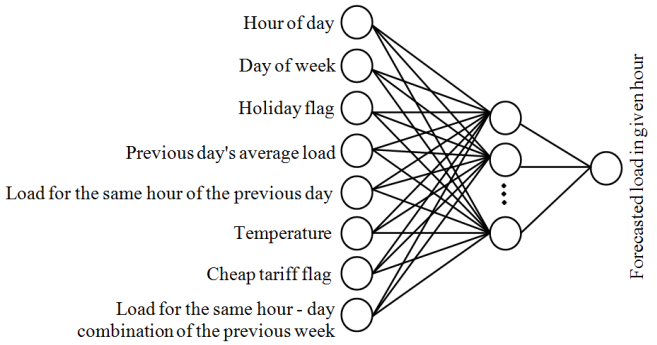


Fig. 1. Multilayer neural network for load forecasting

The main advantage of the artificial neural networks is the ability to learn. For the purpose of forecasting and prediction the supervised learning is used, where the cost function is a function of the difference between the forecasted and the actual data. During the training (learning) phase the goal is to adjust the weights in order to minimize the cost. The cost used is the mean-squared error, given as:

$$E_D = \frac{1}{2}\sum_{i=1}^{N}\{y_i - t_i\}^2 = \frac{1}{2}\sum_{i=1}^{N} e_i^2 \quad (2)$$

where y_i is the actual data and t_i is the forecasted data.

In order to minimize the error, the weight and bias values are updated according to the conjugate gradient backpropagation with Powell-Beale restarts [13]. The weights are updated until a criteria is satisfied, i.e. the performance goal, maximum train time, maximum validation failures, or maximum number of epochs is reached. In order to have a comparable execution time, we have intentionally set the first three parameters such that the training phase is always completed when the maximum number of epochs is reached. Thus the neural network will always be trained with the given number of epochs.

For the generalization (testing) phase [12] we used the most common metric in the electricity load forecasting: MAPE (mean absolute percent error) and MAE (mean absolute error).

B. *Possibilities for Neural Netwoks parallelization*

There are many parameters that influence the neural network training time. One example is the number of training samples. Training samples can be used as a good source for parallelism, since disjoint subset of training samples can be allocated to each parallel process.

Another parameter that is very important while analyzing the parallelization of a neural network is the number of neurons. In our case study we have eight neurons in the input layer and one in the output layer. As there is no general rule that can be applied for the number of neurons in the hidden layer with which the best results will be achieved, this value is commonly variable and it depends on the certain problem that is being solved.

The number of epochs is also a variable that plays an important role when exploring the speed up possibilities for the training phase because it defines the number of times data will be processed. However, the epochs are a limiting factor in parallelization, because each epoch depends on the previous one. Basically, at the end of each epoch the weights are calculated and used at the beginning of the next epoch.

III. Neural Network Parallelization on a Hybrid System Using MATLAB

At the lowest system level the instruction set of a modern microprocessor contains fused instructions for multiplication and addition. This means that multiplication and addition in one neuron can be done at the same time [14]. There is also instruction level parallelism. On upper level, the processing effort may be distributed among different cores of a processor with multicore architecture. Multicore processors are characterized by a small number of complex cores. On the other hand, GPUs are composed of a vast number of simple cores. As part of the training phase, inputs of the neuron are multiplied by the appropriate weights and then summarized. This operation can be efficiently performed by GPUs because it is often used in graphical applications by the multiply-and-add instruction [14].

This means that both types of processing units (CPU cores and GPUs) offer a good opportunity for parallelization, so in this paper we investigate the optimal distribution of the workload among them.

A. *Parallelization in MATLAB*

The problem which is being solved is implemented in MATLAB using the Parallel Computing Toolbox [15]. By using the Parallel Computing Toolbox applications can easily be parallelized at high-level because it provides several constructs like parallel for-loop, spmd (single program, multiple data) command for parallel data processing, message-passing functions and special array types for GPU computing. Thus, it is straightforward to take advantage of computer systems with multicore processors and GPUs, with only few modifications of the code, without low-level programming or knowledge of the hardware architecture.

Using the spmd command, the defined statements are executed on each MATLAB worker simultaneously. Therefore, we create a pool of workers (threads) that run concurrently at the CPU level, using the matlabpool command. We have set the number of workers to be equal to the number of physical cores of the processor. Individual MATLAB session or worker is dedicated to the GPU [16] such that a specific thread is created for each GPU (in our case there is only one GPU). The workload is balanced among the threads so that disjoint sets of training samples are assigned to each of them. This is illustrated on Fig. 2.

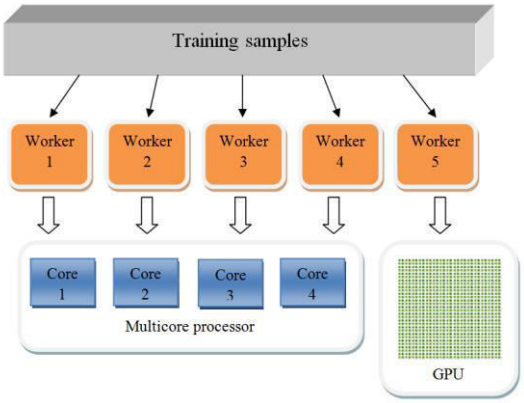


Fig. 2. Multicore processor and GPU approach for load forecasting neural network

In order to be able to distribute the training samples among the CPU cores and the GPU, it is necessary to define the input and the output matrices as Composite. The use of GPU in MATLAB is very straightforward. The only thing that is needed is to define the variables as gpuArray. This transfers the variables from the main memory to the GPU and all the subsequent functions are automatically executed on the GPU.

Using high-level constructs of the Parallel Computing Toolbox, we can control the number of workers used, the number of CPU cores, whether or not to use GPU, simple communication or data transfer between workers and how many training samples will be processed by each of the workers. In order to have greater control of the code execution, such as managing complex computations and data between workers and management of the computing resources, low-level MPI or CUDA programming (which can also be incorporated in Matlab) is needed.

IV. Results

The simulation results presented in this section were obtained on a computer system with Intel(R) Core i7-2630QM CPU at 2.00 GHz and NVIDIA GeForce GT 555M.

Three characteristic cases are analyzed. In the first case the number of training sample is varied. In the second and the third case the number of epochs and the number of neurons in the hidden layer is varied, respectively.

In the first case six simulations sets were executed, starting from 47472 data samples (which correspond to Macedonian hourly electricity consumption data for the period from January 2008 to May 2013) and ending with 32x47472 training samples, as it is shown on Fig. 3. The number of epochs in this case is equal to 1000 and the number of neurons in the hidden layer is equal to 20. Fig. 3 also shows how the optimal workload balancing between the multicore CPU and GPU is affected by the number of training samples. Because the training samples are independent, a part of the samples is assigned to each of the workers.

As it can be seen when there are fewer training samples it is more worthwhile to execute them only on multicore processor. However, by increasing the number of samples the optimal workload is moved to the middle, or half of the samples should be executed on multicore CPU, and the other half on the GPU. In this way maximum exploitation of the available hardware of the computer system is achieved, so that the optimization for serial execution of the multicore processor and the parallelization advantages of the GPU are used. So, by increasing the number of samples the execution time on GPU is reduced. It must be noted that the last simulation has 32x47472 training samples, which in the case of hourly electricity consumption forecasting corresponds to about 170 years of data, is not very realistic in our case. Of course, this example would be reasonable in the case of at least minute based ahead forecasting, when it would correspond to approximately 3 years.

On Fig. 4 the execution times when the NN is trained only on multicore processor, only on GPU and 50% on multicore processor and 50% on GPU is presented. The results show that the workload balance between the multicore processor and the GPU is worthwhile if the number of samples is higher than 90000, or if we analyze data for more than 10 years. In this case, if we have data for less than 10 years, the NN should be trained only on multicore processor.

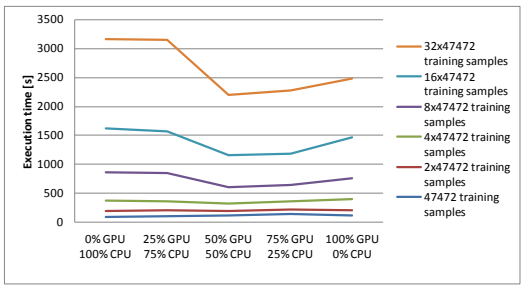


Fig. 3. NN training time with different number of data samples, depending on the workload balancing between the multicore CPU and GPU

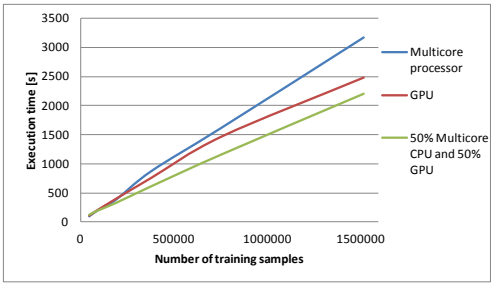


Fig. 4. Totally multicore CPU, totally GPU and 50% multicore CPU and 50% GPU time of NN training, depending on the number of trainig samples

In the second case, the number of epochs is varied, while the number of training samples and the number of neurons in the hidden layer is constant and is equal to 47472 and 20, respectively. Since epochs are dependent from each other, their execution must be serial. Fig. 5 clearly shows that if we have a smaller number of epochs, the execution time on GPU is less than the time on the multicore processor. By increasing the number of epochs, the advantage of multicore processor for efficient serial performance is highlighted. It is important to note that for electricity load forecasting, satisfactory results are obtained with 1000 epochs. By increasing the number of epochs, the forecasting precision is not increased significantly. Fig. 6 shows that in this case, if we have more than 500 training epochs the NN should be trained only on multicore processor.

In the third case, the number of neurons in the hidden layer is varied in the range from 20 to 100. The number of training samples is 47472, and the number of epochs is 1000. The precision of the results for load forecasting is pretty much the same in all cases of the number of neurons in the hidden layer. In Fig. 7, it is clear that by increasing the number of neurons in the hidden layer the parallelization capabilities of the GPUs are becoming more noticeable. It can be seen that GPUs can use the natural parallelization of the NN which arises from their architecture. This is because the calculation of the activation of a single neuron is independent and can be executed in parallel. Therefore, as shown in Fig. 8, if we have more than 40 neurons in the hidden layer, the time of training the NN is much shorter on GPU than on multicore processor.

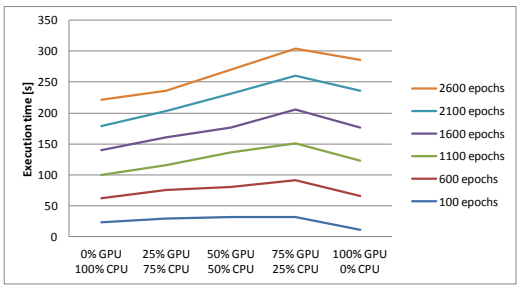


Fig. 5. NN training time with different number of epochs, depending on the workload balancing between the multicore CPU and GPU

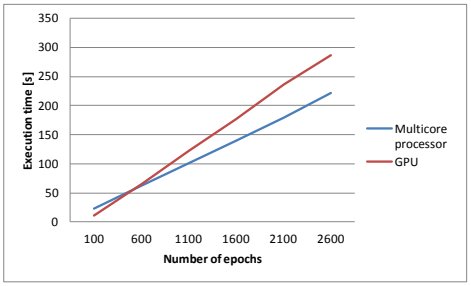


Fig. 6. Totally multicore processor and GPU execution time of the training of NN, depending on the number of epochs

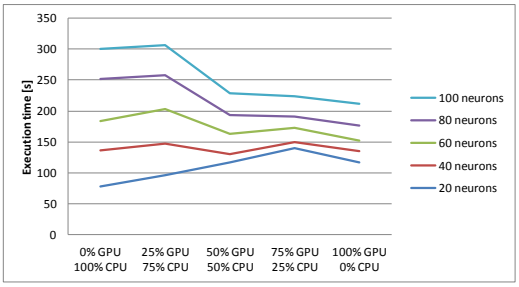


Fig. 7. NN training time with different number of hidden layer neurons, depending on the workload balancing between the multicore CPU and GPU

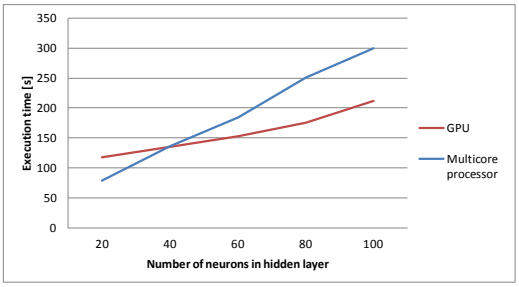


Fig. 8. Totally multicore CPU and GPU time of NN training, depending on the number of neurons in the hidden layer

CONCLUSION

In this paper, parallelization of neural network training using workload distribution on a hybrid system offered by Matlab is analyzed. It can be noticed that by using Matlab, it is straightforward to take advantage of computer systems with multicore processors and GPUs, with only few modifications of the code. A neural network for hourly electricity load forecasting is analyzed as a case study. When analyzing neural networks, the question whether the GPU or the distribution of workload between the processing units can give better results than the multicore processor is complex and depends on many parameters. Therefore, three parameters that influence the neural network training time are varied in this paper: number of training samples, number of neurons in the hidden layer and number of epochs. The results show that by increasing the number of training samples and the number of neurons in the hidden layer, workload balance between the multicore processor and the GPU and only GPU execution is becoming more worthwhile. On the other hand, by increasing the number of epochs, the advantages of multicore processor for efficient serial performance are highlighted.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University.

REFERENCES

[1] T. Bacek, D. Majetic, and D. Brezak, "GPU Implementationof the Feedforward Neural Network with Modified Levenberg-Marquardt Algorithm," International Joint Conference on Neural Networks. China, July 2014.

[2] S. T. A. Gurgel and A. de A. Formiga, "Parallel Implementation of Feedforward Neural Networks on GPUs," 2013 Brazilian Conference on Intelligent Systems, pp. 143-149, 2013.

[3] Y. Tsuchida and M. Yoshioka, "Speed Up Method for Neural Network Learning by Using GPGPU," SCIS-ISIS, Japan, November 2012.

[4] C. Augusto Casas, "Parallelization of Artificial Neural Network Training Algorithms: A Financial Forecasting Application," 2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics, pp.1-6, March 2012.

[5] T. He, Z. Y. Dong, K. Meng, H. Wang and Y. T. Oh, "Accelerating Multi-Layer Perceptron based Short Term Demand Forecasting Using Graphics Processing Units," Transmission & Distribution Conference & Exposition: Asia and Pacific, October 2009.

[6] J.I. Agulleiro, F.Vazquez, E.M.Garzon, and J.J.Fernandez, "Hybrid computing: CPU+GPU co-processing and its application totomographic reconstruction," Ultramicroscopy, vol 115, pp. 109-114, 2012.

[7] G. Teodoro, et al., "Efficient irregular wavefront propagation algorithms on hybrid CPU–GPU machines," Parallel Computing, vol 39, pp. 189-211, 2013.

[8] N. Kandil, R. Wamkeue, M. Saad, and S. Georgeas, "An efficient approach for short term load forecasting using artificial neural networks," International Journal of Electrical Power & Energy Systems, vol. 28, No. 28, pp. 525-530, October 2006.

[9] C. Hsu, and C. Chen, "Regional load forecasting in Taiwan--applications of artificial neural networks," Energy Conversion and Management, vol. 44, No. 12, pp. 1941-1949, July 2003.

[10] A. Dedinec and S. Kalajdziski, " Short-term electricity load forecasting of the Macedonian electric power system using neural networks," International Conference on Informatics and Information Technologies, Macedonia, April 2014.

[11] A. Dedinec and S. Kalajdziski, " The influence of different input variables of artificial neural network on short-term electricity load forecasting in Macedonia, " South East European Conference on Sustainable Development of Energy, Water and Environment Systems, Macedonia, July 2014.

[12] P. Lauret, E. Fock, R. Randrianarivony, and J. Manicom-Ramsamy, "Bayesian neural network approach to short time load forecasting," Energy Conversion and Management, vol. 49, No. 5, pp. 1156-1166, May 2008.

[13] M. J. D. Powell, "Restart Procedures for the Conjugate Gradient Method," Mathematical Programming, vol. 12, No. 1, pp. 241-254, 1977.

[14] D. Verber, "Implementation of Massive Artificial Neural Networks with CUDA, Cutting Edge Research in New Technologies, Prof. Constantin Volosencu (Ed.)", ISBN: 978-953-51-0463-6, InTech, 2012.

[15] Neural Networks with Parallel and GPU Computing, MathWorks, URL: http://www.mathworks.com/help/nnet/ug/neural-networks-with-parallel-and-gpu-computing.html

[16] J. W. Suh and Y. Kim, "Accelerating MATLAB with GPU Computing: A Primer with Examples", ISBN :9780124080805, Morgan Kaufmann, December 2013.

Computational Physics on the Magurele Physics Platform: Some challenges ahead

Alexandru Nicolin
Department of Computational Physics and Information Technologies, Horia Hulubei National Institute for Physics and Nuclear Engineering (IFIN-HH)
Magurele, Romania
alexandru.nicolin@nipne.ro

Virgil Baran
Faculty of Physics, University or Bucharest,
Magurele, Romania

Abstract — **In this paper we discuss some of the challenges in computational physics on the Magurele Physics Platform. Our presentation focuses on Boltzmann-Vlasov simulation for the description of collective modes in nuclear matter and the so-called particle-in-cell (PIC) simulations used in the description of the laser-plasma interaction with emphasis on the ELI-NP computing requirements.**

Keywords — Magurele Platform, distributed and parallel computing, Boltzmann-Vlasov, Particle-in-Cell

I. INTRODUCTION

Scientific computing has become one of the three main pillars of scientific investigations, alongside with theoretical and experimental analyses. The importance of numerical simulations is now unanimously acknowledged and it is quite significant that the Press Release of the 2013 Nobel Prize in Chemistry spoke of "The computer – your Virgil in the world of atoms" [1]. In fact, all large-scale experimental facilities now have substantial computing resources that are used both for data analysis and detailed numerical simulations. The latter are used both to deepen the understanding of the experimental results, by uncovering phenomena that were not transparent from the raw data, as well as to actually guide the experimental investigations by computationally scrutinizing the space of experimental parameters and also identifying the parameter regions that are relevant for a given physical process.

The Magurele Physics Platform emerged in early 1950s and has played a paramount role in consolidating the international stature of the Romanian physics community through continuous forefront scientific research. To give just two examples, this is where the first Romanian computer was design and put into function in 1955 and (one year after Theodore Maiman invented the first working laser in 1960) Ion Agarbiceanu created the first gas laser in Romania. The Magurele Physics Platform now hosts the Faculty of Physics of University of Bucharest, the Horia Hulubei National Institute of Physics and Nuclear Engineering (IFIN-HH), the National Institute for Laser, Plasma and Radiation Physics, the National Institute of Material Physics, the National Institute for Earth Physics, the National Research and Development Institute for Optoelectronics and, finally, the Institute of Space Sciences. In line with the tradition of the Magurele Physics Platform, the ELI-NP experimental facility, which is currently under construction in Magurele within the premises of IFIN-HH, will provide magnificent new opportunities to study the fundamental processes unfolded during light-matter interaction" using the most intense lasers world-wide and will foster an unprecedented interdisciplinary research plan which addresses "frontier fundamental physics, new nuclear physics and astrophysics as well as applications in nuclear materials, radioactive waste management, material science and life sciences" (see [2] for a detailed discussion).

On the computing side, ELI-NP raises a series of significant data and computing requirements that cover high-performance and high-throughput computing, as well as storage of experimental data in the petabyte regime. The envisaged computing activities go from the transfer of legacy codes to new languages and new hardware infrastructures and the optimization and parallelization of current codes, to the development of new codes which address the specific experiments described in the ELI-NP White Book [3].

The rest of the manuscript is structured as follows: in Section II we present some elements concerning the wide set of Boltzmann-Vlasov simulations, while in Section III we overview the computing challenges associated with Particle-in-Cell (PIC) simulations. Both sections emphasize the ELI-NP computing requirements on these computing directions. Finally, in Section IV we present our concluding remarks.

II. BOLTZMANN-VLASOV SIMULATIONS

The Boltzmann-Vlasov partial differential equations discussed in this section represent one of the most advanced semi-classical numerical tools used for the description of nuclear [4-5] and other mesoscopic quantum systems which accounts both for mean-field effects as well as two body collisions in the presence of Pauli blocking. The codes which are most relevant for the investigation of collective modes in nuclear matter are based on a widely used sequential numerical

treatment of the aforementioned equations. In their most classical form the equation take the form

$$\frac{\partial f_p}{\partial t} + \frac{\mathrm{p}}{m} \cdot \nabla_{\mathrm{r}} f_p - \nabla_{\mathrm{r}} U_p \cdot \nabla_{\mathrm{p}} f_p = 0$$
$$\frac{\partial f_n}{\partial t} + \frac{\mathrm{p}}{m} \cdot \nabla_{\mathrm{r}} f_n - \nabla_{\mathrm{r}} U_n \cdot \nabla_{\mathrm{p}} f_n = 0 \quad (1)$$

where *f* are the distribution functions for protons and neutrons, indicated by the *p* and *n* subscripts, while the *U* potentials for protons and neutrons are depend on the level of approximation that is used.

A detailed discussion on the parameters of the Boltzmann-Vlasov equation is outside the goal of our paper, but we show the equation in their full form to convey information about their general structure. The numerical solution of the equations requires quantities that depend quadratically on the number of test particles (pseudoparticles) associated with every physical particle (*i.e.*, nucleon) in the system. As a good spanning of the phase space of the system (*i.e.*, nucleus or nuclei) requires a large number of test particles, the accuracy of the numerical results is strongly limited by the available hardware infrastructure. Grid implementations of the codes take advantage of the internal structure of the numerical scheme which is ideally suited for distributed computing. To put it simple, instead of spanning the whole phase-space of the system with a single CPU core, we assign one phase-space cell to one CPU core. As the test particles that were previously assigned to the complete system now describe only a few physical particles, the number of test particles per physical particle can be substantially higher, and one can therefore simultaneously achieve a significant boost in accuracy and a reduced computation time. The main advantage of the grid implementation is that at a given moment the computation in one phase-space cell is independent from the computations in the other cells which means that, unlike parallel computing, the CPU to which the a given cell is assigned does not need to have a joint shared memory. Moreover, we can further improve the efficiency of the codes by eliminating the slow-downs generated by the loop iterations in each phase-space cell by assigning one CPU per cell and spreading the computing load among several CPU cores, using libraries such as OpenMP or MPI.

The Scientific Case of ELI Nuclear Physic Pillar at the end of the ELI-NP White Book [3] emphasizes that "while the presence of the giant dipole resonance is a well established fact in all nuclei the precise knowledge of the dipole strength on the low energy tail of the giant dipole resonance, especially the properties of the dipole strength at varying number of proton and neutrons, is an open problems in nuclear physics." Similarly, "pygmy dipole resonances (PDRs) in nuclei are frequently interpreted as a collective phenomenon representing an oscillation of the neutrons in excess (to protons) with respect to the *N*=*Z* core" and its properties are still under intensive scrutiny. As numerical investigations are severely limited by the exclusively sequential or parallel nature of the codes, we expect that the grid implementation of the codes developed within the Pygmy and giant dipolar resonances path-finding mini-project will be at least an order of magnitude faster than the existing top-level codes. To give just one example, in recent studies of pygmy dipole resonances in ^{132}Sn nuclei [6] 1500 test particles per nucleon have been used, and this is already one order of magnitude higher than the existing results which currently use 40 to 200 test particles per nucleon. However, to simulate numerically some of the physical phenomena that are relevant for ELI-NP, we would need at least 4000-6000 test particles per nucleon to reduce the numerical noise and this level of numerical accuracy can only be achieved using a hybrid distributed and parallelized versions of the codes. This increase in the quality of the numerical results comes, however, with a substantial computational cost, as the run for 6000 test particles took roughly one year, while that for 1500 took around one month using a standard workstation. Finally, we add that the numerical efficiency of the hybrid distributed and parallelized codes will allow us to investigate the dipolar responses considering more complex, momentum dependent, effective interactions, which have not yet been used for high-precision numerical computations of this type.

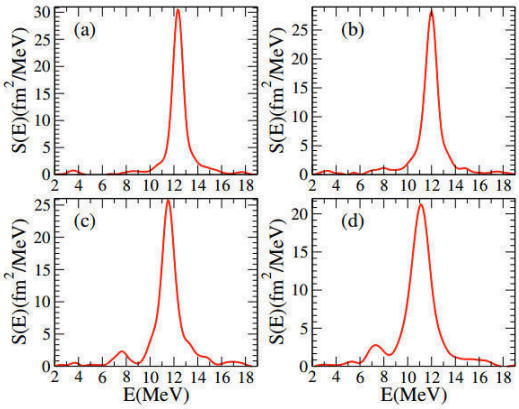


Fig. 1 The dipole strength function for: (a) ^{116}Sn, (b) ^{124}Sn, (c) ^{132}Sn, and (d) ^{140}Sn respectively.

In Figure 1 we show the dipole strength function for various Tin species. These results are obtained through the numerical solution of the Boltzmann-Vlasov equations.

According to a recent report by the National Research Council of the National Academy of Science (USA), the origin of the heaviest elements remains one of the 11 greatest unanswered questions of modern physics. It is therefore obvious that rapid and accurate results will help us to understand the dynamics of nuclei synthesis at the mean-filed level. The proposed hybrid implementations are ideally suited to study the fusion cross-section as a function of energy and the competition between fusion and fast fission phenomena. The physics addressed by the numerical codes pertains not only to the Scientific Case of ELI-NP but also to the experiments at SPIRAL 2 and SPES experimental facilities located in France and Italy. The

experience gained with Boltzmann-Vlasov transport codes will help with the numerical investigations of other physical systems also relevant for ELI-NP as for example the interaction of intense laser fields with metallic clusters and fullerenes, and the dynamics of plasmas in intense fields, thereby allowing us to address in the future the relativistic transport of quarks and gluons which is needed to simulate numerically quark-gluon plasma at CERN conditions.

For the parallel versions of the existing Boltzmann-Vlasov codes it is ideal to use the PCAM (Partitioning, Communication, Agglomeration and Mapping) methodology to obtain high quality parallel programs. As the name states, this approach addresses all the relevant aspects of parallelization. First, we will identify all the basic tasks, by decomposing each computation into small (elementary) tasks and focusing on finding opportunities for parallel execution. Next, the communication required to coordinate task execution is determined, identifying not only the number of messages, but also the volumes of data to be exchanged between nodes. The complexity of the code is then evaluated, paying special attention to computational and communication complexities. Thus, tasks may be combined to reduce data transfers and to group similar activities. In the end, tasks are assigned to CPU cores either in a statistical manner or using load-balancing techniques. As the concurrency and scalability issues are addressed by the first two PCAM phases of this process, the final two phases deal with performance related issues. Moreover, given that numerical codes will also be used for very accurate numerical investigations for ELI-NP, formal specification tools and techniques based on LOTOS (Language Of Temporal Ordering Specification, ISO Standard IS8807) will be used to formally prove all the required characteristics of our new parallel and distributed programs, thus avoiding possible bottlenecks. Thus, not only that we prove the correctness of our codes but we will also prove that our parallel and distributed codes are free of deadlocks and livelocks and that computing times do not diverge.

III. Particle-in-Cell simulations

One of the main computational challenges at ELI-NP is the laser wakefield acceleration, which is a major research field world-wide [7]. The main appeal of this research direction comes from the possibility to accelerate electrons using relatively small-scale laser-based experimental facilities. The main idea is that one can use the huge electric fields generated during laser-plasma interaction to accelerate electrons in a fashion similar to the way a surfer uses the water waves to increase his velocity. In simple words, the mechanism behind the acceleration is the following: as the laser propagates in the neutral gas it dislocates electrons from the neutral atoms (through the ponderomotive force) which then acquire an oscillatory motion around the resulting static ions. The resulting waves propagate with a velocity close to that of the speed of light behind the laser pulse that generated them and can be used to accelerate electrons over short distances using the longitudinal electric fields of the waves. It should be noted that these fields are thousands of times greater than the accelerating fields in conventional accelerators, as they are no limited by the upper bound set in conventional accelerator by the electrical breakdown limit which is around 100 MV/m. This increase of the accelerating gradient is the key ingredient in reducing the size, and therefore the costs, of particle accelerators.

While there are already notable successes for the acceleration of electron beams in the GeV regime [7], it is still necessary to demonstrate beam quality and, moreover, the pulses need to be reproducible in energy. To this end, the experimental study can be efficiently complemented by computational studies that rely on the so-called particle-in-cell (PIC) methods [7].

Particle-in-cell (PIC) methods have a long-honored history that goes back to the mid 1950s, before the first Fortran compilers were available, and have received recurrent interest from the physics community. These methods are basically a set of numerical solvers for determining the Lorentz forces which describe the individual particles in the system and the Maxwell equation which determine the electric and magnetic fields. As most systems are usually too large for elementary particles to be treated individually, the so-called super-particles are commonly used. These are finite-sized large particles which consist of many real particles which act as one in terms of their dynamical properties, thanks to the Debye shielding which make the real particles invisible to one-another with respect to the Coulomb force at distances larger than the Debye length. While most PIC solvers follow the above general structure, there are numerous differences in their numerical implementations. These differences impact directly the physical regime described by the codes (density of plasmas, properties of the electron beams, etc.) and the deployment infrastructure. For the latter, the most important information concerns the maximal size of the HPC cluster up to which the code is known to scale efficiently.

A survey of the parameter space will then show the impact on the properties of the electron beams of distinct physical parameters such as: the type and density of gaseous target, the duration and energy of the laser pulse, the acceleration length, the plasma density, etc. One point of special interest concerns the competition between the laser wakefield acceleration and the direct laser acceleration, as well as the impact of the latter mechanism on the energy spectrum of the electron bunch. Another topic of particular interest is the sensitivity of the final results with respect to the spatial and temporal non-uniformities of the laser pulse. The resulting parametric charts will be of immediate interest to the ELI-NP experimental groups, while at a more general level the experience gained with these codes will help streamline a research direction dedicated to particle-in-cell simulations. The numerical investigations will also be focused on the impact of the dimensionality of the codes, the resolution of the numerical grids, the number of super-particles, etc., on the final results.

Two-dimensional codes are particularly appealing when performing parametric scans due to their lower computational load, but they have to be tested for relevance through detailed comparisons will with fully three-dimensional ones.

IV. Conclusions

In this manuscript we have focused on some of the computing challenges of interest on the Magurele Physics Platform, with particular emphasis on some of the computing requirements of ELI-NP, and have briefly outlined a series of future directions of interest. The main message of the previous sections is that substantial software optimization is needed for the current codes to be used efficiently to uncover the mechanism behind the physics processes summarized in the ELI-NP White Book. To this end, one has to bridge the purely computing aspects of numerical modeling with frontier nuclear physics problem such that the *in silico* experiments can actually guide the experimental investigations foreseen at ELI-NP.

Acknowledgment

For this work the authors were supported by the Romanian Authority for Scientific Research and Innovation, CNCS-UEFISCDI, Project No. PN-II-PCE-2011-3-0972. AIN was also supported by PN 09370108/2015.

References

[1] See the full Press Release at: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/press.html

[2] D. Habs, T. Tajima, and N. V. Zamfir, "Extreme Light Infrastructure – Nuclear Physics (ELI-NP): New Horizons for Photon Physics in Europe," Nuclear Physics News, vol. 21, pp. 23-29, 2011.

[3] The ELI-NP White Book is available at http://www.eli-np.ro/documents/ELI-NP-WhiteBook.pdf

[4] See V. Baran *et al*. Phys. Rev. C 85, 051601(R) 2012.

[5] See Th. Fennel *et al.* Review of Modern Physics 82, 1793 (2010).

[6] See V. Baran *et al.* Phys. Rev. C 88, 044608 (2013).

[7] For a review of the main results see V. Malka, Phys. Plasmas 19, 055501 (2012).

Performance comparison by running job in Hadoop by defining the number of Maps and Reduces

Zirije Hasani
Faculty of Computer Science
Public University of Prizren
Prizren, Kosovo
zh12796@seeu.edu.mk

***Abstract*—Analyzing Big Data is a considerable challenge for every company and institution that deals with large amount of data. On one hand it is faced with problem to store a large amount of data, and on another, to process it in reasonable or even real time. At the present time the best known system for analyzing a Big Data is Hadoop. Hadoop implements MapReduce System whose role is to orchestrate the processes on distributed nodes (servers), to run the various tasks in parallel, to manage all communications and data transfers between the various parts of the system and to provide needed redundancy and fault tolerance features.**

In this paper we investigate the ways for tuning the system, by setting the number of Maps and Reduces in order to compare and optimize performances of the same job execution. Experiments are done on the job that counts the word frequencies in the 1218 Mb Wikipedia text, with automatically assigned Map/ Reduce numbers by Hadoop and our values defined through the SSH or by means of the own java code.

Keywords—Big Data Analytic, Hadoop, MapReduce

I. INTRODUCTION

Analyzing Big Data is a considerable challenge for every company and institution that deals with large amount of data. On one hand it is faced with problem to store a large amount of data, and on another to process it in reasonable or even real time. At the present time the best known system for analyzing a Big Data is Hadoop. For purpose of testing its performances, and possibilities of tuning, it was installed at Faculty working environment. We have analyzed Wikipedia text data. Simple job that was processed count how many times one word appears in the document.

Hadoop system implements MapReduce programming model composed of Maps tasks (like filtering, sorting) executed in parallel, and then Reduce tasks (e.g. summary operations) in order to get the result desired. Hadoop MapReduce System orchestrates the processing on distributed nodes (servers), running the various tasks in parallel, manages all communications and data transfers between the various parts of the system, and provides needed redundancy and fault tolerance features.

Hadoop system offer automatically defined number of Maps and Reduces.

In this work we explore additional possibilities that we have to set the number of Maps and Reduces and investigate which values to choose, in order to compare and optimize performances of the same job execution. We would like to know if by experiments we can find the optimal numbers of Maps and Reduces depending on the job which is executed, the numbers that can outperform the automatic ones. Experiments are done on the 1218 Mb text file with automatically assigned values by Hadoop, and our values defined through the SSH, or values defined by the own java code.

In the section 2 we explained what Hadoop is how it works, the configuration components of Hadoop under which we have executed the experiments. Also the Hadoop pros and cons are discussed. In the section 3 we present related work-general discussions given in literature on the Map and Reduce numbers influencing the Hadoop performance and ways these numbers to be defined. Result of our job execution where number of Maps and Reduces was automatically defined by Hadoop is presented in section 4. In the section 5 we present two methods for manually definition of the number of Maps and Reduces. In the section 6 comparison of the performance of the executed job is given. The last section is a conclusion of this experiment.

II. HADOOP

In our research we have examining different platforms [8] and architectures [9], [10] for analyzing Big Data in real time. Part of our examination was Hadoop which is an open source platform still in the phase of development, where continuously new components are added or improved. Up to now, Hadoop was platform only dedicated for the batch processing, not suitable for real time analytics applications. But now (2015 year), new platform of Hadoop, called esHadoop[1] was introduced. It is integrated with Elasticsearch [7], [10] and can handle Big Data real time analytics. Our future investigation will be this two way connector, that allows to index data in Elasticsearch [10] and query them in real time.

[1] https://www.elastic.co/products/hadoop

Hadoop is a framework for processing, storing and analyzing massive amounts of distributed data. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Hadoop handle the big data by breaking them in many parts, where each part is processed in parallel, meaning each part can be processed and analyzed at the same time.

Hadoop enables a computing solution that is [6]:

- Scalable– we can add new nodes without interrupting the continuous work of other nodes
- Cost effective– Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.
- Flexible– cope with various types of data (structured, non-structured) can admit data from multiple sources, data can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.
- Fault tolerant- if cluster loses a node, the system redirects work to another location of the data and continues processing without missing a beat.

A. *How Hadoop Works*

The sources from where unstructured data came are many: log files, social media feeds, internal data stores and others. Hadoop breaks the data up into "parts," which are then loaded into a file system made up of multiple nodes running on commodity hardware. The data are stored in Hadoop Distributed File System (HDFS). File systems such as HDFS are adept at storing large volumes of unstructured and semi-structured data as they do not require data to be organized into relational rows and columns.

The benefit of HDFS is that each "part" of data is replicated multiple times and loaded into the file system so that if a node fails, another node has a copy of the data contained on the failed node. A Name Node acts as facilitator, communicating back to the client information such as which nodes are available, where in the cluster certain data resides, and which nodes have failed.

Once the data is loaded into the cluster, it is ready to be analyzed via the MapReduce framework. The client submits a "Map" job to one of the nodes in the cluster known as the Job Tracker. The Job Tracker refers to the Name Node to determine which data it needs to access to complete the job and where in the cluster that data is located. Once determined, the Job Tracker submits the query to the relevant nodes. Rather than bringing all the data back into a central location for processing, processing then occurs at each node simultaneously, or in parallel. This is an essential characteristic of Hadoop.

When the each node has finished processing its given job, it stores the results. The client initiates a "Reduce" job through the Job Tracker in which results of the Map phase stored locally on individual nodes are aggregated to determine the "answer" to the original query, then loaded on to another node in the cluster. The client accesses these results, which can then be loaded into one of number of analytic environments for analysis. After the MapReduce job is completed now the data are available for further analysis. We can use this data to transfer into some database platform or to save in data warehouses for further analysis based on the need of the researchers of organization.

B. *Hadoop Technical Components*

Hadopp configuration consists of many interrelated components. Functionality of these components are:

- Hadoop Distributed File System (HDFS): The default storage layer in any given Hadoop cluster;
- Name Node: The node in a Hadoop cluster that provides the client information on where in the cluster particular data is stored and if any nodes fail;
- Secondary Node: A backup to the Name Node, it periodically replicates and stores data from the Name Node should it fail;
- Job Tracker: The node in a Hadoop cluster that initiates and coordinates MapReduce jobs, or the processing of the data.
- Slave Nodes: The grunts of any Hadoop cluster, slave nodes store data and take direction to process it from the Job Tracker.

The above mentioned components are the main components of Hadoop but the Hadoop ecosystem is made up of a number of complimentary sub-projects. NoSQL data stores like Cassandra and HBase are also used to store the results of MapReduce jobs in Hadoop. Open source language Pig helps to the developers to write Hadoop function even if they are not the Java programmers. Hive is an open source data warehouse originally developed by Facebook that allows for analytic modeling within Hadoop.

C. *Hadoop: The Pros and Cons*

Like every other system and Hadoop have its pros and cons.The main benefit of Hadoop is that it allows enterprises to process and analyze large volumes of unstructured and semi-structured data, previously inaccessible to them, in a cost- and time-effective manner. Because Hadoop clusters can scale to petabytes and even exabytes of data, enterprises no longer must rely on sample data sets but can process and analyze all relevant data. Data Scientists can apply an iterative approach to analysis, continually refining and testing queries to uncover previously unknown insights. It is also inexpensive to get started with Hadoop. Developers can download the Apache Hadoop distribution for free and begin experimenting with Hadoop in less than a day.

The downside to Hadoop and its components is that they are immature and still developing. As with any young, raw technology, implementing and managing Hadoop clusters and

performing advanced analytics on large volumes of unstructured data requires significant expertise, skill and training. Unfortunately, there is currently a deficiency of Hadoop developers and Data Scientists available, making it impracticable for many enterprises to maintain and take advantage of complex Hadoop clusters. Further, as Hadoop's many components are improved upon by the community and new components are created, there is, as with any immature open source technology/approach, a risk of forking.

The good news is that some of the brightest minds in IT are contributing to the Apache Hadoop project, and a new generation of Hadoop developers and Data Scientists are coming of age. As a result, the technology is advancing rapidly, becoming both more powerful and easier to implement and manage. An ecosystems of vendors, both Hadoop-focused start-ups like Cloudera and Hortonworks and well-worn IT stalwarts like IBM and Microsoft, are working to offer commercial, enterprise-ready Hadoop distributions, tools and services to make deploying and managing the technology a practical reality for the traditional enterprise. Other bleeding-edge start-ups are working to perfect NoSQL (Not Just SQL) data stores capable of delivering near real-time insights in conjunction with Hadoop.

Hadoop configuration at FINKI

HDFS	
NameNode	Hadoop-ambari-master-1.finki.ukim.mk
SecondaryNameNode	Hadoop-ambari-master-2.finki.ukim.mk
DataNodes	6 DataNodes Hadoop-ambari-slave-2.finki.ukim.mk Hadoop-ambari-slave-2.finki.ukim.mk Hadoop-ambari-slave-2.finki.ukim.mk Hadoop-ambari-slave-2.finki.ukim.mk Hadoop-ambari-slave-2.finki.ukim.mk Hadoop-ambari-slave-2.finki.ukim.mk
HDFS Disk Capacity 393.7 GB	
MapReduce	
JobTracker	Hadoop-ambari-master-2.finki.ukim.mk
TaskTrackers	6 TaskTrackers Hadoop-ambari-slave-1.finki.ukim.mk Hadoop-ambari-slave-2.finki.ukim.mk Hadoop-ambari-slave-3.finki.ukim.mk Hadoop-ambari-slave-4.finki.ukim.mk Hadoop-ambari-slave-5.finki.ukim.mk Hadoop-ambari-slave-6.finki.ukim.mk

III. Partitioning Job Into Maps and Reduces

Selection of the appropriate size for the tasks for execution of a job can radically change the performance of Hadoop. Increasing the number of tasks increases the framework overhead, but increases load balancing and lower the cost of failures. At one extreme is the 1 Map/1 Reduces case where nothing is distributed. The other extreme is to have 1,000,000 Maps/ 1,000,000 Reduce where the framework runs out of resources for the overhead. So one of the performance optimization parameters depends on proper choice of the numbers of Maps and Reduces. Different authors propose different methods to define the number of Maps and Reduce in Hadoop.

A. *Number of Maps*

The number of Maps Hadoop assigns automatically is usually driven by the number of DFS (distributed file system) blocks in the input files. This causes users to adjust their DFS block size to adjust the number of Maps. The right level of parallelism for Maps is usually set to be around 10-100 Maps/node although can increases with cpu-light Map tasks.

It is suitable to control the number of Maps. The hint to the InputFormat for the number of Maps is Mapred.Map.tasks parameter. The default InputFormat behavior is to split the total number of bytes into the right number of fragments. However, in the default case the DFS block size of the input files is treated as an upper bound for input splits. A lower bound on the split size can be set via Mapred.min.split.size. Thus, for the 10TB of input data file with 128MB DFS blocks is expected to have 82k Maps, unless the Mapred.Map.tasks is even larger. Ultimately the InputFormat determines the number of Maps [5].

The number of Map tasks can also be increased manually using the JobConf's conf.setNumMapTasks (int num). This can be used to increase the number of Map tasks, but will not set the number below that which Hadoop determines via splitting the input data [5].

The number of Map task can be set manually by Mapred.Map.tasks=(number of Maps).

B. *Number of Reduces*

If we have small number of data that need to be processed the right number of Reduces proposed by [5] is 0.95 or 1.75 * (nodes_number * Mapred.tasktracker.tasks.maximum). At 0.95 all of the Reduces can launch immediately and start transferring Map outputs as the Maps finish. At 1.75 the faster nodes will finish their first round of Reduces and launch a second round of Reduces doing a much better job of load balancing [5].

Currently the number of Reduces is limited to roughly 1000 by the buffer size for the output files (io.buffer.size * 2 * numReduces << heapSize). This will be fixed at some point, but until then, it provides a pretty firm upper bound

The number of Reduces also controls the number of output files in the output directory, but usually that is not important because the next Map/Reduce step will split them into even smaller splits for the Maps [5].

The number of Reduce tasks can also be increased in the same way, as the Map tasks, with command Mapred.Reduce.tasks=(number of Reduces)

IV. Results of Executed Job with Hadoop Automatically Assigned Number of Maps and Reduces

Following is the result file of the Hadoop job execution that analyzes 1218 Mb Wikipedia text file with automatically assigned number of Maps and Reduces. The job was executed in 9.17 minutes:

From the results in figure 1 we can see that Hadoop has launched automatically 7 Map tasks and 1 Reduce task. There is 0 ms of the time spent by all Reduces and Maps waiting after reserving slots. The numbers of HDFS (Hadoop distributed file system) bytes that are read are 879072470, and the numbers of HDFS bytes that are written are 611391.

```
s

 map 100% reduce 100%
Job complete: job_201404291230_0014
Counters: 29
  Job Counters
    Launched reduce tasks=1
    SLOTS_MILLIS_MAPS=2569053
    Total time spent by all reduces waiting after reserving slots (ms)=0
    Total time spent by all maps waiting after reserving slots (ms)=0
    Launched map tasks=7
    Data-local map tasks=7
    SLOTS_MILLIS_REDUCES=567965
  File Output Format Counters
    Bytes Written=611391
  FileSystemCounters
    FILE_BYTES_READ=1198188810
    HDFS_BYTES_READ=879072470
    FILE_BYTES_WRITTEN=1823748810
    HDFS_BYTES_WRITTEN=611391
  File Input Format Counters
    Bytes Read=879071518
  Map-Reduce Framework
    Map output materialized bytes=625062998
    Map input records=18570332
    Reduce shuffle bytes=625062998
    Spilled Records=54155931
    Map output bytes=587930544
    Total committed heap usage (bytes)=3671085056
    CPU time spent (ms)=290950
    Combine input records=0
    SPLIT_RAW_BYTES=952
    Reduce input records=18566206
    Reduce input groups=9001
    Combine output records=0
    Physical memory (bytes) snapshot=2277122048
    Reduce output records=9001
    Virtual memory (bytes) snapshot=11399880704
    Map output records=18566206
```

Fig. 1. Hadoop Default Execution Output

V. Defining the Number of Maps and Reduce Manually

Following are two methods for determination of the number of Maps and Reduces for a job execution in Hadoop. The first method is to define the number throw SSH (Secure Shell), and second by writing the suitable commands in the java code.

A. *Definition of number of Map and Reducers by the SSH*

Following lines of code submitted through the SSH, define number of Maps and Reduces.

Hadoop jar al3oh.jar al3oh –D Mapred.Map.tasks=4 –D Mapred.Reduce.tasks=2 bufferr.txt out_buffer_7
(1)

In order to analyze the data from the file buffer.txt, we first need to save this file into HDFS with the following command: `Hadoop dfs -copyFromLocal buffer.txt bufferr.txt` (which means copy the file buffer.txt from local directory and put into HDFS and name it as bufferr.txt).

The first part of this code locate the java code al3oh.jar. It is pre-written java code that count how many times a word appears in a document. It will be executed then with -D Mapred.Map.tasks=4 -DMapred.Reduce.tasks=2 by which we define the number of Maps and Reduces. Next we specify which file to read in order to execute the job (buffer.txt) and at last we define the file name where to save the result from finished job (out_buffer_7).

B. *Numbers of Maps defined by java code*

The following commands in java code Fig.2. define the number of Maps tasks which in this case is 10.

```
1  public static void main(String[] args) throws Exception {
2  Configuration conf = new Configuration();
3  String[] otherArgs = new GenericOptionsParser(conf,
4  args).getRemainingArgs();
5      if (otherArgs.length != 2) {
6        System.err.println("Usage: wordcount <in> <out>");
7        System.exit(2);
8      }
9   //    conf.setInt("mapred.map.tasks", 10);
10     Job job = new Job(conf, "word count");
11     conf.setNumMapTasks(10);
12     job.setJarByClass(Test.class);
13     job.setMapperClass(TestMapper.class);
14     job.setCombinerClass(IntSumReducer.class);
15     job.setReducerClass(IntSumReducer.class);
16     job.setOutputKeyClass(Text.class);
17     job.setOutputValueClass(IntWritable.class);
18     job.setInputFormatClass(CustomInputFormat.class);
19     job.setOutputFormatClass(TextOutputFormat.class);
20     FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
21     FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
22     System.exit(job.waitForCompletion(true) ? 0 : 1);
```

Fig. 2. Java code for defining the number of Map

Command (line 11) conf.setNumMapTasks(10) define the number of Map tasks and also in the same way the number of Reduce tasks are defined.

VI. Comparison of the Performances of the Executed Job

In our experimental work we have analyzed performance of the execution of job that counts the word frequencies in the 1218 MB Wikipedia text file. We manually defined Maps and Reduces numbers through the SSH commands (1) given in section 5.1 The results from this experiment are shown in table 2 below. It is important to mention that this analytic is done just on one file with of 1218 Mb data in one execution perh defined Maps/Reduces values. The aim of the experiment was to find out how number of Maps and Reduces influence the time execution of the same job and is it possible to find solution that outperform one proposed automatically by Hadoop. Simple experiment starts with Hadoop default values (7 Maps and 1 Reduces), first by decreasing the number of Maps and keeping default value of Reduces, and than by increasing the number of Reduces and keeping the default number of Maps.

From the results shown in the table 2 below we can conclude that for this kind of job processing, better solution exist than proposed automatically by Hadoop. But, there is no proportional decrease or increase of time in correlation with increasing decreasing the number of Maps/ Reduces task. It is worth mentioning that almost all of the executed times were better than proposed by Hadoop. Reasons can be different but one we have in mind can be that Hadoop needs some time to calculate the best values not needed by manual definition. So experiments we made thus far unfortunately indicate only that we should experiment further in order to take into account other parameters that influence the time execution, as also to find out how different processes, e.g size and number of files influence the performance.

In the table below the measured time to finish the job of one job execution per given choice of Map/ Reduce number of tasks is given.

TABLE I. RESULTS FROM EXECUTED JOB WITH DEFINED MAPS AND REDUCES

	Number of Maps and Reduces		
	Number of Maps	***Number of Reduces***	***Time to finish job***
1	8	1	10 min
2	**7 *(default)***	**1 *(default)***	**9.17 min**
3	6	1	4.13 min
4	**5**	**1**	**3.26 min**
5	4	1	4.34 min
6	3	1	4.23 min
7	2	1	3.35 min
8	1	1	5.14 min
9	7	2	5.31 min
10	7	3	4.33 min
11	7	4	8.01 min
12	7	5	4.08 min

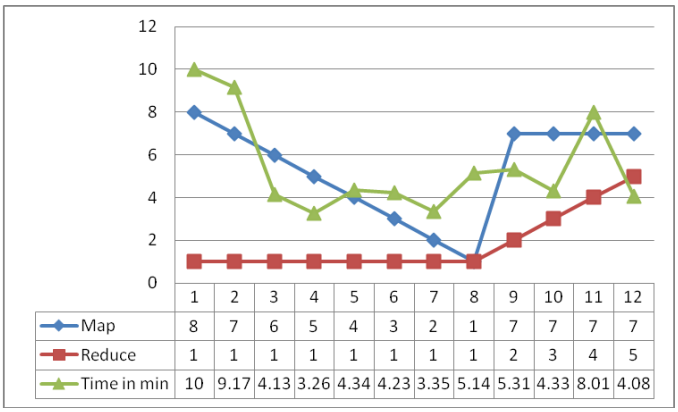


Fig. 3. Chart for results in table 2 (time in min)

VII. CONCLUSION

Analyzing big data in a real time needs careful choice among available technologies, and even more cautious tuning and settings of parameters that will influence and optimise its performance. Our recent research is oriented to investigation of available open source systems suitable for specific real time big data analytics.. In this paper we show some experiments that we have done with Hadoop. To increase Hadoop performances it is possible to define number of nodes in cluster used, to tune several its parameters, memory allocation and other. In this paper we experiment only with two parameters: number of Maps and number of Reduces tasks. Hadoop assigns automatically these numbers, but it is also possible to define them manually. Our experiment showed that better values can be found than offered automatically by Hadoop. Further investigation should be done to determine optimisation model to achieve these optimal values. Even more, other factors (number of nodes, size and number of files, memory allocation, and similar) should be investigated in order to propose related optimization model. In the next research phase we are going to analyze the performance of job execution taking into account also these other factors. From the simple experiments done we can conclude that Hadoop use some methodology to define default parameters, but it does not give optimal performances.

References

[1] Tom White. Hadoop: The definitive Guide. Second edition, O'REILLY 2011.

[2] Pascal Alma. Writing a Hadoop MapReduce task in Java. (August 24, 2013) Retrieved May 7, 2014, from http://www.javacodegeeks.com/2013/08/writing-a-Hadoop-MapReduce-task-in-java.html

[3] stackowerflow. MapReduce count example. Retrieved May 9, 2014, from http://stackoverflow.com/questions/6164230/MapReduce-count-example

[4] Carlo Scarioni. Hadoop Basics - Creating a MapReduce Program. (03.18.2012)Retrieved May 9, 2014, from http://java.dzone.com/articles/Hadoop-basics-creating

[5] Hadoop Wiki. Partitioning your job into Maps and Reduces. (20.09.2009) Retrieved

[6] Big Data: Hadoop, Business Analytics and Beyond. Retrieved May 7, 2014, from http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond

[7] Elasticsearch, Retrieved April 03, 2015, from https://www.elastic.co/products/Hadoop

[8] Zirije Hasani, Margita Kon-Popovska, Goran Velinov. *Survey of Technologies for Real Time Big Data Streams Analytic.* 11th International Conference on Informatics and Information Technologies. April 11-13, 2014 – Bitola, Macedonia.

[9] Zirije Hasani, Margita Kon-Popovska, Goran Velinov. *Lambda Architecture for Real Time Big Data Analytic.* ICT Innovations 2014 Web Proceedings ISSN 1857-7288

[10] Zirije Hasani. Implementation of Infrastructure for analytic big data logs. UKIM Conference, Aprill 2015.

Virtuoso, system for saving semantic data

Zirije Hasani
Faculty of Computer Science
Public University of Prizren
Prizren, Kosovo
zh12796@seeu.edu.mk

***Abstract*— Nowadays the need of describing the data and knowing the relationship between them is very important. This is one of the reasons why the World Wide Web brings the new era of web, known as semantic web. Semantic web is a collection of standard technologies that realize a web of data and tends to transfer the web of data to web of knowledge. The most valuable aspect of semantic web is its assistance to people in their everyday life in order to make it easier, by offering opportunities to find the things they need in faster and better way. According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". We are exploring the system known as Virtuoso Universal System that can store and process semantic data. We use Virtuoso to store data in columns, where web data we make bigrams to compare two or three words to find the semantic and then to store they in Virtuoso.**

Keywords*— *Virtuoso, Semantic web, World Wide Web.

I. INTRODUCTION

Nowadays describing the data and relationship between them in a way that computers can understand is very important [4]. This is one of the reasons why the World Wide Web brings the new era of web, known as semantic web. Semantic web is a collection of standard technologies that realize a web of data and tends to transfer the web of data to web of knowledge. The most valuable and interesting aspects of semantic web is to support people in their everyday life, by assisting them to find the things they need in a faster and better way. For example someone may like to search for places where to pass the vocation, but has limited time to explore places, hotels, restaurants at some specific location. The solution for this is web of data. According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". A common model has to be provided for machines to understand the “labels” and draw some conclusions from that information. The “classification” of the terms can become very complex for specific knowledge areas: this is where ontology, thesauri, vocabularies, etc. take their role. W3C has developed a set of standards and technologies for this:

- RDF – the Resource Description Framework
- OWL – the Web Ontology Language (based on RDF)
- SPARQL – a Query language for the Semantic Web

RDF is a data model for objects (resources) and relations between them that provides the foundation for publishing and linking data. It provides a simple semantics that can be represented in the XML syntax. OWL is a richer vocabulary description language that can depict the classes and their properties, such as relation between classes (ex. disjointness), cardinality (ex. "exactly one"), equality, types and, characteristic of properties and enumerated classes, eexample, Joan is a professor and is part of category staff, also Ana is a professor and is part of category staff. With OWL we describe that Joan and Ana fall in the same category which is staff. With logic we know that in staff category are both Joan and Ana. SPARQL is a semantic query language for databases, able to retrieve and manipulate data stored in RDF format.

"If HTML and the Web made all the online documents look like one huge book, RDF, schema, and inference languages will make all the data in the world look like one huge database". (Tim Berners-Lee, Weaving the Web, 1999).

The Semantic Web is a Web of actionable information—information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of “meaning” in which the logical connection of terms establishes interoperability between systems. A Web of data and information would look very different from the Web we experience today. It would routinely let us recruit the right data to a particular use context—for example, opening a calendar and seeing business meetings, travel arrangements, photographs, and financial transactions appropriately placed on a time line, [5].

We can find good example in [3] illustrated in figure 1. Software Agents will be greatly facilitated by semantic content on the Web. In the depicted scenario, Lucy’s agent tracks down a physical therapy clinic for her mother that meets a combination of criteria and has open appointment times that mesh with her and her brother Pete’s schedules. Ontologies that define the meaning of semantic data play a key role in enabling the agent to understand what is on the Semantic Web, interact with sites and employ other automated services.

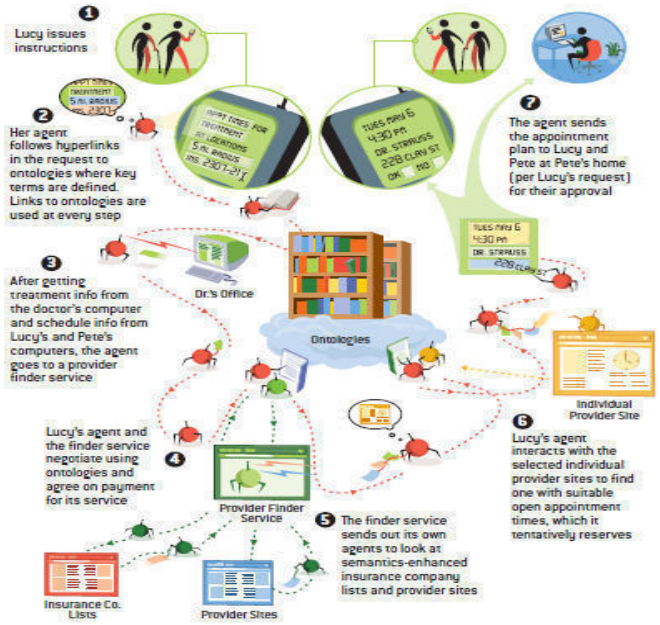


Fig. 1. Software agent, [3].

Properties are a special kind of resources; they describe relations between resources, for example "written by", "age", "title", and so on. Statements assert the properties of resources. A statement is an object-attribute-value triple, consisting of a resource, a property, and a value, [1].

We explore the system known as Virtuoso Universal System. It is a revolutionary, next generation, high-performance virtual database engine for the Distributed Computing Age that can store, process and manipulate semantic data in an optimal manner.

II. System for Saving Semantic Data: VIRTUOSO

Virtuoso is a revolutionary, next generation, high-performance "universal server" for the Distributed Computing Age that enables a single multithreaded server process that implements multiple protocols. Virtuoso Universal Server is a middleware and database engine hybrid that combines the functionality of a traditional RDBMS, ORDBMS, virtual database, RDF, XML, free-text, web application server and file server functionality in a single system. The open source edition of Virtuoso Universal Server is known as OpenLink Virtuoso. It is cross platform Universal Server that implements Web, File, and Database server functionality alongside RDF storage, Native XML Storage, and Universal Data Access Middleware, as a single server solution. It includes support for key Internet, Web, and Data Access standards such as: XML, XPATH, XSLT, SOAP, WSDL, UDDI, WebDAV, SMTP, SQL-92, ODBC, JDBC, and OLE-DB. Virtuoso currently supports several Operating systems - Windows 95/98/NT/2000, Linux (Intel, Alpha, Mips, PPC), Solaris, AIX, HP-UX, Unixware, IRIX, Digital UNIX, DYNIX/PTX, FreeBSD, SCO, MacOS X, [2]. Virtuoso provides transparent access to existing data sources, which are typically databases from different database vendors. Through a single connection, Virtuoso simultaneously connect ODBC, JDBC, UDBC, OLE-DB client applications and services to data within Oracle, Microsoft SQL Server, DB/2, Informix, Progress, CA-Ingres and other ODBC compliant database engines. All databases are treated as single logical unit [2]. Virtuoso exposes all of its functionality to Web Services. This means that existing infrastructure can be used to support Web Services directly [2].

III. Virtuoso SQL Optimization

Virtuoso provides a cost based SQL optimizer which performs the following types of query transformation:

- Join Order
- Loop Invariants
- Opening derived tables
- Migrating enclosing predicates into derived tables or unions
- Dropping unreferenced columns or results
- Detection of identically false predicates
- Index selection
- Grouping of co-located remote tables into single remote statements
- Selection of join algorithm

Virtuoso evaluates various permutations of joined tables against its cost model and determines the best fit, from which it generates a query graph. This query graph can be returned as a result set by the explain() SQL function. The cost model is based on table row counts, defined indices and uniqueness constraints, and column cardinalities, i.e. counts of distinct values in columns. Additionally, histograms can be made for value distribution of individual columns.

Virtuoso automatically maintains statistics about tables in the local database. When tables are attached from known types of remote DBMS's, Virtuoso also attempts to retrieve statistics information if available. The sys_stat_analyze or sys_db_stat stored procedures can be used to force an update of statistics, also recompiling all SQL statements or procedures depending on these statistics. Once this is done, this overrides the automatic statistics. The values of automatic statistics can be seen in the SYS_COL_AUTO_STAT table.

The stored procedure: DB.DBA.SYS_STAT_ANALYZE (in full_table_name varchar, in prec integer); constructs the basic table statistics and feeds it into the DB.DBA.SYS_COL_STAT system table. The DB.DBA.SYS_DB_STAT stored procedure performs this operation on the entire database.

The stored procedure: DB.DBA.SYS_STAT_HISTOGRAM (in full_table_name varchar, in full_column_name varchar, in n_buckets integer,

in prec integer); constructs table column histogram and feeds it into the DB.DBA.SYS_COL_HIST system table. The default value of prec, in both cases, is 5, which implies that a five percent sample of the table will be used. A percentage of 0 means that the whole table will be read.

A. *Demonstration of the STAT_ANALYSE & STAT_HISTOGRAM procedures*

The following script is intended for use with the ISQL program as the user dba, in the DB qualifier. The foreach statement is a special feature of the ISQL utility.

```
create table "DB"."DBA"."DTTEST" ("ID" integer primary key);
foreach integer between 1 10 insert into "DB"."DBA"."DTTEST" ("ID") values (?);
sys_stat_analyze ('DB.DBA.DTTEST');
select * from DB.DBA.SYS_COL_STAT;
sys_stat_histogram ('DB.DBA.DTTEST', 'ID', 2);
select * from DB.DBA.SYS_COL_HIST;
```

That yields:

```
Resultset 1 (from DB.DBA.SYS_COL_STAT)
---------+------------------------------
CS_TABLE         CS_COL           CS_N_DISTINCT CS_MIN  CS_MAX  CS_N_VALUES CS_N_ROWS
VARCHAR NOT NULL VARCHAR NOT NULL INTEGER       VARCHAR VARCHAR INTEGER     INTEGER
-------------------------------------------------------------------------------------

DB.DBA.DTTEST    ID               10           1       10      10          10

Resultset 2 (from DB.DBA.SYS_COL_HIST)
--------------------------------------
CH_TABLE         CH_COL           CH_NTH_SAMPLE     CH_VALUE
VARCHAR NOT NULL VARCHAR NOT NULL INTEGER NOT NULL  VARCHAR
-------------------------------------------------------------------------------

DB.DBA.DTTEST    id               0                 1
DB.DBA.DTTEST    id               5                 6
DB.DBA.DTTEST    id               10                0
```

IV. USING VIRTUOSO

Part of Virtuoso is interactive SQL command from where if we type the command *tables;* all tables in database will be shown. The result is shown in figure 2.

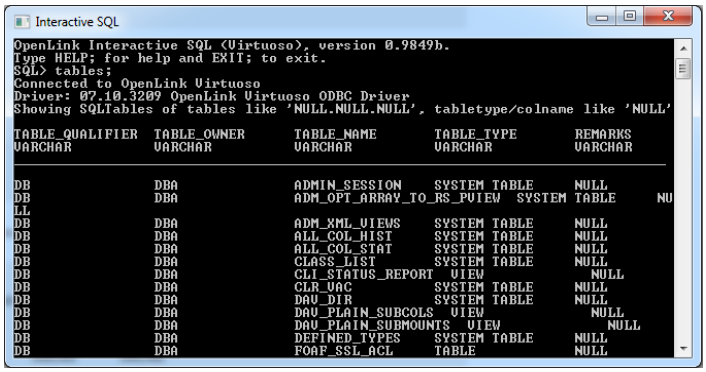


Fig. 2. Result from command tables

Through its menu bar (Database/Create table) Virtuoso has option for the table creation. Textbox opened enable inserting the SQL code for table creation (figure 3).

The SQL command is:

```
CREATE TABLE DB.DBA.AI (
AI INTEGER IDENTITY,
XX VARCHAR,
PRIMARY KEY (AI)
)
```

Example creating new table and then select this table in Virtuoso.

```
CREATE TABLE demo_table (
id INTEGER PRIMARY KEY,
txtdata VARCHAR(20)
)
```

Written in textbox

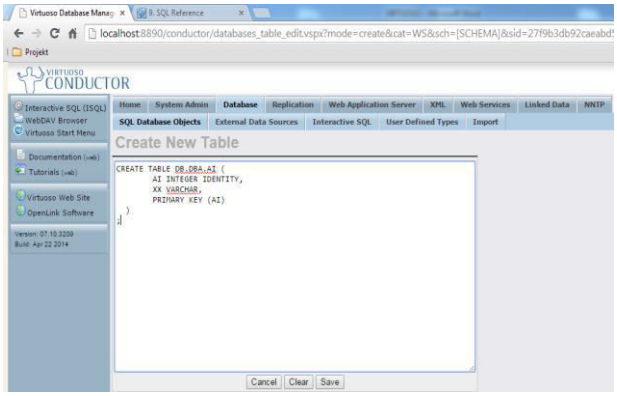


Fig. 3. Option for creating new tables in Virtuoso

Select the demo_tabel

```
SQL> SELECT * from demo_table;
id                  txtdata
INTEGER NOT NULL    VARCHAR

0 Rows. -- 15 msec.
SQL>
```

Following example illustrate creation of the table in Virtuoso and insert data to it.

```
CREATE TABLE test_check (
 name VARCHAR,
 age INTEGER
 CHECK (age > 18),
 email VARCHAR
 CHECK (regexp_like(email, '^([a-zA-Z0-9_.-])+@(([a-zA-Z0-9-
 ])+.)+([a-zA-Z0-9]{2,4})+\$'))
 )
```

```
INSERT INTO test_check (name, age, email) VALUES
('Jack', 18, 'jack@foo.bar');
```

Select the table demo_table in shown below:

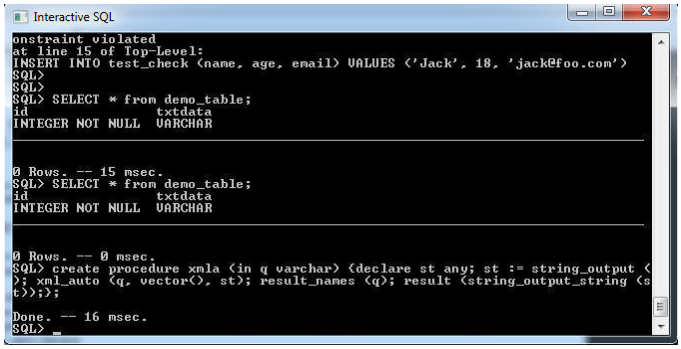
```
SQL> SELECT * from demo_table;
id                  txtdata
INTEGER NOT NULL    VARCHAR

0 Rows. -- 0 msec.
SQL>
```

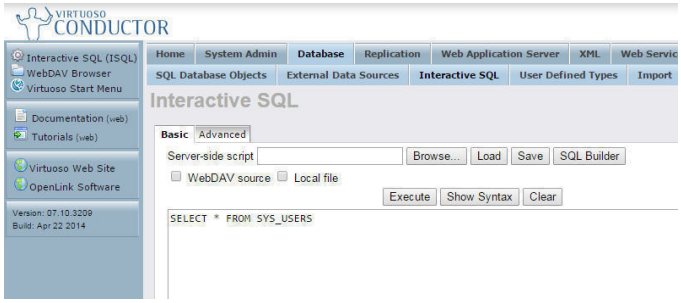
Following command create procedures in Virtuoso:

```
create procedure xmla (in q varchar) {
declare st any;
st := string_output ();
xml_auto (q, vector (), st);
result_names (q);
```

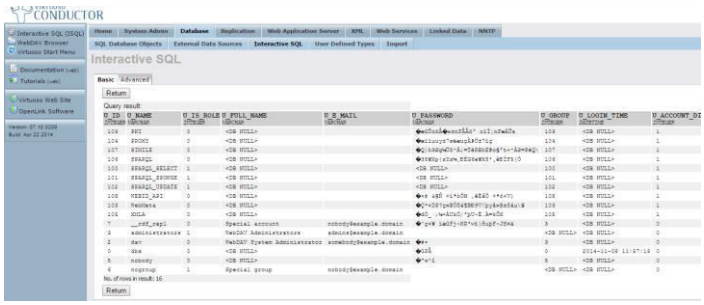
result (string_output_string (st)); };
Result is:



Select all data from table user:



Result is shown below:



Virtuoso supports SPARQL (a language that enable search of semantic web data). SQL is used to query relational data, XQuery is used to query XML data, and however SPARQL is used to query RDF data. At its most basic, a SPARQL query is an RDF graph with variables. For example, consider the following RDF graph:

ex:juan foaf:name "Juan Sequeda".
ex:juan foaf:based_near ex:Austin.

A version of the previous RDF graph that has variables instead of values is given by:

?x foaf:name ?y.
?x foaf:based_near ?z.

Note that variables in SPARQL queries start with a question mark (?).

Example of SPARQL query for the RDF data above is:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
FROM <http://example.com/dataset.rdf>
WHERE {
  ?x foaf:name ?name .
}
ORDER BY ?name
```

Command ***PREFIX*** describes appropriate declarations for abbreviating URI. Command ***SELECT*** returns data matching some conditions and command ***FROM*** defines the RDF dataset which is being queried. Command ***WHERE*** specifies the query graph pattern to be matched. And command ***ORDER BY*** is used to rearrange the query results. In this example we show how we can write a query to search for semantic data.

V. CONCLUSION

Semantic web is changing the way we live and think. The new era of web enable faster web response to our search for needed information. This opportunity is offered with new standards of describing, storing, interconnecting, processing and querying semantic data (RDF – the Resource Description Framework, OWL – the Web Ontology Language, SPARQL – a Query language for the Semantic Web). Technological solution that supports these standards is Virtuoso. **OpenLink Virtuoso** is cross platform Universal Server that implements Web, File, and Database server functionality alongside RDF storage, Native XML Storage, and Universal Data Access Middleware, as a single server solution. It includes support for key Internet, Web, and Data Access standards such as: XML, XPATH, XSLT, SOAP, WSDL, UDDI, WebDAV, SMTP, SQL-92, ODBC, JDBC, and OLE-DB. It currently supports several operating systems and provides transparent access to existing data sources, which are typically nowadays databases of different database vendors. After installation of Virtuoso we examine its basic functionality and conclude that with this system it is easy to manage.

References

[1] Grigoris Antoniou and Frank van Harmelen. "A Semantic Web Primer". The MIT Press Cambridge, Massachusetts, London, England. Second edition. 2008.

[2] OpenLink Software Documentation Team. OpenLink Virtuoso Universal Server: Documentation. OpenLink Software, 1999 - 2009

[3] Tim Berners-Lee, James Hendler and Ora Lassila. *The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* SCIENTIFIC AMERICAN, INC. 2002.

[4] w3schoole.com. http://www.w3schools.com/web/web_semantic.asp, accesed on 29/06/2013.

[5] Nigel Shadbolt , Wendy Hall and Tim Berners-Lee. *The Semantic Web Revisited.* IEEE Computer Society, 2006.

[6] Virtuoso documentation. Link: http://docs.openlinksw.com/virtuoso/CREATETABLE.html, Accessed on 08/11/2014

Concept of Automated Data Transfer by Using Linked Servers

Microsoft SQL Server 2012

Jelena Jovkovska, Adrijan Bozhinovski, Veno Pachovski
School of Computer Science and Information Technologies
University American College Skopje
jelena_jovkovska@hotmail.com, bozinovski@uacs.edu.mk, pachovski@uacs.edu.mk

***Abstract*—The elaborate networks of systems and solutions that drive organizations are becoming increasingly complex and difficult to implement integrate and manage. Workflow automation plays a big role in unifying these systems and streamlining the data flow throughout the business. Automated data transfer solutions give businesses the power to achieve unprecedented levels of automation, freeing the IT staff from repetitive time intensive manual processes, thus increasing productivity and reducing costs.**

To make the process more scalable, auditable and maintainable, there will be an additional server created on top of the current architecture which will act as a gateway between both systems. The server will be accessible by the client and the operations can be performed immediately upon request without including any manual work. The architecture behind the process will make auditing more clear and visible for both parties. The results of the daily data import will be placed in a table and will be accessible to the client at all times.

The main functionality of the new server will be to automate as much manual work as possible and to provide the end user with more flexible methods of processing the data at any time. It is important to provide the end user with a visibility of their daily data requests in order to be sorted quickly. The result will be a significant increase in data processing efficiency, which in turn would yield more productivity and savings in the corporate environment.

Keywords— Automated Data Transfer, Linked Servers, Import, Re-import, Data

I. INTRODUCTION

Automation of processes plays a big role in providing more effective and efficient way of how the organizations work. The structure of the most important processes in a business organization can significantly affect the relationship between the business and its clients. Organizations strive to provide their clients the best support they can by using less of their resources to achieve that and performing no manual work at all. The transformation of IT has been fueled by the promises and deployment of automation across all business functions. The rollout of ERP systems, the adoption of Internet and/or cloud-based software and infrastructure, the adoption of policy-based IT management (e.g. ITIL) are all done with the aim of providing automation, i.e., greater operational efficiency, to the business or agency. The automation can be applied to business functions native to the business's core competency (like data transfer), or to ancillary and support functions (like HR and IT). One of the most used solutions is the automation of the repetitive processes and freeing IT staff from manual work [1]. At its most core level, automation is the elimination of repetitive, manual tasks, rendering high operational value, especially if those tasks have high mission criticality and/or a propensity toward human error.

This paper is based on a research performed in a business organization where the main stream that drives all the processes is data. This organization is in charge of the affiliation marketing systems of online gaming companies worldwide. The Business is providing a reporting system based on data figures. Every day the current system is processing a large amount of data through the backend system to the production environment. Each process is scheduled and customized based on the client's requirements.

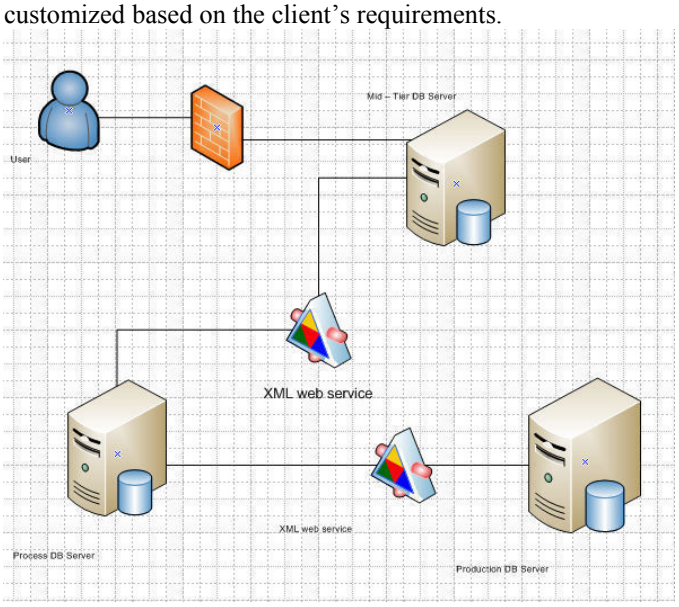


Fig. 1. Linked Servers Workflow [2]

It was stated that 20% of all the manual work done by the IT team is data transfer related. The team is performing repetitive work on a daily basis to ensure the smooth running of the reporting system. This is usually not convenient for either the company or the clients. The client will have to wait for the IT

team to respond to their data related queries, which most of the time is due to the exception failures of the current data process, such as data was sent late (not picked up by the scheduler), data was incorrectly sent and not processed by the system etc.. Based on the aforementioned, the company has come up with a solution to ensure that no data manual work would be done in future. To allow the clients to push the data into the system at any time without using scheduled jobs, the company will introduce another server on top of the current architecture which will be accessible by the client.

The aim of this paper is to present an automated transfer of data by using a complex architecture. To achieve this, Microsoft SQL Server 2012 will be used as the main platform of database technology.

II. THE PROCESS OF SENDING DATA

A process links actions and tasks together to achieve a desired end. In general, the more actions involved in the process, the higher the complexity. Additionally, there may be conditional processing of actions, dependency based processing or even complex logic (if, then, else, loops, nesting, etc.) directing the flow of processing. A process can be a human process, or a human-machine interaction. With automation, the process can be carried out completely by a machine, following a human's design of the process flow and its logic. The higher the complexity of the logic, combined with the frequency of execution and value of the results, the greater the incentive becomes to automate the process.

The main drive of the solution is the request table. Without having a request record pushed to the Mid-Tier along with the data to be imported, the data process will not start. Once the client transfers the data into the transfer tables, they will also need to insert 3 rows for each Data Type (Player Registration, Player Activity, and Exchange Rate) in the above table which will include the request and the status set to 'Received'. When the daily import starts, the status is updated to 'Executing'. If the data import process fails, the status is updated to 'Failed'. On the other hand, if the data import process is successful, the status is updated to 'Success'. A specific column, previously agreed upon will be used when the client requests re-imports of data. When the client requests a re-import, the scripts will delete all players' activity data from the system by using the parameter given and will import the new data accordingly. The Date Requested column will be set automatically by the system when a request record has been added. To have clear visibility of the entire data transfer process, a new table will be introduced in the database. In this table, the system will store audit records for all requests made by the client and the status of every request. Once a request is performed by the system, an audit record will be inserted in this table.

The client will be pushing the data in three tables containing the same fields as the columns in the current daily files. Schedules will be created so that every 3 hours it will be checked if the client sent the data. A process will be put in place that will check if the data has been received before starting the data import job. If there are no data in the transfer tables, a notification email will be sent and the job will be stopped. Once the data gets inserted successfully, the tables are cleared for the next day import.

In this solution three types of automated data import will be covered:

- Daily Transfer
- Exceptions Transfer
- Data Re-import\

A. *Daily Transfer*

The Daily Transfer type of import is the most basic. It caters for the usual daily import where the system receives the new figures for each day and stores them into the reporting system. These figures represent the activity of the players and the players itself for the previous day.

Once the Daily Import starts, a figures reconciliation table will be populated with the received figures grouped by activity date and product. This will allow both parties a clearer and a more visible process of reconciling the figures sent per month. The Daily import scripts will process the data into the production system and will return back the customer exceptions to the MID-Tier Server.

B. *Exceptions Transfer*

This type of data transfer is closely related to the 'Daily Transfer'. Once the system processes the data from the Daily Transfer, the system generates exception records as an output. These exceptions are records that were not successfully imported into the system due to various factors. The system is created to handle exceptions and provide feedback.

The output of exceptions is sent back on the Mid-Tier Server where the client will have full access to the records that did not manage to be imported successfully into the system with a description accordingly. Then the client will resolve the failed records, by providing the correct information, which will be sent as a request for Exception Import to the Mid-Tier Server. This process will be explained in details further in this paper.

The system starts the Exception import process, which processes the data. Once the exception transfer process is completed, the system will resolve any successfully processed customer exceptions. If a customer exception is resolved after the exception transfer process, the record will be removed. If there are still exception customers after the exception transfer process, the system will populate the appropriate table with the new exceptions.

C. *Re-import Process*

This process is the most complex of all due to the steps involved in reaching to the end. First, the system removes the old data based on the parameter given from the client, and then it imports the new figures and refreshes the system.

In order for the Re-import process to begin, the client must specifically request for one on the Middle Tier Server. Based on the request the procedures on the backend will follow the scripts to perform the re-import. Once the re-import process is completed, the appropriate table will be updated with the new imported and received figures.

To ensure a positive customer experience, a process must trigger an automated action, and automation must then manage the experience across multiple channels, personalize the customer's interaction, and predict the best next actions. It is highly important for both parties to create a solution which will not include any manual work. Therefore, to make this process more intelligent, every record sent from the client will be stamped with a unique constraint. The constraint will be unique for each request as well.

III. Automated Exception Handling

After the data has been imported in the system, three files are created with the customer and activity exceptions. Once the middle layer is set up, another table will be created to store the customer registration exceptions. The Exceptions tables will store all records received which failed to be imported into the system. The exception history will store the exceptions for the last 7 days.

IV. Automated Daily Reconciliation

A process of shifting through the corrected exceptions is included on both ends. This will provide a more accurate reconciliation report. The reconciliation numbers will be stored in an appropriate table, which will be populated on a daily basis after performing a request from the client, in order to keep track of the figures more accurately. There will be a script performing checks on the accuracy and to make sure that the table is updated with the latest figures. After importing the exceptions, the reconciliation table will be refreshed so that the latest data becomes available. On a daily basis, a procedure will check that the number of rows received match the number of rows inserts plus the number of rows in exceptions. If there is a mismatch in the figures, an Alert will be sent to both parties.

V. Alerts

The following Alerts will be created to make sure that any issues are picked up promptly.

Current System:

1. Missing data in the transfer tables;
2. If there are mismatches in the row count;

The exact time when the aforementioned alerts will be sent out must be agreed beforehand.

3. Send alert if there is any server problems (CPU, SQL service);
4. Alerts when exporting reports. As soon as the export reports are populated an email is sent to the client.

Client:

1. Alert if the client have problems connecting to their servers.

A. Alerts upon Failure

The Alerts which will inform of any system failures will be part of a stored procedure in SQL.

Step 1 – Alerts due to system failure

A Step in the data transfer job will be created to inform both parties when the data transfer fails due to the following reasons:

1. System Failure during the Data Transfer Process;
2. Incorrect data request in the request table.

The step will have detailed explanation of why the data transfer failed.

Step 2 - Alerts due to missing data

The procedure will check if the respective data is imported into the transfer tables:

a) table_playertransfer
b) table_playeractivitytransfer
c) table_exchangeratetransfer

Apart from pushing the data into the tables above, the client will also need to import a request type.

Request Types are the following:

a) Daily Transfer
b) Re-import
c) Exception

B. Daily Transfer

An alert will be sent to the client if any of the above tables are empty and the data transfer process will be stopped. The tables will be checked again on the next scheduled job. If the player's table is the only one containing data, the procedure will continue with the import but will inform the client that the other two tables are missing. This is the only case when the customer's table has data.

C. Reimport

An alert will be sent to the client if any of the above tables are empty and the data re-import will not continue based on the transfer type specified in the request table:

a) Player Registration
b) Player Activity
c) Exchange Rate

The procedure will process this request if the above parameters provided are correct.

D. Exceptions

It is important that the client specify the request type of the Exceptions. The procedures will check the transfer tables based on the request parameters passed in the request table and will inform the client if the Exceptions Transfer fails.

VI. MAINTENANCE PLAN

Backup Policy:

1. 1 full backup daily locally on the same server;
2. Transaction log backup every 15 minutes.

Backup retention period: 1 week

The average size of each full backup will be approximately 20 % of the database size. This backup plan may be fine-tuned when the project evolves. Both a base agent and an SQL Server Backup Agent will be used to generate and manage the backup schedules. If a problem is experienced with the database and it is found that the database had been corrupted and the tables are inaccessible, a backup will be restored in line with the SLA agreements.

VII. SERVER SPECIFICATIONS

Server	Virtual Server
Operating System	MS Windows 2012 R2
Database Management System	MS SQL 2012 Standard Edition
Processor	vCPU 4 (4 virtual processors running on a physical hex core)
Memory	8 GB RAM
Storage	500GB The remit of the server is to store 'data in transit' and not to be a permanent storage of historic data so the amounts of data storage required will not be big.
Network	1000 Mbps
Access	Client's staff will be granted an access to the SQL Server instance on the server.
Server Management	The side creating the solution will have the responsibility to maintain the server's uptime and its smooth running.

VIII. RESULTS

This solution will help both businesses optimize their current data processes. The Mid-Tier Server is a customized solution created to answer all clients' requirements. As previously mentioned, with the current data transfer process the client is not able to perform re-imports of data. Re-import is usually performed when the data are not complete or some figures are not correct. This process was done manually in the past by the IT team, by removing the old data from the system and importing the new figures sent from the client. It is a repetitive work that needs to be automated.

Both the client and the provider of the solution agreed that the queuing mechanism will significantly help both IT teams in delivering the data more accurately and having visibility of all the records pushed to the system.

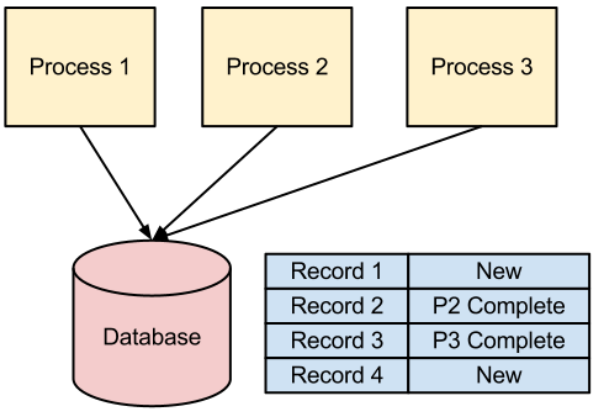


Fig. 2. Data Queuing Mechanism [3]

True process of automation should be deployable to line functions or staff/support functions. Inefficiencies in either area impact the bottom line negatively. Automation can live and deliver benefits in the IT organization or to the business units. Below there is a paragraph where it presents value of automation based on two parameters Frequency of Execution and Complexity of Process.

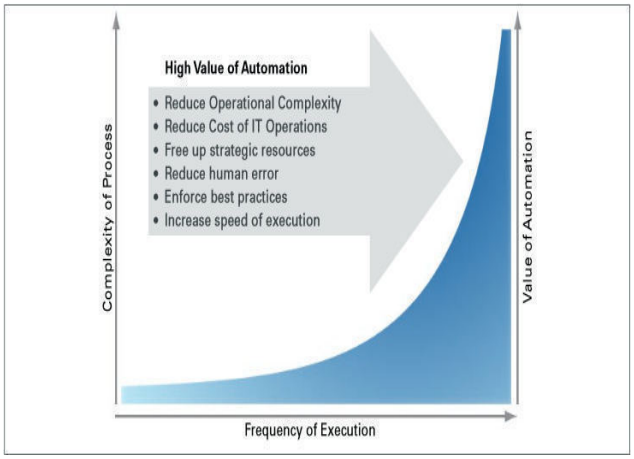


Fig. 3. Increased Value of Automation [4]

IX. CONCLUSION

An essential part of a business is to have its processes in place, to use least of their most crucial resources and provide excellent support for their clients. Businesses are exponentially growing when their clients are satisfied with the support that they receive on a daily basis. It is important to give clients a stable structure and stable processes to follow.

The end users want to order applications in a self-service manner and take delivery within minutes. IT needs to deliver application workloads on-demand in an automated and repeatable manner, eliminating manual provisioning or de-provisioning of resources.

The data transfer automation will increase the speed and agility of the business. The data center is a critical part of a business strategy. The faster it produces results, the more competitive the business can be in a world that values speed. To reach the speed a business need, the data center must be agile – it must be automated. The solution presented in this paper will significantly reduce the manual work of the IT team and will enable full access from client's side. The Middle Layer Server will play a big role in any future modifications of the data center because it is scalable and efficient. The process explained is complex and requires a customized logic in order to please client's requirements. However, it is scalable and other future requested functionalities can be added easily.

X. REFERENCIES

[1] Network Automation, *Process Automation for the Twenty-Ten Decade,* 2010, Available at: http://www.networkautomation.com/, Accessed 6th of March 2015

[2] Arnold Media Group, *Netrefer - Affiliate Program Software,* 2015, Available at: http://www.netrefer.com/, Accessed 6th of March 2015

[3] Network Automation, *Process Automation for the Twenty-Ten Decade,* 2010, Available at: http://www.networkautomation.com/documents/Process-Automation-for-the-2010-Decade.pdf?noRw, Accessed 6th of March 2015

[4] M. Hadlow, *The Database as Queue Anti-Pattern*, 2012, Available at: http://mikehadlow.blogspot.com/2012/04/database-as-queue-anti-pattern.html, Accessed 7th of March 2015

Cloud application for water resources based on open source software

Blagoj Delipetrev PhD Assistant professor
University Goce Delcev
Faculty of Computer Science
Shtip, Republic of Macedonia
Blagoj.delipetrev@ugd.edu.mk

***Abstract*—The cloud application for water resources based on open source software is deployed on a hybrid model of public – private cloud, running on two separate virtual machines (VMs). The first one (VM_1) is operating on Amazon web services (AWS) and the second one (VM_2) is working on a Xen cloud platform. The cloud application is developed using several programming languages (PHP, Ajax, Java, JavaScript), libraries (OpenLayers, JQuery) and open-source software components (GeoServer, PostgreSQL, PostGIS) and OGC standards (WMS, WFS, WFT-T). The cloud application has four web services for 1) data infrastructure (DI), 2) support for water resources modelling (WRM), 3) user management and 4) water resources optimization (WRO). The cloud application is: available all the time, accessible from everywhere, it is scalable, works in a distributed computer environment, it creates a real-time multiuser collaboration platform, the programming languages code and components are interoperable, and it is flexible in including additional components. This research shows the capability to scale and distribute the cloud application between several VMs. The cloud application was successfully tested in the Zletovica case study in a distributed computer environment with concurred multiple access.**

Keywords— Cloud computing; open source; application development component;

I. INTRODUCTION

Presently, most of the existing software for water resources is desktop-based, designed to work on a single computer. Often scientists are complaining about running model time, that can take days, even weeks, making research quite difficult. Another limitation is a multiple user support. The classical desktop application do not support multiple users to work together, sharing data, models, and results. Data and model portability is an issue, and depending on the software type or version, sometimes even the same software on another computer can make problems with data and model portability. The software availability and accessibility are often restricted by licenses that must be used inside the organization building, or the user must be physically there to use the desktop software. These are all major limitations of any desktop software. Which brings us the main research question of this article: Is it possible to develop water resources application that alleviates most of the issues mentioned before that provides seamlessly endless storage and computing power, is available everywhere and anytime, provides multi-user platform, diminish data portability issue, and it is flexible for upgrading the existing and connecting/adding other software components.

The only feasible solution lies in the web and cloud. There are various examples of web applications [1-4], cloud web services [5, 6] and mobile applications [7] in the water domain. Previously mentioned systems often need `only a web browser_ to be used, hiding all implementation details from the end-users. These advantages can significantly increase the number and the diversity of users.

The cloud application for water resources is continuation of previous research [4] that presents the development of a web application for water resources based on open source software. The web application for water resources has three web services. The cloud application has an additional web service and in total has four web services for 1) data infrastructure (DI), 2) support for water resources modelling (WRM), 3) user management, and 4) water resources optimization (WRO). The cloud application enhancement are the following:

- The web application is distributed / deployed on two VMs. The VM_1 is running as a micro instance of Amazon web services (AWS) public cloud, and the VM_2 is running on a Xen cloud platform at the University Goce Delcev in the Republic of Macedonia.
- The web service for support of WRM that runs on VM_1, and the DI web service that runs on VM_2, are communicating with WFS-T (web feature transactional protocol), XML messages over the internet, demonstrating distributed computer environment.
- Additional optimization algorithms nested stochastic dynamic programming (nSDP) and nested reinforcement learning (nRL) are included into the web service WRO, improving its capabilities.
- Clear demonstration of hybrid cloud is presented, where VM_1 is part of the AWS public cloud, and VM_2 is running in the private cloud. The advantage of this distributed computer environment is that the data security and protection can reside in the private cloud (VM_2), while the web services can work in the public cloud (VM_1).

The cloud computing application was tested using data from the Zletovica river basin located in the north-eastern part of the Republic of Macedonia. The test is a proof of concept that this application can be a foundation for a modern cloud based solution. The application url www.delipetrov.com/his/ provides video presentation and explanation of the system components, guides how to use the services etc.

II. Design and Implementation

The cloud application has four web services:

- DI.
- Support of WRM.
- User management.
- WRO

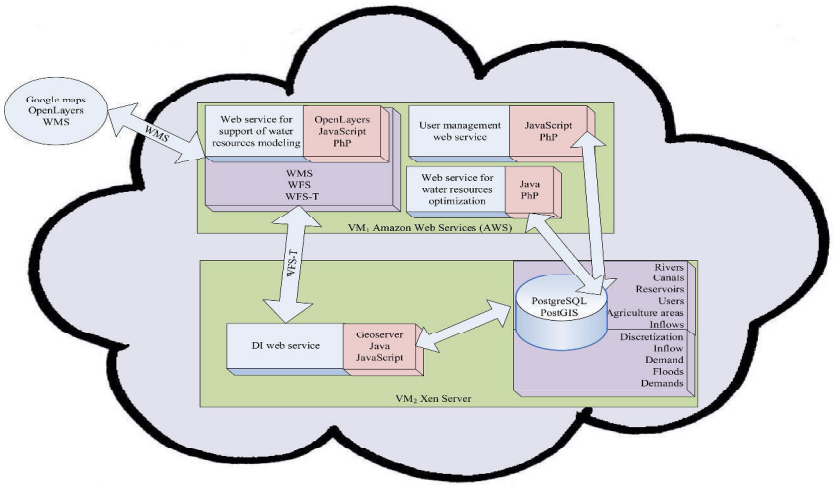


Fig. 1. Design of the cloud computing application for water resources

Fig. 1. shows the web services and the data communication links represented by arrows. The communication between the web service for support of WRM and the DI web service is asynchronous, or on demand.

A. DI web service

The DI web service is composed of two components 1) HMak database created in PostgreSQL and PostGIS and 2) GeoServer application. The HMak database is the central database in this cloud application storing all data. The DI web service is running on VM_2 that resides in the private cloud created by the Xen cloud platform. The HMak stores geospatial data, including topographic, hydro-geological, rivers, roads, municipal, etc. The GeoServer shares the HMak geospatial data on the internet. The HMak stores six geospatial vector layers: rivers, canals, reservoirs, users, inflows, and agriculture land, and their attribute tables, that are used by the web service for support of WRM. Additionally, the HMak stores around 40 time series data tables used by the WRO web service.

GeoServer is an open source server for sharing geospatial data. It is interoperable and can publish data from any major geospatial data sources using open standards. More importantly, GeoServer is as a middle tier application that connects the HMak database on one side and provides OGC web services (WFS-T) on the other side.

B. Web service for support of WRM

The web service for support of WRM interface provides tools to work with the geospatial objects from the six layers described before, that represent water resources components and its infrastructure, shown in Fig. 2. The main difference between the web service for support of WRM and desktop interfaces is that the web service for WRM is accessible by multiple users simultaneously over the internet using a web browser. Generally, the web service for support of WRM is a customized web GIS service designed for water resources.

The web service for support of WRM is developed using PHP, Ajax, JavaScript and the most importantly OpenLayer library. The OpenLayer library support OGC standards: Web Map Services (WMS) and Web Feature Services - Transactional (WFS-T). The web service for support WRM using WMS connects to two different basic map providers: 1) Google Maps and 2) OpenLayers WMS. This allows users to select the background map from the menu as shown in Fig. 2, where OpenLayers WMS is used. Important is that other users can select different background maps e.g. Google Maps without interfering between each other.

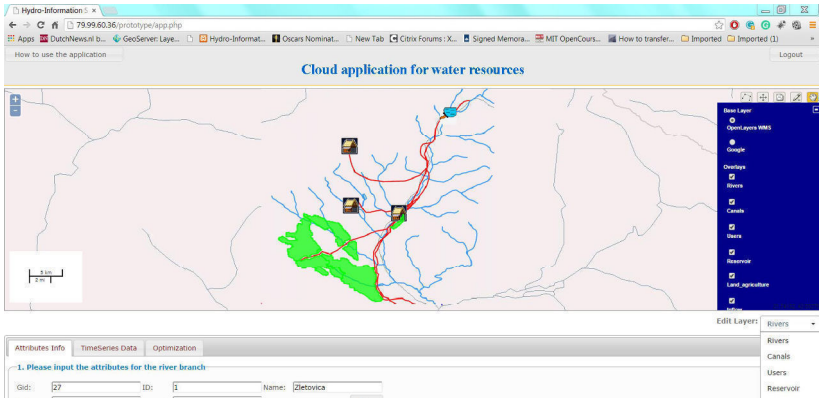


Fig. 2. Cloud application for water resources interface

The OpenLayer library creates WFS-T communication between the web service for support of WRM user interface running on VM1, and the geospatial data stored in HMak where the GeoServer acts like a middle tier running on VM_2. The WFS-T communication provides a framework to create, update, and delete geospatial data over the internet and the two VMs.

C. *Web service for water resources optimization*

The WRO web service implements three algorithms for the optimal reservoir operation named 1) nested dynamic programming (nDP) 2) nested stochastic dynamic programming (nSDP) and 3) nested reinforcement learning (nRL). Further discussion about the algorithms is not explained, because it goes beyond the scope of this article. The three algorithms are coded in Java and, developed as prototype applications.

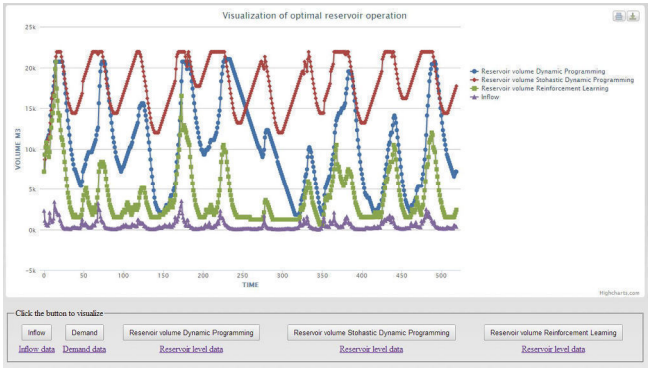


Fig. 3. Optimal reservoir operation using WRO web service

The WRO web service has several components: PHP web form for data input and upload into the HMak database running in VM1, the three prototype Java applications for each algorithm residing in VM_2, and a separate page for results presentations again in VM_1. The web application directory on VM_2 stores the three algorithms executable JAR files. The `Optimization_ tab provides user interface for the WRO web service. The JavaScript library `highcharts.js_ presents the results.

D. *Web service for user management*

The web service for users˘ management is simple with a main purpose to control the cloud computing application access. The user profiles stores the cloud application usage time. Further development of this service will include users˘ computer power and storage usage. Using these information˘s, the administrator can effectively manage the cloud application users.

III. PRELIMINARY RESULTS AND TESTS

Important milestone is the deployment of the cloud application between the two VMs running on separate physical servers. Thet VM_1 is a micro instance on the AWS, and the second VM_2 is running on the Xen cloud platform. The VM_1 has 8 GB HDD, 1 GB RAM and Ubuntu 13 as an operating system. The VM_2 has 30 GB HDD, 1 GB RAM and Fedora 16 as an operating system. The VM_2 is running on a physical server IBM x3400 M3 with four-core Intel Xeon E5620 2.40 GHz with 12 MB of cache per processor. The AWS management console, and the Citrix XenCenter, respectively manage the VM_1 and VM_2. Both management consoles provide tools to control the cloud application and server environment i.e. control the CPU usage, the memory, the disks and the network connections on all of the virtual machines.

Fig. 2 shows the hydro system Zletovica model created by the web service for WRM. The hydro system contains the reservoir Knezevo, river network, canal network, towns as users and agricultural areas. The WRO web service was tested using data from an existing study of the Zletovica river basin. Three nDP, nSDP and nRL datasets were uploaded as CSV files into the HMak using the WRO web service interface. Fig. 3 presents the optimal reservoir storage results of the three algorithms.

The test general conclusion is that the cloud application is stable and functional, supporting multiple users with increased workload. The system performance measures, the workload and other characteristics were not performed, although both AWS management console and Citrix Xen control provides that information. The main test was to demonstrate that the cloud computing application works as designed.

IV. DISCUSSION AND CONCLUSIONS

The cloud computing paradigm `Only a web browser is needed to use the application_ is accomplished. Furthermore, the presented cloud application is deployed on two VMs working in a distributed computer environment. The presented application demonstrates all cloud advantages, such as diminishing concerns about working platform, software versions, data portability, and other implementation details.

Further, the NIST definition [8] of cloud computing is considered to evaluate the presented cloud application. The first two essential characteristics of the cloud application are `on-demand self- service_ and `broad network access._ The cloud application is available and accessible all the time and from anywhere and it only requires a web browser. Moreover, our interaction with the application is on-demand and driven by user needs. The web application is available on any device that has a web browser (e.g. mobile phones, laptops, etc.).

The third and fourth essential cloud application characteristics are the capability for `resources pooling_ and `rapid elasticity._ The basic adjustment concerning the workload can be performed by increasing the current VMs computational power. The VMs workload can be monitored over AWS console and XenCenter and adjust appropriately. The cloud application components, standards and programming languages are interoperable and can be deployed on an unlimited number of servers and connect appropriately. The issues concerning scalability and resource pooling can be resolved by creating many data repositories similar to the HMak deployed on a number of VMs, storing large quantities of geospatial and other types of data. Furthermore, several GeoServer instances can connect to multiple data repositories, creating a giant DI system. The web service for support of WRM can connect to the multiple GeoServer instances, and finally the web services can be replicated on several VMs. Depending on the number of users, the workload, the storage capacities, the processing power, the number of servers available, etc., the optimal cloud application environment can be adapted.

The last essential characteristic of cloud computing is `measured service_ which is rudimentary supported by measuring the time of each user˘s usage of the system. This satisfies cloud computing criteria, but needs to be vastly improved (e.g. with measuring processing power consumption, storage capacity utilization, etc.).

Concerning service models, the presented cloud computing application belong to software as a service (SaaS). Users with a web browser access the cloud application and do not care about underlying cloud infrastructure. The current deployment model is hybrid of public - private cloud because the VM_1 is running in public cloud AWS, and VM_2 is on private Xen cloud platform.

The most valuable feature of the cloud application is its real time collaboration platform capabilities. Multiple users using only a web browser can work jointly with the web services and collaborate in the same working environment in real time. An example is the web service for supporting WRM is when a user saves the current work. After that moment all other distributed users with just refreshing the web browser window can see the change (new/modified rivers, users etc.). All of the data and models are stored in HMak and users do not have to be concerned about hardware and software support infrastructure.

Another important concern about implementing cloud solutions is the data protection and safety. Often many companies and organization dismiss implementation of cloud computing solutions just because their data will be stored somewhere on the internet raising major concerns about its safety and protection. This prototype cloud application makes an elegant solution where services are residing in the public cloud, while the data is stored in the private cloud. If for instance an attack happens on the cloud application, to protect the data the private cloud VM_2 can be disconnected from the public cloud AWS VM_1. Another key point is that data resides inside the institution, and only the web services are `outsourced._ This concept can be applied in many organizations where the data needs to be stored internally.

This research demonstrates that there are available open source software and technologies to create complex and robust cloud application for water resources. The application is a SaaS hybrid cloud solution running on two virtual machines VMs, from which VM_1 runs on AWS and VM_2 on Xen cloud platform. The cloud application was tested in the Zletovica case study and with real-time multiple users that collaborated and jointly modelled and optimized water resources.

The cloud application can be further improved by connecting to new data repositories. Additional modelling, optimization and other decision support services have already been envisioned and can be added to the existing platform, so that it can evolve into a fully cloud based water resources modelling system.

REFERENCES

[1] J. Choi, B. Engel, and R. Farnsworth, "Web-based GIS and spatial decision support system for watershed management," *Journal of Hydroinformatics,* vol. 7, pp. 165-174, 2005.

[2] B. Delipetrev, D. Mihajlov, M. Delipetrev, and T. Delipetrov, "Model of the Hydro-Information System of the Republic of Macedonia," *Journal of Computing and Information Technology,* vol. 18, pp. 201-204, 2010.

[3] J. Horak, A. Orlik, and J. Stromsky, "Web services for distributed and interoperable hydro-information systems," *Hydrology and Earth System Sciences,* vol. 12, pp. 635-644, 2008.

[4] B. Delipetrev, A. Jonoski, and D. P. Solomatine, "Development of a web application for water resources based on open source software," *Computers & Geosciences,* vol. 62, pp. 35-42, 2014.

[5] C. M. Bürger, S. Kollet, J. Schumacher, and D. Bösel, "Introduction of a web service for cloud computing with the integrated hydrologic simulation platform ParFlow," *Computers and Geosciences,* vol. 48, pp. 334-336, 2012.

[6] V. M. Quiroga, I. Popescu, D. Solomatine, and L. Bociort, "Cloud and cluster computing in uncertainty analysis of integrated flood models," *Journal of Hydroinformatics,* vol. 15, pp. 55-70, 2013.

[7] A. Jonoski, A. Almoradie, K. Khan, I. Popescu, and S. Van Andel, "Google Android mobile phone applications for water quality information management," *Journal of Hydroinformatics,* vol. 15, pp. 1137-1149, 2013.

[8] P. Mell and T. Grance, "The NIST definition of cloud computing," *National Institute of Standards and Technology,* vol. 53, p. 50, 2009.

Business Analysis for Hosting an e-Ordering System in the Cloud

Filip Dimitrievski, Sashko Ristov, Marjan Gushev and Goce Armenski
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering
Rugjer Boshkovikj 16, P.O. Box 393,
1000 Skopje, Macedonia
dimitrievski@outlook.com, {sashko.ristov, marjan.gushev, goce.armenski}@finki.ukim.mk

***Abstract* - Until about a decade ago, a lot of companies run their software on computers in the buildings (on premises), rather than at a third party. Recently, the new paradigm – cloud computing has emerged. The cloud removes the need for the user to install any software on premises. Running software remotely can be more cost effective because of the reduced man hours and days for maintenance and administration, power consumption, etc. In order to provide these cost savings, the cloud providers charge their customers on a subscription basis. As a result, companies can spread their budget more efficiently and focus on growing the business. In this paper we conduct a financial analysis whose aim is to compare the capital expense (CAPEX) and operating expense (OPEX) of an e-Ordering system as on-premise solution on the one hand and Software as a Service (SaaS) on the other hand.**

Keywords - eOrders; Business Analysis; Business Model; Cloud Computing;

I. INTRODUCTION

Today, the cloud services are becoming more and more attractive to small and start-up companies that are seeking to reduce licensing costs and IT staff and focus on growing the business [1]. This is the case with the e-Ordering system; a system that allows client companies to order some services from the sellers. Cloud computing is an attractive proposition for companies that are seeking to make the most of the today's technology at lower costs and lower risk. They don't need to invest in software or hardware, and they only pay for what they use. In general, the services in the cloud are organized in three categories [2]: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Each layer of the cloud architecture uses the services of the layers below. The IaaS layer is the lowest layer in the cloud stack, while the SaaS layer is the highest. Going from IaaS to SaaS, the customers have less work and responsibility as the layers go up. Customers subscribed to the IaaS service have improved IT agility and can migrate their applications and data on virtualized servers without the need for network and equipment administration. Virtualization as a method is widely used at the IaaS layer. Unlike IaaS, the PaaS layer is a development platform that allows customers to develop and deploy cloud services and applications on the cloud, without having concerns about maintenance and security. The SaaS layer gets all the benefits that IaaS and PaaS offer as cloud service layers. Customers do not have to deploy their own software nor employ IT staff for administration. They can subscribe to any SaaS service. Most common services sold as SaaS are email applications, CRMs, ERPs, TMSs, etc. Unlike the consumers, the service providers are constantly managing their resources. Most common IaaS providers are: Amazon AWS, Google Compute Engine, Microsoft Windows Azure, Rackspace Open Cloud, etc. The PaaS providers have their own resources, while some of them rent resources from other IaaS providers. Most common PaaS providers are: Amazon AWS, Engine Yard, Google App Engine, Heroku, Red Hat OpenShift, etc. The SaaS providers also have their own resources and when necessary they rent resources from other IaaS or PaaS providers. Today, the SaaS providers are specialized in almost every field, from productivity and project management, to security, monitoring, customer service, help desk and more. A typical deployment of a SaaS application does not require any hardware, because the vendor provides the entire infrastructure. The SaaS model is designed to deliver applications anywhere and anytime. To do this continuously, globally and securely, the vendors need to have reserve capacity in order to handle any spikes in usage, outages or network mishaps, and they also need to employ support teams that are available to customers on short notice.

There are three key cost drivers for any software implementation; the cost of the software application, the hardware required to run the application; and the people required to design, develop, deploy and maintain the application. The pricing of traditional software is limited to the cost of the software application and it is up to the customer to determine the cost of the hardware and cost of the people services. While these applications are based on a model with large upfront licensing costs and annual costs for upgrades and support, the hosted applications are based on recurring subscription fees. A traditional on-premise solution requires backup, networking, hardware and server deployments in order to accommodate the number of users on and off campus. An increased number of users may raise the customer's costs due to the need for additional hardware deployments and servers. These types of software applications tend to be highly customizable and they come at a price. The customer is responsible for providing maintenance, management and security to the application, as well as support and training for the end-users. On the other hand, hosted applications are based

on a recurring subscription fees and usually are based on a monthly or annual fee for using the service. The service fee usually includes maintenance, support, training and upgrades and is inclusive of all hardware, networking, storage, database, administration and other costs. The cost may increase as the usage of the application increases and vice versa. This depends on the number of users, number of transactions, etc. As a result, a hosted application is much friendlier to the company's growth model when compared to on-premise solution.

II. Case Study

Almost every SaaS provider rents virtual machines (VMs) from other IaaS providers [3], usually on hourly basis. The cost for a VM is calculated by the hours the rented VM(s) has been activated. The revenue of the SaaS application, in our case the e-Ordering system, depends on the number of users and transactions made. Most of the services provided by the sellers of the e-Ordering system allow client companies to make orders only at the end of the week, with exceptions of some sellers. So the system has increase in the number of orders (users and transactions), every Thursday and Friday, as presented in Fig. 1. Since both, renting the VMs and selling the e-Ordering service to the client companies, are happening at the exact time of the week, the company must create a business model in order to maximize the profit and grow its business. This will save the company's capital and operating expenses, and also resources needed for administration and management [4]. This study will contain a detailed analysis of the current pricing models for renting and selling resources on both markets, for cloud and on-premise as well. After a successful analysis, we expect to choose the most cost effective model for the solution.

Further, we plan to divide the e-Ordering system on two subsystems, i.e. static and dynamic. The first one will be active during a normal or a reduced system usage, while the other will be active at higher peaks of usage [5]. As mentioned above, when the system has increase in the number of orders, every Thursday and Friday, the dynamic subsystem will be active in order to scale the resources in the cloud, either to utilize more smaller VMs (with less resources, or scale horizontaly) or less but greater VMs (with more resources, or scale vertically) [6].

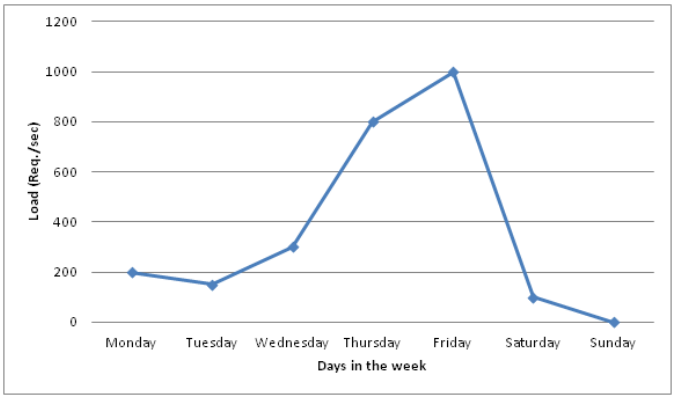


Fig. 1. Expected weekly distribution of number of orders.

A number of challenges of how to become a profitable provider by using the resources are discussed by Schwanengel and Hohenstein [7]. Few papers reported that the price for scaling the resourcses in a VM is linear [8, 9], i.e. using one VM with two cores is two times more expensive than one with one core, and so on. Later in this study, we conduct a business analysis where we examine the feasibility of the proposed architecture.

III. Packing Algorithm

The e-Ordering system is designed to handle a large number of small and start-up companies [17]. The number of employees those companies have is usually up to 50, but there are exceptions. Some companies can even have up to hundred of registered employees. We need to group as more companies as we can and put them on a single server or instance. We are assuming that a compute optimized server can handle a hundred of users. Most cloud providers (CPs) offer servers with 1, 2, 4 and 8 cores. If one CPU can handle a total of 20 users, then a server with 8 cores can handle 8 times more users than a server with one core, thus handling a total of 160 users. In order to open the e-Ordering service to the consumer companies (every Thursday and Friday), we need to deploy N number of servers that are optimized to fit and serve all companies. To solve this packing problem, we suggest a greedy algorithm.

First fit is a greedy algorithm that processes items in arbitrary order. For each item or company, the algorithm is going through all the bins, in our case instances, until it finds a space that is large enough to place the iterated item. It puts the item in the first instance that can accommodate that item. It doesn't matter if the space is a bit larger than the company's size. We just want to allocate the first space that the item can fit. If no bin can accommodate the item, then a new one is opened and the item is placed within the newly created instance. The following paragraphs present an example.

Here is a list of 20 items waiting to be packed by the first fit algorithm: 12, 20, 43, 7, 10, 38, 55, 60, 5, 10, 25, 40, 57, 53, 50, 8, 15, 23, 20 and 16.

Each item represents a size of a given company or a number of users that are allowed by that company to use the e-Ordering service on Thursdays and Fridays. Using (1) we can find out the number of instances needed to be deployed.

$$\frac{\sum_{i=1}^{n} ai}{V} \quad (1)$$

First, we need to sum the sizes of all the companies (a_1, a_2, …, a_{20}) and then divide the total by the size of the servers V. Previously, we mentioned that we want a greedy algorithm, that is, to place as more companies as we can on a single server. Therefore, knowing that a single server with 8 virtual cores can handle hundred of users, we let V to be 160. The outcome of (1) rounds up to 4 (3.54), thus giving us the exact number of instances needed to be deployed in order to accommodate all the consumer companies from the list. In order to prove the outcome result, we need to run the previously suggested greedy algorithm - first fit. The visual placement of all items and bins is given in Fig. 2.

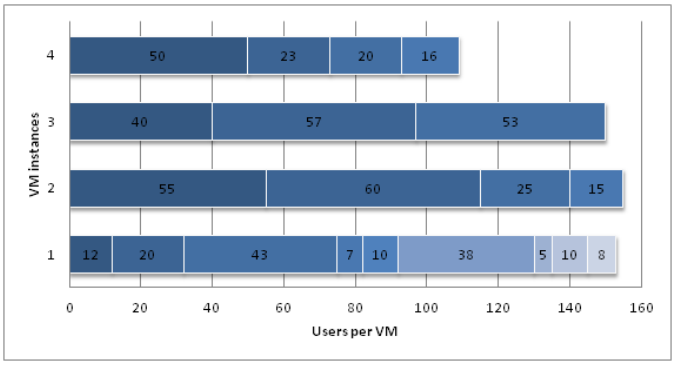


Fig. 2. Packing companies with the first fit algorithm.

The previous outcome result is confirmed, i.e. the algorithm opened as much bins as it was calculated by (1). After defining the consolidation algorithm, the next section explains the costs for renting VM instances from the top three CPs Amazon [10], Google [11] and Microsoft [12].

IV. Cloud Hosting

The CPs offer different types of VM instances to the consumers. That is, VM instances can be optimized for computing, memory, storage or GPU processing. Each CP offers different configurations for its VMs in terms of CPU, RAM, storage, operating system and hosted region.

Table I presents the current prices for renting VM instances from three most common CPs, indicated in USD per hour. The prices are strictly for the EU region, and for both Linux and Windows platforms. Also, we present only two types of VM instances: VMs that are optimized for computing and VMs that are used for general purpose. Since we need a computing power to serve all the companies that are packed in a single VM, we are not presenting the pricing models of GPU processing, memory and storage optimized machines. The presented machines have up to 8 virtual cores, 30GB of RAM and 400GB of storage.

Working with the price lists for renting VMs is surprisingly complex. We cannot simply choose the cheapest VM instance and feel reliable about it. In the following subsections we discuss all VMs that are presented in Table I.

A. *Microsoft Azure*

Microsoft Azure offers few series of VMs, but we have taken only those from the A and the D series. We present 4 VMs from the basic tier of the economical A-series and 4 VMs from the D-series. The A-series VMs range in sizes from extra small (A0) to extra large (A4), while the more powerful D-series has VMs that range from small (D1) to extra large (D4). The A-series VMs are not meant for memory-intense applications. They feature hard disk drives (HDD), while the D-series machines feature solid state drives (SSD). The D-series VMs have double RAM memory and 60% faster processors than the A-series and they are ideal for applications that demand faster processors or higher memories. The amount of computing power and RAM memory in all Azure VMs doubles at each stepping size.

TABLE I. Prices for VMs

	Virtual Machine	Virtual Cores	RAM (GB)	Storage (GB)	Price per Hour	
					Linux	*Windows*
Amazon EC2	m3.medium	1	3.75	4	$0.083	$0.146
	m3.large	2	7.5	32	$0.166	$0.292
	m3.xlarge	4	15	80	$0.332	$0.584
	m3.2xlarge	8	30	160	$0.665	$1.169
	c3.large	2	3.75	32	$0.129	$0.212
	c3.xlarge	4	7.5	80	$0.258	$0.424
	c3.2xlarge	8	15	160	$0.516	$0.848
Google Compute	n1-standard-1	1	3.75	/	$0.053 + $0.11	$0.053 + $0.04
	n1-standard-2	2	7.5	/	$0.106 + $0.11	$0.106 + $0.08
	n1-standard-4	4	15	/	$0.211 + $0.11	$0.211 + $0.16
	n1-standard-8	8	30	/	$0.421 + $0.11	$0.421 + $0.32
	n1-highcpu-2	2	1.8	/	$0.066 + $0.11	$0.066 + $0.08
	n1-highcpu-4	4	3.6	/	$0.132 + $0.11	$0.132 + $0.16
	n1-highcpu-8	8	7.2	/	$0.263 + $0.11	$0.263 + $0.32
Microsoft Azure	A1	1	1.75	40	$0.047	$0.075
	A2	2	3.5	60	$0.094	$0.15
	A3	4	7	120	$0.188	$0.30
	A4	8	14	240	$0.376	$0.60
	D1	1	3.5	50	$0.094	$0.159
	D2	2	7	100	$0.188	$0.318
	D3	4	14	200	$0.376	$0.636
	D4	8	28	400	$0.752	$1.272

In Table I, we excluded Microsoft's smallest A-series VM (A0) with 0.75GB of RAM. The A1 VM has 1 vCPU, 1.75GB of RAM and costs $0.047 per hour for Linux and $0.075 per hour for Windows. The biggest A-series VM (A4) has 8 vCPUs, 14GB of RAM and costs $0.376 per hour for Linux and $0.60 per hour for Windows. One of the interesting facts is that Microsoft charges almost a double for a Windows VM. Apart from the A-series, we present four more powerful VMs from the D-series. The smallest D-series VM (D1) has 1 vCPU, 3.5GB of RAM and costs $0.094 per hour for Linux and $0.159 per hour for Windows. The biggest D-series VM (D4) has 8 vCPUs, 28GB of RAM and costs $0.752 per hour for Linux and $1.272 per hour for Windows, making it the most expensive VM in our selection.

B. *Google Compute*

Google follows the same mechanism for doubling the price while doubling the size of the VM. Via its Compute Engine service, Google offers standard VMs, as well as VMs with higher and lower amounts of RAM and CPU cores. We present 4 standard and 3 highcpu VMs from Google's Compute Engine service. The number of cores these VMs have varies from 1 to 8. The smallest standard VM (n1-standard-1) has 1 vCPU, 3.75GB of RAM and costs $0.053, while the biggest (n1-standard-8) has 8 vCPUs, 30GB of RAM and costs $0.421. Unlike the standard VMs, the highcpu VMs have more powerful processors and lower amounts of RAM. The smallest (n1-highcpu-2) has 2 vCPUs, 1.8GB of RAM and costs $0.066, while the biggest (n1-highcpu-8) has 8 vCPUs, 7.2GB or RAM and costs $0.263.

Google has special pricing policy and charges additional $0.11 per hour for Linux based VMs that are running a SUSE operating system. For example, the total price of using an "n1-highcpu-8" instance on Linux would be the sum of VM's cost and SUSE's cost $0.373 per hour. On the other hand, Windows VMs are charged based on the number of virtual cores, i.e. Google charges additional $0.04 per core/hour. For example, the total price of using the same "n1-highcpu-8" instance on Windows platform would be $0.583 per hour.

If a VM is used for a significant portion of a month, the consumer can get a monthly discount for every hour used for that instance. For example, a discount is applied when an instance is used for more than 25% of a month. In our case, the VM instances will be used for almost 30% of a month or 2/7 of a week, i.e. every Thursday and Friday. Table I already has these discount prices for Google's Compute Engine service. The discount increases with usage of the instance. The consumer can get up to a 30% discount for instances that are used a 100% of the month.

As Table I shows, all Google's VMs do not have storage. Google offers several types of disk for its VMs and charges its consumers for the amount of provisioned space per disk. I/O operations are included in the charges for provisioned space. The drives are available as both standard hard-disk drives (HDD) and solid-state drives (SSD). They can be attached and detached to any VM instance in the same zone. The cost for HDD is $0.04 per GB/month, while the cost for SSD is $0.17 per GB/month. Therefore, 40GB HDD would cost $1.6 for the whole month, while 40GB SSD would cost $6.8 for the whole month. If the consumer provisioned 40GB for half a month, then the disk would cost half the price, i.e. 40GB HDD would cost $0.8 and 40GB SSD would cost $3.4 for half a month.

C. Amazon EC2

The most interesting options come from Amazon's EC2 service. Amazon offers various number of VMs and pricing options. Same as the previous two providers, Amazon charges double for twice as much capacity. Amazon via its EC2 service offers on-demand and reserved instances. The latter can be pre-purchased for 1 or 3 years with a significant discount compared to on-demand pricing. A reserved instance can be purchased either by paying an upfront payment (All Upfront or Partial Upfront) or not paying at all (No Upfront). The All Upfront option provides the largest discount for a reserved instance, while the other two provide a discounted hourly rate for the duration of the reserved term. By purchasing a reserved instance, a consumer has the freedom to turn the VM ON and OFF according to its needs. This model has sense if the VM needs to be active the entire time for a period of 1 or 3 years, thus helping consumers to save more money.

As mentioned earlier, Amazon offers on-demand instances, as well. There are various types of VM instances (T2, M3, C4, C3, R3, I2), but our research needs only M3 and C3. The M3 instances are general purpose VMs, while the C3 instances are compute optimized VMs. Both types have solid-state drives. In Table I, we present four M3 VMs and three C3 VMs. The M3 VMs range in sizes from medium to extra extra large (2xlarge). The M3 medium has 1 vCPU, 3.75GB of RAM, 4GB SSD and costs $0.083 per hour for Linux and $0.146 per hour for Windows. The M3 2xlarge has 8 vCPUs, 30GB of RAM, 160GB SSD and costs $0.665 per hour for Linux and $1.169 per hour for Windows. Amazon, same as Microsoft, charges almost a double for a VM that is running a Windows operating system. The C3 VMs are less expensive and have lower amounts of RAM memory. They start with the large model. The C3 large has 2 vCPUs, 3.75GB of RAM, 32GB SSD and costs $0.129 per hour for Linux and $0.212 per hour for Windows. The C3 2xlarge has 8 vCPUs, 15GB of RAM, 160GB SSD and costs $0.516 per hour for Linux and $0.848 per hour for Windows.

D. Data transfer

The price of the VM is just part of the customer's total cost. Another important factor is the data transfer, which refers to incoming and outgoing network traffic towards customer's VMs. All surveyed CPs charge only outgoing traffic.

As presented in Table II, Google is the most expensive CP for network egress. Both Amazon and Microsoft offer free monthly usage for the first 1GB and 5GB, respectively. On the other hand, Google charges $0.12 per GB for the first TB. For monthly usage up to 10TB, Amazon charges $0.09 per GB, Google charges $0.11 per GB, and Microsoft charges $0.087 per GB.

The outgoing data transfer can have a significant impact on the costs in 3 years. Therefore, we must calculate how much outgoing traffic per month will be transferred from the CPs' data centers. When *N* companies (buyers) are grouped on a single VM, it is normally to expect a minimum number of *N* sellers, because the buyers usually order services from multiple sellers. Thus, we set the number of sellers per VM (NS) to be 10. Most sellers offer multiple menu items; some offer more and others less. Sellers like supermarkets offer hundreds of items, but local fast food restaurants offer several. Thus, we set the number of items per seller (NI) to be 50 and the size of each item (S) to be 0.5MB (including text and images). In average, users will probably review only a quarter of the menu, so we let the user's view rate (VR) to be 0.25. In Section 3 we concluded that a single VM can accommodate maximum of 160 users, so let the number of users per VM (NU) to be 160. The system will be active in cloud only on Thursdays and Fridays, i.e. 2 days in a week or 8 days in a month. Thus, we set the number of active days per VM in month (ND) to be 8.

TABLE II. PRICES FOR OUTGOING DATA TRANSFER

	Monthly Usage	**Price per GB**
Amazon EC2	First 1 GB	$0.00
	1 GB - 10 TB	$0.09
	10 - 50 TB	$0.085
	50 - 150 TB	$0.07
	150 - 500 TB	$0.05
Google Compute	0-1 TB	$0.12
	1-10 TB	$0.11
	10+ TB	$0.08
Microsoft Azure	First 5 GB	$0.00
	5 GB - 10 TB	$0.087
	10 - 50 TB	$0.083
	50 - 150 TB	$0.07
	150 - 500 TB	$0.05

In order to determine the value, (2) defines the total monthly outgoing traffic *MOG*.

$$MOG = NS*NI*S*VR*NU*ND \quad (2)$$

Using (2) we obtained *MOG* = 78GB per month

Therefore, a rough calculation for data transfer of 78GB per month for a single VM will cost around $7 per month, $84 for 1 year and $252 for 3 years. If we multiply the total cost by 4 (number of VMs in Fig. 2), we will get a total of $1008.

V. Choosing a Hosting Solution

In this section, we present the most optimal VM configuration of Table I, and offer self hosting solution for the dynamic part of the e-Ordering system. We conclude the analysis by choosing the cheapest solution for hosting the dynamic subsystem.

A. Most optimal VM configuration

From the previous sections, we concluded that we need a bigger VM in order to handle a larger number of users. Therefore, the most suitable solution for customers' needs will be a VM with 8 virtual cores. Table I contains 6 VM instances that have 8 vCPUs, i.e. 2 VMs by each CP.

Amazon offers two models with 8 cores: m3.2xlarge and c3.2xlarge. The M3 costs $0.665 per hour for Linux and $1.169 per hour for Windows, while the C3 costs $0.516 per hour for Linux and $0.848 per hour for Windows. The M3 VM costs more because it has twice more RAM than the C3. But we aren't in pursuit of a memory. Instead, we need better computing power and CPU performance, like the one C3 has: a higher frequency Intel Xeon processor E5-2680 v2 with 25M Cache and 2.80 GHz versus the M3's Intel Xeon processor E5-2670 v2 with 25M Cache and 2.50 GHz.

Google also offers two 8 core VMs: n1-standard-8 and n1-highcpu-8. The former costs $0.531 per hour for Linux and $0.741 per hour for Windows, while the latter costs $0.373 per hour for Linux and $0.583 per hour for Windows. The n1-series are equipped with all three generations of Intel Xeon E5 processors [13], i.e. they either have a 2.6 GHz E5 processor (Sandy Bridge), a 2.5 GHz E5 v2 processor (Ivy Bridge) or a 2.3 GHz E5 v3 processor (Haswell). These VMs only look cheaper than Amazon's because they aren't attached with storage. As we mentioned in Section 4, the storage cost is calculated based on the amount of stored data. Therefore, we omit Google as an option.

Like the previous two providers, Microsoft has also two VMs with 8 cores: A4 and D4. The A4 VM costs $0.376 per hour for Linux and $0.60 per hour for Windows, while the D4 costs $0.752 per hour for Linux and $1.272 per hour for Windows. They are both powerful and both can be analyzed. The A4 VM is much cheaper than the D4 VM. The differences between them are in the amount of memory and storage. The D4 VM has twice more memory and storage than A4, i.e. 28GB of RAM and 400GB SSD versus 14GB of RAM and 240GB HDD. Therefore, we choose A4 (Linux) as an optimal solution for hosting the e-Ordering system because of its computing power and price.

TABLE III. Total Costs for one Azure's A4 VM

Year	Server	Data Transfer	Other Costs	Total
Year 1	$956.54	$76.21	$0.00	$1,032.75
Year 2	$956.54	$76.21	$0.00	$1,032.75
Year 3	$956.54	$76.21	$0.00	$1,032.75
			TOTAL:	**$3,098.25**

A single VM will be active 24 hours a day, and 106 days in a year (Thursdays and Fridays), thus costing $9.02 per day and $956.54 per year. In the previous section, we concluded that the total data transfer per VM is 78GB per month, as well as that Microsoft does not charge for the first 5GB. Therefore, 73GB of data transfer cost $6.35 per month and $76.21 per year. Support and I/O operations are free of charge and are included in "Other Costs" column in Table III. The total cost for renting Microsoft's A4 VM for 1 year is $1,032.75 and $3,098.25 for 3 years. Therefore, renting four VMs for 3 years will cost the e-Ordering company $12,393.00.

B. Self hosting (on-premise) solution

The consumer can choose between tower, rack, or blade server for self-hosting solution. Tower servers look more like desktop computers than servers, so they can be positioned on or underneath a desk. They offer a basic level of performance and cost same as desktops. Unlike tower servers, rack servers are designed to be positioned in data centers, stacked on top of each other. The benefit of having all components located in one place makes the system maintenance much easier. Rack servers are very expandable, thus they have more disks, more processors and large amount of RAM. Their size is measured in Us (units), from 1U to 8U. The latest development in the history of servers is the blade servers. Blade servers have great processing power and take up less space. Also, they share their hardware with other blade servers in the chassis, thus providing cost reduction and efficiency.

Choosing the tower servers is the best solution only if one or two servers are needed. We choose the rack type of server because it is cheaper than the blade and meet our needs about massive scalability. Because all three CPs are equipped with Intel Xeon E5 family of processors, we propose few similar hosting solutions from HP [14], Dell [15] and IBM [16].

From Table IV, we choose Dell PowerEdge R430, because it is the cheapest rack server with 8 cores. The lowest cost for the server is $2,870.00. We also need to take into account the costs for electric power and maintaining the equipment:

3 years * 106 days * 24 hours * 4kW * $0.1 = $3,052.8

The average power consumed by a rack server is about 2kW, so the additional 2kW are for cooling. The server will be active 24 hours a day, and 318 days in 3 years, thus costing $3,052.8. By far, the R430 costs $5,922.8 and we have not added the expenses for maintenance yet, i.e. technical support, employment, etc. In addition, if a single core of a VM is equal to a single thread of a server, then deploying 2 VMs on a single rack server with 8 cores will not be a problem, because the R430 has 2 threads per core. Therefore, to accommodate all users from Fig. 2 we need 2 PowerEdge R430 servers from Dell, so the cost will double to $11,845.6.

TABLE IV. PRICES FOR RACK SERVERS

Server	CPU	RAM	Price
HP ProLiant DL160 Gen9	Intel Xeon E5-2603 v3 (6 core, 1.6 GHz, 15MB)	8GB	$1,719.00
HP ProLiant DL160 Gen9	Intel Xeon E5-2609 v3 (6 core, 1.9 GHz, 15MB)	16GB	$2,569.00
HP ProLiant DL180 Gen9	Intel Xeon E5-2603 v3 (6 core, 1.6 GHz, 15MB)	8GB	$1,829.00
HP ProLiant DL180 Gen9	Intel Xeon E5-2609 v3 (6 core, 1.9 GHz, 15MB)	8GB	$2,309.00
HP ProLiant DL360 Gen9	Intel Xeon E5-2603 v3 (6 core, 1.6 GHz, 15MB)	8GB	$2,799.00
HP ProLiant DL360 Gen9	Intel Xeon E5-2630 v3 (8 core, 2.4 GHz, 20MB)	16GB	$4,059.00
HP ProLiant DL380 Gen9	Intel Xeon E5-2609 v3 (6 core, 1.9 GHz, 15MB)	8GB	$2,819.00
HP ProLiant DL380 Gen9	Intel Xeon E5-2620 v3 (6 core, 2.4 GHz, 15MB)	16GB	$3,889.00
Dell PowerEdge R430	Intel Xeon E5-2630 v3 (8 core, 2.4GHz,20MB)	8GB	$2,870.00
Dell PowerEdge R730	Intel Xeon E5-2609 v3 (6 core, 1.9 GHz, 15MB)	8GB	$3,420.00
Dell PowerEdge R730	Intel Xeon E5-2630 v3 (8 core, 2.4 GHz, 20MB)	16GB	$4,760.00
Dell PowerEdge R730	Intel Xeon E5-2650 v3 (10 core, 2.3 GHz, 25MB)	16GB	$7,900.00
IBM System x3550 M5	Intel Xeon E5-2609 v3 (6 core, 1.9 GHz, 15MB)	8GB	$2,722.00
IBM System x3550 M5	Intel Xeon E5-2630 v3 (8 core, 2.4 GHz, 20MB)	16GB	$3,475.00
IBM System x3650 M5	Intel Xeon E5-2630 v3 (8 core, 2.4 GHz, 20MB)	8GB	$3,286.00
IBM System x3650 M5	Intel Xeon E5-2650 v3 (10 core, 2.3 GHz, 25MB)	16GB	$4,191.00

Finally, we can add the expenses for maintenance. The cost for employing a server administrator for 3 years is $18,000.00, assuming the average monthly salary of a server administrator in our country is $500. Thus, the total cost for on-premise solution is $29,845.6. Therefore, we can conclude that the self hosting solution is not very cost effective, because it costs twice more than the cloud hosting solution which turned out to be $12,393.00 in the previous subsection.

VI. CONCLUSION AND FUTURE WORK

Small and start-up companies are often faced to choose whether they should go for a hosted or an on-premise solution. There is really no right or wrong choice. The main idea is that not all solutions work best in the cloud or on-premise. The choice often requires a trade between costs and security to optimize returns. Costs are the primary concern for a company, but we cannot ignore the issues related to security and compliance. The cloud often appears to be the best choice for cost savings, but most security products are best implemented on premise and there are cases where an on-premise solution is more beneficial in a long-term. We need to choose what makes business, security and sense from technical point of view. The cloud services are most attractive to those consumers that need big blocks of computing power for short period of time. If the VMs are truly used for short period, then pay as you go model can cut the company's costs significantly. In our case, the workload of the e-Ordering system is in short bursts, so we do not need severs that just sit there and consume electricity.

In this paper, we analyzed the public cloud computing market and determined that the most appropriate VM is the Azure's A4 VM as the most optimal VM configuration from the CPs to host the system in cloud. Additionally, the analysis showed that Dell's PowerEdge R430 is the most suitable rack server as an on-premise solution. Comparing the costs of both solutions, the evaluation showed that hosting the dynamic part of the e-Ordering system every Thursday and Friday for 3 years is at least twice cheaper in cloud than on-premise. This conclusion leads towards designing such a scalable solution in the cloud and prove its scalability.

ACKNOWLEDGMENT

The work presented in this paper was partially financed by the Faculty of Computer Science and Engineering, "Ss. Cyril and Methodius" University, Skopje, Macedonia.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. (2010). "A view of cloud computing," *ACM Comm.*, 53(4), Apr 2010, pp. 50-58.

[2] NIST. The NIST Definition of Cloud Computing, SP 800-145, Apr. 2015. [Online]. Available: http://csrc.nist.gov/publications/.

[3] Buyya, R., Yeo, C. S., and Venugopal, S. (2008). Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. *High Performance Computing and Communications*, 2008. *HPCC'08. 10th IEEE Int. Conference on*, pp. 5-13.

[4] Rana, O. (2014). "The costs of cloud migration," *Cloud Computing, IEEE, vol. 1, no. 1,* pp. 62–65.

[5] Ristov, S.; Gusev, M.; Armenski, G.; Bozinoski, K.; Velkoski, G., "Architecture and organization of e-Assessment cloud solution," *IEEE EDUCON 2013*, pp.736-743, 13-15 March 2013, best paper award.

[6] Gusev, M., Ristov, S., Koteska, B., Velkoski, G. (2014). "Windows Azure: Resource Organization Performance Analysis", in *Service-Oriented and Cloud Computing*, Villari, Zimmermann, and Lau (eds), LNCS, Springer Berlin Heidelberg, Manchester, UK, pp. 17-31.

[7] Schwanengel, A., and Hohenstein, U., (2013). "Challenges with Tenant-Specific Cost Determination in Multi-Tenant Applications," in *Proceedings of CLOUD COMPUTING 2013 : The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization,* pp. 85–95.

[8] Simjanoska, M., Gusev, M., Ristov S., and Velkoski, G. (2013a). "Scaling the Performance and Cost for Elastic Cloud Web Services," *J. of Computing and Information Technology, vol 21, no. 2,* pp. 85–95.

[9] Simjanoska, M., Ristov, S., Velkoski, G., Gusev, M., (2013b). "Scaling the performance and cost while scaling the load and resources in the cloud," *36th Int. Convention MIPRO 2013*, pp.151-156.

[10] Amazon, EC2 Pricing, Apr. 2015. [Online]. Available: http://aws.amazon.com/ec2/pricing/.

[11] Google, Compute Engine Pricing, Apr. 2015. [Online]. Available: https://cloud.google.com/compute/#pricing.

[12] Microsoft, Azure Pricing, Apr. 2015. [Online]. Available: http://azure.microsoft.com/en-us/pricing/.

[13] Intel Xeon Processor E5 Family, Apr. 2015. [Online]. Available: http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e5-family.html.

[14] HP Rack and Tower Servers, Apr. 2015. [Online]. Available: http://www8.hp.com/us/en/products/servers/proliant-servers.html#tab=TAB2.

[15] Dell PowerEdge Rack Servers, Apr. 2015. [Online]. Available: http://www.dell.com/uk/business/p/poweredge-rack-servers.

[16] IBM Rack Servers, Apr. 2015. [Online]. Available: http://www-03.ibm.com/systems/x/hardware/rack/index.html.

[17] Dimitrievski, F., Ristov, S., Gusev, M., Armenski, G.. "Scalable Architecture of e-Ordering System in Cloud," in Proc. of *38th International Convention MIPRO 2015*, Croatia, 2015, 1817-1822.

Overview of Cloud Computing Security Standards and Challenges

Elena Markoska, Sasko Ristov, Katerina Zdravkova, Marjan Gusev
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
Email: elena.markoska@live.com, {sashko.ristov, katerina.zdravkova, marjan.gushev}@finki.ukim.mk,

***Abstract*—Cloud security is one of the primary concerns that cloud customers have when it comes to cloud computing. Not only are there security risks present within every Cloud system, but there are also numerous risks which arise with the implementation of various components, as well as solutions of security risks. Multiple efforts have been made to produce standards which provide controls which promote security. In this paper we provide an overview of the security risks within the Cloud, as well as some of the efforts at standardization and guidance.**

***Keywords*-Cloud Computing, Security risks, Cloud security, Standardization.**

I. Introduction

With the increased usage of cloud technologies and their infiltration in the technological advancement come many advantages which exploit the cloud as a technology, but also numerous setbacks which have become a point of concern for both the cloud providers as well as the customers [1].

As a technology that offers redundancy, scalability, performance and multiple services, nowadays, cloud computing is used by smaller-scale companies and organizations instead of only the world IT leaders. The security risks that the users previously had to eliminate are now increased with the allowance of multiple tenants, virtualization as well as data outsourcing that the cloud technologies offer [2]. In order to overcome the security challenges that cloud computing imposes, numerous international standards have been created [3]. This paper aims to provide a general overview of these standards and assess whether they cover the majority of security threats that come with cloud computing. Fig. 1 represents a general overview of the types of standards that concern the Cloud service providers.

II. Security challenges in the cloud

The numerous benefits cloud computing offers come with a set of drawbacks that are to be assessed and resolved. Detaching an application from its native environment and migrating it into a new, cloud based environment most often places the service into an environment which most often is not ideal when observed from the aspect of security. This kind of migration and outsourcing can lead to the increased risks and the rise of new ones. In addition, the introduction of multiple tenants within a single cloud brings increased vulnerability, especially in the public cloud [4]. Below is an overview of some of the main risks that cloud computing brings, with the single aim to identify them and attempt to offer solutions.

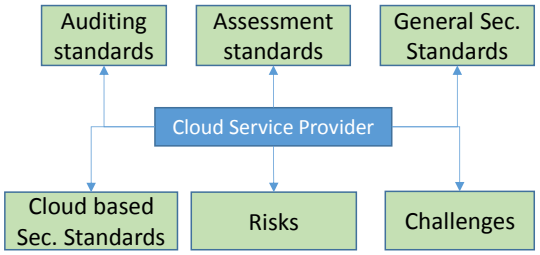


Fig. 1. Cloud application architecture.

A. Multi-tenancy

The appearance of multi-tenancy is incredibly frequent in cloud computing. While in most cases this is a principle that offers many benefits, the threat of malicious insiders is always present. More often than not, cloud service providers do not clearly reveal how a tenants access is granted within the physical or virtual environment, or how their behavior is monitored. This threat additionally increases if tenants are hosted on the same piece of hardware. This kind of circumstances offer great opportunities for the adversary [5] .

B. Heterogeneous environment in interoperability

The migration of large and complex systems within the cloud can often pose more threats than benefits. These systems are most commonly made up of a large number of subsystems which require individual migration within the cloud environment. Due to this situation, the number of subsequent interactions between the subsystems increases, which in turn adds to the overall vulnerability of the system [6]. The aforementioned situation can be remedied with the usage of predefined security standards which enforce adapters or wrappers, or even the usage of various data transformations. The cloud

service provider must meet international standards that a cloud customer needs [7].

C. Loss of data

There are numerous ways in which loss of data can occur. Data breaches are most often the result of a malicious and intrusive actions. Still, often data can be lost because of human error or hardware failure. Accidental deletion or alteration of data, without the usage of backups may lead to permanent loss of valuable information. In addition, the loss of cryptographic keys used for either encryption or decryption of data may lead to the inability to access data. This also does not exclude the possibility of hardware failure, which can be detrimental to a system if redundancy has not been achieved to a sufficient extent [5] [8].

D. Loss of control

The companies which have the role of Cloud customers must transfer a part of the control over their assets to the Cloud service provider. Therefore, the Cloud service provider must meet a service-level agreement (SLA), where the extent to which each bit of allowed access and control is clearly stated and formally defined. If a Cloud service provider does not offer any form of a SLA or any similar proof of fidelity and security, then the Cloud customer must be aware of all the present risks, which in turn would allow for improved contingency plans.

E. Data protection and privacy

Replication of data offers multiple benefits in the events of disasters, but it also allows for a gap in the security of the system. Not only this, data within the cloud systems is typically stored in a shared environment alongside data that belongs to another cloud customer. In these cases, the only mean by which the data can be protected and visible to only its respective customer is by encrypting it. Another relevant point when it comes to data storage in a shared environment is the location on which the Cloud customers data is stored. In some cases, the data can be stored in different countries, which in turn implies a difference in the law regarding data storage and processing. Cloud providers must comply to the jurisdiction and laws of the country in which they store the Cloud customers data. Regardless of whether the Cloud costumer is aware of where geographically his data is stored, a cloud provider is always obliged to prepare a plan as to what will happen to the Cloud customers assets in the event of a disaster. The Cloud customer must be aware that any offering that does not supply the replication of data in events of disasters is ultimately vulnerable to complete failure [9].

F. Regulatory compliance

When a Cloud customer signs a contract with a cloud service provider, the Cloud customer has the responsibility to choose a provider which holds the best offer. This implies that the Cloud customer is the one that is ultimately responsible for the security and integrity of their own data. The Cloud service providers are subjected to external audits and demands of certificates regarding the security of the services they offer. These external audits can also be performed by the Cloud customers themselves, in order to assess the risks and take them into account in future planning. This is a key point in the further development of the Cloud customer; If a Cloud customer finds the security risks are unacceptable, they should consider a different form of hosting. Some Cloud service providers reject this and refuse to undergo such scrutiny, which in turn implies that they are unprepared to provide high quality services for tasks more complicated than the basic tasks of the cloud. Still, acquiring certification can be as simple as complying to the ISO 27001:2013 standard which offers controls that cover the issues regarding the security of a cloud [10].

G. Decrease of performance

With the numerous attempts to increase the security within the cloud as well as eliminate the majority of the threats, come various means by which this can be achieved whether to full or partial extent. Frequently, the usage of these means leads to an overall decrease of the clouds performance [11]. The implementation of management of identity, monitoring of systems, web and email security as well as the employed mechanisms against intrusion, all produce a negative overhead which must be taken into account as a negative effect [12].

III. Cloud security standards

This section provides an overview of the present existing international security standards and guidelines, with particular aim to their application within cloud computing security. The analysis provided is intended to reveal the scope of these standards, which will in turn answer relevant questions regarding the minimum standard compliance level a given Cloud service provider must meet.

A. ISO/IEC 27000

ISO/IEC 27000 is a family of ISO/IEC Information Security Management Systems (ISMS) standards, commonly referred to as the ISO/IEC 27000 series [13]. This series provides recommendations on information security management, risks and controls within the context of an overall ISMS. This series is generally broad in its scope with the intention to cover more than just privacy and confidentiality, but to also security issues. It is deliberately applicable to numerous types of organization, with the aim to globally encourage the implementation of appropriate security controls according to their needs.

1. ISO/IEC 27001:2005 is a standard that covers many types of organizations and companies. It specifies the requirements for establishing, reviewing and maintaining a documented ISMS within the organization, from the aspect of the organizations overall business risk. This standard implies that the developed ISMS is developed in such a way that is easy to change and improve [14]. There ISO/IEC 27001:2013 puts a greater emphasis on measuring and evaluating the ISMSs performances. While the ISO/IEC 27001:2005 emphasized the

Plan-Do-Check-Act cycle, this standard avoids the promotion of this principle.

2. ISO/IEC 27002:2005 offers practical guidelines for achieving satisfactory levels of security through effective management practices. Following these guidelines will also lead to increased confidence in activities within the organization that attempts to apply these guidelines [15]. The revision in 2013 produced the ISO/IEC 27002:2013, which offers a more focused standard that also brought new aspects to light, putting an emphasis on cryptography, communications security as well as system security testing.

3. ISO/IEC 27005:2011 is a standard which provides guidelines for information security risk management. The general concepts of ISO/IEC 27001 are supported, however the initial design and purpose is to assist the implementation of information security based on an approach that satisfies risk management. This standard is applicable to all organizations of various types, ranging from commercial enterprises to government agencies as well as non-profit organizations, which intend to manage the risks that could compromise the organizations security of information [16].

4. ISO/IEC 27017 is a standard for the information security management for cloud systems which is currently in preparation [17].

B. NIST 800-53 R4 Security Controls

NISTs special publication 800-53 R4, Security Controls for Federal Information Systems and Organizations is a guideline which provides a plethora of security and privacy controls primarily intended for federal information systems as well as a process for selecting controls to protect organizational operations.

The controls within the publication are easily customizable and are implemented as a part of a process that manages information security as well as privacy risk. The publication of security controls addresses security from multiple perspectives, both from a functionality perspective, as well as assurance [18].

IV. Auditing and Assessment Standards

Acquiring a certification is done by undergoing an internal external auditing process which is intended to assess whether the organization fully complies to the standard its attempting to acquire. For this purpose, there are several certifications which additionally provide guidance.

COBIT 5 stands for Control Objectives for Information and Related technology. Developed by the Information Systems Audit and Control Association (ISACA), COBIT 5 provides a total of 37 control objectives which are grouped into two areas: Governance of Enterprise IT and Management of Enterprise IT. The first area covers a single domain, while the second one covers the remaining four, each of them containing a certain amount of processes which cover all aspects of information as well as the technology that supports it. Addressing these 37 high level control objectives will allow the company to ensure appropriate control system within its IT environment, as well as assure the Cloud customers the provided service is secure and robust [19].

SAS 70 (Audit) Type II is developed by the American Institute of Certified Public Accountants (AICPA). Unlike COBIT 5, SAS 70 does not specify a set of control objectives and activities that a Cloud customer must meet, but instead it provides guidance which enabled independent auditors, sometimes even Cloud customers, to form opinions regarding the descriptions of the services of Cloud service providers offer. More specifically, SAS 70 Type II certifies that a Cloud service provider has had a thorough audit of all of its controls which relate to the performance of the functionalities as well as the level of security it provides. Allowing a Cloud Customer to view the certification of a Cloud provider allows for a build of trust with the customer. This is key, since such documents contain valuable information regarding the effectiveness and reliability of the Cloud service provider [20].

Health Insurance Portability and Accountability Act, HIPPA, is a standard that deals with the security and privacy of health data, intending to improve and promote the efficiency and error rate of healthcare systems by the encouragement of the usage of software systems which promote communication between entities using electronic data. In addition, it offers a set of rules which ensure the protection of individually identifiable health information which is stored within the systems. All health care providers must comply to the act, as well as health care clearing houses and health plans, which might involve businesses which handle and process fragile health information, or individuals and groups which provide or pay for the medical care [21].

PCI DSS V3.0 (Payment Card Industry Data Security Standard), is developed with the intention to encourage cardholder data security, promoting enhanced security measures which are widely used. The usage of such standards facilitates the broad adoption of consistent and mutually-compatible data security measures globally. The structure of the PCI DSS V3.0 standard is organized in 12 high level requirements. These requirements cover the building and maintenance of secure network systems which in turn protect the cardholder data.[22].

The CSA Security, Trust and Assurance Registry (STAR) Program is another guidance developed by the Cloud Security Alliance. It offers a set public registry for cloud provider trust and assurance, to a broad spectrum of customers and providers. This program is made up of three levels of assurance, all of which are based upon a list of high level cloud control objectives.

As portrayed in the above paragraphs, some standards are general purpose and can be applied to a multitude of topics and organizations. Some of them, on the other hand, are more specific and deal with problems typical for the fields for which they are intended.

V. Cloud security efforts on standardization

Despite the fact that numerous standards and guidances exist, covering general scenarios or perhaps incredibly specific ones, there still exists the question whether all of these cover

all the security gaps a Cloud system might possess. CSAs study in [5] explicitly identified the top threats to cloud computing, with the aim to support the creation of solutions and additional standards. ENISA, on the other hand, assessed the risk level within the Cloud environment as a function of the business impact and the likelihood of the scenario supporting an incident.

The Cloud Security Alliance produced an initial report V 2.1 which contains a taxonomy based on 15 different security domains and processes that need to be followed when undergoing the process of Cloud deployment. New candidate domains, which have been proposed in the new version 3, have been of great interest to experienced industry consumers as well as security professionals. Each domain addresses core functionalities, features, services, threats that have been addressed, as well as challenges that the domain focuses on [12].

NIST has also made additional efforts in which there are many aspects of cloud computing discussed. In [23], the threats, technology risks as well as security for public cloud environments are thoroughly discussed.

VI. Discussion

With the identification of the security threats that arise with the introduction of the cloud computing technologies, comes a vast array of standards which provide evidence that a specific organization, such as a cloud service provider, has met a satisfactory level of security. Still, in the presence of numerous standards and guidelines, companies are faced with the question of which standard is most appropriate for cloud computing technologies. Naturally, the most intuitive answer is that meeting both the ISO standards as well as NISTs guidance controls will provide the best solution. However, there are fundamental differences between the two standards which imply neither one of them can be subsumed within the other.

For instance, NISTs 800-53 is applicable for information systems which support executive agencies of the federal government, whereas the ISO 27001:2005 is a general purpose standard to all types of organisation.

Another fundamental difference lies within the mismatch of numerous security controls, in the sense that controls with similar topics are addressed in both standards, but cover entirely different contexts or perspectives. Furthermore, some controls from one standard are covered by several others within the other [24].

VII. Conclusion

The very nature of the cloud offers many benefits, but also introduces several security setbacks. In order to provide means to overcome these security setbacks, various organizations, such as ISO and NIST, have produced standards and guidelines which comprise numerous security controls.

Much of the task regarding the defining of general purpose security standards, as well as the defining of cloud-based security standards has been completed. However, due to the development of new services within various levels of the cloud, these standards must evolve and improve in order to satisfy these changes. New standards that specify particular security points of these changes must also be defined, in order to satisfy the ongoing changes within the cloud computing technology.

References

[1] "Defining cloud services and cloud computing." [Online]. Available: http://blogs.idc.com/ie/?p=190

[2] K. C. Sean Carlin, "Cloud computing security," 2011. [Online]. Available: http://scisweb.ulster.ac.uk/~kevin/ijacivol3no1.pdf

[3] C. S. C. Council", "Cloud security standards what to expect and what to negotiate," 2013. [Online]. Available: http://www.cloud-council.org/Cloud_Security_Standards_Landscape_Final.pdf

[4] S. Ristov, M. Gusev, and M. Kostoska, "A new methodology for security evaluation in cloud computing," in *MIPRO, 2012 Proc. of the 35th Int. Convention, IEEE Conference Publications*, 2012, pp. 1808–1813.

[5] C. S. Alliance, "Top threats to cloud computing v1.0," 2010. [Online]. Available: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[6] NIST.

[7] ENISA, "Cloud computing security risk assessment," 2009.

[8] C. Babcock, "9 worst cloud security threats." [Online]. Available: http://www.informationweek.com/cloud/infrastructure-as-a-service/9-worst-cloud-security-threats/d/d-id/1114085

[9] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," 2011.

[10] J. Brodkin, "Seven cloud-computing security risks," 2008. [Online]. Available: http://www.infoworld.com/article/2652198/security/gartner--seven-cloud-computing-security-risks.html

[11] S. Ristov, M. Gusev, M. Kostoska, and K. Kiroski, "Virtualized environments in cloud can have superlinear speedup," in *Proceedings of the Fifth Balkan Conference in Informatics*, ser. BCI '12. New York, NY, USA: ACM, 2012, pp. 8–13.

[12] "Preparing to take the Certificate of Cloud Security Knowledge (CCSK)," 2013. [Online]. Available: https://cloudsecurityalliance.org/wp-content/uploads/2013/02/CCSK-Prep-Guide-V3.pdf

[13] "ISO Freely Available Standards." [Online]. Available: http://standards.iso.org/ittf/PubliclyAvailableStandards/

[14] ISO/IEC, "27001:2005 Information technology - Information security management systems." [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42103

[15] ——, "27002:2005 Information technology - Code of practice for information security management." [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50297

[16] ——, "27005:2011 Information technology - Information security risk management." [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56742

[17] ——, "DIS 27017 Information technology - Code of practice for information security controls based on ISO/IEX 27002 for cloud services." [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43757

[18] "NIST Special Publication 800-53 Revision 4, Security and Privacy Controls for Federal Information Systems and Organizations," 2013. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf

[19] P. Bernard, "Cobit 5 - A Management guide," 2012.

[20] Christopher G. Nickell and Charles Denyer, "An Introuction to SAS 70 Audits," 2007.

[21] Centers for Disease Control and Prevention, USA, "HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services," *MMWR: Morbidity and mortality weekly report*, vol. 52, no. Suppl. 1, pp. 1–17, 19, 2003.

[22] Security Standards Council, "Payment card industry data security standard - requirements and security assessment procedures," 2013.

[23] "NIST SP800-144 Guidelines on Security and Privacy in Public Cloud Computing," 2015. [Online]. Available: http://csrc.nist.gov/publications/PubsSPs.html

[24] S. Ristov, M. Gusev, and M. Kostoska, "Cloud computing security in business information systems," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 4, no. 2, pp. 75–93, 2012.

Translating BPMN to WS-BPEL

Elena Markoska, Sasko Ristov and Marjan Gusev
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
1000 Skopje, Macedonia
elena.markoska@live.com, {sashko.ristov, marjan.gushev}@finki.ukim.mk

***Abstract*—Business processes have been the primary tool for modelling complex activities within companies, enveloping a linked network of tasks needed to complete a particular goal. There are numerous means by which business process can be expressed and modelled, such as BPMN and WS-BPEL, which are thoroughly examined in this paper. These two languages are used in the design and implementation phases respectively and as such there is a degree of conceptual mismatch between them. In order to achieve an easier modelling using BPMN and WS-BPEL through an automated mean, the question of translating between them arises. In this paper, we observe several approaches to achieve the translation between both languages and assess their completeness, effectiveness and readability.**

***Keywords*-BPMN, WS-BPEL, BPMN2BPEL.**

I. Introduction

The business processes are becoming more complex, which increases the need for a flexible, yet automated mean by which business processes can be modeled and explored. Various means have been developed and implemented with the aim to aid and simplify the processes of software design and implementation.

Recent research on the translation from Business Process Modeling Notation (BPMN) [1] to Web Services Business Process Execution Language (WS-BPEL) [2] clearly outlines a conceptual mismatch between both languages which then results in decreased quality of the WS-BPEL code as a product of the translation.

Depending on the original model described using BPMN, certain models are easier and simpler to map into WS-BPEL. Furthermore, depending on whether the final aim is to either achieve a block-structured or a graph-structured BPEL format, the outcome may differ in readability and completeness.

This paper overviews the strategies that have been used to map BPMN into WS-BPEL and explores the possible drawbacks of each. We do this by providing information regarding the general structure of the languages in order to elaborate the strategies used for the mapping between both languages. Section II sheds light into the scope and usage of the languages, elaborating on the respective phases of development they are most commonly used in. Section III discusses the nature of the BPMN and WS-BPEL languages, exploring their components and elaborating on how they can be used. Section IV discusses existing strategies for mapping between BPMN and BPEL and explores their strengths and weaknesses, along with the levels of completeness and readability they provide. Section V provides examples of translation between the two languages. Section VI offers a discussion of the state-of-the-art and makes suggestions regarding future development. Finally, Section VII draws the conclusions.

II. Scope and usage

The modelling of the business processes is a task that has been the topic of research for many scientists, developing tools, standards and specifications, which aid the process of development in its respective stages. The BPMN modelling language is used in the design phase, where the aim is to clearly outline the business process in a graphical representation which is suitable for all parties of the entire development process. WS-BPEL, however, is a language that is used within the latter stages of development, and as such provides a structured tree of predefined tags which describe the behaviour of the process, closer to programming code. Achieving automated mapping between these two languages will offer a faster development phase [3]. These languages find various applications in other spheres as well. For instance, there have been efforts made into using BPEL with TOSCA (Topology and Orchestration Specification for Cloud Applications), which is used to describe the structure of portable services in a topology. With an analysis of this specification and its compatibility with BPMN, BPMN4TOSCA has been developed as a BPMN extension with the ability to describe TOSCA specified orchestration using the BPMN modelling language[4]. The TOSCA specification also clearly states that it is compatible with WS-BPEL and there are many efforts and implementation that work towards an efficient translation between both of them.

III. Business Process Modelling and Execution Languages

This section presents the standards in business process modelling and business process execution languages. Both BPMN and BPEL are examined, outlining their usage and structure. The most common elements within both languages are explained in greater detail.

A. Business process modelling notation

BPMN is a widely recognized standard for business process modelling that uses graphical elements in order to specify business processes in a Business process diagram. Essentially it is based on a flowcharting technique not dissimilar to the one used within the Unified Modelling Language. The intention behind the creation of this graph-oriented language is

to provide a notation which could both be used by technicians, software designers and software developers, providing an understandable and sufficiently informative model of the entire business process. All in all, BPMN provides several types of process models among which the most crucial can be defined as:

1) *Private (internal) business processes.* Private business processes are usually within the scope of the internal affairs within a single company and are not to be published in order to maintain confidentiality within the company.
2) *Abstract (public) business processes.* Abstract business processes are created when there is a need for interaction between a private business process with an external participant. These participants can be other processes, individuals or companies.
3) *Collaboration (global) business processes.* Collaboration business processes are intended to provide an external view of the process of interaction between two or more participants.

Regarding to the language itself, it is consisted of various elements that dictate the control flow of the business processes. Such elements are grouped within several categories of flow objects, connectivity objects, swim-lanes, as well as artifacts. The category of flow objects is then divided into events, activities and gateways which group various item within them. Some of these groups are depicted in Fig. 1. In the following paragraphs these elements are explained in greater detail with the intention to better explain their mapping into WS-BPEL [5]:

Events. An event is something that occurs during the execution of the business process. Based on what kind of an action it involves, events can be start, intermediate and end events. A start event indicates when a business process will start and does not have any incoming arcs. An intermediate event is an event thats somewhere between a start and end event, while end events represent the termination of a business process and have no outgoing arcs.

Activities. Activities are actions within a company which represent the tasks or processes which are meant to be completed. They may take the form of atomic tasks, or non-atomic tasks. A task is the lowest level of an activity and cannot be broken down, however it can be used to make up a process.

Gateways. A gateway is used to create the control flow within a business process. As such, it is necessary to offer branching, forking and joining of paths. Often these are insufficient, which leads to the introduction of different types of decisions which can be used to accurately model the flow of a business process. Such decisions are exclusive, inclusive or complex. An exclusive decision is based on known data or other events and its outcome isnt known at design time. An inclusive decision, on the other hand, is based on a condition which is known during design time. Complex decisions are situations which cannot be placed under other gateways. Finally, parallel gateways are present to offer forking and joining of concurrent activities.

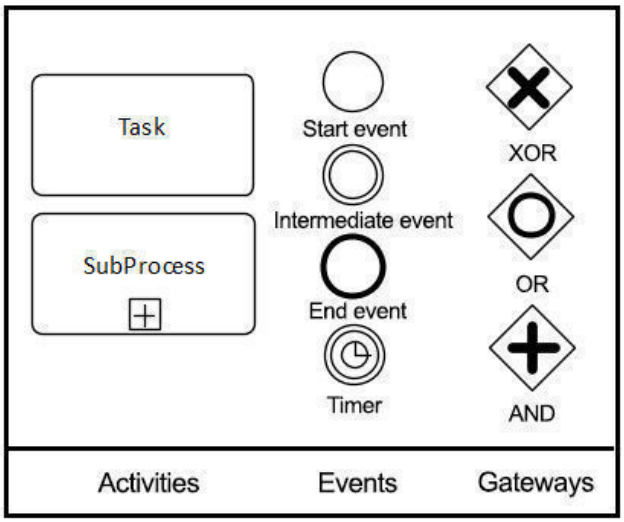


Fig. 1. Essential elements to BPMN models

Connectivity objects. The connectivity objects are crucial to establish the flow of the business process. They may differ depending on whether a sequence flow, a message flow or an association is being shown. Sequence flows simply show the order in which the defined activities are executed within a process. A message flow shows the direction of messages from the sender to the receiver, while the associations present the link between the information and the actual flow object.

Swimlanes. Generally there are two type of swimlanes - lanes and pools. Lanes have the ability to organize and categorize activities, while pools indicate the different sub processes and participants within a larger process.

Artifacts. There are three types of artifacts within the BPMN specification - data objects, groups and text annotations. The data objects provide information regarding the requirements for executing certain activities and their outcome. The groups are used to document and display processes, while the text annotations are only meant for displaying additional information to the reader. All artifacts do not interfere with the flow of the business process.

B. Business Process Execution Language

On the other hand, WS-BPEL, or commonly known as BPEL, is an executable language intended to specify the actions within business processes that contain web services. Within web-service interactions, there are two ways of applying the BPEL concepts to the business processes:

- *Abstract business process.* An abstract business process holds a descriptive role, and aims to portray the observable behaviour within the process itself, which is most commonly executable in nature. It often does not contain sufficient operational data which then prevents the abstract process from its future execution.

```
<sequence>
    <scope name="ApplicationUniversity">...</scope>
    <if>
        <condition>
            $POApprovalResult="approved"
        </condition>
        <sequence name="continue">
            <scope name="AcquiringStudentBook">...</scope>
                <flow>
                    <scope name="ElectronicStudentBook">...
                    </scope>
                    <scope name="HardCopyStudentBook">...</scope>
                </flow>
            <scope name="StudentCard">...</scope>
        </sequence>
    <else>
        <exit/>
    </else>
    </if>
</sequence>
```

Fig. 2. An example of WS-BPEL

- *Executable business process.* This type of a business process models the simple behaviour of a participant within a business interaction.

The specification itself contains many elements, some of which created with a specific purpose to satisfy a particular situation. The complete list of all BPEL elements can be found in [2]. Three types of classes categorize the most commonly used elements: basic activities, structured activities and partner links. Fig. 2 shows an example of WS-BPEL, in which some of the mentioned tags are used.

Basic activities. These activities have the ability to perform the logic of the process in BPEL, describing the elemental steps of the process behaviour. Many elements belong to this group, several of which are: Invoke, Receive and Reply, Assign, Throw and Rethrow, Wait, Empty, Exit. The Invoke, Receive and Reply elements handle the communication with web services, invoking them when necessary or receiving the responses from them. The latter elements can also be used in composite business processes when a sub process needs to terminate and deliver its outcome to ongoing elements. The Assign element has the ability to change the values of variables during the course of a business process. Throw and Rethrow, on the other hand, handle the faults within the business process. These two generally have the functionality of try-catch blocks within multi-purpose modern programming languages. The Wait element is used to achieve sufficient time management within a business process. For example, if an activity needs to wait for a specific time, the tags <for>and <until>specify a fixed duration or a deadline it must comply to. An Empty element specifies an activity that performs no action and is generally used to generate valid BPEL code through the means discussed in Section IV Finally, the Exit activity terminates all activities within the process instance.

Structured activities. The structured activities describe the control flow logic of a business process, meaning they specify the order in which the activities are executed, most of which can also be seen within programming language to achieve a desired flow of execution. To achieve this, this group contains the following elements: Sequence, If, While and repeatUntil, Pick, Flow, ForEach. The sequence control activity contains activities which are performed one by one in a predetermined order. For behaviour that requires the usage of conditions, the <if>tag is used to initiate conditions, while the <else >, <elseif>as well as <condition>tags further specify the body of the <if>tag. While and repeatUntil are activities that provide a flow which allows for repetition of controls. The <while>and <repeatUntil>tags are used to initiate the loops, while the aforementioned <condition>tag is used to specify the moment of termination. The <flow>activity provides synchronization and concurrency, allowing different groups of activities to be initiated simultaneously. It uses <link>tags which can be specified within <source>and <target>elements. The <pick>activity, on the other hand, waits for the occurrence of one event, and then executes all activities related to this event. Finally, the forEach element allows for branching of processes or subprocess, offering sequential or parallel execution [6].

Partner links. The partner links describe interactions between businesses. These interactions are usually done through web service interfaces. They are concrete references to services that a BPEL business process interacts with.

IV. BPMN2BPEL Mapping Strategies

While both BPMN and BPEL excel in their own respective purposes, there is a very vast difference between the two languages, which increases the difficulty of mapping one into the other. BPMN is a language most commonly used in the design phase in which graphical notation is frequently used by business analysts to design and discuss the nature of the business process itself. BPEL, on the other hand, is used during the phase of implementation, which is a more advanced phase throughout the development process. BPEL is usually intended for either technical analysts or programmers. In addition, the nature of the two languages implies a mismatch of a certain extent, given the fact that BPMN is a graph-oriented language while BPEL is a block-oriented language. While many attempts to map the two exist, the conceptual difference is still the source of many problems that arise in the implemented mappings.

The mapping of BPMN to BPEL has been the topic of much research in the past years. A plethora of different strategies and implementations have been discovered, all of which assuming different methods. Generally, these methods can be grouped into several subsets which contain the properties of the strategies used. These subsets include strategies based on:

- *Element preservation.* This strategy seeks to map all the process graph elements to constructions of flow, map connectors to empty elements and arcs to links. One of the advantages to this strategy is its simplicity of implementation, along with the fact that both the resulting BPEL code as well as the original process are graph-based and easy to compare, therefore increasing the readability. Still, this strategy is only correctly usable is the process is entirely acyclic, disallowing the completeness

of the translation with the implied restriction of the flow-based approach. [7]

- *Element minimization.* The general idea behind the strategy of element minimization is to remove empty activities that have previously been generated, replacing them with transition conditions that describe splitting behaviour or join conditions to represent joining behaviour. In essence, this strategy simplifies the processes by removing the empty elements. Similarly to the element preservation strategy, this strategy also implies the usage of acyclic graphs in the BPMN code, which reduces the completeness of the approach. In addition, the removal of empty activities leads to a decreased readability and greater difficulties with understanding the initial graph. However, due to the reduced number of activities that require processing, this strategy offers better performances [8].
- *Structure identification.* The gist of the transformation using a structural identification is to identify structured activities within the processes and to apply certain predefined rules for the reduction of the structure. This transformation if applied correctly most often results in highly readable BPEL code. However, the relation of the BPEL code with initial graph of the process might be less intuitive since BPEL is block-structured. Despite its increase in readability, this strategy offers a drawback in its completeness. This strategy cannot be applied to all the graphs, and the graphs must be structured in order to be able to be translated, since every component within its composition must be mapped into structured activities [9].
- *Structure maximization.* Structure maximization uses different techniques in order to determine the maximum structure. It can be applied onto processes which don't have a clearly defined structure. Contrary to the structure minimization strategy, this strategy also supports unstructured processes, as long as the loops within them can be reduced by the structure-identification strategy. This in turn implies the requirement of the two strategies being implemented and applied, which reduces the performance of the translation. Despite the increase of completeness and its applicability to most processes, this strategy still requires the unstructured processes to have a single point of entry as well as a single point of exit [10].
- *Petri Nets approach.* The Petri Nets have a very strong theoretical foundation, and can be used as one of the strategies to translate BPMN into BPEL code. The greatest disadvantage of this approach is the fact that the original BPMN must be first mapped into a Petri Net. This mapping is not covered by this strategy, and therefore greatly contributes to the incompleteness of the approach. Still, the generated BPEL code is correct and easily readable, and it can be applied to different kinds of processes without restrictions, which additionally adds to the usability of the approach [11].
- *Event-condition action-rules strategy.* Similar to the structure-maximization strategy, the idea behind this approach is to apply the reduction rules of the structure-identification strategy as frequently as possible in order to identify a maximum structure. The remaining graph is translated according to the Event-Condition-Action rules presented in [12]. This strategy offers the highest level of completeness, allowing the translation of any process graph, including unstructured loops. Still, contrary to its high level of completeness, the generated code is most often difficult to understand and highly unreadable.
- *Graph structure approach.* This approach offers a translation which maintains the graph-like structure of BPMN models. With regard to the BPEL requirements of the global process, namespace as well as variable information, these have to be supplied in addition since the BPMN graph-like structure does not normally contain this information. One of the disadvantages of this approach is the generation of BPEL code which might be difficult to understand, therefore reducing its readability.
- *The BPMN to BPEL way.* Unlike some of the aforementioned strategies, this strategy allows the translation of arbitrary cycles within the BPMN code, by using an event-based translation. This approach can also deal with all parallel gateways as well as XOR gateways with an arbitrary topology. Due to the usage of block structured components which are mapped before the flow-based translation, the generated BPEL code is readable. Because of the usage of events for unstructured processes, the generated BPEL code can sometimes be difficult to read and understand. In addition, the combination of three strategies within a single one implies greater difficulties in implementation, along with decreased performance during translation. Still, every component present in the BPMN graph can be mapped in an order [13].

While some strategies attempt to combine several of the categories in an attempt to achieve correctness, efficiency and completeness, often they result in decreased readability stemming from many conflicting criteria.

V. Mapping Examples

In this section we provide several examples in order to illustrate the translation process. The three examples portray the translation of the beginning activities of the process, the situation when an activity branches into several other activities, as well as the situation when a loop occurs within the process diagram. Each of these are shown with their respective BPMN diagram, which does not consist of the full BPMN process, but only the part that we translate into BPEL within the scope of the example [14] [15].

A. Mapping the beginning of a process

In this example we illustrate the translation of a fragment of a BPMN process diagram as depicted in Fig. 3.

On the diagram, the business process begins with the enrolment of the student within the university, which in turn further in the diagram invokes additional activities. The process begins with a Receive Message Start Event which is the mechanism

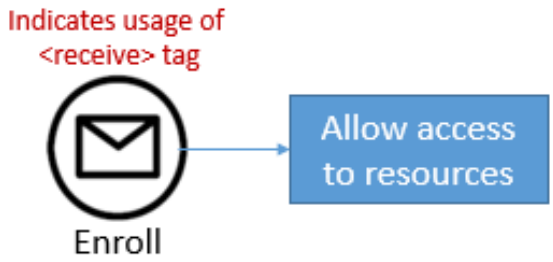


Fig. 3. BPMN diagram example 1.

that invokes the process through the receiving of a message, in this case, the enrolment of a student to the university. This component will be mapped into a receive element in BPEL. The resulting BPEL code is shown on Fig. 4.

```
<receive createInstance="yes" operation="book"
name="Enroll" wpc:displayName="Enroll" portType=
"wsdl0:enrollPort" variable="input" wpc:id="2">
        <source linkName="link1" />
</receive>
```

Fig. 4. BPEL example 1.

The flow and link elements of the BPEL code are assumed to have previously been defined, hence the <link1>reference within the<source>tag. All the properties within the<receive>tag are based on previously specified parameters.

B. Mapping a parallel flow of activities

Fig.5 shows a fragment that continues the BPMN process diagram, relevant to this example.

After the activity Allow access to resources, two main activities can occur. These two involve the enrolment to courses the university provides, as well as viewing the study programmes, which can be done in parallel. The parallelism begins with the existence of two outgoing Sequence Flow elements from the Allow access to resources activity. Both of these Sequence Flow will result in two BPEL link elements (link1 and link4). These will be included within source elements in the Allow access to resources activity. The mapping will result in an<assign>tag which will precede and <invoke>tag. A<link>element (link3) will be added in order to create a sequential dependency between the<assign>and<invoke>elements. The translated code for only one of the activities, in this case Enroll for courses, is depicted in Fig. 6.

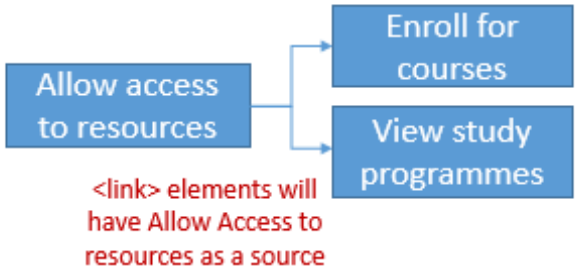


Fig. 5. BPMN diagram example 2.

```
<assign name="Resources" wpc:displayName="Resources" wpc:id="21">
    <target linkName="link1"/>
    <source linkName="link2"/>
    <copy>
      <from part="programmes" variable="input"/>
      <to part="programmes" variable="viewStudyProgrammesRequest"/>
    </copy>
</assign>
<invoke inputVariable="viewStudyProgrammesRequest" name=
"viewStudyProgrammes"
    operation="doFlightReservation" outputVariable=
    "viewStudyProgrammesResponse"
    partnerLink="viewProgrammesService" portType=
    "wsdl3:viewProgrammesService"
    wpc:displayName="View Study Programmes" wtx:id=10>
    <target linkName="link2"/>
    <source linkName="link3"/>
</invoke>
```

Fig. 6. BPEL example 2.

C. Mapping a loop of activities

This final example translates the fragment of BPMN process diagram shown in Fig. 7.

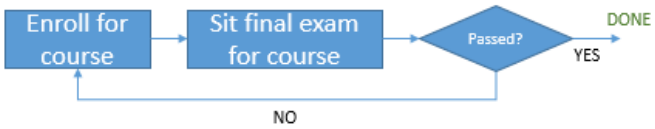


Fig. 7. BPMN diagram example 3.

The two tasks, Enroll for course and Sit final exam for course can be done a number of times, each time the process passing through a decision Gateway which splits the flow based on whether the passing condition is met or not. According to the diagram, if a student passes the final exam, then the control flow continues onto other parts of the diagram, otherwise, it returns to the initial enrolment of the course. The resulting BPEL code is shown in Fig. 8.

```
<while condition="Passed" name="While" wpc:id="11">
    <condition> $passed &eq; "NO"</condition>
    <invoke partnerLink="EnrollForCourseService" />
</while>
```

Fig. 8. BPEL example 3.

VI. Discussion

Both BPMN and BPEL are languages that are used within the process of production of business models, each used in its respective phase of the entire process. Their usage is most commonly found in the design and implementation phase, yet there have been various instances of these two language being used for modelling within external implementations.

For instance, Topology and Orchestration Specification for Cloud Applications (TOSCA) is a relatively new standard that specializes in the portability and migration of cloud applications along with maintaining their topologies in the process through the usage of various methods including workflows. BPMN and BPEL are languages that can be used to model these workflows [4].

Numerous effort exist that attempt to provide an automated mean by which a successful translation from BPMN to WS-BPEL can be achieved. Despite the abundance of approaches, many of which use exotic data structures and strategies, none of them provide a flawless translation, offering instead numerous drawbacks on performance, completeness, effectiveness, correctness or readability.

Still, given the large amount of drawbacks that BPMN implies when it comes to its mapping to BPEL, resorting to other globally accepted means of expression might be more suitable. Given the graphical nature of the BPMN language, it is often a source of inefficiency and unreadability of the majority of the business processes in their transformation to other languages. Ristov et al. [16] propose the usage of XML instead of using BPMN for application modelling and description. XML implies a flexible and configurable mean by which business processes can be expressed without the additional need for translation and encountering problems of the conceptual type. There have also been several successful attempts at demo implementations by using P-TOSCA, which successfully bridge TOSCA and XML as communication and orchestration specifications [17], [18], [19].

VII. Conclusion

BPMN is a widely recognized multi-purpose language for modeling of business processes. It is primarily a graphical language that is intended for the usage in the design phase. WS-BPEL, on the other hand, is an executable language intended for the phase of implementation, which is a later phase and intended for technical analysts and programmers. Various attempts have been made to combine both languages and provide an automated bridge between the two. Each of these attempts uses a strategy or a combination of strategies with the purpose of achieving better results based on predefined criteria. Despite the attempts to produce means by which the resulting WS-BPEL code will be readable and complete, many of which have numerous drawbacks in various aspects. The research that involves BPMN and BPEL has also been extended into other spheres, such as TOSCA, which uses BPMN, as well as to use XML.

Acknowledgment

This work was partially financed by the CloudArchitect-Man project (Cloud Architecture Management) at Faculty of Computer Science and Engineering at the ”Ss. Cyril and Methodius” University, Skopje, Macedonia.

References

[1] Object Management Group, “Business Process Model And Notation (BPMN),” 2011. [Online]. Available: http://www.omg.org/spec/BPMN/2.0/

[2] OASIS, “Web Services Business Process Execution Language Version 2.0,” 2007. [Online]. Available: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[3] J. Mendling, K. B. Lassen, and U. Zdun, “On the Transformation of Control Flow between Block-Oriented and Graph-Oriented Process Modeling Languages,” Tech. Rep., 2006.

[4] O. Kopp, T. Binz, U. Breitenbücher, and F. Leymann, “BPMN4TOSCA: A domain-specific language to model management plans for composite applications,” in *Business Process Model and Notation*, ser. Lecture Notes in Business Information Processing, vol. 125. Springer, 2012, pp. 38 – 52.

[5] “BPMN Tutorial,” 2015. [Online]. Available: http://www.bpmn-tool.com/en/tutorial/

[6] “Oracle - Structured Activitiesl,” 2005. [Online]. Available: http://docs.oracle.com/cd/E19182-01/820-6323/giaih/index.html

[7] L. Bai and J. Wei, “A Flexible Transformaion Scheme between the ’OR’ of BPMN and ’Link’ of BPEL.” IEEE, 2008, pp. 1300–1307.

[8] S. Stein, S. Kuhne, and K. Ivanov, “Business to IT Transformations Revisited,” Germany, 2009.

[9] L. Garcia-Banuelos, “Pattern Identification and Classification in the translation from BPMN to BPEL,” 2008.

[10] T. Kuster and A. Hesler, “Towards Transformations from BPMN to Heterogeneous Systems,” 2009.

[11] N. Lohmann, E. erbeek, and R. Dijkman, “Petri Net Transformations for Business Processes - A Survey,” Rostock, Germany, 2009, pp. 46 – 62.

[12] J. Recker and J. Mendling, “On the translation between BPMN and BPEL: Conceptual mismatch between process modeling languages,” in *Proceedings 18th International Conference on Advanced Information Systems Engineering*, 2006, pp. 521–532.

[13] C. Ouyang, M. Dumas, A. H. ter Hofstede, and W. M. van der Aalst, “From Business Process Models to Proces-oriented Software Systems: The BPMN to BPEL Way,” vol. 5, no. 1, Netherlands, 2008, pp. 46 – 62.

[14] S. A. White, “Using BPMN to Model a BPEL Process,” IBM Corp., United States, Tech. Rep.

[15] D. Schumm, D. Karastoyanova, F. Leymann, and J. Nitzsche, “On Visualizing and Modelling BPEL with BPMN.” IEEE, 2009, pp. 80–87.

[16] S. Ristov, M. Kostoska, and M. Gusev, “P-TOSCA portability demo case,” in *2014 IEEE 3rd International Conference on Cloud Networking (IEEE CloudNet’14)*, Luxembourg, Oct. 2014, pp. 269 – 271.

[17] M. Gusev, M. Kostoska, and S. Ristov, “Cloud P-TOSCA porting of N-tier applications,” in *Proceedings of the 22nd International TELFOR Forum, IEEE Conference Publications*, 2014, pp. 935–938.

[18] M. Gusev, M. Kostoska, S. Ristov, and A. Donevski, “P-tosca portability of soa applications,” in *Proceedings of 5th International Conference on Cloud Computing and Service Science (CLOSER)*, 2015, in press.

[19] M. Kostoska, A. Donevski, M. Gusev, and S. Ristov, “Porting an n-tier application on cloud using p-tosca: A case study,” in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, IEEE Conference Publications*, 2015, p. in press.

Cloud Computing – Challenges and Demands of the Business Sector

Kocho Hrisafov, Adrijan Bozinovski, Ivana Stojanovska, Biljana Stojcevska,
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
khrisafov@hotmail.com, {bozinovski, ivana.stojanovska, stojcevska, }@uacs.edu.mk

***Abstract*—Technological advancements and innovations in a world ruled by information and their delivery to end recipients showed the need for a change in the way the business sector is managed and operated. Usually the organizations, companies and corporations who don't follow the latest developments in new technologies are left over in the information "ice age" and are usually left behind. This cloud model has several essential characteristics which will be discussed in this paper.**

In cases where a big part of modern society (business sector, government, citizens) functions over internet, there is a demand for this type of services. While in the past there was a need to install software and applications on real machines in datacenters, today cloud computing changes this situation in the Internet domain. It saves time, space and, most importantly for the business sector, it reduces costs. The cloud computing popularity is rapidly increasing; this trend is the most noticeable in European countries and the USA. Companies under market pressure have to reduce costs and maximize profits, so the cloud technology offers an opportunity in that direction.

The following concepts of cloud computing will be analyzed: specifics, advantages, use in business sector etc. The analysis of the situation in the business sector will be made - specifically the usage of cloud computing in developed countries. The gathered data suggests that enormous number of companies had already switched to cloud services. It is also expected that companies will increase their spending in this area.

Keywords—cloud computing; challenges; demand; bussines sector

I. INTRODUCTION

Technological advancements and innovations in a world ruled by information and their delivery to end recipients showed the need for a change in the way the business sector is managed and operated. Usually the organizations, companies and corporations who don't follow the latest developments in new technologies are left over in the information "ice age" and are usually left behind.

There are different definitions which describe the scope and characteristics of cloud computing, but what is common for most of them is that they describe it as a consumption of online computer resources. A part of them describe it as a use of Internet technologies to service the needs of end users. Cloud computing is becoming a social reality and a growing technology. It's very hard to predict how it will further develop and what will be the consequences of its development. Cloud computing has brought in a change in the way users depend on services offered by cloud providers. The economic, legal and ethical implications generated from that change remain to be seen [1].

Today's national economies cannot function separately and independently from the global world market. Globalization and Internet technologies have made a strong impression not only on individuals but also on organizations. The ability to create databases, the use of Web 2.0, deployment of various software solutions, applications and tools, all designed for cloud computing platforms, provide an opportunity to produce solutions and guidelines that make business forecasting real and possible.

Cloud computing service models can be divided, according to their use, generally in three main categories [2]:

- Software as a service (an example is Cisco WebEx, which is used by agencies to facilitate communication and cooperation of individuals located in different locations through the use of video, voice, and sharing over the web);
- Platform as a service (some agencies share their infrastructure which is used for web applications, video etc.);
- Infrastructure as a service (Companies often develop software solutions on server platforms which they do not possess).

When the software is offered as a service, it is very similar to using a local application. The second type of a service platform refers to the delivery of an "environment" e.g. environment for programming. The third of the aforementioned services refers to the use of an infrastructure (e.g. resources of a supercomputer) which is given to the use of the ultimate recipient of the service [3].

II. Benefits and Challenges of Cloud Computing

In terms of the advantages offered by the cloud, it primarily provides a reduction in capital investments. Until recently, the opening of data centers meant a venture without exact forecasts about how much of it will be used and what benefits will it bring. In the cloud there are no unnecessary costs – finances are spent only when necessary using accurate assessment. The capacity of real servers in corporations and enterprise organizations is primarily limited; in the cloud, server capacity is no longer an issue. Applications give an opportunity for global action, access is just a few clicks away and certainly it is the most important to note that there is no continuous investment in large data centers [4].

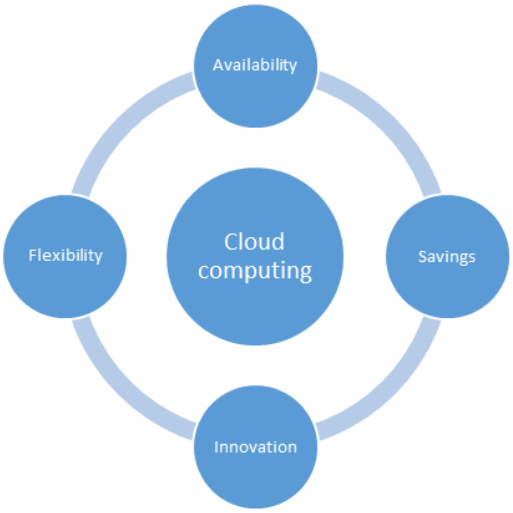


Fig. 1. Advantages of cloud computing

III. The Impact of Cloud Computing on the Business Sector

The financial crisis that originated in the last decade in the US first and then spread throughout the world was initiated by the real estate market. Later, it engulfed the entire financial and real sectors of developed countries and also those in development. Its impact is reflected in different ways and with varying intensities across countries. The crisis has caused a decline in all macroeconomic aggregates and indicators. Its consequences can be defined by a decline in production, dramatically reducing exports, insolvency, increased unemployment, growth of poverty and reduced standard of living [5]. In such a critical situation it is very important for the business sector to manage the costs, savings and, most importantly, how the technological innovations that have flourished in the last decade will be used.

The first and perhaps the most important advantage of cloud computing is cost reduction. This is achieved by reduced costs for computer centers, a larger capacity and of course payment as needed – without paying unnecessary and unexpected maintenance costs of data centers. Using platforms and infrastructure as a service significantly reflects on reducing costs proportional to the size of the company. As a large data center infrastructure is not required, it opens an opportunity for the management to use the knowledge and capacities of their employees to other fields for additional revenue and higher profit [6]. The way that cloud computing is changing the business sector and business processes is through reduction of costs, automation of the use and preservation of the necessary data, flexibility, improved mobility, increased capacities and certainly better allocation and utilization of workforce [7].

One of the advantages that can be observed in real time is the creation and development of applications. Cloud platforms provide an opportunity to create applications on existing infrastructures as quickly as writing the code itself. This way, companies have strict control of expenses. Availability and continuous operation of the virtual servers is very important, regardless of the size of the business. Greater availability, utilization and continuous operation of the virtualized servers is possible due to their continuous development [8].

With the use of cloud services, it is becoming clearer that the risks in terms of investment in data centers are reduced. Planning and investment in the purchase of hardware and software is not needed nor does it require the engagement of a large workforce to maintain the data center.

In technical terms, the benefits of cloud computing can be located in the ability to access the necessary resources or information from anywhere. Work is no longer associated with a particular computer or a specific location; the only thing that is needed is a web browser [9].

A. *Cloud Models and Business Organizations*

In the digital era, especially in the era of cloud computing, companies have the challenge to continually adapt to innovation, to improve the quality of service offered and to maintain their competitiveness in the market. It requires an integrated strategy that includes marketing, infrastructure, IT service management and project management. It is expected that in the future not only companies will adapt to the benefits that cloud computing has brought to them, but also to change the way they operate. It is expected that the companies that will flourish will be the ones with employees located in different locations and flexible levels of its staff. This model is also called the Hollywood model, where the goal is the task and its execution. All of this is enabled by the cloud. The merging of the funds is no longer linked to the employment community space [10].

The following case is an example how businesses are transforming under a surge of new technological cloud innovations. Pearson is an American company that uses the cloud to literally change education in the world. It is the largest provider of educational services in the United States that offers a million tests and 111 million essays a year. Through partnerships with educators, leaders, entrepreneurs, policymakers etc., Pearson is creating people with developed skills prepared for the market [11]. Thanks to the cloud, Pearson has established a global hybrid infrastructure

associated with interconnected systems and processes that allow a diversity of resources to web-oriented products, resulting in measurable results for students. The purpose of this company is a fast growth by covering major markets such as China and certainly to increase the number of offered products that are digital in nature [12].

It can be seen from this example that the possibility of starting a new business with significantly lower costs has increased considerably. In times of economic crisis, small businesses and their flexibility can be a bright point of stability and later growth.

B. Risks and Hazards

When using these services, the first concern of the consumers and companies certainly is security. Many companies provide security for themselves by using private cloud environments. The public cloud market has grown 20% per year with about 2,200 enterprise buyers worldwide earlier this year, but the private cloud, however, is expected to grow between 40% to 50% a year for the next several years [13].

Providers also face pressure because of the preservation and privacy of data. One could say that the cloud is one of the many side segments of the process of globalization. Policy makers make great efforts to distribute the responsibility for the use of cloud services between end users and all parties involved. At the same time, the pressure varies in the direction of increasing the safety [14].

Safety problems arise when a large number of unrelated users share access to the same resources. Therein lays the risk of leaks and the risk of disclosure of identities and access to the data. Authentication and access control have great impact on security and they are in a direct correlation with the provider, as they determine the approach and the way in which the authorization is performed. Who will have access to information, to what extent it will be allowed, and where the data is physically stored, are just some of the issues [15].

When any individual is a recipient of public cloud services, all things are not under her/his control. Although most services speak about their perfection and continuous work, it often happens for hours or days to have service interruptions. A large problem is the loss of data. Very often, though providers argue that the data is protected, permanent loss of data occurs. That's why the most important information is stored locally in the companies. In terms of ownership there is also a problem, because the real end-users are not the only owners of the data. Even some of the biggest cloud providers have clauses in their contracts which specify that the data they store is in their possession [15].

IV. CONCLUSION

Technological developments and the progress of information and communication systems have led to major changes in the business sector. Not only do existing enterprises need to adapt to the new situation, but an opportunity is opened to start a new business with minimal funds. In terms of economic stagnation or crisis, small businesses and their flexibility is an important factor that can lead to development. Cloud computing gives the opportunity to save significant resources that can be redirected to the other needs of companies and also to the employees of the IT sector to be used for more productive purposes than maintaining hardware and software. The availability of information from more places does not bind employees with a specific physical space or computer. The easy accessibility and the innovation of the cloud give opportunities to start businesses where employees are not located in one place. The target is certainly achieved by work-oriented execution of tasks.

There are security risks however, such as information leakage and loss of data. While responsibility in this regard falls on the final recipients, policymakers have lately exerted strong pressure on providers to increase the safety and reliability of the data. All of this happens because of the fact that rarely any company today can afford absence from using cloud services.

REFERENCES

[1] D. C. Marinescu, Cloud Computing: Theory and Practice. Newnes, 2013.

[2] "Cloud Computing Definition from PC Magazine Encyclopedia." 2015 [Online]. Available: http://www.pcmag.com/encyclopedia/term/57964/cloud-computing. [Accessed March 4, 2015]

[3] N. Antonopoulos and L. Gillam, Cloud Computing: Principles, Systems and Applications. Springer Science & Business Media, 2010.

[4] Amazon Web Services, "AWS | What is Cloud Computing - Benefits of the Cloud," 2015 [Online]. Available: http://aws.amazon.com/what-is-cloud-computing/. [Accessed: 25-Feb-2015].

[5] World-crisis.net, "Real Estate Crisis - Housing Bubble - World Crisis webpage." 2008 [Online]. Available: http://www.world-crisis.net/realestate-crisis.html. [Accessed: 02-Mar-2015].

[6] E. A. Marks and B. Lozano, Executive's Guide to Cloud Computing. John Wiley and Sons, 2010.

[7] G. Lucio, Pervasive Cloud Computing Technologies: Future Outlooks and Interdisciplinary Perspectives. IGI Global, 2013.

[8] A. Shadi, Cloud Computing Advancements in Design, Implementation, and Technologies. IGI Global, 2012.

[9] M. I. Williams, A Quick Start Guide to Cloud Computing: Moving Your Business into the Cloud. Kogan Page Publishers, 2010.

[10] G. Winfrey. "How the Cloud Will Transform Business by 2020," Inc.com., 2014 [Online]. Available: http://www.inc.com/graham-winfrey/why-the-cloud-will-transform-small-business-by-2020.html. [Accessed: 26-Feb-2015].

[11] "About Us." [Online]. Available: https://www.pearson.com/about-us.html. Accessed: 02-Mar-2015].

[12] P. Guido, "IBM Smarter PlanetVoice: Three Companies That Transformed Their Businesses Using Cloud Computing," Forbes. 2014. [Online]. Available: http://www.forbes.com/sites/ibm/2014/11/03/three-companies-that-transformed-their-businesses-using-cloud-computing/. [Accessed: 26-Feb-2015].

[13] S. Gaudin, "Enterprises increasingly look to the private cloud," Computerworld, 17-Jul-2014. [Online]. Available: http://www.computerworld.com/article/2490138/private-

cloud/enterprises-increasingly-look-to-the-private-cloud.html. [Accessed: 02-Mar-2015].

[14] S. Pearson and G. Yee, Privacy and Security for Cloud Computing. Springer Science & Business Media, 2012.

[15] R. A. Grimes, "The 5 cloud risks you have to stop ignoring," InfoWorld, 19-Mar-2013. [Online]. Available: http://www.infoworld.com/article/2614369/security/the-5-cloud-risks-you-have-to-stop-ignoring.html. [Accessed: 26-Feb-2015].

IN-MEMORY VS TRADITIONAL DATABASES – ADVANTAGES AND CHALLENGES

Petar Tagasi
School of Computer Science and Information Technologies
University American College Skopje
Skopje, Macedonia
petar.tagasi@gmail.com

Adrijan Bozinovski, PhD
School of Computer Science and Information Technologies
University American College Skopje
Skopje, Macedonia
bozinovski@uacs.edu.mk

Veno Pachovski, PhD
School of Computer Science and Information Technologies
University American College Skopje
Skopje, Macedonia
pachovski@uacs.edu.mk

***Abstract*—Because of the growing robustness and complexity of today's systems, traditional relational databases often cannot adequately respond to the increased demands. Therefore, there is a continuous search for new solutions that can optimally meet these requirements. In recent years, new database paradigms have been developed.**

In-memory databases are a relatively newly introduced solution. The main benefit is the speed of access, since the data reside in primary memory. By contrast, in traditional databases, the overall data structure is located in a data storage device, and access to them demands that a significant number of I/O operations be executed.

In-memory databases have much higher hardware requirements, especially concerning RAM and CPU. However, reduced cost and wide availability of these hardware components justify their use. Another challenge is ensuring data persistency, which becomes an issue when storing large quantities of data in primary memory.

Keywords—Database, In-memory, I/O, SAP HANA, Microsoft Sql Server, Relational data, Cloud Computing, TPC-C benchmark, RAM memory, CPU, scalability, OLTP, OLAP

I. INTRODUCTION

In today's information age, computers and automation are not only present in every aspect of people's lives, but are becoming necessary constitutes of their everyday living. People are going online to shop for their favorite products, make stock transactions or even negotiating airplane tickets in order to safely reach the designated destination. Completing all of the aforementioned actions means processing of huge amounts of data, produced by various enterprise applications or services and used by large numbers of users, and returning corresponding results.

Because of this demanding and expanding requirement for fast processing of various data, new concepts of database paradigms are developed. In-memory databases are a relatively new concept contrasting the traditional way of storing data. In a Main Memory Database (MMDB) all data reside permanently in the main memory, compared with conventional data systems where data are resident on the disk. In a Disk Resident Database (DRDB) system data may be cached in RAM memory for access, whereas Main Memory Databases can also keep copies of data on disk [1].

The lowered prices of RAM and CPU units, make storing of large database contents in memory achievable. This enables the Main Memory Databases to become reality, and there is a variety of this type of databases on the market.

Direct access to the Main memory provides the In-memory databases with better response times and transactional throughputs, as opposed to disk resident databases, that are executing a significant number of I/O operations. This is especially important for all applications that are demanding processing of huge amounts of data in real time. Examples are some robust e-commerce solutions, which demand substantial scalability and where processing of a lot of application data efficiently is a must.

The difference in the way of storing data between these two types of databases implies specific design and performance. The main differences can be listed as follows [1]:

- The access time for main memory databases is much less than for disk storage databases;
- The main memory is normally volatile, while disk storage is not, but it is possible to construct nonvolatile main memory as well;
- Disks have high and fixed costs per access, which do not depend on the amount of data that is retrieved during the access. For this reason disks are block-oriented storage devices, while the main memory is not block oriented;
- The layout of data on a disk is much more critical than the layout of data in main memory, since sequential access to a disk is faster than random access. Main memory is normally directly accessible by the processor(s), while disks are not. This may make data in the main memory more vulnerable, than disk resident data, to software errors.

This paper will explain the architecture of both concepts via their represents SAP HANA (in-memory database) and Microsoft SQL Server (disk resident database).

II. The SAP HANA In-Memory Database

The general goal of the SAP HANA database is to provide a main memory-centric data management platform to support pure SQL for traditional applications, as well as a more expressive interaction model specialized to the needs of SAP applications [2].

It can be deployed as on-premise or in the cloud. Used as an on-cloud instance, SAP HANA provides all the benefits that cloud computing offers. Cloud computing is becoming an adoptable technology for many of the organizations with its dynamic scalability and usage of virtualized resources as a service through the Internet [3].

Taking a robust e-commerce application as a example, the main advantages of cloud computing in that case would be [3]:

- Backup of data;
- Storage – the cloud allows its users to store almost all types of content;
- Accessibility – any data stored in the cloud can easily be accessed from almost any device;
- Collaboration – the cloud allows multiple users to work and edit documents at the same time;
- Unlimited Accessibility;
- Security and Stability;
- Full Scalability;
- Minimal Costs or No Costs.

SAP HANA offers a platform which is suited the best for performing real-time analytics, developing and deploying real-time applications. Major databases that are offered on the market are good at either transactional or analytical workloads, but not both. When transactional Database Management Systems are used for analytical workloads, they require separating the data in different databases, one for OLTP (Online transaction processing) and one for OLAP (Online analytical processing). Due to its hybrid structure for processing both transactional and analytical workloads fully in memory, SAP HANA consolidates the two landscapes (OLAP and OLTP) in a single database. Since the data in the SAP HANA database reside completely in memory, complex calculations, functions and data-intensive operations are processed directly on the data, thus reducing the time consuming and costly movements of data between the database and the applications [4], as well as additional I/O

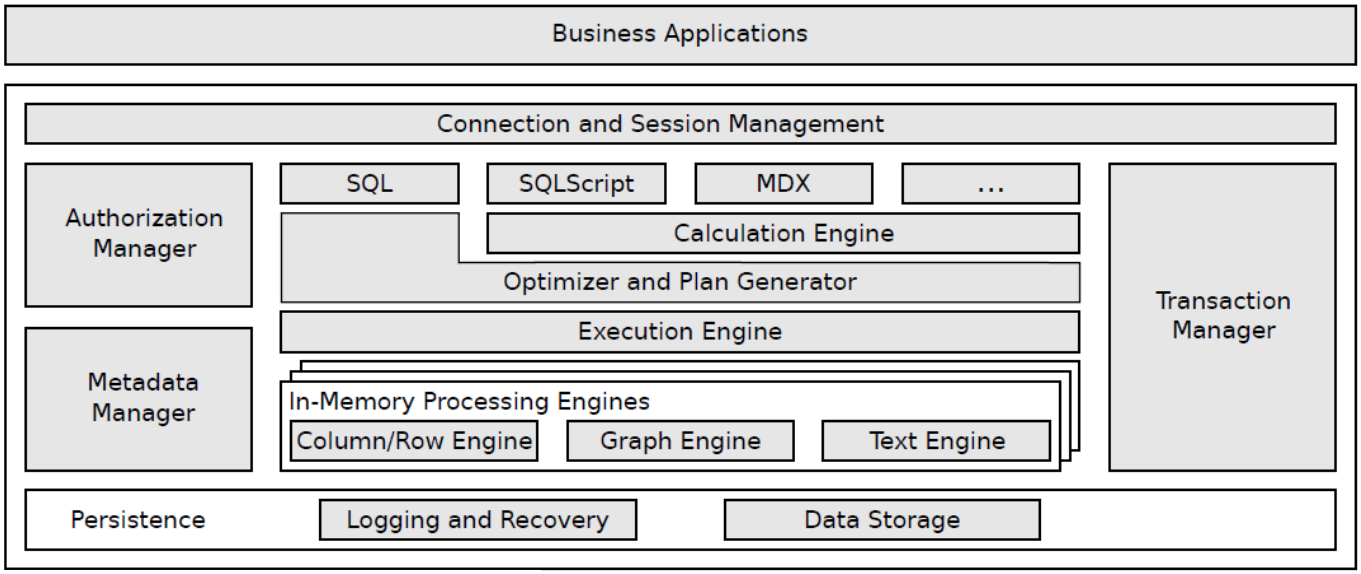


Fig. 1. The SAP HANA Architecture [2]

operations.

The core of the SAP HANA database consists of three in-memory processing engines: the Column/Row engine, the Graph engine and the Text engine (Fig 1). Relational data resides in tables in column or row layout in the combined column and row engine, and can be converted from one layout to the other. The graph and text data reside in the Graph engine and the Text engine respectively [2].

All data are kept in main memory as long there is enough space available. One of the most important features is that all data structures are optimized for cache-efficiency, compared with the organization in disk blocks in the traditional database architecture. The data are also compressed using various compression schemes. When the limit of the main memory is reached, the entire data objects are unloaded from the main memory and reloaded back in when required [2].

SAP HANA supports both row and column table layouts. Column oriented tables provide a greater level of compression of data and improve query efficiency since only the columns that are needed for a certain query are accessed.

The following can be listed as the main advantages of column tables [2]:

- OLTP scenarios can greatly benefit from the compression schemes available in column tables;
- Real-world transactional workloads have larger portions of read operations than standardized benchmarks. Hence, the read-optimized column-oriented storage layout may be more appropriate for OLTP workloads than suggested by the benchmarks;
- Column-stores usually follow the simple "append-only" scheme: when an existing row is updated, the current version is invalidated and a new version is appended. This scheme is simpler than in-place updates, as it requires neither reordering nor encoding of the values;
- Column-tables greatly reduce the need for indices: the high scan performance of column-stores on modern hardware permits the users to have indices only for primary keys, columns with unique constraints and frequent join columns. In all other cases, the scan performance is good enough without indices, especially in small tables or small partitions with up to a few hundred thousand rows.

III. Microsoft SQL Server

The core application service in Microsoft SQL Server is the Database engine. Its main task is to maintain, secure and process data in relational format. A transaction log keeps a record of every change that is made to the database.

In SQL Server the data are stored in pages that contain rows of data, and every page is 8KB in size. Eight pages make up an extent, and the database engine keeps track of which extent is allocated to which tables and indices [5].

The database engine in SQL Server manages the speed of access to data in several ways. Fast access to data in SQL Server can be achieved with the creation of indices. Another way is to keep frequently accessed data in memory and the excess of memory for an SQL Server Instance is used as data cache. The data pages can be stored in the cache and when a query is requested from the database, the database engine first checks if the requested pages are already in the memory cache. If not, it reads them off the disk [5].

IV. OLTP Comparison Tests

A. *The TPC-C Benchmark*

To compare the performances of both database concepts, series of OLTP tests were conducted on targeted databases SAP HANA and MS SQL SERVER. For this purpose the TPC-C Benchmark was used.

The TPC-C Benchmark (which has obtained its name from the Transaction Processing Performance Council, which has devised it) is OLTP workload which is composed from a set of read-only and updates intensive transactions. Their purpose is to simulate a complex OLTP application environment, which is done by exercising a breadth of system components associated with such environments, which are characterized with several aspects like simultaneous execution of multiple transaction types, on-line and deferred transaction execution modes, multiple on-line terminal sessions, moderate system and application execution time, significant disk input/output, transaction integrity etc.

A TPC-C Benchmark performance report contains measures of the number of orders that are processed per minute. Multiple transactions simulate activities for processing orders, and each transaction is subject to a response time constraint [6].

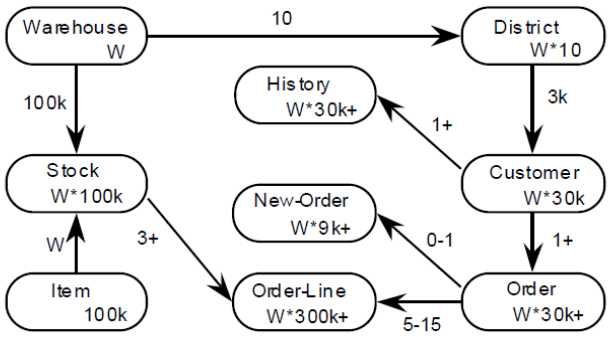


Fig. 4. The TPC-C Data Model [6]

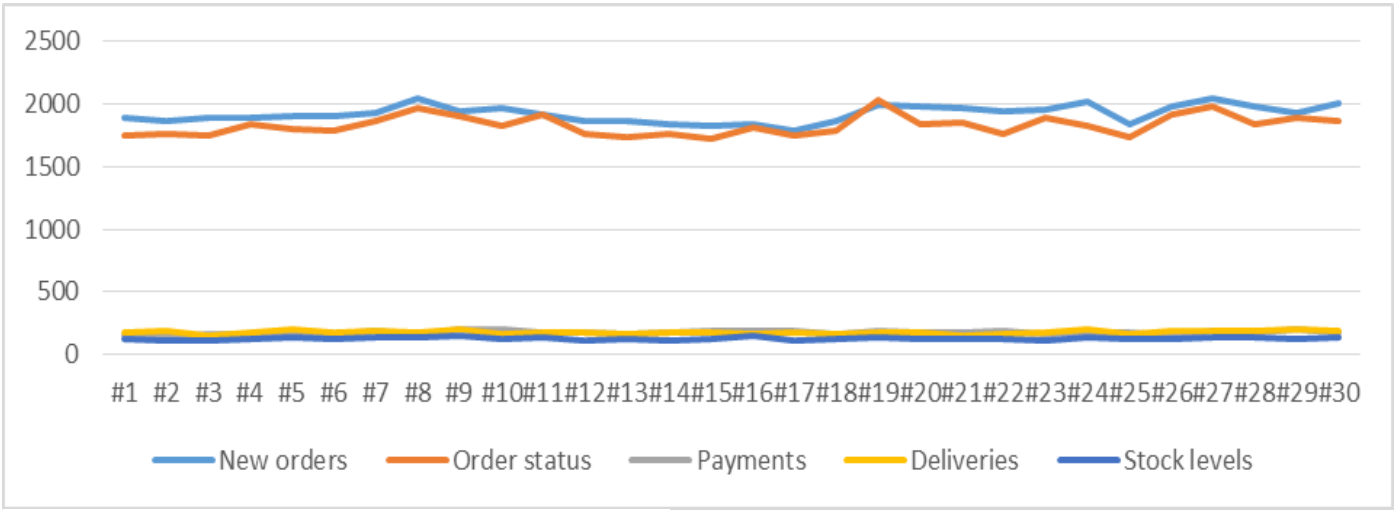


Fig. 3. SAP HANA TPC-C Benchmark Results

The benchmark performed in this paper used a defined data model (Fig 2) and a set of operations that were triggered in a unit of time. They simulate life-like whole sale scenario in which the portrayed Company was a wholesale supplier with distributed sales districts associated with warehouses Each Warehouse covered 10 Districts and accordingly each District served 3000 Customers. All warehouses maintained Stocks for 100000 Items.Several operations were used to simulate a real life scenario: Place a new order from the Customer, Check the status of existing order, Receive payment from a Customer, Process orders for delivery and Examine specific Items stock levels.

TABLE I. TPC-C BENCHMARK AVERAGE OPERATIONS IN 1 MINUTE

	SAP HANA	MSSQL Server
New Orders	1923	731
Order Status	1832	698
Payments	173	65
Deliveries	175	65
Stock Level	126	49
Total Operations	4229	1608

30 iterations of one minute tests were conducted with the aforementioned operations in the following proportion: 45% new orders, 43% check order status, 4% check stock levels, 4% payments, 4% delivery. The TPC-C data model was filled with data for two Warehouses in documented volumes and the database was reinitialized to the initial state before every iteration. Every test iteration was started with 100 parallel connections, which simulated 100 concurrent users.

An open source client was used, programmed in C# which was developed to perform the TPC-C Benchmark for MS SQL Server [7]. This Client was translated and adapted to conduct the same tests for SAP HANA.

For SAP HANA testing, a cloud instance with 64 GB of RAM memory was used. The tests on MS SQL Server where done on version 2008 R2 installed on 64-bit virtual Windows Server 2008 R2, with the following hardware configuration: Intel Xeon CPU E5620 @ 2.40 GHz and 20 GB of RAM memory.

B. Test Results

In its initial state, the size of the data in the databases used for the TPC-C benchmark was 236 MB on the disk for MS SQL Server and 94 MB in the main memory for SAP HANA. During the conducted test, the count of created new orders, checks of order statuses, processed payments, delivery and check stock level processes were recorded.

The test results for TPC-C benchmark conducted on SAP HANA can be seen on Fig 3. The result values for all operations measured in one minute tests, conducted in 30 iterations, are presented. The lowest value for the New orders operation is 1791 and the highest is 2050. For the Check Order Status operation the lowest value is 1723 and the highest is 2030. The lowest value for the Payment process operation is 133 and the highest is 198. For the Deliveries operation the lowest result is 153 and the highest is 199. For the Check Stock Level operation the lowest value is 112 and the highest is 143.

On the SQL Server side (Fig. 4) the lowest amount of completed New Order operations is 615 and the highest is 791. The lowest value for the Check Order Status operation is 609 and the highest value is 761. For the Payments operation the lowest result is 50 and the highest 79. For the Delivery operation the highest value is 81 and the lowest is 47. For the Check Stock Level operation the lowest value is 38 and the highest is 57.

With an average 4229 completed operations in one minute, SAP HANA is approximately 2.5 times faster than SQL

Server with an average completed 1608 operations. From the perspective of TPC-C separate benchmark operations, SAP

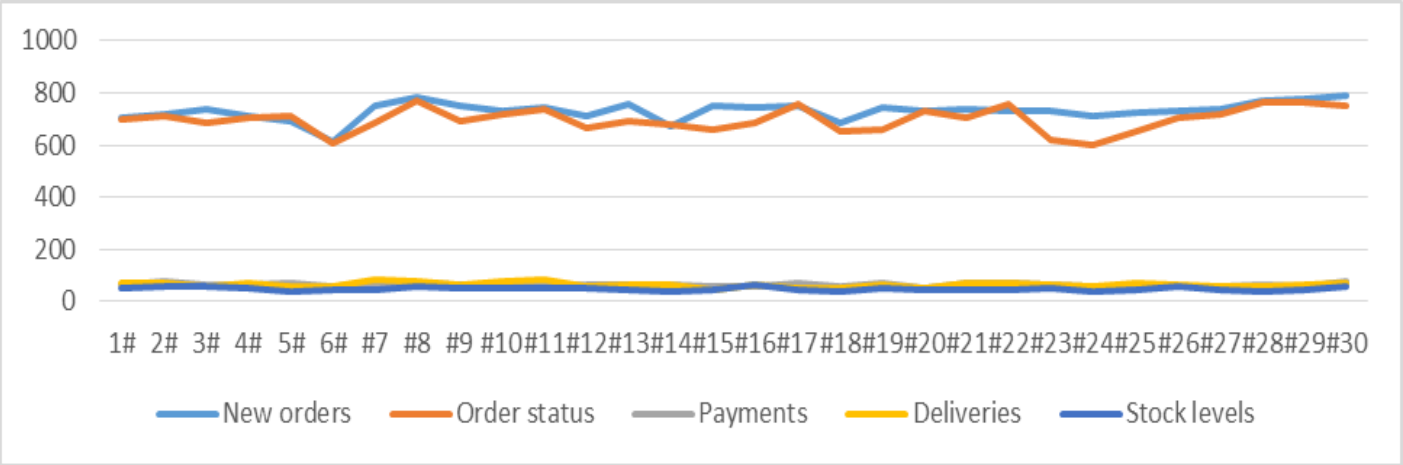


Fig. 4. MS SQL SERVER TPC-C Benchmark Results

HANA is faster than SQL Server in every segment by more than 250% (Table 1).

V. Conclusion and Future Work

In this article we've tried to cover the main characteristics of In-memory and traditional disk resistant databases. To compare the performances of the two types of databases, TPC-C test were performed, using two representatives from both concepts, namely SAP HANA (In-memory database) and Microsoft SQL Server 2008 R2 (disk resident database). This test included the operations for Place new order, Check for order status, Payment process, Delivery and Check stock level. The test results showed that SAP HANA is approximately 2.5 times faster in completing these operations.

The large speed of In-memory databases is a result of reduced I/O operations, since all data contents are kept in main memory. However, this type of database requires more complicated hardware architecture and thus increased cost. Considering the fact that the prices for hardware components needed for operation of this kind of databases are constantly decreasing, though, as a result of increasing competition in that branch of manufacturers, In-memory databases become increasingly more affordable.

The choice of a type of a database which will be implemented by a company depends on its needs. For a small or medium company, with a limited number of users, an appropriate choice would be a disk resident database such as SQL SERVER, Oracle or any other commercial database available on the market. In case when the budget is limited, using an open-source solution such as MySQL is a valid option.

However, if a company wants to announce its products catalog online, available for a huge number of users, or it is a multinational company with dispersed units around the world, it is necessary to provide a sustainable solution that can respond to this challenge. The scalability and availability that allows the hardware environment and software architecture, including all the benefits that In-memory databases provide, certainly make them a worthy solution for the execution of such tasks.

Our future work will be focused on developing more test benchmarks, including other databases from both concepts available on the market. The main goal will be to measure the various database performances on an identical hardware environment, which will present a clearer picture about the differences among them.

References

[1] Hector Garcia-Molina, Kenneth Salem, "Main Memory Database Systems: An Overview" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL 4. NO.6, DECEMBER 1992.

[2] Franz Farber, Norman May, Wolfgang Lehner, Philipp Große, Ingo Muller Hannes Rauhe, Jonathan Dees, "The SAP HANA Database – An Architecture Overview", SAP AG, 2012.

[3] A.M.Mansuri, Manish Verma, Pradeep Laxkar, "Benefit of Cloud Computing for Educational Institutions and Online Marketing", Department of Computer Science and Engineering, MIT Mandsaur, Mandsaur, India, Department of Computer Science and Engineering, ITM Universe, Vadodra, India, 2014.

[4] SAP AG, "What is SAP HANA?", September 2012, http://scn.sap.com/docs/DOC-60338, Date of last access 20.02.2015.

[5] Ray Rankins, Paul Bertucci, Chris Gallelli, Alex T. Silverstein, "Microsoft SQL Server 2008 R2 Unleashed", SAMS, Pearson Education, Inc, 2011.

[6] Transaction Processing Performance Council (TPC), "TPC BENCHMARK C Specification", February 2010, http://www.tpc.org/information/current_specifications.asp, Date of last access 15.02.2015.

[7] Wes Brown, SQLServerIO/TPCCBench, Jun 2011,

[8] https://github.com/SQLServerIO/TPCCBench, Date of last access 07.02.2015.

Simplifying parallel implementation of algorithms on Hadoop with Pig Latin

Eftim Zdravevski
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: eftim.zdravevski@finki.ukim.mk

Petre Lameski
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: petre.lameski@finki.ukim.mk

Andrea Kulakov
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: andrea.kulakov@finki.ukim.mk

Sonja Filiposka
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: sonja.filiposka@finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: dimitar.trajanov@finki.ukim.mk

***Abstract*—In this paper we present a general technique for parallelizing regular algorithms with the tools the Hadoop ecosystem offers: MapReduce, HDFS, HBase and Pig. This framework can be applied for parallelizing algorithms for feature selection, clustering, machine learning etc. It consists of several steps: load the datasets in HDFS, apply some transformations if they are needed, store the datasets in HBase, and implement the algorithm in Pig with the help of User Defined Functions.**

***Keywords*—Hadoop, MapReduce, HBase, Pig, parallel algorithms, distributed algorithms**

I. Introduction

In the recent years companies, organizations and governments collect, process and analyze enormous volumes of data. For most of them the data is not only generated from their normal work, rather it a prerequisite for their success. As a result many companies have followed different ideas on how cope with the Big Data challenge. One idea was to scale-up hardware so it has more processing power that can handles the larger volumes of data, and it has proven to work up to a certain point. However, after this point is reached, this idea can not work. That lead to the other idea of distributing the computation and data storage to clusters. Even though this is not so new idea in general, it was not until about ten years that it started to gain popularity. Inspired by Google's approach described in the 2004 MapReduce [1] and 2006 Big Table [2], many other companies and open-source projects followed similar pathways developing different distributed systems. One of the most popular such systems is Apache Hadoop. It is open-source software that contains a set of algorithms for distributed processing, scheduling and storage of large datasets on computer clusters. It is well established framework and Hadoop Wiki [3] lists some of its prominent users like Yahoo, Facebook, Ebay, Adobe etc.

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia

The MapReduce programming paradigm [1] [4] is essential to the distributed computation and storage that Hadoop achieves. It consists of two phases: map and reduce. The first phase, map, threats the data processing problems as embarrassingly parallel by splitting the data into distinct subsets that can be processed in parallel. The reduce phase is second and final aggregates the output from the map phase and produces the final result. In other words, the map procedure can perform variety of operations like: reading, projecting, filtering and sorting data. The output from this phase is an intermediate result usually comprised of a list of keys and values. These are mandatory for the reduce phase. Hadoop makes sure that the output gets to the reduce procedures in proper order so it can perform some summary or aggregate operation. Even though the MapReduce model is fairly restricted, its simplicity is making it very suitable and efficient for extremely large-scale implementations across thousands of nodes.

Hadoop with its different services schedules, distributes, orchestrates and monitors communications, data transfers, while providing redundancy and fault tolerance. There are many services (i.e. subsystems) in Hadoop that aid accomplishing the previous goals, but three of them are most notable: YARN (MapReduce2), HDFS and HBase [5] [6] [7].

The fundamental idea of YARN (i.e. Yet Another Resource Negotiator) [8] is to take care of resource management and job scheduling/monitoring, by splitting these responsibilities into separate daemons: a global ResourceManager and per-application ApplicationMaster. The ResourceManager is the ultimate authority that arbitrates resources among all the applications in the system. The per-application ApplicationMaster is responsible for negotiating resources from the ResourceManager and working with the NodeManagers to execute and monitor the component tasks. In other words, YARN is responsible for allocating resources to the MapReduce jobs, distributing them to the most appropriate nodes, etc.

Hadoop Distributed File System (HDFS) [9] is a file

system that provides scalable, fault-tolerant, distributed storage system that works closely with MapReduce which was designed to span large clusters of commodity servers. The combined resources of the servers within the cluster can easily grow with the demand. An HDFS cluster is comprised of a NameNode which manages the cluster metadata and DataNodes that store the data. The file content is split into large blocks (typically 128 megabytes), and each block of the file is independently replicated at multiple DataNodes. The blocks are stored on the local file system on the DataNodes.

HBase is an open source, non-relational, distributed database modeled after Google's BigTable. It runs on top of HDFS (Hadoop Distributed Filesystem), providing BigTable-like capabilities for Hadoop [10] [11] [12]. In other words, it provides a fault-tolerant way of storing large quantities of sparse data. HBase is a NoSQL (Not Only SQL) database and is not a direct replacement for a classic Relational SQL databases [13]. Unlike traditional databases where normalization of data and splitting it into related tables is the substance of the design, designing HBase tables takes a different approach which analyzes the usage patterns. Motivations for this approach include simplicity of design, horizontal scaling, and finer control over availability. The data structures used by NoSQL databases, including HBase, differ from those used in relational databases, making some operations faster in NoSQL and others faster in relational databases. The particular suitability of a given NoSQL database depends on the problem it must solve. Tables in HBase can serve as the input and output for MapReduce jobs run in Hadoop. In the parlance of Eric Brewers CAP theorem, HBase is a CP type system (i.e. Consistent and Partition tolerant). [14].

Because of its simplicity the MapReduce programming model has become popular. By some users this model is preferred over the traditional SQL which is a high-level declarative approach. Be that as it may, the extreme simplicity of MapReduce leads to much low-level hacking to deal with the many-step, branching dataflows that arise in practice. Furthermore, users must repeatedly code standard operations such as join by hand. These practices increase development time, introduce bugs, harm readability, and may obstruct optimizations [15]. A group at Yahoo motivated by these repeatable tasks on daily basis, has developed a scripting language called Pig Latin. Pig is a high-level dataflow system that is a compromise between SQL and MapReduce. Pig offers constructs for data manipulation similar to SQL, which can be integrated in an explicit dataflow. Pig programs are compiled into sequences of MapReduce jobs, and executed in the Hadoop MapReduce environment [16].

II. Related work

This section describes some of the most recent work on parallelizing different algorithms with MapReduce.

The authors in [18] address the problem of efficient feature evaluation for logistic regression on very large data sets. Here they present a new forward feature selection heuristic that ranks features by their estimated effect on the resulting model's performance. They test the method on already available datasets from UCI, but also generate artificial datasets for which they know the logistic regression coefficients. They use that to evaluate the selected features.

By using the MapReduce paradigm in [21] a data intensive parallel feature selection method is proposed. In each map node, a method is used to calculate the mutual information and combinatory contribution degree is used to determine the number of selected features.

In [26] an implementation based on the MapReduce programming model of Naive Bayes is proposed. During the map phase all counts needed for calculating the conditional probabilities are emitted, and during the reduce phase they are aggregated.

A parallel implementation of the SVM algorithm for scalable spam filtering using MaprReduce is proposed in [22]. By distributing, processing and optimizing the subsets of the training data across multiple participating nodes, the distributed SVM reduces the training time significantly.

In [23] the authors propose a method for reducing the dataset to a small but representative subset that can be later on used for faster machine learning. Also here the speedup is being calculated against a cluster with 1 node or with minimal number of nodes so that the computation finishes in reasonable time.

In [24] an approach based on MapReduce for distributed column subset selection is proposed. In this approach each node has access to a random subset of features.

A wrapper approach for parallel feature selection is proposed in [25]. Here with features are added to the selected set if after their addition, the performance of the classifier does not degrade. Then in a second phase from the subset obtained in the previous step, features are removed if their discarding does not degrade the classifier performance.

III. Framework design

Writing parallel computer programs is more difficult than writing sequential ones, because parallelization introduces several new types of potential software bugs of which most common are race conditions, communication and synchronization between the different subtasks. Choosing Hadoop as environment for parallelization of algorithms overcomes many of those challenges without needing the programmer to put much effort for solving those kinds of challenges.

The framework that we propose in this paper consists of several phases, as shown on Fig. 1, and each of them is described in the following subsections.



Fig. 1. Data flow phases based on MapReduce and Hadoop

A. Load data into HDFS

This is the first and most simple phase. This phase should be performed once or multiple times, depending on how the dataset is structured. The most common formats for datasets are:

- CSV (comma separated values). This format is usually used to store dense datasets.
- ARFF (Attribute-Relation File Format). Also used to store dense datasets.
- EAV (Entity Attribute Value). Used to store sparse matrices that have a lot of zeros and some non-zero elements.

If the dataset is only one file then this it will be copied from the Linux File System to HDFS using a simple command. This means that for this step cannot have parallelism. However if the dataset is dispersed into multiple files, then all of them can be copied simultaneously to HDFS. Be that as it may, this step usually is very fast compared to the following steps for machine learning, so its parallelization is not necessary at all.

B. Transformation and loading data into HBase

After the previous step III-A is finished the dataset files reside on HDFS. As it is extensively described in [9], each file in HDFS is replicated accross several nodes for reliability. A typical file in HDFS is gigabytes to terabytes in size, splitted in blocks of 128 MB by default. If the files are too small than that could degrade the performance of the system and limit the level of parallelism. Map tasks usually process a block of input at a time. If the file is very small and there are a lot of them, then each map task processes very little input, and there are a lot more map tasks, each of which imposes extra bookkeeping overhead. Ideally the dataset that we have loaded to HDFS is one large file dispersed on multiple blocks so while we load it, transform it and store it in HBase we can have greater parallelism. Nevertheless, this step again is usually very fast especially compared to the step that performs the actual machine learning algorithm, so we do not recommend to spend too much time on optimizing the file sizes for better parallelism.

Even though we can achieve parallelism while processing files stored on HDFS, the control of degree of parallelism is difficult, more involved and at very low-level. On the other hand, HBase offers many other services built on top of HDFS, among which is a much better control of the degree of parallelism. This is due to the fact that the data in HBase is stored in a structured manner, while having various mechanisms that simplify random reads and writes from rows and columns. Namely, HBase tables are divided into potentially many regions, while one or more regions are serviced by a region server. The tables can be horizontally and vertically segmented while they are physically stored in HBase. Because many machine learning applications access the data by rows, in this paper we will continue to discuss only horizontal segmentation. As HBase was designed with very large tables in mind, a common use case is the following. A table at creation has only one region, which is serviced by one region server (a physical node in the Hadoop cluster). When this table is loaded with data it gets bigger and at some point it will become too big, so HBase will split its region into two regions. Then the new region will be assigned to the same region server or can be moved to another region server. The default splitting threshold is 10 GB. There are numerous reasons why HBase was designed that way, and we will not go into details about that. From parallelization perspective, this can pose a challenge, because for the automatic splits there are no guarantees that every region will contain equal amount of data, when are the splits going to occur exactly, are the regions going to be served by different region servers (nodes) etc. Further more, if one is using Hadoop for research purposes only then the dataset may not be that large, thus never overcoming the threshold for splitting. To overcome this challenge we can pre-split the tables on creation. This in turn means that the table can be configured at creation time to be stored on as many-regions as needed. Usually the number of regions is a multiple of the number of HBase region servers. The logic for having more region servers than acutal nodes is because the nodes are multi-core machines, so different threads on the same node can service different regions.

Before loading the dataset in HBase, we need to define the table structure and create it. Column names and data types are provided when storing data in each row, so at creation time we need to only specify a table name and a column family. There are some advanced configuration features that can be specified, but they are not topic of this discussion. Be that as it may, there is one very important decision that we need to make before loading data in the table. Because HBase tables, unlike SQL tables, cannot have secondary indexes, the primary key (row key) needs to be designed according to the usage patterns of the table. There are many considerations when designing the row key and they are very important for production use of HBase tables. However, for scientific use and for parallelizing machine learning algorithms, we need a simple design that allows uniform data distribution across nodes. In most scientific datasets the data instances (i.e. rows) do not have ids for their instances, or if they do they are not used for the actual machine learning. Nevertheless, in order to store a row in a HBase table, it needs a row key. For flat flies like CSV, ARFF the row key can be the line number of the instance. However, sequential row keys are very bad choice for HBase tables because the inserts will always be on the last region, therefore having no parallelism during the load, a problem called Region Server hotspotting. There are multiple ways of overcoming this problem, and one of them is a technique called salting. With this technique each sequential id is salts row keys with a prefix. The prefix is usually the modulo number between the original sequential id and the number of regions. Even though, this is very important topic, the step of loading the datasets in HBase is not the primary field of interest in this paper.

Once the dataset files are loaded into HDFS we need to transform them if needed and store them in HBase. If we have totally N rows in the dataset, and M regions, then we would like to distribute the data uniformly so each region gets N/M rows. This in turn means that we need to specify $M-1$ split points when creating the table. If we use sequential ids for the row key (like the line number in the file), than these split points would be: $N/M, 2N/M, 3N/M, ..., (M-1)N/M$. If we use a more sophisticated row key design, then the split points should reflect that design. For instance, if we take the modulo number of the id and the number of regions, then each region would get almost the same number of rows. This design of the row key allows fast random reads and writes, and additionally it facilitates addition of new data to the table at a later time without needing to redesign the table for equally dispersed load across regions. the following example shows

how a table can be pre-splitted on creation. The row key design is described with the function in listing 1. It returns a tuple in which the first element is the padded modulo number and the second part is the padded sequential id.

```
(pad(seq_id % num_regions), pad(seq_id))
```

Listing 1. Row key design

The numbers are padded with zeros so that they are lexicographically sorted. For instance if the ids vary from 1 to 100000 and we have 5 regions, then the Id 123 would be encoded to the following row key: $123\%5 = 3, pad(3) = 3, pad(123) = 000123 => row_key = (3, 000123)$. The create statement for a table that has one column family 'r' and has 4 split points is shown in listing 2.

```
create 'dataset5', {NAME=>'r',
    COMPRESSION=>'gz', VERSIONS=>'1'},
{SPLITS=>[ "(1,", "(2,", "(3,", "(4," ]}
```

Listing 2. Crating HBase table with pre-split regions

Once the HBase table that will contain the dataset is created with appropriate split points for even data distribution across the cluster, the data can be loaded. One can write pure MapReduce jobs in Java or Python. If we choose that path, then we need to write a separate map and reduce function for each task. However, by using a scripting language called Pig Latin [15] we can write scripts from a higher-level perspective. These Pig scripts generate MapReduce tasks in the background so the programming effort is simplified and the development time is reduced. The downside of using Pig is that when Pig scripts are compiled into MapReduce jobs, there is some overhead but for longer running MapReduce tasks is insignificant because it adds up to 1 minute to the executing time. The listing 3 shows how we can load EAV files with Pig Latin. All variables starting with $ are parameters that are passed to the script on execution time.

```
register '$udf_path' using jython as
    paddingUDF;
eav_data = LOAD '$hdfs_data_path' USING
    PigStorage(',') as (id:int, feature:int,
    value:int);
eav_data_pad = FOREACH eav_data GENERATE
  paddingUDF.generate_rowkey(id,
      $padding_digits_id, $modulo_number) as
      idPad,
  paddingUDF.pad_number(feature,
      $padding_digits_feature) as featurePad,
  value;
eav_data_final = FOREACH eav_data_pad GENERATE
  idPad,
  [featurePad,value] as values;
STORE eav_data_final INTO '$table_dataset'
    USING
org.apache.pig.backend.hadoop.
    hbase.HBaseStorage('r:*');
```

Listing 3. Loading EAV files in HBase with Pig

The above script when compiled into MapReduce jobs would have only a map phase which will read the data from the HDFS file and store it in HBase. Because there is no grouping of keys needed, a reduce phase will not be generated. It uses two user-defined functions (UDFs) that are written in Python, but in order to be compatible with Pig and the rest of Hadoop, are compiled into Java byte code using Jython. The *generate_rowkey* function, shown in listing 4 calculates the row key according to the design shown in listing 1. With the function *pad_number*, shown in listing 5, the numbers are padded with zeros.

```
@outputSchema("padded:chararray")
def pad_number(number, numZeros):
    f = '%0' + str(int(numZeros)) + 'd'
    paddedNumber = f % int(number)
    return paddedNumber
```

Listing 4. Python UDF for generating row keys

```
@outputSchema("rowkey:(id_mod:chararray ,
    id:int)")
def generate_rowkey(id, id_num_digits,
    mod_number, mod_number_digits=3):
    prefix = int(id) % int(mod_number)
    return (pad_number(prefix,
        mod_number_digits), pad_number(id,
        id_num_digits))
```

Listing 5. Python UDF for padding numbers with zeros

During this step we can add various methods for data preprocessing like discretization, transformation and other methods that rely only on the values. But if the transformations need something like normalization that need the mean value for the whole dataset, then a separate step would be needed.

C. Processing HBase tables

After the dataset is loaded in a HBase table we can continue implementing machine learning algorithms. In general, this phase can be comprised of several substeps of data processing, depending on the nature of the algorithm that is being implemented. In this subsection we show how the mean value of each feature can be calculated. This is a simple task, but nevertheless, it illustrates this step of the framework. Listing **??** shows the Pig Latin script that calculates mean values. All parameters that start with *$* can be passed to Pig script when invoking it. Such parameters are the table names, number of features, index of the class value, number of padding digits etc. Lines 2 through 6 load the data from the table '$table_dataset'. In line 3 *'r:*'* denotes that all columns in the column family *r* will be loaded and line 6 denotes that they will be available as a dictionary (map) in the Pig script. Then at line 8 the UDF *expandFeatures* is invoked which accepts these arguments: *r* -the dictionary (map) of pairs *(featureIndexPadded, featureValue)*, the number of features in the dataset *$num_features*, maximum number of padding digits *$num_features_digits*, and the index of the class value *$label*. The UDF code is shown in listing 7. It will process the passed dictionary of feature indexes and values and will generate a list of triplets *(featureIndex, featureValue, class)*. The great thing about HBase is that it doesn't need to store the empty features from the dataset and each row can have different number of

columns. However when a row is loaded, like in the UDF *expandFeatures*, we can generate the zero-valued cells so they can be used for calculation of the mean value of the feature. In lines 10 and 11 the rows are grouped by feature index and we calculate the average of their values. Lines 15 through 17 store the mean values in a table that can be created in a similar manner as explained in III-A. We could easily export the calculated data in CSV files to HDFS.

What happens in the background for the script 6, is very peculiar. Pig will determine the number of regions of the table '$table_dataset' and it will start that number of map tasks. The number of reduce tasks is by default 1, but this can be also manually specified and does not depend of the table structures. During the map phase happen these statements: loading of data (lines 2-6), expanding each row to a list of tuples, merging (union) the list of tuples generated from each row into a final list (the *FLATTEN* operator in line 8) and it will emit the grouping key *featureIndexPadded* in line 10. Then in the reduce phase the rows will be grouped by the *featureIndexPadded* key, the mean value will be calculated (lines 12-14) and finally the result will be stored in the table *$table_feature_mean_value*. In this table the row key is *featureIndexPadded* and the mean value is stored in column family *r* and column *featureValueMean*. We can easily change this statement to export the results in a CSV file to HDFS.

```
1  register '$udf_path' using jython as
       paddingUDF;
2  dataset = LOAD '$table_dataset' USING
       org.apache.pig.backend.hadoop.
3  hbase.HBaseStorage('r:*',
4         '-loadKey=true') AS
5       (rowkey:tuple(prefix_padded:chararray,
           id_padded:chararray, id:int),
6        r:map[]);
7  dataset_expanded = FOREACH dataset GENERATE
8    FLATTEN(paddingUDF.expandFeatures(r,
         $num_features, $num_features_digits,
         '$label'));
9
10 feature_value_class_group = GROUP
       dataset_expanded BY (featureIndexPadded);
11
12 feature_value_class_mean = FOREACH
       feature_value_class_group GENERATE
13   group as featureIndexPadded,
14   AVG(dataset_expanded.featureValue) as
         featureValueMean:double;
15 STORE feature_value_class_counts INTO
       '$table_feature_mean_value' USING
16      org.apache.pig.backend.hadoop.
17      hbase.HBaseStorage('r:featureValueMean');
```

Listing 6. Pig script for calculating mean value of features

The code in listing 7 shows how UDF *expandFeatures* is implemented. Lines 1 and 2 describe the output format of the data that is being returned. In this the function will return a bag (Pig Latin equivalent of a unordered list) of tuples. Each tuple is a triplet containing the featureIndex (padded with zeros), the featureValue and the class. For the calculation of the mean values of the features the class is not needed, but for many other algorithms and metrics it will be. Line 5 calls a function for decoding the input arguments from byte representation to appropriate types - string and dictionary. The in the loop in lines 8 through 11 from the dictionary of pairs feature index and feature value, we generate a list of triplets, which the function returns at line 12. The interesting thing that this function does is that from a row of the HBase table it generates a list of tuples, so in a way it's transposing a row into multiple rows with one composite column (i.e. the triplets). Then in the Pig script in listing 6, with the FLATTEN operator merges those rows into a common set of rows of triplets, thus transposing the whole dataset. Having all features, values and class as triplets in separate rows is useful and allows us to leverage the SQL-like capabilities of Pig Latin. Further on in separate scripts we can normalize the values of the dataset by joining to the set of mean values and doing appropriate normalizations.

```
1  @outputSchema("feature_value_class:bag{
2  t:(featureIndexPadded:chararray ,
       featureValue:double, class:chararray)}")
3  def expandFeatures(featureValuePairs,
       numFeatures, numDigits, classKey):
4      currentClass =
           featureValuePairs.get(classKey, 0)
5      (currentClassFinal,
           featureValuePairsFinal) =
           decodeBytes(currentClass,
           featureValuePairs)
6
7      feature_value_class = []
8      for featureIndex in range(1,
           int(numFeatures) + 1):
9          featureIndexPadded =
               pad_number(featureIndex, numDigits)
10         featureValue =
               featureValuePairsFinal.get(
               featureIndexPadded, float(0))
11         feature_value_class.append(
               (featureIndexPadded, featureValue,
               currentClassFinal))
12     return feature_value_class
```

Listing 7. UDF for expanding a HBase row into a list of tuples

D. Exporting results

After the main work is performed during the previous step III-C, the results need to be exported. Using Pig Latin, the output from the Pig scripts can be stored in HDFS files or HBase tables. In the Hadoop ecosystem there are advanced services like Flume or Sqoop that facilitate connectivity with RESTful web services, various SQL databases etc. The most simple way is however to export the results in HDFS files in a common format like CSV, and then to export the HDFS files to the Linux file system.

IV. Conclusion and Future Work

In this paper we have proposed a framework for parallelization of machine learning algorithms by using the Apache Hadoop platform including its services HDFS, Yarn MapReduce and HBase, and Pig Latin as a scripting language. We have demonstrated how can we manually set the degree of parallelism by pre-splitting the HBase tables so they have

optimal number of regions and even data distribution across regions. We have also provided exemplary user-defined functions written in Python for transforming and formatting the data while it's being loaded, and also a way to efficiently decode the sparse matrix for which we only keep the non-zero elements. We have provided an exemplary script that calculates the mean values of each feature.

In order to affirm the proposed framework, we will implement various machine learning algorithms with it, measure the impact of the parallelization with different cluster configurations on different datasets. The performance should be evaluated in terms of speedup vs the sequential versions of the algorithms.

REFERENCES

[1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, ser. OSDI'04. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251254.1251264

[2] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," in *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7*, ser. OSDI '06. Berkeley, CA, USA: USENIX Association, 2006, pp. 15–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267308.1267323

[3] "Hadoop wiki: List of institutions that are using hadoop for educational or production uses, howpublished = https://wiki.apache.org/hadoop/poweredby, note = Accessed: 2015-01-29."

[4] D. Miner, *MapReduce design patterns*. Sebastopol, CA: O'Reilly, 2013. ISBN 9781449327170

[5] C. Lam, *Hadoop in action*. Greenwich, Conn: Manning Publications, 2011. ISBN 9781935182191

[6] A. Holmes, *Hadoop in practice*. Shelter Island, NY: Manning, 2012. ISBN 9781617290237 1617290238

[7] T. White, *Hadoop: the definitive guide*, 3rd ed. Beijing: O'Reilly, 2012. ISBN 9781449311520

[8] "Apache hadoop nextgen mapreduce (yarn), howpublished = http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/yarn.html, note = Accessed: 2015-01-29."

[9] "Hdfs architecture guide, howpublished = http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, note = Accessed: 2015-01-29."

[10] L. George, *HBase the definitive guide*. Sebastopol, CA: O'Reilly, 2011. ISBN 9781449315771 1449315771. [Online]. Available: http://public.eblib.com/choice/publicfullrecord.aspx?p=769368

[11] Y. Jiang, *HBase administration cookbook master HBase configuration and administration for optimum database performance*. Birmingham: Packt Publishing, 2012. ISBN 9781849517157 1849517150 1849517142 9781849517140. [Online]. Available: http://site.ebrary.com/id/10598980

[12] N. Dimiduk and A. Khurana, *HBase in action*. Shelter Island, NY: Manning, 2013. ISBN 1617290521 9781617290527

[13] A. Awadallah and D. Graham, *Hadoop and the Data Warehouse: When to Use Which*. Coudera, Teradata, 2011. [Online]. Available: http://www.teradata.com/When-to-Use-Hadoop

[14] E. A. Brewer, "Towards robust distributed systems (abstract)," in *Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing*, ser. PODC '00. New York, NY, USA: ACM, 2000. doi: 10.1145/343477.343502. ISBN 1-58113-183-6 pp. 7–. [Online]. Available: http://doi.acm.org/10.1145/343477.343502

[15] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, "Building a high-level dataflow system on top of map-reduce: The pig experience," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1414–1425, Aug. 2009. doi: 10.14778/1687553.1687568. [Online]. Available: http://dx.doi.org/10.14778/1687553.1687568

[16] A. Gates, *Programming Pig*. Sebastopol: O'Reilly Media, 2011. ISBN 9781449317690 1449317693 9781449317683 1449317685. [Online]. Available: http://public.eblib.com/choice/publicfullrecord.aspx?p=801461

[17] E. Zdravevski, P. Lameski, A. Kulakov, and D. Gjorgjevikj, "Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*, Sept 2014. doi: 10.15439/2014F500 pp. 387–394.

[18] S. Singh, J. Kubica, S. Larsen, and D. Sorokina, "Parallel large scale feature selection for logistic regression." in *SDM*. SIAM, 2009, pp. 1172–1183.

[19] Z. Zhao, R. Zhang, J. Cox, D. Duling, and W. Sarle, "Massively parallel feature selection: an approach based on variance preservation," *Machine Learning*, vol. 92, no. 1, pp. 195–220, 2013. doi: 10.1007/s10994-013-5373-4. [Online]. Available: http://dx.doi.org/10.1007/s10994-013-5373-4

[20] "Mpi: A message-passing interface standard. version 3.0, howpublished = http://www.mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf, note = Accessed: 2015-01-29."

[21] Z. Sun and Z. Li, "Data intensive parallel feature selection method study," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, July 2014. doi: 10.1109/IJCNN.2014.6889409 pp. 2256–2262.

[22] G. Caruana, M. Li, and M. Qi, "A mapreduce based parallel svm for large scale spam filtering," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, vol. 4, July 2011. doi: 10.1109/FSKD.2011.6020074 pp. 2659–2662.

[23] I. Triguero, D. Peralta, J. Bacardit, S. García, and F. Herrera, "Mrpr: A mapreduce solution for prototype reduction in big data classification," *Neurocomputing*, vol. 150, pp. 331–345, 2015.

[24] A. K. Farahat, A. Elgohary, A. Ghodsi, and M. S. Kamel, "Distributed column subset selection on mapreduce," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 171–180.

[25] A. Guilln, A. Sorjamaa, Y. Miche, A. Lendasse, and I. Rojas, "Efficient parallel feature selection for steganography problems," in *Bio-Inspired Systems: Computational and Ambient Intelligence*, ser. Lecture Notes in Computer Science, J. Cabestany, F. Sandoval, A. Prieto, and J. Corchado, Eds. Springer Berlin Heidelberg, 2009, vol. 5517, pp. 1224–1231. ISBN 978-3-642-02477-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02478-8_153

[26] L. Zhou, H. Wang, and W. Wang, "Parallel implementation of classification algorithms based on cloud computing environment," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 10, no. 5, pp. 1087–1092, 2012.

Automatic Injection Of Web Semantics To Static And Dynamic Web Sites

Gorjan Jovanovski* and Dimitar Jovanov†
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, Macedoina
*jovanovski@gorjan.info, †dimita.jovanov.92@gmail.com

***Abstract*—With the rapid expansion of the Internet, people sought to create web sites in a human readable format. In term, this led to problems with machine readability of the available information on those sites. The large amounts of data make it hard for search engines, such as Google and Bing, to determine which data is relevant.**

Our solution allows webmasters to inject semantic tags and transit their web sites to the Web 3.0 era . It is web based and can be run on any server, independent of its platform. The automatic nature of our implementation doesn't require user interaction or any kind of previous knowledge. The information that is detected includes titles, images and dates, and for each of those, a special semantic tag is added.

From our findings we can infer that the automatic module has a 76% success rate and with further work can be improved, while the manual module allows for greater accuracy.

***Keywords*—automatic injection, web semantics, search engine optimization, HTML parsing**

I. Introduction

The amount of data on the Internet is growing by the second. According the Internet Live Stats [1], the total number of live websites exceeds over 1 billion and the average amount of data on one page is around 1.2 MB. But most of this data is intended for easy human consumption and presents a challenge for automated systems to read, understand and extract important information from.

The World Wide Web Consortium has also realized this problem, and is pushing for a better structured web, which they call the Semantic Web [2]. This term defines the web as an inter-connected field of data where everything is referenced across sites using links [3]. Their form of the semantic web is known as Linked Data, which is based around the usage of ontologies and interconnected semantic resources that help describe the objects which it refers to. This concept is built on top of so-called triplets that consist of a subject, predicate and object, and are in the Resource - Relationship - Resource or Resource - Relationship - Literal form. Every object must be a URI, and is recommended to be a URL so humans and machines alike will be able to follow that link, thus discovering more data on the given subject [4].

Our motivation came from the fact that the process is mostly manual and people spend a lot of time and resources optimizing their sites to be search engine friendly. During our research, we came across Microdata, a markup language used by most major search engines and aggregators to easily find valuable data from web pages [5]. It consists of 3 basic keywords: `itemscope`, `itemtype` and `itemprop`. The `itemscope` keyword is placed in the outer most element that contains all the structured data the user plans to highlight. In that same element, the `itemtype` keyword is placed with a value that is a URL leading to a Schema.com definition of the subject that is being described. From there on, any element that needs to be highlighted is decorated with the `itemprop` keyword that has a value describing that particular element according to the schema that was referenced in the parent element. A code snippet of how a simple Microdata implementation looks like is shown in Figure 1.

```
<div itemscope itemtype="http://schema.org/
    Person">
    Hello, my name is
    <span itemprop="name">John Doe</span>,
    I am a
    <span itemprop="jobTitle">undergraduate
        student</span>
</div>
```

Fig. 1: Sample Microdata code implemented in a HTML fragment

After this data is implemented into a web page, the next time a search engine crawler access the site, it will pick up and store this new data, and later use it to show more relevant information to users.

This paper is organised in 4 sections which include related work which explains other research conducted on the topic, detailed explanation of our solution to the problem, evaluation which consists of an experiment over a local data set of sites and a conclusion which summarizes the results of our research.

II. Related Work

Although linked data and ontologies are not a new concept, webmasters have only recently started to implement such information on their existing pages, allowing for search engines and applications alike to more easily access the sites data.

A. Best Buy Semantics

One clear example of this is **Best Buy**. Back in 2012, Best Buy was one of the first retailers to incorporate semantic data into their online catalog [6]. They used the GoodRelations ontology to indicate items, their categories, prices and

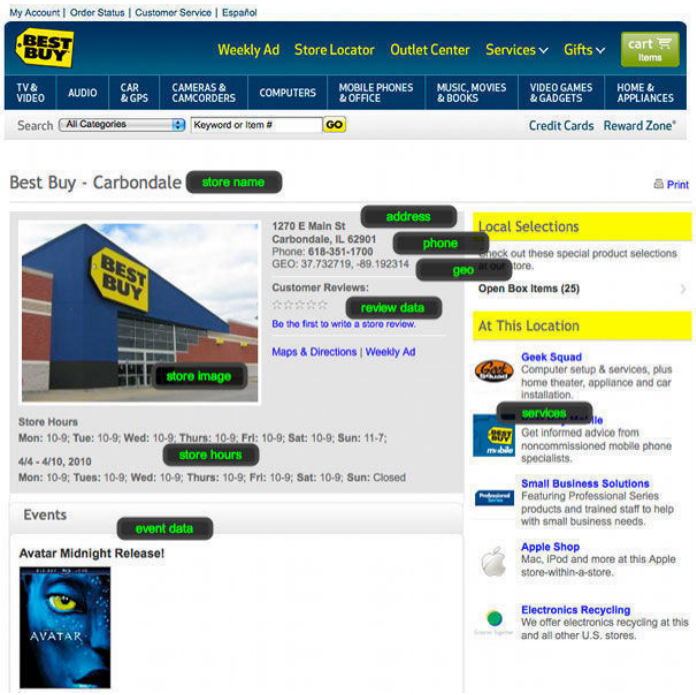


Fig. 2: Best Buy's semantics on a store info page

technical details, but also important information about their store locations. This, in term, led to Google being able to extract and show information about a store's location, work hours and contact information as shown in Figure 2. Support and implementation of semantics from large chains like Best Buy shows that the industry is starting to accept and realize the potential of semantics.

B. Google Structured Data Markup Helper

But the addition of RDF-a tags into the existing HTML has been done by hand and is not easy to distinguish where these tags need to be placed in the original markup of the site. Realizing the problem stated above, Google come up with a simple solution of how to help web developers add semantics to their sites with just pointing and clicking. The idea is that after a URL is provided to the **Structured Data Markup Helper** tool of the site in question and after selecting the data type of the page (article, event, product etc.), a render of the page will be shown to the user. From there, selections on the page can be made of where the title, date of publication, main image and other data type related attributes are located, and tags to those HTML elements will be added accordingly. In the end, the full HTML code of the site is displayed to the user, with the added tags highlighted, specifying the exact position of placement. This is a good solution for static pages where the code is the same, but if the page is dynamically generated, this may be of little to no help to the webmaster.

C. Drupal RDF-a Implementation

Another example of RDF-a implementation on the web is via the **Drupal** content management system. Every installation of Drupal comes with a predefined template that incorporates semantics in the user's posts [7]. Titles, pictures, dates and comments are all marked with RDF-a tags so that the user has no need to add them. Although this may look like a good solution, it has it's setbacks. The user is unable to modify the tags and has to add other tags manually if he wishes to specify additional data from a page. The tags that are automatically added are in a RDF-a format, and not in Microdata, which is the prefered format of search engines.

D. Semi-Automatic Semantic Annotations for Web Documents

In this research paper conducted by researchers from Italy and Canada titled **Semi-Automatic Semantic Annotations for Web Documents** we see an implementation which has a similar approach. The main concept consists of three phases which are a lightweight robust parse to get the basic structure, vocabulary and patterns of the document, externalization of the facts to a database and a use of the transformation rules inferences and structural patterns to deduct semantic markup of design facts [8].

Their concept is similar to ours but the main difference is that the scope of their annotations is larger and ad based, while our main focus is on parsing news articles and the addition of annotations to the elements.

III. Our Solution

Search Engine Optimization is a process which includes the addition of attributes to HTML elements that, in term, help add machine readable meaning to the human readable web page. The duration for its addition and the knowledge needed to properly implement it are usually long and difficult.

Our solution consists of two modules, one module that automatically adds semantics to the web pages, and a second one that needs the user's definition of certain tag ids to help utilize them more properly.

The first, and more preferred way, goes through the entire **Document Object Model** tree using a database of predefined keywords to identify a possible place for semantic injection.

- Firstly these keywords are looked up in the *id* and *name* attributes. With this method we try to use some of the descriptive properties that the web developer may have used during the making of the website.
- Secondly, we search for them in the *class* attributes of the HTML elements. This relies on the web designers naming convention and the semantics that it contains.
- Thirdly, we try to use some of the old and newly added HTML 5 tags like *article*, *h1*, *h2* and others that contain semantic properties in their original W3C definition [9].

The second module is still automatic, the only manual part is the addition of the keywords that the primary way would not be able to recognise or find.

This is achieved through a separate **API** that enables the user, with a small knowledge of his web page's internal architecture, to choose the right terms, which are then used to inject the RDF-a tags with.

A pseudo-code sketch of our algorithm is given in Algorithm 1. Our solution loads the web page's DOM tree. After

Site Name	Found Article Title	Found Article Picture	Found Article Date
Telma	Yes	Yes	Yes
24 Strumica	Yes	Yes	Yes
Utrinski	No	Yes	Yes
Fokus	No	No	Yes
Prizma	Yes	Yes	Yes
MRT	Yes	Yes	Yes
Nova Makedonija	Yes	Yes	Yes
Kanal5	No	No	Yes
Sitel	Yes	Yes	Yes
Press24	Yes	Yes	Yes
Ohrid News	No	Yes	Yes

TABLE I: Partial results from the algorithm evaluation

Data: URL of the web pages
Result: Search engine optimized HTML
initialization;
load keywords in memory;
load supported HTML 5 tags in memory;
while *not at the end of the DOM* **do**
  read current tag;
  if *id or name matches keywords* **then**
    inject RDF-a to current tag;
  else if *class matches keywords* **then**
    inject RDF-a to current tag;
  else if *tag is an HTML 5 tag* **then**
    inject RDF-a to current tag;
  else
    ignore current tag;
  end
end

Algorithm 1: Injection of semantics to the tags

the algorithm processes all of the tree's tags it returns an HTML that contains the RDF-a added as attributes to all of the tags. This injection happens realtime while the page is loaded by search engine bots, most of which allow the execution of JavaScript before crawling the page.

IV. Evaluation

In order to test out our solution on a real life scenario, we took a total of 68 web pages from a 17 Macedonian news portals and ran our algorithm on them. Some of the results can be seen in Table I.

Out of the total 17 web sites tested, the article title was the most difficult to find, failing on 4 of the test candidates. An extra two had a problem locating the article image, since that function is dependant of finding the title.

All in all, 13 sources passed the test in which our solution was able to correctly detect the article titles, main images and publication dates. It found the correct HTML tags and was able to inject the correct Microdata tags. A before and after sample can be seen in Figure 3 and Figure 4.

The test results suggest that with more work, our solution may be able to correctly detect and inject Semitics into existing pages.

```
<body>
    <h1>Zasadivme li uste eden "Pelister"?</h1>
    <img src="/mk/file/show/Images/Den␣na␣drvoto
        ␣logo.jpg">
    03. dek. 14
</body>
```

Fig. 3: Sample HTML snippet before running trough the algorithm

```
<body itemscope itemtype="http://schema.org/
    NewsArticle">
    <h1 itemprop="name">Zasadivme li uste eden "
        Pelister"?</h1>
    <img itemprop="image"src="/mk/file/show/
        Images/Den␣na␣drvoto␣logo.jpg">
    <span itemprop="datePublished">03. dek. 14</
        span>
</body>
```

Fig. 4: Sample HTML snippet after running trough the algorithm

The main problems that we ran into are that many sites are not standardized and are structured in many different ways, thus making it very hard to detect the correct HTML elements. Another problem that we ran into was the date format. Almost every site used a different date structure, ranging from `MM/DD/YYYY`, `dd MM yy` and `dd Month. YYYY`. This called for complex regular expressions that were used to detect the publishing date.

In order to test if our solution adds the correct context and semantic tags to the sample websites, we ran the results trough

type:	http://schema.org/newsarticle
property:	
name:	Zasadivme li uste eden "Pelister"?
image:	Den%20na%20drvoto%20logo.jpg
datepublished:	03. dek. 14

Fig. 5: Snippet ran trough Google's Structured Data Validator

Google's Structured Data Validator tool, where the HTML snippet is crawled by Google and any linked data is extracted and presented to the user. The results in Figure 5 show that the inserted semantic tags are correct and search engines will be able to correctly parse them.

V. Conclusion and future work

With the vast amount of web sites that are created every day, it is difficult to stand out from the million's of search results that search engines return. Our automatic approach brings a new perspective and an easier way to make web pages more understandable for search engines.

Through our experiment we can see that our solution passed on 13 out of 17 web sites. This means that the success rate is 76% which shows that the use of an automatic web semantic injection script is relatively reliable. Even if it doesn't inject the whole page with the proper semantic tags, it can help with the detection of some of them.

Furthermore the second module comes in place if the user has knowledge of the source code from the web page. If the ids and classes are correctly entered and because this is a more manual approach the success rate is pushed to around 99%.

In the future we would like to add a JavaScript API for easier implementation. This would allow for the script to change the source code during runtime on the client's side. This way the JavaScript has instant access to the DOM allowing it to easily manipulate the elements. Furthermore it doesn't require the site to be changed at all, with the addition of the script as part of the page it will change the code directly.

On another note, we would like to expand the database of keywords and add multilingual support to it, in order to make the script more globally available. This approach would require an expansion of the two already available modules and a definition of more databases for different language types.

Moreover there is room to optimize the algorithm. One of the methods would be to add partitioning support. In term, this speeds up our solution by concentrating on the important parts as the content and ignores parts like navigation menus, irrelevant parts of code like scripts and user comments. The other one would be optimizing parts of the code to lower the complexity of the code.

There is room to further our solution and make it an Open Data generator. This is practical for news articles and general knowledge sites. The use of this script would allow to convert the human readable content to an RDF or TTL document. This document would later be referenced and added as part of the Linking Open Data project. Linked Data is used to connect data to previously linked documents, by using the Web to make data more open and accessible to the world [10].

References

[1] "Total number of websites" *Internet Live Stats*. 2014. Web. 14 Dec. 2014

[2] "W3C Semantic Web Activity" Sandro Hawke *W3C*. 19 Jun. 2013. Web. 14 Dec. 2014

[3] Grigoris Antoniou, Paul Groth. *A Semantic Web Premier*. Third Edition. London, England. MTS Press. 2012. Print.

[4] David Wood, Marsha Zaidman, Luke Ruth, Michael Hausenblas. *Linked Data*. New York, USA. Manning Publications. 2014. Print.

[5] "HTML Microdata" Ian Hickson *W3C*. 29 Oct. 2013. Web. 14 Dec. 2014

[6] "Announcing: Best Buy Product Catalog via Semantic Endpoints" Eric Franzon *SemanticWeb*. 18 Dec. 2012. Web. 14 Dec. 2014

[7] "How Does RDF Work in Drupal 7?" Karen Stevenson *LullaBot*. 31 Aug. 2011. Web. 15 Dec. 2014

[8] Nadzeya Kiyavitskaya, Nicola Zeni. *Semi-Automatic Semantic Annotations for Web Documents*. Print.

[9] "HTML5 element list" Ashley Wright, *Mozilla Developer Network*. 4 Nov. 2012. Web. 15 Dec. 2014

[10] "Linked Data - Connect Distributed Data across the Web" Tom Heath, *Linked Data community*. 2013. Web. 15 Dec. 2014

Statistical Approach to Meta-optimization of Classification Models Applied on SVM Models

Kliment Mahoski, Žaneta Popeska
Faculty of Computer Science and Engineering
Skopje, Macedonia
kliment.mahoski@gmail.com, zaneta.popeska@finki.ukim.mk

***Abstract*—We present a novel approach to meta-optimization based on performance data. The performance data is obtained through grid search on a given parameter space, on a number of datasets. The grid search is applied on classification models generated by the SVM algorithm. Finally, the dataset formed by the obtained performance data can be used to train regression models that predict the performance for given parameter models for given datasets described through a number of descriptive statistics and information theory metrics. We examine this prospect for statistical meta-optimization and evaluate the results of this procedure conducted on a dataset compiled from the performance data obtained by grid search on five datasets.**

Keywords*—meta-optimization; optimization; support vector machines; supervised learning; classification; data mining; machine learning

I. Introduction

The no free lunch theorems state that there is no universally optimal model, but the performance of the model depends on the input. [1] [2] This concept applies to the task of classification — there is no classification model that is optimal for every possible problem, but its performance depends on the given problem, i.e. dataset given as input. Certain classification models, usually those produced by popular and prominent machine learning algorithms (support vector machines, neural networks, random forests, etc.), very often work well on many popular, real life problem instances, but these do not represent well the entire space of possible problem instances. Therefore, there exists an underlying relationship between the dataset and the fitness[1] of the classification model.

The aforementioned phenomenon raises the question of the possibility to predict the model performance based on the description of the dataset. To test this, we present a way of describing a dataset through certain descriptive statistics and information theory metrics, in a way that is *relevant in the context of performance of the different models for the given dataset*. The relevance of those statistics and metrics with the purpose of describing the dataset is given by the correlation between them and the model parameters in the context of the performance of the models.

In this paper, we use *meta-optimization* to estimate the model performance according to given criteria. Meta-optimization is the process of optimizing another optimization method. In some literature, it is also known as meta-evolution, super-optimization, automated parameter calibration, or hyper-heuristics — although not always with the same meaning, since the existence of different approaches to achieving the goal. In our case, the meta-optimization algorithm finds the model that is statistically most likely optimal for the given dataset. This is achieved via multiple regression, based on a training set consisting of empirical performance data for many different models for different parameters, obtained by grid search on a defined parameter space for many different datasets. Grid search is considered to be traditional and the most widely used strategy for meta-optimization. [3]

Our procedure is generalizable to all algorithms that have the same purpose. Therefore, it is applicable to all classification models, including classification models produced by different algorithms. In this paper, the procedure will be applied on homogeneous models, i.e. models generated by one same classification algorithm. The purpose of this simplification is to more clearly demonstrate the idea proposed in this paper without the need of handling technicalities that are not in our main focus and avoid the possible complications that arise with the use of heterogeneous models.

The classification algorithm we chose is the support vector machine (SVM) algorithm. The main idea behind the SVM algorithm was first proposed by Vladimir Vapnik in 1963. With the popularization of the SVM algorithm during the 90s, a few very important ideas have been incorporated into the initial framework — some of the most notable being the application of the kernel trick (proposed in [4]) to the maximum-margin hyperplanes [5], and the soft-margin classification [6]. As a widely used machine learning algorithm, it encompasses the most frequent machine learning tasks, including classification, which is the purpose of the models that are optimized in this paper.

We employ the experimental mathematics methodology to empirically examine the validity of our proposed meta-optimization strategy.

II. Related Work

The area of meta-optimization encompasses a number of different strategies, which despite sharing a common goal, may have little in common in regard to the approach to reaching their goal. Often, evolutionary and genetic algorithms are used for the task of meta-optimization. [7] [8] [9] [10]

There is a class of optimization strategies related to the meta-optimization strategies — the *metaheuristics*, which are heuristics that choose a solution from a given space of solutions, given certain criteria of optimality. They do not optimize only the hyperparameters of one model, but are able to work

[1] The notion of fitness is goal-dependent.

with many heterogeneous heuristics for solving a given task. [11]

It can be argued that the area of meta-optimization is still relatively unexplored. There is a lot of potential for exploration and real-life use in optimization tasks involving, among other things, choice of models that perform a given machine learning task.

III. Meta-Optimization Strategy

In this section we cover a novel meta-optimization stategy. Our strategy is a statistical approach to meta-optimization of SVM classification models that utilizes the description of a dataset and performance data about previously built models in order to make inferences about the performances for a new dataset given as input for models generated by the SVM algorithm for certain parameter vector.

The procedure of statistical meta-optimization can be considered a procedure independent of the type of models (the algorithm that builds those models). In this phase of our research, it is most feasible to conduct the tests on one certain type of models. Our models of choice are the SVM models.

A. Dataset Description

For dataset representation purposes, we consider a number of additional features that are potentially able to describe a given dataset in a context relevant to the performance of the classification models. These features are a number of descriptive statistics and information theory metrics that will be given shortly. We should note that in order for the following features to be meaningful, the attributes of the datasets should be standardized. The following formulations are under the assumption that the data is standardized ($\mu = 0$, $\sigma = 1$).

Let us first define a how we combine the descriptive statistics for each of the attributes of the dataset represented by the random variables that generate their values, $X_j, j = 1, ..., k$. The combining statistic is a function $F_C : \mathbb{R} \to \mathbb{R}$, such that for a statistic related to the attributes of the dataset, F, it takes the value $F_C(F_i(X_1), ..., F_i(X_k))$. A combining statistic that we take into consideration can be the mean, the minimum, the maximum, the median, or the entropy value of the values for a given descriptive statistic for each of the attributes. For example, if F_C is the mean, then for each per-attribute statistic $F_i(X_j)$, we have:

$$\overset{*}{F}_i = \frac{1}{k}\sum_{j=1}^{k} F_i(X_j) \tag{1}$$

where by '$*$' we represent that the mean is the combining statistic applied to the per-attribute descriptive statistic F_i.

Following this definition, let us look at the descriptive statistics formulated as the means of the per-attribute descriptive statistics and information theory metrics for a given dataset of size N. As potentially representative features, we choose the measures of spread, variability, distribution shape, and entropy.

- Mean of the attribute i-th quartiles, $i \in \{1, 2, 3\}$:

$$\overset{*}{Q}_i = \frac{1}{k}\sum_{j=1}^{k} Q_i(X_j) \tag{2}$$

- Mean of the attribute interquartile ranges (IQR's):

$$\overset{*}{\mathrm{IQR}} = \frac{1}{k}\sum_{j=1}^{k} \left[Q_3(X_j) - Q_1(X_j)\right] \tag{3}$$

- Mean of the attribute sample skewness coefficients (unbiased):

$$\overset{*}{\widehat{G}}_1 = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{\sqrt{N(N-1)}}{N-2}\right. \left.\frac{\frac{1}{N}\sum_{i=1}^{N}(X_j^{(i)} - \overline{X_j})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_j^{(i)} - \overline{X_j})^2\right]^{3/2}}\right] \tag{4}$$

- Mean of the attribute Pearson mode skewness coefficients:

$$\overset{*}{G}_{\mathrm{mode}} = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{\overline{X_j} - \mathrm{mode}(X_j)}{S_{X_j}}\right] \tag{5}$$

- Mean of the attribute Pearson's first skewness coefficients:

$$\overset{*}{J}_1 = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{3(\overline{X_j} - \mathrm{mode}(X_j))}{S_{X_j}}\right] \tag{6}$$

- Mean of the attribute Pearson's second skewness coefficients:

$$\overset{*}{J}_2 = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{3(\overline{X_j} - Q_2(X_j))}{S_{X_j}}\right] \tag{7}$$

- Mean of the attribute L-skewness coefficients:

$$\overset{*}{\mathcal{T}}_3 = \frac{1}{k}\sum_{j=1}^{k}\mathcal{T}_3(X_j) = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{\Lambda_3(X_j)}{\Lambda_2(X_j)}\right)$$
$$|\ \Lambda_r(X_j) = r^{-1}\binom{N}{r}^{-1}\cdot$$
$$\sum_{X_j^{(1)}<\ldots<X_j^{(i)}<\ldots<X_j^{(r)}} (-1)^{r-i}\binom{r-1}{i}X_j^{(i)} \tag{8}$$

- Mean of the attribute sample kurtoses (unbiased):

$$\overset{*}{\widehat{G}}_2 = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{N(N+1)}{(N-1)(N-2)(N-3)}\right. \left.\frac{\sum_{i=1}^{N}(X_j^{(i)} - \overline{X_j})^4}{\left(\frac{1}{N-1}\sum_{i=1}^{N}(X_j^{(i)} - \overline{X_j})^2\right)^2}\right] \tag{9}$$

- Mean of the attribute L-kurtosis coefficients:

$$\overset{*}{\mathcal{T}}_4 = \frac{1}{k}\sum_{j=1}^{k}\mathcal{T}_4(X_j) = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{\Lambda_4(X_j)}{\Lambda_2(X_j)}\right) \tag{10}$$

- Mean of the attribute Shannon entropy values — continuous (real-valued attribute) case (histogram estimator, applied on the attribute discretized in 10 bins), for each attribute except the class attribute:

$$\overset{*}{\widehat{H}} = \frac{1}{k}\sum_{j=1}^{k}\left[-\sum_{i=1}^{N} f_{X_j}(X_j^{(i)})\ln\left[\frac{f_{X_j}(X_j^{(i)})}{w(X_j^{(i)})}\right]\right] \tag{11}$$

- Mean of the attribute Shannon entropy values — discrete (nominal attribute) case, for the class attribute:

$$H_C = \frac{1}{k}\sum_{j=1}^{k}\left[-\sum_{i=1}^{N} f_{X_j}(X_j^{(i)})\ln f_{X_j}(X_j^{(i)})\right] \quad (12)$$

Now that we defined these features with the mean as a combining statistic, we can define the other features accordingly. For all of the above per-attribute statistics and information theory metrics F_i, we define the combining statistics:

- Minimum:

$$\min(F_i) = \min\{F_i(X_1), ..., F_i(X_k)\} \quad (13)$$

- Maximum:

$$\max(F_i) = \max\{F_i(X_1), ..., F_i(X_k)\} \quad (14)$$

- Entropy:

$$H_{F_i} = -\sum_{j=1}^{k} f_{F_i}(F_i(X_j))\ln\left[\frac{f_{F_i}(F_i(X_j))}{w(F_i(X_j))}\right] \quad (15)$$

We will from now on view a given dataset in terms of its description given by these combining statistics, which combine all of the per-attribute statistics, and separate for the class attribute (whenever applicable). Therefore, the dataset can now be represented as a vector of features of the following form:

$$\mathbf{D} = \big(\overset{*}{F}_1, ..., \overset{*}{F}_m, \min(F_1), ..., \min(F_m), \max(F_1), ..., \max(F_m), \widehat{H}_{F_1}, ..., \widehat{H}_{F_m}\big) \quad (16)$$

B. Performance Data Acquisition

We use classification SVMs of type 1 (C-SVMs). We define the SVM model parameter space as

$$R_{(K,C,\mathrm{Coef}_0,\Gamma,D)} = R_K \times R_C \times R_{\mathrm{Coef}_0} \times R_\Gamma \times R_D \quad (17)$$

where

- R_K is the range of the kernel function parameter:

$$R_K = \{f_{\mathrm{linear}}, f_{\mathrm{polynomial}}, f_{\mathrm{radial}}, f_{\mathrm{sigmoid}}\} \quad (18)$$

- R_C is the range of the cost (of constraint violation) parameter:

$$R_C = \{2^a \mid a = \overline{-4, 5}\} \quad (19)$$

- R_{Coef_0} is the zero-coefficient parameter (bias):

$$R_{\mathrm{Coef}_0} = \{\pm 2^a \mid a = \overline{-2, 2}\} \quad (20)$$

- R_Γ is the range of the gamma parameter:

$$R_\Gamma = \{2^a \mid a = \overline{-4, 5}\} \quad (21)$$

- R_D is the range of the degree parameter:

$$R_D = \{0, ..., 4\} \quad (22)$$

We employ a grid search-like strategy for obtaining the performance metrics for models built with a given parameter vector $(k, c, \mathrm{coef}_0, \gamma, d) \in R_{(K,C,\mathrm{Coef}_0,\Gamma,D)}$.

For the SVM algorithm implementation, we use the interface to 'libsvm' in the 'e1071' R library [12].

C. Statistical Meta-optimization Dataset Compilation

The performance metrics given in the previous subsection make up a dataset that can be used to make predictions for new, unseen datasets. In order to make predictions, we treat the database consisted of the test outputs as a regular dataset on which we can perform regression tasks — the independent attributes are the dataset description attributes and the model parameters attributes, while the dependent attributes are the performance metrics. Thus, in that manner one can make predictions, via regression, for a certain perfromance metric.

For each point in the model parameter space, which can be represented as a vector:

$$\mathbf{M_p} = \big(K, C, \mathrm{Coef}_0, \Gamma, D\big) \quad (23)$$

we obtain performance metrics, that can be represented as a vector

$$\mathbf{P} = \big(P_1, P_2, ..., P_l\big) \quad (24)$$

In the dataset of performance metrics, we would have the performances that are obtained for a given dataset described by the vector $\mathbf{d}_1$, for a certain point of the parameter space $\mathbf{p}_1$. If an element in that dataset is a vector, the part of the vector that contains the independent features is a concatenation of the dataset description vector and the model parameter vector:

$$\mathbf{D}^\frown\mathbf{M_p} \quad (25)$$

Therefore, an element of the performance dataset is a concatenation of the vector of the independent features as above, and the vector of the the performance metrics, which are be the target attributes:

$$(\mathbf{D}^\frown\mathbf{M_p})^\frown\mathbf{P} \quad (26)$$

We can then select any attribute $p_i, i = \overline{1, l}$ and consider it the target attribute of a multiple regression task with the goal of predicting its value, i.e. the performance metric of a model build for certain parameters, for a certain dataset.

IV. Experiment

This section covers an experiment which is conducted as a step toward gaining more insight about our proposed meta-optimization strategy.

A. Datasets

The datasets that we will use for the experiment are acquired from the UCI repository [13]. Table I contains basic quantative descriptions of the datasets given by their:

- number of instances — N;
- number of attributes — k; and
- number of classes — n_c.

Let us now take at look in Table II at the part of the description the datasets by the constructed features — as an example, those obtained with the mean as a combining statistic.

TABLE I. BASIC QUANTITATIVE DESCRIPTIONS

Dataset	N	k	n_c
Ecoli	335	5	8
Fertility	100	9	2
Glass	214	9	6
Ionosphere	351	34	2
Iris	150	4	3

TABLE II. DESCRIPTIVE STATISTICS

Dataset	$\overset{*}{Q}_1$	$\overset{*}{Q}_2$	$\overset{*}{Q}_3$	$\overset{*}{\mathrm{IQR}}$	$\overset{*}{\widehat{G}}_1$	$\overset{*}{G}_{\mathrm{mode}}$
Ecoli	-0.73	-0.15	0.77	1.5	0.27	0.21
Fert.	-0.64	-0.14	0.87	1.5	-0.044	0.09
Glass	-0.51	-0.1	0.31	0.82	1.7	0.43
Ion.	-0.52	0.13	0.66	1.2	-0.52	-0.31
Iris	-0.97	0.073	0.7	1.7	0.067	0.94

Dataset	$\overset{*}{J}_1$	$\overset{*}{J}_2$	$\overset{*}{\mathcal{T}}_3$	$\overset{*}{\widehat{G}}_2$	$\overset{*}{\mathcal{T}}_4$	$\overset{*}{\widehat{H}}$
Ecoli	0.63	0.44	0.075	-0.25	-2.9	2.8
Fert.	0.27	0.43	-0.0074	-0.27	-3	1.6
Glass	1.3	0.3	0.2	9.9	-2.7	1.9
Ion.	-0.93	-0.38	-0.12	0.23	-2.9	2.8
Iris	2.8	-0.22	0.011	-0.75	-3	2.9

B. Optimal Model Performance Results

We perform 100 repetitions of 10-fold CV for estimating the error rate as a performance metric. In the following Table III we have the results for the optimal models obtained by grid search in the previously defined parameter space $R_{(K,C,\mathrm{Coef}_0,\Gamma,D)}$.

TABLE III. OPTIMAL MODELS

Dataset	k	c	coef_0	γ	d	Error rate
Ecoli	sigmoid	1	-0.5	0.125	R_d	0.13576
Fertility	radial	2	R_{C_0}	0.5	R_d	0.11930
Glass	polynomial	8	4	0.5	2	0.27674
		32	2	0.25		
Ionosphere	sigmoid	8	R_{C_0}	0.0625	R_d	0.04792
Iris	polynomial	0.25	2	0.125	3	0.02413
		2	1	0.0625		

C. Utilization for Making Inferences

The compiled performance dataset consists of 10^5 elements. We will now examine the performance on the obtained performance data for a few regressors. We will look at the correlation coefficient of the learned regression model, the mean absolute error (MAE), the root mean squared error (RMSE), the relative absolute error (RAE), and the root relative squared error (RRSE). The predictions are made for the error rates of the models. The results are obtained by 10-fold CV.

It is important to note that the results are not the optimal ones. This regression procedure should be optimized as well, but that is beyond the scope of this experiment. The parameters for these regressors that are chosen are those that are most commonly used, i.e. the default ones in the Weka [14] (version 3.7.12) implementations. Additionally, other regression algorithms should be tried out, as well as different preprocessing procedures (choices of parameters, PCA, etc.).

TABLE IV. MODEL PERFORMANCE INFERENCE RESULTS

Regressor	Corr.	MAE	RMSE	RAE	RRSE
Bagging (REP tree)	0.9992	0.0038	0.0080	2.2271%	3.9606%
REP tree	0.9989	0.0039	0.0095	2.3075%	4.7045%
M5P	0.9970	0.0089	0.0159	5.2047%	7.9153%
Random forest	0.9926	0.0118	0.0260	6.9498%	12.9175%
Random tree	0.9590	0.0161	0.0575	9.4396%	28.5633%
Decision table	0.9367	0.0367	0.0705	21.5210%	35.0073%
Multilayer perceptron	0.9315	0.0474	0.0734	27.8450%	36.4624%

V. CONCLUSION

More experiments are needed in order to get a more accurate picture about the performance and feasibility of the proposed meta-optimization strategy. The grid search should be performed on a greater number of datasets in order to have a more representative performance dataset that can be used to make more accurate predictions. Additionally, instead of cross-validation as a evaluation method, one should try using an entirely new dataset with its performance metrics in order to obtain a better estimate of the accuracy of the regressors.

If the conjecture about the ability to make predictions in the presented manner is empirically shown to be valid with a reasonable level of confidence, more formal mathematical examinations will be performed for it to be proven.

REFERENCES

[1] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[2] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.

[3] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[4] M. A. Aizerman, E. A. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," in *Automation and Remote Control*, no. 25, 1964, pp. 821–837.

[5] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: ACM, 1992, pp. 144–152.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] R. Mercer and J. Sampson, "Adaptive search using a reproductive meta-plan," *Kybernetes*, vol. 7, no. 3, pp. 215–228, 1978.

[8] V. Nannen, "A method for parameter calibration and relevance estimation in evolutionary algorithms," in *Genetic and Evolutionary Computation Conference, GECCO 2006, Proceedings*. ACM, pp. 8–12.

[9] S. K. Smit and A. E. Eiben, "Comparing parameter tuning methods for evolutionary algorithms," in *Proceedings of the Eleventh Conference on Congress on Evolutionary Computation*, ser. CEC'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 399–406.

[10] P. Krus and J. Ölvander, "Performance index and meta-optimization of a direct search optimization method," *Engineering Optimization*, vol. 45, no. 10, pp. 1167–1185, 2013.

[11] M. E. H. Pedersen, "Tuning & simplifying heuristical optimization," 2010.

[12] D. Meyer and T. U. Wien, "Support vector machines — The interface to libsvm in package e1071," 2014.

[13] M. Lichman, "UCI machine learning repository," 2013.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

Multiple Gene Ontology Tools Analysis in Discovering Enriched Colorectal Cancer Signature

Monika Simjanoska and Ana Madevska Bogdanova
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Email: {monika.simjanoska, ana.madevska.bogdanova}@finki.ukim.mk

***Abstract*—In this paper we perform multiple gene ontology (GO) analyses to explore the molecular functions and the biological processes of a predefined set of colorectal cancer (CRC) biomarkers. The main objective of this research is to discover whether there exists a common CRC biomarkers signature that appears to be enriched by all the GO tools we explore. Furthermore, we investigate their function and tight association with common cancers with the aim to test their capability of indicating the direction of the possible progression of the CRC and even more, to test their potentiality to predict distant metastasis.**

***Keywords*—Bioinformatics, Colorectal Cancer, Biomarkers, Gene signature, Gene Ontology**

I. Introduction

In the past the analyses of single markers have been in the focus of the genome-wide association studies. However, this approach often lacks the power to uncover the relatively small effect sizes conferred by most genetic variants. Therefore, using prior biological knowledge on gene function, pathway-based approaches have been developed with the aim to examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest [1].

The goal of the Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing [2]. The Gene Ontology (GO) project since 1998 is a collaborative effort to provide consistent descriptors for gene products in different databases and to standardize classifications for sequences and sequence features. The GO project provides ontologies to describe attributes of gene products in three non-overlapping domains of molecular biology [3]:

1) **Molecular Function** describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities that perform the actions, and do not specify where, when or in what context the action takes place.
2) **Biological Process** describes biological goals accomplished by one or more ordered assemblies of molecular functions.
3) **Cellular Component** describes locations, at the levels of subcellular structures and macromolecular complexes.

In this paper we analyze the colorectal cancer problem whose serious prevalence and incidence information is reported in the latest research on cancer's statistics [4]. This work relies on results from our previous research [5] that referred to the detection of statistically significant gene expression obtained from CRC tissues and adjacent normal controls. The discriminative power of the group of genes discovered as significant was tested on both carcinogenic and healthy patients and proved to be very accurate for distinction between the two different classes (CRC and healthy) of patients.

Since the results were very promising for predicting CRC, we continued our research to test whether they can be an indicator for the possible direction of the CRC progression and metastasis. In order to achieve our goals we established and followed a procedure described in Section III that involved multiple GO analysis to investigate:

1) The biomarkers that are closely related to CRC;
2) The gene signature that is involved in the enriched biological and molecular functions addressed by all GO approaches we used in the analysis;
3) The relation between those biomarkers and other types of cancer.

The rest of the paper is organized as follows. In Section II we present cancer research where the authors analyze gene expression data to obtain gene profiles by which metastasis of a particular cancer can be predicted. The GO tools for the analysis are presented in Section III and the obtained results are given in Section IV. In Section V we present the conclusion from our work and our plans for further research.

II. Related Work

In this section we present some researches where gene expression profiles have been discovered to predict distant metastasis for various types of cancers.

Wang et al. [6] analyzed the expression of 22 000 transcripts from total RNA of frozen tumor samples from 286 lymph-node-negative patients who had not received adjuvant systemic treatment. In a training set of 115 tumors, we identified a 76-gene signature consisting of 60 genes for patients positive for oestrogen receptors (ER) and 16 genes for ER-negative patients. This signature showed 93% sensitivity and 48% specificity in a subsequent independent testing set of 171 lymph-node-negative patients. The gene profile was highly informative in identifying patients who developed distant metastases within 5 years and also represented a strong prognostic factor for the development of metastasis in the subgroups of premenopausal and postmenopausal patients.

Budhu et al. [7] show that a unique inflammation response-related signature is associated with noncancerous hepatic tissues from metastatic Hepatocellular carcinoma (HCC) patients. A refined 17 gene signature was validated as a superior predictor of HCC venous metastases in an independent cohort, when compared to other clinical prognostic parameters. They suggest that a predominant humoral cytokine profile occurs in the metastatic liver milieu and that a shift toward anti-inflammatory responses may promote HCC metastases.

Finak et al. [8] state that although it is increasingly evident that cancer is influenced by signals emanating from tumor stroma, little is known regarding how changes in stromal gene expression affect epithelial tumor progression. They used laser capture microdissection to compare gene expression profiles of tumor stroma from 53 primary breast tumors and derived signatures strongly associated with clinical outcome. They cross-referenced genes differentially expressed in each stroma outcome subtype against gene ontology annotations to identify overrepresented GO categories.

O'Donnell et al. [9] assume that metastasis via the lymphatics is a major risk factor in squamous cell carcinoma of the oral cavity (OSCC). They sought to determine whether the presence of metastasis in the regional lymph node could be predicted by a gene expression signature of the primary tumor. A total of 18 OSCCs were characterized for gene expression by hybridizing RNA to Affymetrix U133A gene chips. Metastatic primary tumors could be differentiated from nonmetastatic primary tumors by a signature gene set of 116 genes. They showed that lymph node metastasis could be predicted by gene expression profiles of primary oral cavity squamous cell carcinomas.

III. The Methods

In this section we describe the methods and tools we use for GO analysis of a CRC biomarkers set that contains 138 genes that showed significant gene expression in presence of CRC [5].

Given the set of biomarkers for particular study, we expect that for the biological processes that are abnormal, the enrichment analysis should increase the potential of some genes to be selected as a relevant group. Since the analytic conclusion is based on a group of biomarkers instead of on an individual gene, it increases the possibility the correct biological processes of interest to be identified. The enrichment can be quantitatively measured by some common and well-known statistical methods as Chi-square, Fisher's exact test, Binomial probability and Hypergeometric distribution. The GO tools map the input gene list to the associated biological annotation terms and then statistically examine the enrichment of gene members for each of the annotation terms by comparing the outcome to the reference background [10].

Huang et al. [10] in their research collected 68 bioinformatics tools and no matter the differences they described the general procedures of all approaches as having three major layers depicted in Figure 1:

- Layer 1: Data support (the annotation database);
- Layer 2: Data mining (the algorithm and statistics);
- Layer 3: Results presentation (interface).

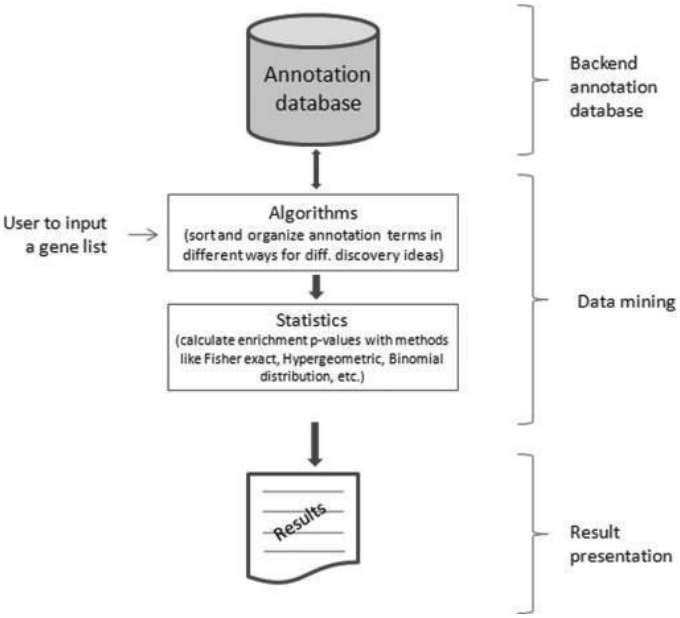


Fig. 1. Three layer model of the GO tools [10]

Even though there are many different bioinformatics tools for GO analysis; however, many of them require local installation and specific platform. In this research we use the following freely accessible online tools.

A. GOEAST

Gene Ontology Enrichment Analysis Software Toolkit is a web based tool which applies appropriate statistical methods to identify significantly enriched GO terms among a given list of genes. Beside the other functions, GOEAST supports analysis of probe set IDs from Affymetrix microarrays - the technology used to observe the gene expression of our biomarkers. It provides graphical outputs of enriched GO terms to demonstrate their relationships in the three ontology categories: molecular function, biological process and cellular component [11]. GOEAST belongs to the class of singular enrichment analysis tools. It takes a predefined list of biomarker genes, and then iteratively test the enrichment of each annotation term one-by-one in a linear mode. The enriched annotation terms passing the enrichment P-value threshold are ordered by the enrichment probability. The enrichment P-value is the number of genes in the list that match a given biology process as compared to pure random chance [10]. To calculate the P-value, GOEAST uses the Hypergeometric test. To perform an analysis using the Hypergeometric tests, one needs to define a 'gene universe' which usually is the whole genome. In our case, for microarray data, the 'gene universe' can be the unique gene identifiers assayed in the experiment. The next step is to identify the subset of the universe that is considered interesting, i.e., the biomarkers [12]. The tool 'knows' how many of the genes of the 'gene universe' are associated with the particular biological processes and how many of those genes are in our biomarkers list.

Let N be the size of the 'gene universe', n be the number of biomarkers, K be the number of genes associated with particular biological process and k be the number of genes

associated with the same biological process in the biomarkers set. The enrichment P-value is calculated as presented in (1):

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \quad (1)$$

B. WebGestalt

Zhang et al. [13] have developed a web-based integrated data mining system, WebGestalt, to help biologists in exploring large sets of genes. WebGestalt is composed of four modules: gene set management, information retrieval, organization/visualization, and statistics. The management module uploads, saves, retrieves and deletes gene sets. The information retrieval module retrieves information for up to 20 attributes for all genes in a gene set. The organization/visualization module organizes and visualizes gene sets in various biological contexts, including GO, tissue expression pattern, chromosome distribution, metabolic and signaling pathways, protein domain information and publications. The statistics module recommends and performs statistical tests to suggest biological areas that are important to a gene set and warrant further investigation. WebGestalt also belongs to singular enrichment analysis class of tools and uses Hypergeometric test to calculate the enrichment P-value (1).

C. DAVID

The DAVID Gene Functional Classification Tool uses a novel assembling algorithm to condense a list of genes or associated biological terms into organized classes of related genes called biological modules. This organization is accomplished by mining the complex biological co-occurrences found in multiple sources of functional annotation. It is a powerful method to group functionally related genes and terms into a manageable number of biological modules for efficient interpretation of gene lists in a network context [14].

DAVID is both singular and modular enrichment analysis tool. It inherits the basic enrichment calculation found in singular enrichment analysis and incorporates extra network discovery algorithms by considering the term-to-term relationships. The benefit of the modular enrichment analysis approach is that it considers the term-term relationships in which joint terms may contain unique biological meaning for a given problem, not recognizable by individual terms [10]. To determine the enrichment probability DAVID uses EASE score which is a modified version of the one-tailed Fisher's exact test. This test is adopted to measure the gene-enrichment in annotation terms. The one-tailed Fisher's exact test is identical to the Hypergeometric test (1). However, the EASE score is a rigorous adjustment to the Fisher's probability. It is calculated by penalizing, i.e., removing one gene from the list within a given biological process and calculating the resulting Fisher's exact probability for that process [15].

The tools we use for our analysis all belong to the same class of tools and use nearly the same statistical approaches to identify enriched biological processes and genes. However, it is not unusual researchers to use similar tools in order to obtain more reliable results [16].

The results of our experiments are reported in the next Section IV.

IV. Experiments and Results

In this section we present the results for the genes in the enriched biological processes and molecular functions reported by three different GO tools.

The total number of biomarkers used as input for all the GO tools is 138 and the total number of Affymetrix probes is 50,000. Therefore, the 'gene universe' is represented by the pool of Affymetrix probes and the subset of significant genes are the 138 biomarkers.

All the tools express the ontology results in both graphical and textual manner for all the three domains previously discussed in Section I. Once we compared the biological and molecular functions that appeared commonly enriched by all tools, we identified the genes that were involved in those processes. Considering all the genes in the common enriched processes, we used GeneCard [17] to investigate their role in the human organism and also their relation to the CRC. As a result we obtained lists of genes associated with CRC, presented in Table I.

In order to obtain the common genes that are included in the significant processes discovered by all the tools, we made an intersection and shortened the list to 11 genes out of 138 for which it is confirmed to be related to CRC. In our further research, this list of 11 genes is referred to as gene signature.

Once we found the gene signature, we inspected the other types of cancer that also identify with those genes. The information presented in Table II is retrieved from GeneCards where we investigated the substantial relation to cancer for each biomarker of the gene signature distinctively.

V. Conclusion and Future Work

This research relies on 138 significant genes whose expression showed to be disrupted in presence of CRC. The genes were used to reveal even more remarkable subset of important genes with the potential to predict distant metastasis. To achieve this goal, we analyzed and used bioinformatics tools that are specialized for gene ontology analysis. All the

TABLE I. Genes associated with CRC from enriched processes

GOEAST	WEBGESTALT	DAVID
EPHA7	EPHA7	EPHA7
MMP7	MMP7	MMP7
AQP8	AQP8	AQP8
GUCA2B	GUCA2B	GUCA2B
UNC5C	UNC5C	UNC5C
KLK10	KLK10	KLK10
MFAP5	MFAP5	MFAP5
C2orf40	C2orf40	C2orf40
EPHA3	EPHA3	EPHA3
SLC26A3	SLC26A3	SLC26A3
DPEP1	DPEP1	DPEP1
MEIS1	MEIS1	/
FABP6	/	FABP6
CA7	/	CA7
NEUROG1	/	/
LONRF2	/	/
CNKSR2	/	/
/	CLDN8	CLDN8
/	SCGN	SCGN

TABLE II. GENE SIGNATURE ASSOCIATION WITH CANCER

Gene	Type of cancer
EPHA7	Hepatocellular cancer Breast cancer Colorectal cancer Leukemia
UNC5C	Colorectal cancer Kidney cancer Hepatocellular cancer
DPEP1	Colorectal cancer Breast cancer
MMP7	Colorectal cancer
KLK10	Ovary cancer Breast cancer Prostate cancer Stomach cancer Testicle cancer Colorectal cancer Tongue cancer Lung cancer
SLC26A3	Colorectal cancer
AQP8	Hepatocellular cancer Breast cancer Colorectal cancer
MFAP5	Colorectal cancer Breast cancer Ovary cancer
EPHA3	Breast cancer Colorectal cancer Prostate cancer
GUCA2B	Colorectal cancer Pancreatic cancer
C2orf40	Oesophageal cancer Colorectal cancer Breast cancer Leukemia

tools used in the research use similar statistical procedure and present the results in a similar manner and thus we could easily compare the outcomes and derive the most probable conclusion.

Assuming the gene ontology analysis has accented the enriched abnormal biological processes and molecular functions, we collected the genes that are directly involved in those processes. For each gene we inspected its association with the CRC and the genes that were reported as such were distinguished as distinct subset of special biomarkers that promises further advances in the field of research.

Since our aim is to find gene signature according to which we will be able to predict the CRC progress and distant metastasis in the future, we inspected the relation of the 11 special biomarkers with the rest of the most common carcinogenic disorders. Beside the colon and rectal cancers, the special genes were also related to breast cancer, hepatocellular cancer, leukemia, prostate cancer, etc.

In our future work, we will examine gene expression data from CRC and from the other types of cancer discussed above, that evolved from CRC as primary cancer. Particularly, we will observe the expression of the 11 special biomarkers and will test their power of predicting the CRC progression.

ACKNOWLEDGMENT

Special thanks to Vera Kurtevska, Angela Spirkoska, Filip Bogdanoski and Kiril Cvetkov who worked on this problem during the "Introduction to Bioinformatics" course taught from February to June 2014.

REFERENCES

[1] K. Wang, M. Li, and H. Hakonarson, "Analysing biological pathways in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 12, pp. 843–854, 2010.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[3] G. O. Consortium *et al.*, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D258–D261, 2004.

[4] WCRFI, "Colorectal cancer statistics," 2012. [Online]. Available: http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/colorectal-cancer-statistics

[5] M. Simjanoska, A. Madevska Bogdanova, and Z. Popeska, "Bayesian posterior probability classification of colorectal cancer probed with affymetrix microarray technology," in *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on*. IEEE, 2013, pp. 959–964.

[6] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[7] A. Budhu, M. Forgues, Q.-H. Ye, H.-L. Jia, P. He, K. A. Zanetti, U. S. Kammula, Y. Chen, L.-X. Qin, Z.-Y. Tang *et al.*, "Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment," *Cancer cell*, vol. 10, no. 2, pp. 99–111, 2006.

[8] G. Finak, N. Bertos, F. Pepin, S. Sadekova, M. Souleimanova, H. Zhao, H. Chen, G. Omeroglu, S. Meterissian, A. Omeroglu *et al.*, "Stromal gene expression predicts clinical outcome in breast cancer," *Nature medicine*, vol. 14, no. 5, pp. 518–527, 2008.

[9] R. K. O'Donnell, M. Kupferman, S. J. Wei, S. Singhal, R. Weber, B. O'Malley, Y. Cheng, M. Putt, M. Feldman, B. Ziober *et al.*, "Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity," *Oncogene*, vol. 24, no. 7, pp. 1244–1251, 2005.

[10] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[11] Q. Zheng and X.-J. Wang, "Goeast: a web-based software toolkit for gene ontology enrichment analysis," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W358–W363, 2008.

[12] S. Falcon and R. Gentleman, "Using gostats to test gene lists for go term association," *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.

[13] B. Zhang, S. Kirov, and J. Snoddy, "Webgestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W741–W748, 2005.

[14] G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists," *Genome Biol*, vol. 8, no. 9, p. 183, 2007.

[15] S. A. Krawetz and T. Werner, *Bioinformatics for systems biology*. Springer, 2009.

[16] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nature Reviews Genetics*, vol. 9, no. 7, pp. 509–515, 2008.

[17] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. I. Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug *et al.*, "Genecards version 3: the human gene integrator," *Database*, vol. 2010, p. baq020, 2010.

Plant images classification based on the angles between the leaf shape-contour points

Petre Lameski, Eftim Zdravevski, Andrea Kulakov, Dejan Gjorgjevikj
University of Sts. Cyril and Methodius in Skopje
Faculty of Computer Science and Engineering
Skopje, Macedonia
{petre.lameski, eftim.zdravevski, andrea.kulakov, dejan.gjorgjevikj}@finki.ukim.mk

***Abstract*—In this paper we compare two shape-based descriptors for plant leaf image classification. The leaves in the dataset are already segmented from the background only the contour detection algorithm is applied to extract the contour points and generate the shape-based descriptors. We propose a reduced size descriptor based on the angles between three points of the leaf contour and compare this descriptor with other similar descriptors based on their classification quality. The classification quality is measured both with 1-nearest neighbor comparison and with RBF-SVM model trained on the generated descriptors.**

Keywords— image processing; leaf image classification; machine learning

I. INTRODUCTION

Plant identification based on plant images receives an increased attention from the image processing and machine learning community in recent years. Especially interesting is the plant identification based on images of their leaves. One of the problems that need to be asessed towards creating an efficient system for plant identification is the plant classification. There are several datasets available that allow the benchmark of algorithms for plant identification based on their leaves. Also several approaches can be identified that are mainly used for leaf classification. These approaches can be based on the leaves textures; the leaves shape exterior contours or a combination of both of them. The main drawback of the approaches for plant leaf classification based on the leaf shape is that they depend on the successful prior segmentation of the leaf from the image. Texture based descriptor don't have this drawback. Texture based descriptors, however, requre higher resolution images of the leaves for a successful extraction of the texture. Higher resolution images are not always available and their processing requires more computational resources. This is why in this paper we focus our attention towards the classification of plant leaf images based on the contour of their leaves. One of the first approaches for contour based matching is described in [1] and is also widely used in combination with other shape features. In [2] the shape context descriptor [3] is used in combination with the SIFT descriptor [4] to classify leaf images. The approach in [6] uses the shape of the leaves to extract points and generate descriptor based on the triangles that the points on the leaves contour make. Time series analysis can also be used if the points of the margins of the leaf are considered as data series [7]. In [5] the authors suggest a probabilistic combination of both texture based and shape based approaches. Many other approaches exist in the literatures that are used to classify leaf images. The available contour based descriptors are mainly used with k nearest neighbors (k-NN) classifiers and with SVM classifiers.

In this paper we present a simplified version of the descriptor described in [6] that gives comparable results with the compared descriptors. First we describe the descriptors that are extracted from the images. Then we describe the algorithms used to build the classification model and the parameters that are used for those algorithms, then we give the obtained results and the conclusion and finally we refer to future improvements and additions to the work described in this paper.

II. ANGLE BASED DESCRIPTOR

Our descriptor is based on the descriptors described in [6]. The authors prpopose several descriptors based on the triangular area that are created by the geometric properties of triplets of chosen points on the exterior contour of the leaf image. They use four types of descriptors based on the relationship between the three points. They use Triangle Area Representation or TOA descriptor that uses only the area that the triangle between three points of the contour take. The next descriptor is the Triangle Size Length representation or TSL descriptor that uses normalized side lenghts between the central and the two neighboring points. Then they use the Triangle represented by two side lengths and an angle or TSLA descriptor that uses the size of the sides and the angle between the points and finally the Triangle represented by two oriented angles or TSA descriptor represented by two consecutive angles between the points. All of the proposed descriptors are scale invariant since the neighboring points are taken with different distances. First the leaf shape contour is represented by N consecutive points that are evenly distributed on the contour. Next, for each point of the leaf contour we select two neighbor points that are on each side of the selected point. To achieve scale invariance multiple neighboring points are selected for each point. Each neighbor point is distanced equaly from on both sides of the chosen point by D points. We

select *Ns* such neighbors on each side that represent the angle between the neighbor points on *Ns* scales. The distance between the points on each scale increases by *D* so that on the normal scale we have distance of 1 between the points and on each next scale we increase the distance by *D*. We are using the linear increment for the distance between points on each scale. The other option is to use logaritmic increment where the distance between points on each scale increases logaritmicly but this option was not taken into consideration in our work.

In the proposed descriptor in this paper we neglect the information for the distance between the points of the TSLA or the consecutive angle of the TSA descriptor. We only calculate the angle that each selected point make with its 2 neighbors. The obtained angle between each point is illustrated on Fig 1.

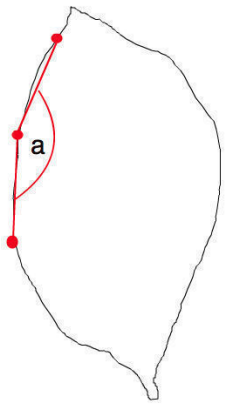


Fig. 1. Angle 'a' between the selected point and its connection with its neighbors

The main motivation for ignoring the distance is to check the significance of the distance between the points of the contour for the quality of the leaf classification.

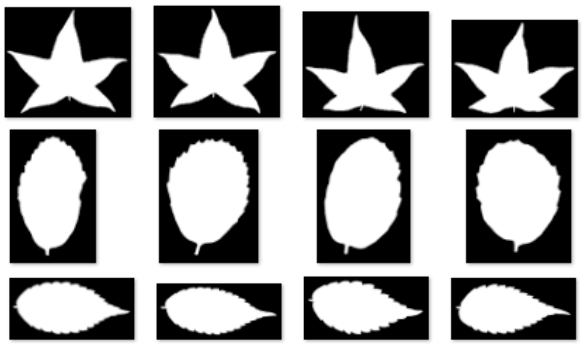


Fig. 2. Example shape images from the 100 plant dataset used in [5] from three different classes of leaves

What we are trying to verify is the significance of the angle relationship between the points of the shape contour in the shape classification process. Based on the observation of the dataset used for the comparison of the descriptors, which can be seen in Fig. 2, the leaf images have the same or close enough orientation. Which means that the descriptors and the obtained results do not account for the rotation of the leaves.

III. Experimental setup

To test the classification accuracy of our descriptor we are using the 100 leaves dataset [5]. The dataset is divided on 11 train and 5 test images from each of the 100 classes. For each of the 16 images for each plant we first generate the TSLA descriptor and our descriptor. In this work are only investigating the significance of the angle between the points on each scale and wether we can ignore the distances in the TSLA descriptor to obtain comparable results with the original descriptor. The authors in [5] use leave-one-out approach or 16-fold cross validation to evaluate the classification performance while in this paper we use 2/3 of the dataset for training and 1/3 for testing.

We are using our own implementation of the TSLA descriptor. The method used to train the classification model also defers from the reported method in [6] where the authors use Locality Sensitive Hashing (LSH) [11] to compare each of the *N* generated descriptors for each image with each descriptor of every other image and use an evaluation function to calculate the similarity between the leaf shapes. Locality sensitive hashing is used to speed up the comparison between the descriptors, because comparing the *N* descriptors from a single image with each of the *N* descriptor of every other image is a time consuming task due to the large number of floating-point operations needed to calculate each distance.

In our approach we generate a bag of visual words (BOW) descriptor for each of the images. The BOW uses a visual word dictionary that is generated by k-means clustering of all descriptors in the train set. Each of the centroids of the k-means clustering represents a single word in the dictionary. Based on this, we can calculate which descriptor represents which word by calculating the distance between the descriptors and each of the k centroids. In this experiment we use Eucledian distance to measure the distance between the centroids and the descriptors. The Eucledian distance is used also for the k-means clustering of the training data when the centroids are generated. Other distances can also be used for this purpose but their evaluation is out of the scope of this paper. In our experiment k is set to 6000, so the visual dictionary will contain 6000 words. This number was chosen empirically based on manual evaluation of the training set. We chose 500, 1000, 3000, 6000 and 10.000 length dictionary. The best initial results on the validation on the train set by using three-fold cross-validation with 1-NN classification were obtained for 6000 and this number was used for the generation of the full model for the testing phase. The performance of the BOW based classification, as expected, increased with the increase in the number of visual words in the dictionary with a very slight degradation of performance for k=10.000 when compared to k=6000.

After this step, for each image, a term frequency – inverse document frequency (tf-idf) approach similar to [8], is used to

generate the descriptor. The term frequency represents the number of times a single visual word is found inside the image and the inverse document frequency represents the number of images the same term is found in. If the visual word is found in large number of images, that visual word brings less information about the general characteristics of the images and has small contribution in uniqueily classifying the image. So for each visual word we calculate the number of appearences in each image and we normalize this number with the number of images in which the word appears, giving more significance to visual words that appear less frequently in images. These visual words give the biggest contribution in uniquely describing the characteristic of the images and the class they represent.

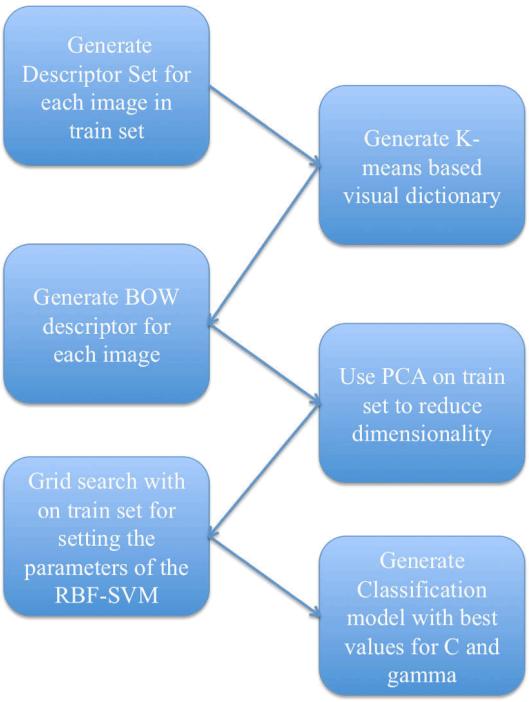


Fig. 3. Data flow for the RBF-SVM based classification model generation process

To reduce the dimensionality of the descriptor Principal Components Analysis (PCA) [9] is used. The PCA is a statistical method that reduces the dimensionality of the descriptors by analyzing and grouping corelated attributes of the feature vector.

The generated descritors are then used in two ways. First, we use the descriptors of each image in the training set as a model for 1-NN classification of the test images. Second, the model for each plant is generated with the set of the descriptors with training an Radial Based Function kernel Support Vector Machine (RBF-SVM). For the TSLA descriptor a grid search is performed for the RBF-SVM and the best results were obtained for C=1.000.000 and gamma=0.00001. For our descriptor the best results are obtained for C=100.000 and gamma=0.0001. The grid search is done for C=10^{-2} to 10^{6} and gamma=10^{-5} to 10^{2}. Both searches are done with exponential increasing steps for the parameters as recommended in [13]. WEKA [10] is used for the PCA and the RBF-SVM grid search and for the k-means clustering we are using the OpenCV [12] implementation of the algorithm. After the descriptors are generated we use the build RBF-SVM model to classify the images from the test set. We use WEKA as a data mining tool that contains many different preprocessing and analsys tools which are suitable for fast model generation and evaluation of the obtained results.

IV. Results and Conclusion

We evaluated the classification performance based on the two classification models. The obtained results are shown in Fig. 4 and Fig. 5. We used 1-NN approach, for the first evaluation, by comparing each image from the test set with each image of the training set and assign the class of the leaf image of the test set based on the class of its nearest neighbor in the training set. The average precisions obtained are 40.4% for TSLA descriptor and 40.2% for our descriptor. The weighted average Area under Curve of Receiver operating characteristic (AUC of ROC) [15][16] for the TSLA descriptor is 0.699 and for our descriptor 0.698. For the second evaluation we used the approach ilustrated in Fig. 3 for generating the model with an RBF-SVM. The average precision obtained is 54.2% for TSLA with SVM and 50.4% for our descriptor and the weighted average AUC of ROC is 0.769 for the TSLA descriptor and 0.751 for our descriptor. Altough the AUC of ROC is used for binary classification, the WEKA implementation uses an approach suggested by [16] which uses the Mann-Whitney statistic for calculating weighted average AUC of ROC for evaluation of multiclass classification problems according to the WEKA online documentation.

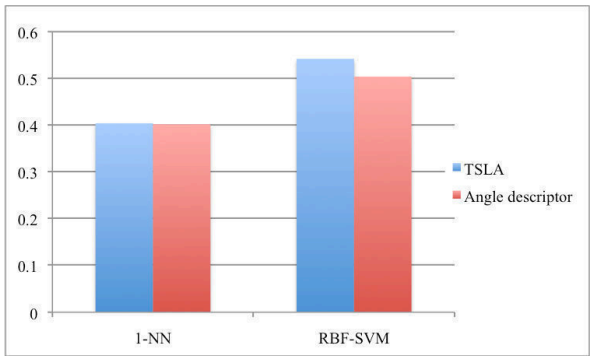


Fig. 4. Classification precision for TSLA and Angle descriptor for both 1-NN and RBF-SVM

As it can be seen from these results we sacrifice little performance of the descriptor but obtain three times smaller shape based descriptor that can be used for leaf image classification and used for the task of plant identification. Also from the results we can conclude that there is a significant improvement when using the RBF-SVM model based classification in comparison to the 1-NN classification when using the obtained descriptors on the given dataset which was expected.

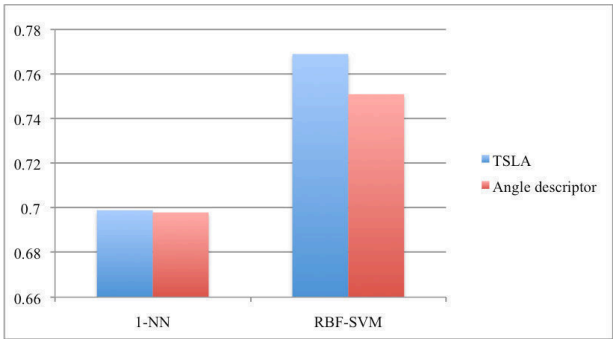


Fig. 5. Region under ROC for TSLA and Angle descriptor for both 1-NN and RBF-SVM

V. FUTURE WORK

We plan to verify the classification quality of the proposed descriptor by using other available leaf image datasets and compare the results between the two approaches. One of the challenges would be to see if the reduction of dimensionality of the initial descriptor, by ignoring the distance, would repeat the results even on datasets where the rotation of the leaf images is present. Further, additional improvements might be expected by modifying the way we obtain the BOW descriptor for the images with using other techniqes for generating the visual word centroids. We also plan to further verify the results by comparing the performance with other comparison techniques that would exclude the dictionary approach but will compare the descriptors as described using LSH. We expect that including other shape information that can be combined with our descriptor would increase the classification performance.

[1] M.-K. Hu, “Visual pattern recognition by moment invariants,” IEEE Trans. Inf. Theory, vol. 8, no. 2, pp. 179–187, Feb. 1962

[2] Zhiyong Wang; Bin Lu; Zheru Chi; Dagan Feng, "Leaf Image Classification with Shape Context and SIFT Descriptors," Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on , vol., no., pp.650,654, 6-8 Dec. 2011

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 509-522, Apr 2002.

[4] Lowe, D. G. "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[5] Mallah, C., Cope, J. and Orwell, J. “Plant Leaf Classification Using Probabilistic Integration of Shape, Texture and Margin Features” at SPPRA 2013

[6] Mouine, S., Yahiaoui, I., Verroust-Blondet, A.: A shape-based approach for leaf classification using multi scale triangular representation. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval. pp. 127– 134. ICMR ’13 (2013)

[7] Cope, James S., and Paolo Remagnino. "Classifying plant leaves from their margins using dynamic time warping" Advanced Concepts for Intelligent Vision Systems. Springer Berlin Heidelberg, 2012.

[8] Sivic, J., Zisserman, A. “Video google: a text retrieval approach to object matching in videos”, In: IEEE Conference on Computer Vision. pp. 1470–1477, 2003

[9] Jolliffe, I.T., “Principal Component Analysis, Series: Springer Series in Statistics”, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4

[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.. “The WEKA Data Mining Software: An Update” SIGKDD Explorations, Volume 11, Issue 1, 2009

[11] Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. , “Locality-sensitive hashing scheme based on p-stable distributions” in Proc. 20th Annu. Symp. Comput. Geometry, 2004, pp. 253–262.

[12] OpenCV – Open Computer Vision library, http://opencv.org (last accessed 15.03.2015)

[13] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[14] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

[15] Fawcett, T., “An introduction to ROC analysis”, Pattern Recognition Letters, Volume 27, Issue 8, June 2006, Pages 861-874, ISSN 0167-8655, http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[16] Hand, D.J., Till, R.J., A simple generalization of the area under the ROC curve to multiple class classification problems. Mach. Learning 45 (2), 171–186. 2001

Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education

Vlatko Nikolovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
vlatko.nikolovski@finki.ukim.mk

Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
riste.stojanov@finki.ukim.mk

Igor Mishkovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
igor.mishkovski@finki.ukim.mk

Ivan Chorbev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
ivan.chorbev@finki.ukim.mk

Gjorgji Madjarov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
gjorgji.madjarov@finki.ukim.mk

***Abstract*— reducing the student dropout rate in higher education is one of the challenges that universities are dealing with. By providing enriched study programs and qualified teams of professors, universities aim to enroll more students. Improving working conditions at the laboratories and other university resources aims to attract both ambitious students as well as high quality staff. After enrollment, the main goal of the faculty is to guide students into successful completion of their studies with the appropriate knowledge and skills acquired. Nowadays however, the development and deployment of Student Information Systems at the universities provides an appropriate infrastructure for student's data organization and storage as well as data acquisition and deeper analyses. This data can help model the behavior of dropouts, and predict future dropouts, therefore giving chance to counselors to advise and guide students into success.**

This paper presents various data mining experiments and results obtained from data for the students from one of the faculties at the University Ss. Cyril and Methodius in Skopje. Initially, we give an overview of several data mining algorithms suitable for analysis of students' data and dropout prediction. Furthermore, we explain modifications and applications of the algorithms over the existing student data. Finally, we provide a predictive model which will identify a subset of students who tend to drop out of the studies after the first year. The classification task aims to identify a pattern among students who tend to drop out.

Keywords*— *Educational Data Mining, Student Dropout Prediction, Machine Learning Algorithms, Classification

I. INTRODUCTION

The use of Data Mining techniques in Educational data sets, known as Educational Data Mining [1] is a relatively new field of research. Educational Data Mining provides tasks for clustering, prediction, relationship mining (subset of sequential mining, association and correlation), social related mining for human behavior and discovery with pattern models [2]. In addition, the methodology of Educational Data Mining is not yet clearly defined and there are no clear standards about which data mining algorithms are preferable in this context. Clustering and classification techniques of data sets for building a predictive model are used in [3]. Another variation of improved data classification with cost-sensitive learning is presented in [4]. Popular association analysis is presented in [5], while neural networks and Bayesian networks analyses are used in [6]. Classification and data retention techniques similar to ours are presented in [7, 8, and 9] to predict the dropout rate of the students in their first year of study.

In our case study, we used data collected over the period of three years, starting from 2011 to 2014, containing details about the students, their course retention and grade evaluation. Divided into three subcategories, based on the study year of enrolment in the faculty, each subset contains about 680 students. The classification techniques were applied over the three datasets, based on different attributes, such as: nationality of the students, sex, city of living, high school grades, study program enrolled, number of earned credits in the first year of study, an average grade in the first year of study. In addition,

we divided the first year courses into two categories, a subset of mathematics courses and a subset of programming courses. So, the attributes pool for data classification was enriched with details for average grade and number of exams applied for each subset of courses.

II. Background and Related Work

Data mining techniques are becoming an essential way of transforming data into linked usable information, extracting unexpected knowledge and discovering numerous patterns among large data sets. Due to the omnipresent implementation of various Information Systems by the faculties for logging and processing students` data, Data mining techniques can be applied to estimate unanticipated relationships among attributes of students, correlation between learning strategies and assessments. Aside from the traditional statistical methods for extracting and processing most valuable information from the large datasets, Data mining techniques provide a huge potential for knowledge discovery since they embrace numerous disciplines such as machine learning and artificial intelligence into an advanced technique for estimating large datasets. Furthermore, these combined techniques produce predictive analysis for identifying interconnections and variables regarding the context of the study [10].

Back at the beginning of this research area, Tinto [12] proposed a model of a theoretical framework for considering factors in academic success. Tinto made a correlation between the students and the faculties, considering the process of student enrolment as a sociological interplay between the characteristics of the student and the experience at the faculty. Furthermore, this interaction between the students` past and present environment leads to a degree of integration of the student into the faculty environment. Based on this model, the integrity of the institution directly depends on the quality of teachers and studies, providing an environment for the new students.

Since the methodology of educational Data mining is not yet transparent, researchers have used various techniques for estimating preferable algorithms in this context. Clustering the datasets in a manner of classification and transformation techniques to provide a considerable predictor is presented in [3]. As presented in [6], using neural networks and Bayesian networks over small datasets are outperformed by decision tree algorithms.

Beside the traditional approach, Diego [13] proposed a meta-algorithm for pre-processing the data before classification, which improves the accuracy of the model. Different techniques are presented in [11] for reducing inaccuracies in prediction of the students` dropout. In addition, the comparison of the three techniques that were used, namely neural networks, support vector machines and probabilistic fuzzy ARTMAP maintained that the most successful technique in predicting students` dropout is the decision scheme.

III. Methodology

This paper presents the usage of classification data mining techniques over the students` data to analyze and extract the important attributes affecting the dropout of students in higher education. In addition, two classifier algorithms were used, J48 [15] and Native Bayes [15] implemented in Weka (Data Mining Software) [14].

The Naïve Bayes classifier is based on the Bayes rule of conditional probability. It analyzes all the contained attributes individually as though they are equally important and independent of each other. In the process of classification, each attribute works independently from the other attributes contained in the model.

Besides the Naïve Bayes independent treatment of the attributes, J48 is a predictive machine-learning model that predicts the attribute as a dependent variable from the values of all other attributes. In order to classify a new item, J48 first creates a decision tree based on the attributes of the training data in order to gain balance, flexibility and accuracy.

IV. Data Mining process

Figure 1 shows the process of acquisition of the students` data, the process of transformation and evaluation of the attributes extracted from the data. It also shows the estimation and evaluation of the classified data and the improvement of the model for prediction and decrease of the dropout rate.

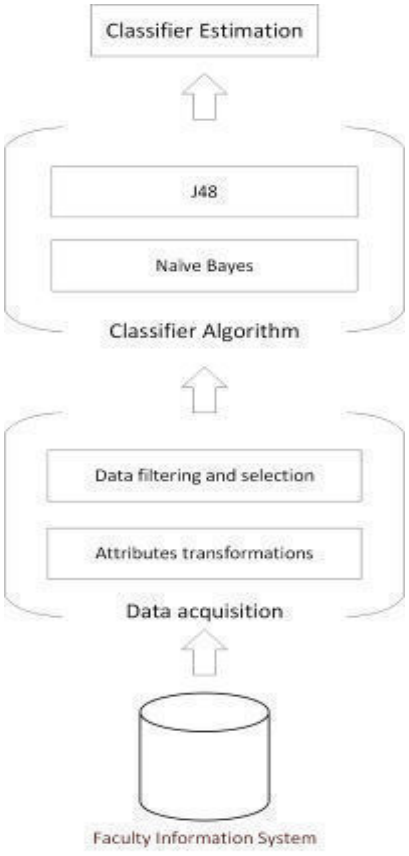


Fig. 1. Data retention process

A. Data acquisition

The dataset used in this case study was collected from the Students` Information System at one of the faculties at the Ss. Cyril and Methodius University – Skopje. In addition, the collected data over the period from 2011 to 2014 contains information about all students being enrolled. In addition, a target dataset of 2029 students was selected from three different generations, with attribute values only for their first year of study at the faculty.

B. Data preparation and attribute selection

In this step, three main datasets were considered shown in table 1: a dataset with students who were enrolled at 2011, containing 665 instances; a dataset with students who were enrolled at 2012 with 679 instances and a dataset with students enrolled at 2013 with 685 instances. Each dataset contains same attributes described in table 2.

TABLE I. DATASETS OF STUDENTS

2011	2012	2013
665 students	679 students	685 students

[a.] Total of 2029 instances

Estimation assumed that the first year of study is the most critical for the students' dropout, so the attribute values are based only on the first year of study of the students. Table 2 shows the attribute retention mechanism for grouping students into categories based on several criteria.

TABLE II. ATTRIBUTES IN THE DATASET

Attribute	Type	Possible values
PreviousGrade	Numeric	{2,3,4,5} – pre-university education curriculum
AvgGrade	Numeric	{5,6,7,8,9,10} – average grade from exams passed in first year of study (FYoS)
SumCredits	Numeric	{0-70} – number of credits enrolled in FYoS
MathematicsAvg	Numeric	{5,6,7,8,9,10} – average grade from mathematics courses enrolled in FYoS
MathematicsCount	Numeric	{0-36} – number of exam applications for mathematics courses enrolled in FYoS
ProgrammingAvg	Numeric	{5,6,7,8,9,10} – average grade from programming courses enrolled in FYoS
ProgrammingCount	Numeric	{0-36 } – number of exam applications for programming courses enrolled in FYoS
LivingPlace	Nominal	{east,west,skopje} – estimated regions based on a geographic basis of the cities in Republic of Macedonia
Nationality	Nominal	{mk,notmk} – estimated values based on nationality of the student
StudyProgramme	Nominal	{Kni,Pet,Mt,Knia,Ke,Iki,Asi,Info,Pit} – study programme of enrolment
Sex	Nominal	{m,f}

[b.] Total of 2029 instances

- Mandatory courses from the first year of study are divided into two categories, mathematical and programming. Based on this classification, each student dataset contains values for average grade and number of exam applications for each category of courses.
- Living place of the students provides another group of students, organizing the geographic locations of the cities in Republic of Macedonia to east region, west region and the region of Skopje – the capital.
- Another group of classification defines the nationality of the students, divided into students who have Macedonian nationality and those who are not ethnic Macedonians.
- The last group defines the study programme in which the student was enrolled at the first year of study at the faculty.

C. Implementation of mining model

The mining model was created based on the attributes described earlier. Initially, a training set was created from the subsets of students enrolled in 2011 and 2012. After the model was evaluated in Weka, the subset of students enrolled in 2013 was supplied as a test set. After the model was completely evaluated and predictions were estimated, we compared the predicted data with the real data obtained from the Students` Information System. The estimation of the model and predictions were made with both J48 and Naïve Bayes classifier algorithms.

D. Result analysis and discussion

Classification accuracies for the dataset containing all three subsets of student enrolment retentions are shown in table 3. The presented results indicate that the Data mining with J48 algorithm is more accurate than the Data mining with Naïve Bayes classifier algorithm.

TABLE III. ACCURACY AND RATES OF TOTAL DATASET

Classifier	J48		Naïve Bayes	
Accuracy	81.1679 %		76.7833 %	
Class signedOut	Yes	No	Yes	No
TP Rate	0.959	0.771	0.918	0.727
FP Rate	0.229	0.041	0.273	0.082
Precision	0.534	0.986	0.479	0.97
Recall	0.959	0.771	0.918	0.727
F-Measure	0.686	0.865	0.629	0.831
ROC Area	0.933	0.933	0.915	0.915

The same classification techniques were applied to the datasets containing both the real data and predicted implications. This comparison aims to achieve the reasons for students' dropout and to predict the circumstances in which the student needs attention.

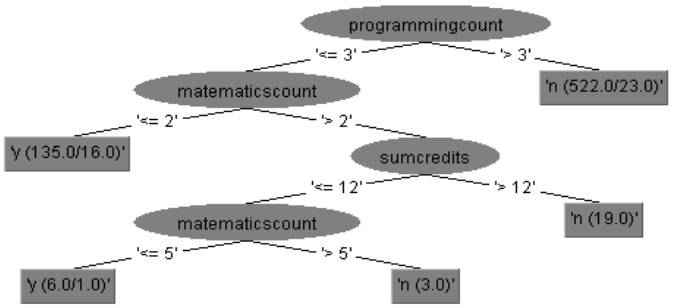


Fig. 2. Decision tree of critical variable values for students` dropout

The presented results in fig 2, emphasize the significant fact that the most critical variables for estimating the students` dropout in their first year of study are the number of exam applications for both mathematics and programming courses.

To get a better insight over the dropout analyses, Appendix A shows the correlation between the nominal attributes that we had extracted from the data and the social and demographic characteristics of the student. From the results, the dropout rate is higher among the students from other nationalities than Macedonian. Also, the dropout rate is higher among the students that are from the western regions of Macedonia.

V. CONCLUSION AND FUTURE WORK

Student dropout prediction is an important and challenging task, yet not clearly defined. This paper presents the accuracy of both J48 and Naïve Bayes classifier algorithms for data mining over the educational data collected from the Students` Information system at one of the faculties at the Ss. Cyril and Methodius University – Skopje.

Our case study shows that the accuracy of the different classifier algorithms notably depends on the quality of the attributes extracted from the data. The accuracy of the classifiers is tied closely with the quality and sophistication of the data model.

According to the results, the most valuable attributes for the prediction are the number of exam applications for both mathematical and programming courses. Since the results reveal a pattern among the number of exam applications between the mathematical and programming courses, equally important are the demographic characteristics of the students.

The model evaluation techniques of classification presented in this paper point to two major improvements that can be noticed. First, the courses grouping into subsets based on the field of study brings a big improvement into the data mining process. In addition, a better encoding grades scheme is required for the students not involved into a specific course. The second and final remark points out the quality and size of the students` dataset. Furthermore, the results are better if the dataset is bigger and well organized.

REFERENCES

[1] Pechenizkiy, M., Calders, T., Vasilyeva, E., De Bra, P. Mining the student assessment data: Lessons drawn from a small scale case study. In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187-191, 2008.

[2] Romero C and Ventura S, "Educational Data Mining: A survey from 1995 to 2005" expert system with Application 33(2007) 135-146

[3] Luan, J. Data mining and its applications in higher education. New Directions For Institutional Research, p. 17-36, Spring 2002.

[4] Romero, C.; Ventura, S.; Espejo, P.G.; Hervas, C. (2008) Data Mining Algorithms to Classify Students. In Proceedings of the First International Conference on Educational Data Mining, (pp. 8-17).

[5] Pechenizkiy, M., Calders, T., Vasilyeva, E., De Bra, P. Mining the student assessment data: Lessons drawn from a small scale case study. In Proc. of the 1st Int. Conf. on Educational Data Mining (EDM'08), p. 187-191, 2008.

[6] Herzog, S. Estimating student retention and degreecompletion time: Decision trees and neural networks vis-avis regression, New Directions for Institutional Research, p. 17-33, 2006.

[7] Gerben W. Dekker , Mykola Pechenizkiy, Jan M. Vleeshouwers, Predicting Students Drop Out: A Case Study, Educational Data Mining, 2009

[8] Dr. Saurabh Pal, Mining Educational Data Using Classification to Decrease Dropout Rate of Students, INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, VOL. 3, NO. 5, MAY 2012

[9] Mario Jadrić, Željko Garača, Maja Ćukušić, STUDENT DROPOUT ANALYSIS WITH APPLICATION OF DATA MINING METHODS, 2010

[10] Su, J.-M.; Tseng, S.-S.; Wang, W.; Weng, J.-F.; Yang, J.T.D. and Tsai, W.-N. (2006). Learning Portfolio Analysis and Mining for SCORM Compliant Environment. In Educational Technology & Society, 9(1), (pp.262-275).

[11] Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G. and Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. In Computers & Education, 53(3), (pp. 950-965).

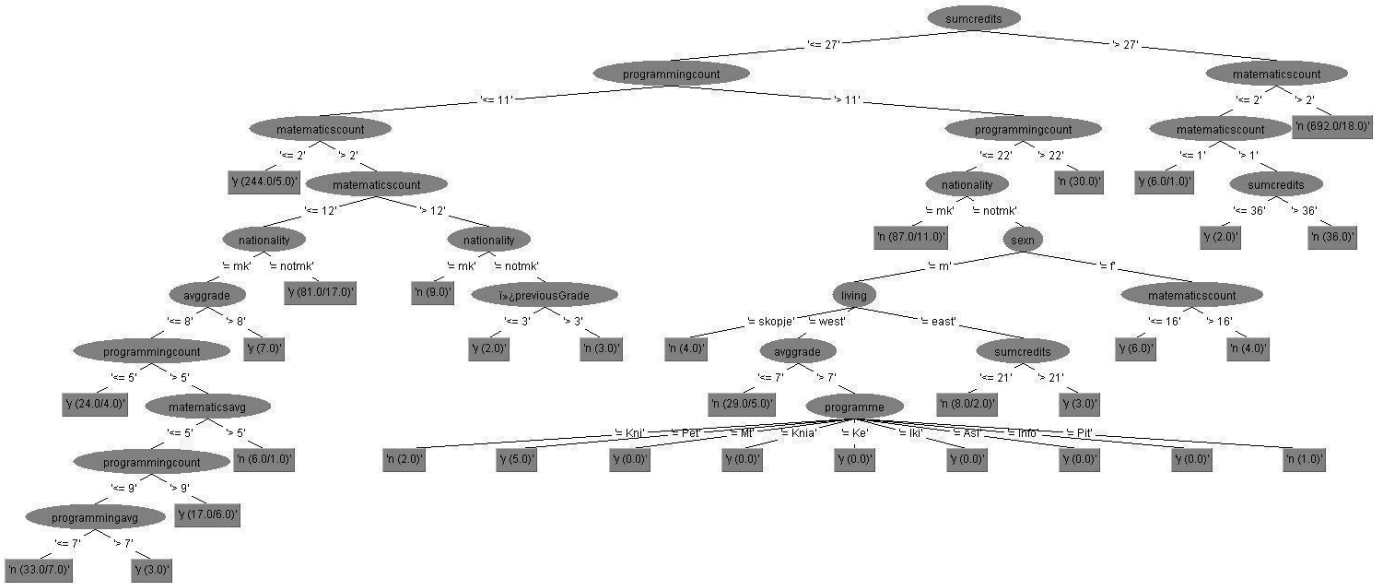
[12] Tinto, V. Limits of theory and practice in student attrition, Journal of Higher Education 53, p. 687-700, 1982.

[13] D. Carcia-Saiz and M.E. Zorrilla, Comparing Classification Methods for Predicting Distance Students' Performance, Journal of Machine Learning Research – Proceedings Track, Vol.17, 2011, pp.26-32.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[15] J. Han and M. Kamber, Data Mining: Concepts and Techniques, second edition, Morgan Kaufmann, 2006.

APPENDIX A. DECISION TREE



Finding Similar Answers Using Some Text Mining Techniques

Bojan Ilijoski
Faculty of Computer Science and Engineering
Skopje, Macedonia
Email: bojan.ilijoski@finki.ukim.mk

Zaneta Popeska
Faculty of Computer Science and Engineering
Skopje, Macedonia
Email: zaneta.popeska@finki.ukim.mk

***Abstract*—The idea of this paper is to describe the efficiency of the spherical k-means algorithm in the process of grouping short texts, in this case, answers of questions in Macedonian. Actually, we test the capability of this algorithm to group a set of answers of certain questions in groups in which each answer would be an answer of a certain question. Also, we suggest a way in which the answers would be presented as vectors and then grouped. At the end we give an evaluation of the efficiency of this method.**

I. Introduction

The evaluation of student performance represents plays a significant role in the educational system. Through regular evaluation students get proper certification for the gained knowledge and a certification for mastering certain material and/or skill. In a dynamic, fast developing society the education is one of the main factors. The society itself forces the citizens to take their education on a higher level and that causes the number of students to increase. The enormous increase of the number of students in a society besides having a positive effect, which is the higher level of education in general, also has negative sides. One of the problems that appear when the number of students increases rapidly is the lack of teaching staff. One of the teachers' assignments is to teach a certain number of students and now that number increases from 4 to 5 times on average. Because the amount of work is bigger, and the resources remain the same, in order not to lower the quality of the education, the implementation of new methods is necessary.

There are techniques of text mining that can help the process of checking textual answers of students. The techniques for finding similarities between short texts can be used in the checking of the students' textual answers and in a way they can make this process semi automatized. The idea is to find the answers close in syntax if there is a large set of answers of a certain question, because there is a possibility to grade these answers as they were equally or almost equally correct. In this way, the process of grading will be easier for the professors because all of the similar answers which are supposed to be given the same number of points will be put in the same group. This method can also be useful for detecting those who copy because, regardless of the size of the set of answers it will put the equal answers in the same group, and will give them large mutual similarity. This will give the evaluator an indication that maybe one of the answers is copied. This would be more difficult to discover with the standard way of checking, especially when it comes to a large set of answers. If one of the answers is read at the beginning of the checking and the other in the end, between those two, hundreds of other answers are read, so it is difficult to pay attention to and notice that there has already been found a similar or an equal answer. The disadvantage of this access is that the process cannot be completely automatized (for now). What this technique can offer to the evaluator are the questions ordered by similarity. The process of grading is still an assignment for the evaluator. In addition, after the evaluation (the grading)of some of the questions, for the questions which follow based on their similarity with the previous ones, the evaluator can be offered certain grade for those questions. Again, we are talking about a process which is not completely automatized and we cannot absolutely rely on it.

In this paper we will discuss a method which can be used in the grouping of answers written in Macedonian and we will show how it functions on a set of answers given by students.

In the next section we present some papers which motivated us to do this research. They describe the methods for resolving the same or similar problems. The section "The method" describes the techniques we use in order to find and cluster similar answers, some preprocessing techniques and Spherical K-means algorithm as main clustering method. The results of our work are presented in the "Results" section and in the last section we talk about our plans to find out a way to improve out method in order to get better results.

The evaluation of the efficiency of this method is performed by using the textual answers from FINKI's electronic test system.

II. Related work

Taken in consideration the need for quality grouping of answers, it is not strange at all that a great number of scientists do a research on this topic. In the paper called "A Semi-supervised Topic-driven Approach for Clustering Textual Answers to Survey Questions" [1] they suggest an algorithm for grouping answers, so that the answers from one group are related to one topic, and the answers from another group to other topics. They use modified k-means algorithm with cosine similarity and Single-link Based Clustering Algorithm for determining the groups. Then, they extract the topics

and Topic to Answer Clustering (T2A)and Answer to Topic Clustering (A2T). In "Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading" [2] the authors present a method for easier grading of the students' textual answers based on the grouping of answers. They say that this enables easier grading and, for the evaluator, multiple answers and feedbacks with just one action. They also suggest a measure for measuring the similarity and distance between the answers. They use PAM algorithm and LDA Clustering for the grouping of answers. Some of the authors try to group open answers mostly for marketing purposes [3] and suggest an analytical system based on rule learning and correspondence analysis. In another paper [4]they give a method for extracting and discovering relations between open and closed answers through the evaluation of the techniques for text clustering. They also indicate the differences in the characteristics that the first and the second type of answers have. In [5] they show the utility of the light-weight Language Engineering techniques for text clustering. They, as well, keep their attention on the pedagogical value of the short textual answers. They show that there is a little difference between the techniques used, and the difference appears because of the way the questions are formed.

III. The method

The algorithm's work is based on the similarity of answers depending on the amount of similar or equal words in each of them. There are several phases in the process of finding similar answers and grouping them.

The first challenge we faced and we had to solve is the script in which the answers were written. Since the Cyrillic script is official in the Macedonian language, some of the answers were written in Cyrillic. On the other hand, the default script on the computers is the Latin. The Macedonian language also uses this script, so there were also answers written in Latin. The first phase consists of putting the answers in the same script. That is the reason why those answers should be changed into Cyrillic.

The next phase includes the selection of the words, finding their similarities and grouping them. This is necessary because of the many forms in which Macedonian words could appear but have the same meaning or a similar one (ex.добриот, подобриот). To be able to find the same or similar words in different answers, the words should be unified in some way. This phase has several sub-phases. First, with a help of a dictionary the lemmas for each word are singled out. Then, their lemmas help their grouping (determining whether they are equal or not). In this part of the process we used the multex [6] dictionary. This enables us to find the lemma of the word, which makes the process of finding equal words easier. If two words have the same lemma, they belong to the same group, i.e. these two words are equal. This way of grouping perhaps will not give a completely equal results because of its several defects. The synonyms (words with different forms but the same meaning) are not grouped together i.e. it cannot find similarity between them (ex. поаѓа - тргнува, лекар-докор, мома-девојка, говори-зборува). The homonyms (words with the same form but different meanings)have the same lemma and are grouped together, although they have different meaning (ex. Книгата има 4 глави, Тој е глава на семејството, Главата е дел од телото). Also, since great number of the answers contained terms that were not Macedonian, spelling mistakes, dialects, this give us just a little help.

After grouping the words (bringing them to the same lemma),the next step is the creation of a vector of words for each answer. Then we create tf–idf (measure that determines how important a word is to a document)according to its presence in a corpus of documents (in this case answers) [7].

In the next step, we determine the answers' similarity. The similarity of the answers is calculated as cosine similarity between their vectors. This is the way in which a similarity matrix is built.

After we build the similarity matrix, the following step is grouping of the answers for which the algorithm Spherical K-means is used [8]. According to this, answers are grouped in the cluster if they have high similarity.

IV. Results

The testing is performed on different number of questions (clusters) and answers (instances). We took the questions and the answers from tests from the IT systems for learning course in a period of three years. The efficiency is measured by determining the value of precision macro, recall macro and the percent of correctly classified answers [9]. If the answer a is the answer of the question q, the answer is correctly clustered if it is in the cluster c, and at the same time the majority of the answers of the question q are in that cluster. The testing of the method is performed on 1982 answers of 30 different questions.

On the figure 1 we present the values for precision macro, recall macro and TP/all through different number of clusters. For small number of clusters this method gives good results. But, as the number of clusters grows, the number of answers also grows and the probability to have two similar questions increases. Because of this, the number of wrong answers and the similar answers that should be clustered in different clusters rises. So, for a larger number of questions the method gives pure results.

On the figure 2 you an see a number of answers and distinct words. These numbers increase as the number of questions increases. But, after a certain number of clusters we can see that the number of distinct words increases more slowly. This probably happens because all of the words on that subject are already included.

Figure 3 presents answers per question and distinct words per question. The values are increasing and decreasing simultaneously. On figure 4 there is a histogram of number of answers per question.

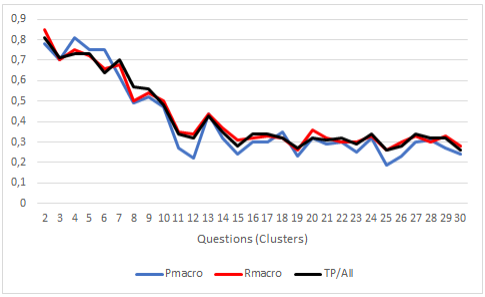


Fig. 1. Pmacro, Rmacro and TP/all values for different number of clusters

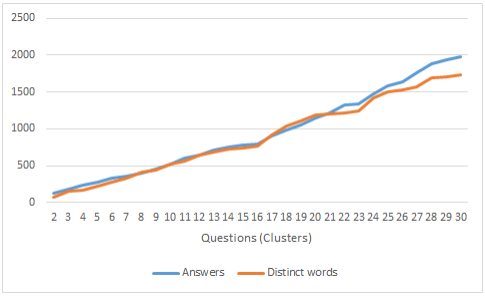


Fig. 2. Number of answers and distinct words for different number of questions

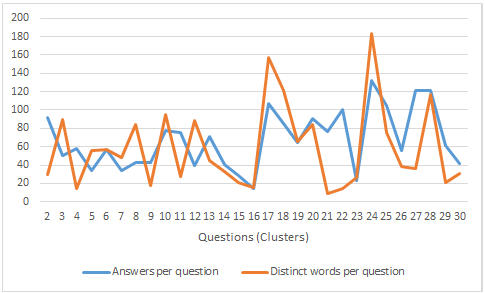


Fig. 3. Answers and distinct words per question

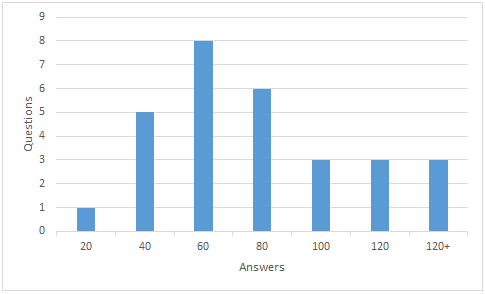


Fig. 4. Histogram of number of answers per question

V. Conclusion and Future work

This method gives good results for a small number of questions. We can see that there is a small number of mistakes. In general, these mistakes are done because of those answers which are really different from the other answers in the cluster they are in. In the following text we will explain why this happens. The results are worse when using greater number of clusters i.e. questions. This happens due to several factors. One of the factors are the spelling mistakes in the answers.(ex. management, managment). This is the cause of a wrong presentation of the answers through their vectors. Because of the way this method is set, it does not offer an access for effective recognition of the words. Those words are considered different. Further on, this is the reason for a great difference in the answers. Another factor which significantly influences the results is the way in which the questions are formed. Namely, there are a lot of similar questions for which similar or the same answers are given. In this way, the answers which naturally belong to one cluster are close with the answers of another cluster. This can also help the formation of a cluster from the answers of two similar questions, which should exist anyway. But, because the algorithm needs a certain number of clusters, one cluster appears as a spare and that cluster gives only wrong results. Additionally, when checking the cluster in which the answers of two questions are merged, the algorithm considers correct only the answers of the one question (those prevailing) while it considers the rest as wrongly clustered. This significantly affects the efficiency. The wrong answers also present a great problem in the process of grouping. The wrong answer of a question is far from the correct answers of the same question. Because of this, the answer is clustered as another question's answer that is considered wrong.

As a future work we plan to eliminate this problems, as much as possible. We will use more effective method for similar word clustering in order to avoid the problem with misspelling. Also, we will manually find the similar or the same questions and put their answers in the same cluster. Or, if the same question is repeated several times, we can remove the rest and keep only one. The answers scoring should contribute to a better clustering. Every answer should get points depending on its correctness (if the question is all correct or not). The better answers will get more points. We can use those points of the answers as weight of answers. The highest weight suggests that an answer really belongs to that cluster(the cluster where all the answers to that questions belong). This technique will allow us to eliminate the importance of wrong answers and this will cause the clusters to gravitate towards the correct answers.

Acknowledgment

The work has been supported by the Faculty of Computer Science and Engineering

References

[1] Hui Yang, Ajay Mysore, and Sharonda Wallace, A Semi-supervised Topic-driven Approach for Clustering Textual Answers to Survey Questions, Advanced Data Mining and Applications: 5th International Conference, ADMA, 2009

[2] Sumit Basu, Chuck Jacobs, Lucy Vanderwende, Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading, TACL, 2013

[3] Yamanishi, K., Hang Li, Mining open answers in questionnaire data, Intelligent Systems, IEEE, 2002

[4] Magnus Rosell, Sumithra Velupillai, Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation

[5] McGee Wood, M. Light-weight clustering techniques for short text answers in human computer collaborative (HCC), CAA, 2006
[6] Tomaz Erjavec, "MULTEXT - East Morphosyntactic Specifications", Department for Knowldege Technologies, Jozef Stefan Institute
[7] Salton, G., Buckley, C. Term-weighing approache sin automatic text retrieval. In Information Processing and Management, 24(5): 513-523.s, 1988
[8] Shi Zhong, Efficient online spherical k-means clustering, Neural Networks, 2005. IJCNN '05. Proceedings. 2005
[9] James Petterson, Tibério S. Caetano, Reverse Multi-Label Learning, Advances in Neural Information Processing Systems 23, NIPS, 2010

Public health facilities spatial distribution in Macedonia

Stefce Aksoski, Dejan Brjakoski, Nikola Koteli, Andreja Naumoski, Kosta Mitreski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Republic of Macedonia
{stefce.aksoski, dejan.brjakoski, nikola.koteli, andreja.naumoski, kosta.mitreski}@finki.ukim.mk

***Abstract* - Health is the first and essential prerequisite for the successful execution of all life activities. The quality of the health system of one country is in correlation with the health of the population. The health care system and the medical facilities, represent one of the most important and most significant factors in medical care. The information regarding the network of health care facilities, or information related to their activity and locations is of great importance. In this direction, this paper aims to present the spatial distribution of the Macedonian public health facilities using ArcGis and HeatMap. The paper presents ArcGis map and heatmap layers, to give a clear picture of the distribution of health facilities in Macedonia. The data that is presented, was taken from the Health Insurance Fund of Macedonia. The presented map shows a clear image of the spatial distribution and thus coverage of the health facilities that are currently available in the country. Further development of the work, should include a mobile application for the smartphone mobile users.**

Keywords— Health heatmap, hospitals, dentist, laboratories, Macedonia

I. INTRODUCTION

GIS or geographic information systems is a collection of various tools that are used to manage geographic relations across space and integrate numerous types of information. This software package provides practical tools for visualizing and analyzing health data, revealing trends, dependencies and inter-relationships. This information can be used to learn about a geographical area, to manage a health care project, choose an ideal site for something or choose a delivery route for faster and best available service. It can acquire, store, manage, and geographically integrate large amounts of information from different sources of data, other programs from different sectors. In this direction, public health hospitals and facilities, specific diseases and other health events can be mapped in relation to their surrounding environment and existing health and social infrastructures. Standardized geo-referencing of epidemiological data facilitates structured approach to data management. This can serve as a common platform for the convergence of multi-disease surveillance activities.

In the recent years, GIS is actively used to represent the health information regarding the health of the people in a given spatial area. Research in the relevant literature shows numerous and important case studies about the applicability of GIS regarding the health issue [1, 2, 3, 4, 5]. In this reaserch area, the implication is obvius exspecaly in the public health sector in many important institutions around the world [6]. There also web based decision support systems based on the geographic analysis to represent not only the data, but to further improve the functunality of such help tool for decision makers [7]. This type of applications are wide and broad, and can be very informative for the general public. However, no such system still exist to conduct such analyses of the public health system in Republic of Macedonia.

The goal of this paper is to develop geocoding map of all health facilities, in total 4040 facilities, that have contracted with the Health Insurance Fund of Macedonia. The data used in this research comes in form of tables, which contains the list of all public health facilities. The list of medical institutions was taken from the website of the Health Insurance Fund of Macedonia. Then we converted the list in Microsoft Excel document for easier accessing data, their changes, and we used and archive the document for further work. Each health facility is presented as a separate row with their data, additionally we added supplementary attributes Longitude and Latitude (x, y) coordinates. The longitude and latitude are used to determine the location of the facilities and to display on the map. The generated file with the georeferenced hospitals is used as a database to generate GIS layers, which is a major step toward building a decision making system for general public health care information.

The rest of the paper is organized as follows: Section 2 provides information about the ArcMap module in the GIS software and its usability. In Section 3 we present the procedure of adding the data on heatmap and automatically representing the map on a web page. The Section 4 build various scenarios where this data is shown, while the Section 5 concludes the paper and given further direction on how we plan to develop our decision support system in the future.

II. ARCMAP

One of the central application of the ArcGIS package is the ArcMap. ArcMap is a software module in the GIS, where user can present and explore georeferenced data. On the map, the user can place symbols and the user can use the software to create a map for publication. This module is also used to create and change the data. ArcMap represent geographic information as a combination of layers and user inputted elements of the map. It is the primary application in ArcGIS and used to perform many common GIS tasks. The GIS layers are actually a mechanism that is used to represent geographic data in

ArcMap. Each layer referenced data set and specifies how that data set is represented using symbols and text labels. When the layer is added to the map it's specify the data set and set of map symbols. Layers are one of the primary ways to work with geographic data in ArcGIS

In order to represent the public health facilities in Republic of Macedonia, we used the geographical data of the health care facilities and their location. This procedure was repeated for each of a total of 4040 facilities. The view of the health facilities given by using ArcMap (See Fig. 1). As a basic map we chose topographical. It can also be changed predefined symbols that mark the location. This can be done by double clicking on the symbol in the Table of Content menu and choose something else. Since our theme of the paper is in relation to health facilities will change the symbol to signify hospitals, as it was shown in Fig 1. If any user wants to get the information related to a given health institution, all he needed to do is the right click on one of the icons shown on the map and receives a list from which we select "What's here?", and we get the required coordinates.

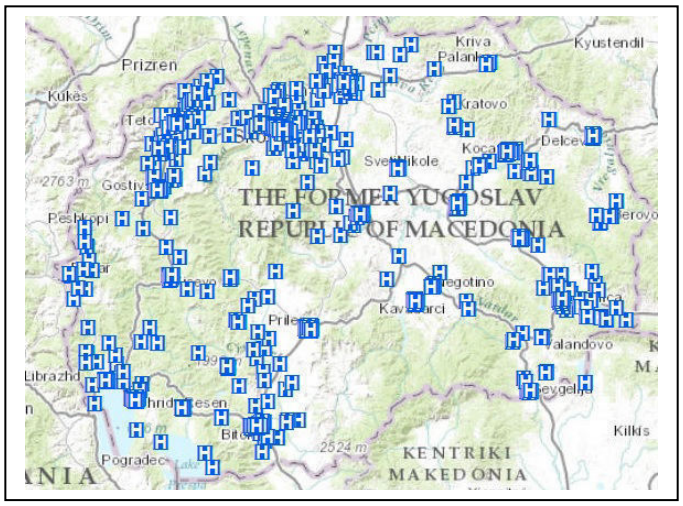


Fig. 1. ArcMap representation of all the public health institution in Republic of Macedonia

The presentation of the above features combines a multi-layered map in ArcMap. Also, the ArcGIS can display different layers that can be combined with each other to provide a clear representation of what is of interest to the user.

III. REPRESENTING THE DATA WITH HEATMAP

First we prepare the table with the data from the ArcMap and then we export the data in order to present the data on a heatmap.

The collected information on the public health facilities in Macedonia are used to display on a heatmap. For that purpose, we had Google Maps V3 Layer and heatmap.js. Via the coordinates generated heat maps for some categories of public health facilities. Using an html editor we have entered the necessary Java Script files and set some of the basic things Google Maps layer.

Before we process the data using the java Script, we separated the coordinates by type and added the list in JavaScript. Then through the appropriate API the system draws information from the collected data and then it represents on the heat map.

```
<script type="text/javascript" src="heatmap.js"></script>
<script type="text/javascript" src="heatmap
gmaps.js"></script>
<script type="text/javascript">

var map;
var heatmap;


window.onload = function(){
var mylatlng = new google .maps .latlng 48.3333, 16.35);
var myOptions = {
zoom: 8 ,
center: new google.maps.latlng 41.681118,21.686153),
mapTypeid: google.maps.MapTypeid.SATELLITE,
disableDefaultUI: false,
scrollwlheel:true,
draggable:true,
nav igationControl:true,
mapTypeControl:false,
scaleControl:true,
disableDoubleClickZoom : false
};
map = new
google.maps.Map(document.getElementByid("heatmapAre
a"), myOptions);
heatmap = new HeatmapOverlay(map,{"radius":15,
"visible":true, "opacity" :60});

document.getElementByid("tog").onclick      function(){
heatmap.toggle();
};
var testData ={...
var labData = {...
var opstaData = {...
```

Fig. 2. JavaCode for the HeatMap

Various operations can be performed using the script on the maps in order to obtain more informative scenarios. The procedure is the same for the other types of health care facilities that we have data.

IV. PUBLIC HEALTH SCENARIOS

To explore different information contained in the data, we have obtain several different heatmap scenarios. The aim of these scenarios is to represent the geographical information of different type of public health facilities.

In this direction on Fig. 2, we present a heatmap of all public health emergency facilities in Republic of Macedonia.

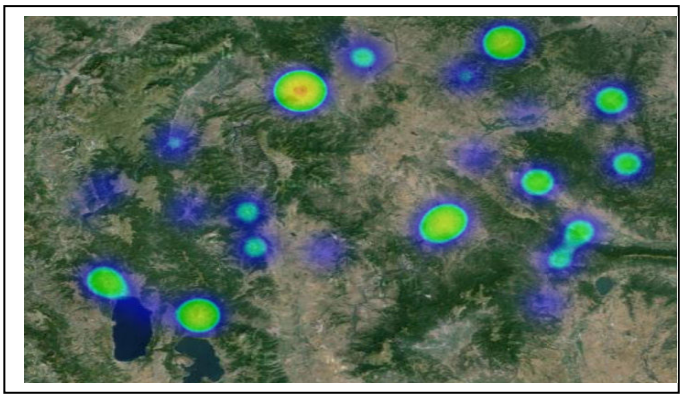


Fig. 3. GIS representations of the Emergency Health Facilities in Republic of Macedonia

As can be see, the density of health facilities for emergency care is quite small. However, they are strategically placed in order to respond to the need of the citizens.

Another interesting informative map is the heatmap of all public health facilities on Fig 3. From the map, it is easy to note the high density of the public health instructions are near the largest cities in Republic of Macedonia. With other words, this claim can be further investigate, there is a correlation of the population density and the density of this map. Further analysis of the population maps could also improve the need of more public health hospitals.

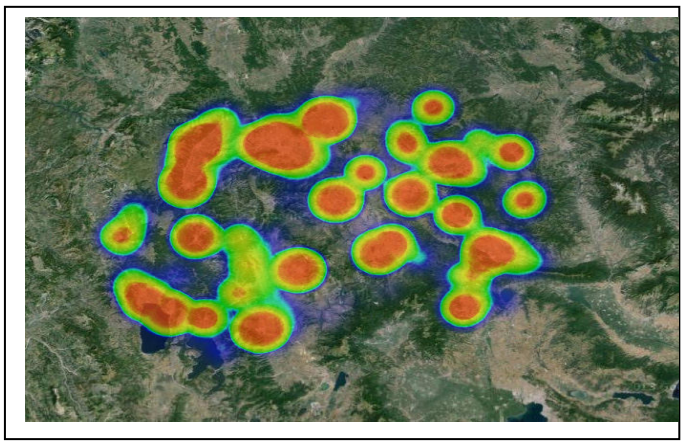


Fig. 4. GIS representations of the Public Health Facilities in Republic of Macedonia

Next, we represented the dentist health facilities (see Fig. 4). We can note that the dentist facilities are less present in every city, but they are far more than the emergency health facilities presented in Fig. 2.

These maps can be used as decision support tool in every aspect of the informative health system as an analytical tool and based on the map provided in this paper. This process can further improve the understanding of the oral health/disease process, and to explore the relationships between cities, rural areas, population and the public health. Further detailed analysis of the pattern of inequality and spatial distribution of oral diseases or any given type of public health need is fundamental for the allocation of resources to areas with the greatest social privation, leading to greater efforts to address the problems.

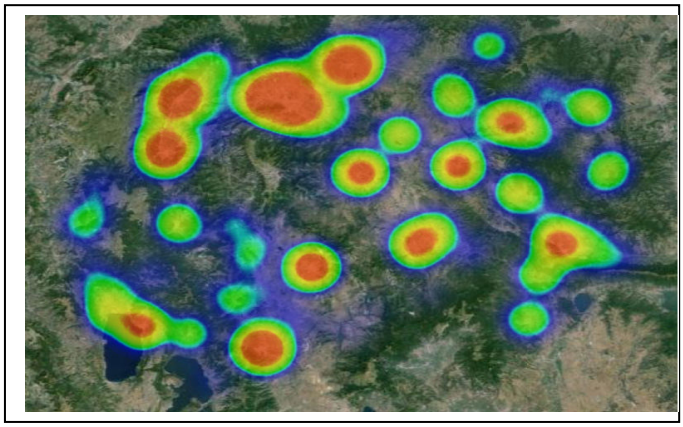


Fig 4. GIS representations of the Public Dental Clinics in Republic of Macedonia

V. CONCLUSION

In this paper we presented two interconnected applications. First, we use the ArcGIS map to show the existing public health facilities in Republic of Macedonia and then a HeatMap to further improve the visual representation of the available data. The entire process, starting with the data collection through the process of making the maps and later through data geo-processing is presented in this paper. The GIS application provided the four groups of aptitudes to work with geo-referenced data input, management, manipulation and analysis. Later this data was inputted into the heatmap java script. The presented maps in this paper included several important information about the public health facilities: the major hospitals, emergency health facilities and public dental clinics in Republic of Macedonia. The HeatMap use interpolation in order to automatically present the data on the map or on a google maps web service using a java script. This provides a useful decision-support tool with a substantial potential to be harness in process of planning public health care services.

In conclusion, our findings demonstrated the obvious limitations in distribution of Emergency Health Facilities in Republic of Macedonia for serving the rural regions, thus making inhabitants lived in places far from the major cities more disadvantaged. Nevertheless, as a government rapidly invest in the public health facilities, they should eliminate these gaps in emergency facilities. In this direction the maps presented in this paper can be a helpful tool in the process of carefully planning the development. Therefore this study highlights the ongoing need for monitoring the distribution of the public health services to ensure they match the evolving socio-economic development.

Further improvement of this system can be done by implementing methods and techniques to analysis other type of data. Such systems could use data from the social network cites [8, 9] and then develop system that interconnects the existing

knowledge and the services that this industry provides. An interconnected system that will update on daily level and stream the map information to every day users is our major future focus. This couldn't be done better then using the latest smartphone technology. In this direction our major future research is directed in developing smartphone applications.

ACKNOWLEDGMENT

This paper work was partially supported by the Faculty of Computer Science and Engineering.

VI. REFERENCES

[1] A. Briney, "Medical Geography – A History and Overview of Medical Geography," [Online]. Available: http://geography.about.com/od/culturalgeography/a/medicalgeograph.htm . [Accessed 27 March 2014].

[2] E. K. Cromley and M. L. Sara, GIS and Public Health, 2nd, Ed., New York: The Guilford Press: New York, 2012.

[3] H. S. o. P. Health, "Geographic Information Systems (GIS) in Public Health Research.," Harvard University, 2 January 2013. [Online]. Available: http://www.hsph.harvard.edu/gis/. [Accessed 2 May 2014].

[4] C. I. Nykiforuk and L. M. Flaman, "Geographic information systems (GIS) for Health Promotion and Public Health: a review," *Health Promot Pract 2011,* vol. 12, no. 1, pp. 63-73, 2011.

[5] E. J. Holowaty, T. A. Norwood, S. Wanigaratne, J. J. Abellan and L. Beale, "easibility and utility of mapping disease risk at the neighbourhood level within a Canadian public health unit: an ecological study," *Int J Health Geogr,* pp. 9-21, 2010.

[6] C. J. Nykiforuk and L. M. Flaman, " Geographic Information Systems (GIS) for health promotion and public health: A review," *Health Promot. Pract.,* vol. 12, p. 63–73, 2011.

[7] P. Fu and J. Sun, "GIS in the Web Era," in *In Web GIS: Principles and Applications*, Redlands, CA, USA, ESRI Press, 2011, pp. 1-24.

[8] C. St Louis and G. Zorlu, "Can Twitter predict disease outbreaks?," *BMJ,* p. 344, 2012.

[9] A. C. Robinson, A. M. MacEachren and R. E. Roth, "Designing a web-based learning portal for geographic visualization and analysis in public health," *Health Inform. J,* vol. 17, pp. 191-208, 2011.

Case-based retrieval aided with image data

Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski
and Suzana Loskovska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia

***Abstract*—In this paper we present a strategy for retrieval of medical articles, which is also referred as case-based retrieval. The proposed strategy is using both text and visual data to perform the retrieval. The aim is to utilize multimodal data to hopefully get better retrieval results. In the text part we used Terrier IR search engine for indexing and retrieval. The medical articles are indexed based on their entire text content (including the caption of the images within). In the visual part we only focus on modality detection using state-of-the-art features to describe the images and then train classifiers to distinguish between different image modalities. Furthermore, the retrieval phase is consisted of a text and visual query i.e. narrative text and accompanying images. We use the text query to perform the retrieval and then based on the modality of the accompanied images we boost the articles that contain this kind of medical data, whether it is the text or in the figures they used to complement it. Our approach shows that combining different modalities in this fashion can provide a boost in the retrieval performance.**

Keywords—case-based retrieval; text retrieval; medical articles;modality classification

I. INTRODUCTION

The main research topic of this paper is medical articles (cases) retrieval. Medical information retrieval (IR) is defined as task related to analysis, organization and retrieval of medical information [1]. This activity is an important part of normal medical practice. Since, it helps medical professionals by providing them with relevant medical cases or other medical data. The data can help the medical professional in the current case he/she is working on. It can help him/her make a better diagnosis for the current patient, or provide more appropriate treatment based on other similar scenarios [2]. Hence, the relevance of the retrieved cases is very important and poses an interesting research problem.

One can find wokring online systems which provide medical articles retrieval, such as Pubmed[3], Pubget[4], eTBLAST[5] etc. These systems essentially treat the articles as a set of words and do the indexing and retrieval based on the keywords passed as a query by the user. A more real-world practice would be a scenario where the queries would be of narrative nature, where the user can explain the situation in more details [6].

The task of case-based retrieval is defined by the Cross-Language Evaluation Forum (CLEF). CLEF is a self-organized body whose main mission is to promote research, innovation and development of information access systems [6]. Case-based retrieval focuses on retrieving the most relevant medical cases in respect to a give text and/or visual query. It is considered as a complicated problem, but that is why it is meant to be more similar to the everyday workflow of the medical professionals. In this problem definition, the input is a case description, with patient demographics, limited symptoms and test results including imaging studies, but the not the final diagnosis. The object of retrieval in the task is medical cases along with the images depicted in them. Hence, the end goal of the task is to find out how much a given medical case is relevant to a given query.

In this paper we want to investigate whether to include image data in the text-based retrieval of medical articles and if that can provide a performance boost. Our approach treats the medical articles like a set of terms and uses techniques for indexing and retrieval implemented by many search engines. After this initial retrieval phase, we want to fine-tune the results by introduction modality classification. The modality classification is performed over the images contained within the cases and based on that and the modalities mentioned in the text query, we re-rank the retrieved cases. The results show that this strategy provides a slight boost in overall retrieval performance.

This paper is organized as follows. The section two presents the related work. Section three contains the details of the proposed approach. Section four presents the experimental setup and design. The results and discussion are presented in section five. The conclusion and future work directions are presented in section six.

II. RELATED WORK

Retrieval of medical article is an area of active research and there are many attempts at solving this complex issue, despite the fact that there are existing commercial systems which offer the same or similar services. Existing approaches usually are based on an off-the-shelf search engine and on top of that other algorithms are added or combined.

The system presented by Vahid et al [7] uses Terrier IR search engine as a basis for indexing and retrieval. The text is preprocessed in the following manner: special signs and stop words removal, token normalization and stemming with Porter stemmer. The retrieval stage is characterized by the TF-IDF

weighting model which is used as one the most frequent models used in IR systems. Their proposed system provides performances among the best in its category.

Vanegas et al. [8] presented a more hands-on approach by implementing their own custom version of the Okapi BM25 weighting model with the help of the Python NLP toolkit. This technique was proposed for text-based image retrieval and it was reported that it achieved outstanding results on the CLEF 2012 ad-hoc retrieval dataset. Hence, the fact that the solution is generic allows for potential use in the case-based domain as well.

The Lucene search engine was used to index and retrieve the medical cases by the Medgift group [9]. Their approach does not have any other optimization and is reported as the best on the CLEF 2012 ad-hoc retrieval dataset.

Choi et al [9] index different fields of the cases i.e. they index the title, abstract and fulltext field of the document. They stop the queries at query time using the 418 INQUERY stopword list, case-folded, and stemmed using Porter stemmer. Furthermore, they provide query expansion with the help of an external corpus (MEDLINE) for robust and effective expansion term inference. They used the 2013 MEDLINE/PubMed Journal Citations from the U.S. National Library of Medicine, which has around 22 million MEDLINE citations. This method provides good results, but only relies on the text part of the cases, not utilizing the visual aspect as well.

Mourao et al [10] proposed two separate approaches i.e. text and visual retrieval for medical cases. The text approach relied on the Lucene search engine. They indexed and retrieved on the full text (including image captions), abstract and title. They applied pseudo-relevance feedback using the top 3 results retrieved using the initial query. Also, added a maximum of 25 new query terms to the original query. Any candidate expansion word had to appear in minimum of 5 documents. Their visual approach was based on generating multiple low level features (FCTH, LBP histograms and color histograms) and retrieving the cases based on the images within them. Although, the separate approaches provide good results, there was no fusion of both approaches to try and provide even better performance.

Simspon et al [11] use the Essie search engine for indexing and retrieval. On the text side, they use the Essie's embedded feature of query expansion using UMLS concepts. On the visual side, they compute the distances between the descriptors (they generate low level features) of the images contained with the cases with Euclidean distance. The results from both types are merged with early and late fusion techniques. Despite the utilizing of both modalities, their case-based retrieval approach did not provide high results on the CLEF 2013 case-based retrieval dataset.

An interesting approach for ad-hoc retrieval is presented in our previous work [12]. The work shows that text-based retrieval of medical documents can be boosted by re-ranking the image based on their modality with respect to the given text query. The provided results are state-of-the-art and are the best reported results on the dataset. Although the approach is directed at solving image retrieval, the idea can be used to boost the performance of cased-based retrieval too.

The related work shows many approaches to solve the problem of case-based retrieval and some of the presented approaches do not try to solve this problem directly, but present ideas that can be used for this approach as well. Although, most approaches use the Lucene search engine, some the most efficient retrievals are done using the Terrier IR search engine. It provides efficient and effective search methods for large-scale document collections. We try to use Terrier IR search engine for indexing and retrieval of the medical cases and use modality classification to re-rank the cases based on the modality of the images they contain with respect to the given text query.

III. PROPOSED APPROACH

The diagram of the proposed approach is presented on Fig. 1. The approach consists of an on-line and off-line phase. The off-line phase deals with the modality classification and data indexing structure. The on-line phase is responsible for the text retrieval and re-ranking the cases based on the modality of images contained within them.

The off-line stage starts with preprocessing and indexing of the medical cases. The preprocessing is consisted of several stages. First, special signs are removed from the text. Next, all words which have no semantic value, but are very frequent are removed from the text (stop words removal). The third stage is token normalization. In this stage, all words are converted to lower case. The last stage of the preprocessing is stemming. At this stage, the words are reduced to their root form. We use Porter stemmer as it is used for removing the commoner morphological endings of words in English [13].

After the preprocessing is done the indexing stage executes. This stage is a standard two-pass indexing [14] and as result an inverted index is created, which afterwards is used for efficient retrieval.

The off-line stage also implements the modality classification algorithm. This part is built by first extracting visual features for a given set of training images for which the modality is known beforehand. The generated features are used to train a classier. The classifier is optimized against a set of validation images for which the modality is also known beforehand (our classifier has a predictive performance of 76.21%). After, the final classifier is obtained, it is used to label all images that are depicted in the medical cases. Of course, prior to that, the same visual features are extracted for all images as well. The result of the modality classification part is an image-modality index i.e. an index where for each image in the database we store it's according modality.

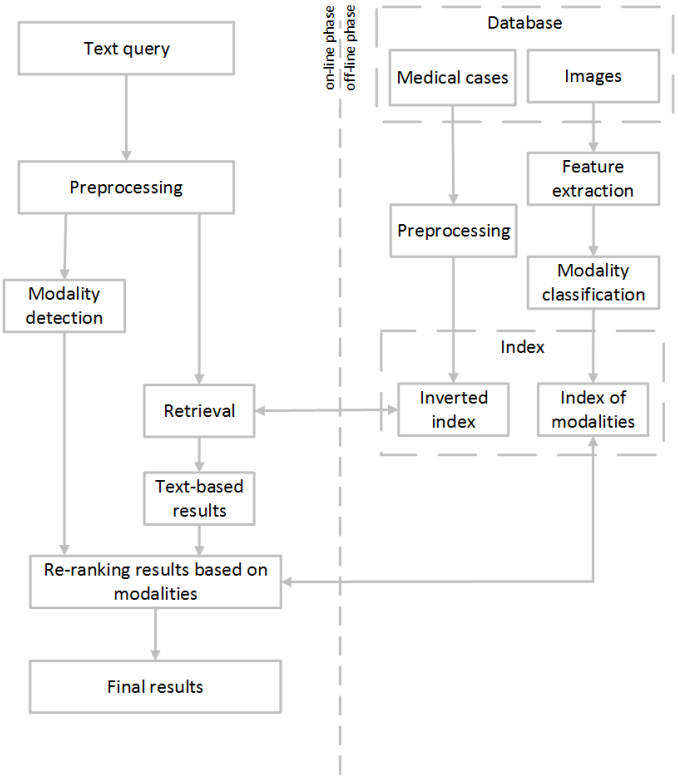


Figure 1: Diagram of the proposed approach

The input is provided the on-line phase. In this situation the input are the text queries. The queries are preprocessed in the same manner as the medical cases and the retrieval is performed using a weighting model. The queries often contain terms related to the modality. In this stage, we extract the modalities of the query by simple term matching. After the retrieval is done, the retrieved cases are re-ranked based on the modality of images within the cases with respect to the modalities of the query.

IV. Experimental Setup and Design

The experiments are performed over the ImageCLEF 2013 [16] dataset. The dataset contains text and visual data. The subset related to case-based retrieval contains 74 654 medical articles (cases), which are mainly journal articles from PubMed [16]. Each article consists of several fields: title, abstract, main text (referred as full-text) and the captions of the images contained within the article. The queries for the case-based retrieval are narrative case descriptions usually containing 3-4 sentences. The dataset has 35 queries. A few examples are provided below:

Query 1. A 50-year-old man with severe right flank pain and hematuria. Renal ultrasound shows a markedly echogenic lesion with a posterior acoustic shadow measuring about 8x10mm in the right kidney.

Query 2. A 49-year-old woman with a prolapsed mass in the opening of her urethra. Pelvic CT shows a heterogeneously enhanced mass on the female urethra. Pathology shows ramifying papillae, high nuclear/cytoplasmic ratio, and brisk mitotic activity.

Query 3. A 66-year-old man with weakness, severe fatigue, bleeding from the gums, weight loss, and easy bruising. MR imaging shows replacement of the fatty bone marrow. CT scan shows enlarged, hyperplastic lymph nodes in the iliac and para-aortic regions.

As a tool for indexing and retrieval we turn to Terrier IR [14]. Terrier is a flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents [14]. Terrier implements numerous state-of-the-art indexing and retrieval functionalities, and provides an easy platform for the rapid development and evaluation of large-scale retrieval applications. From the medical articles we index all fields i.e. title, abstract, full-text and image captions. In the retrieval stage we use BM25 [15], which appears as one the best performing models in this context [16]. In the modality classification part, we use opponentSIFT [17] as visual features as it reported appropriate for modality classification of medical images [12] and Support Vector Machines (SVMs) [18] for the classification part.

For performance evaluation of the retrieval we use different mathematical statistics: mean average precision (MAP), average of precision at 10 documents retrieved (P10), average of precision at 20 documents retrieved (P20), average precision after R (= number of relevant for topic) documents retrieved (Rprec) [19].

V. Experimental Results and Discussion

The goal of the experiments is to answer the question: *Does including modality classification of the images in the medical articles improve the retrieval results over the standard text retrieval?*

We did two experiments to answer our question. The first one (*text*) is a pure text-based retrieval. In this experiment only the text part of the query is used and articles are retrieved based on the inverted index. The second experiment (*text+mc*) is performed on top of the first one. For each image of each article in the list of retrieved articles, we check in the image-modality index to see whether there is an overlap between the modalities of the images and the ones extracted from the query. If there is an overlap we multiply the initial score of the article (provided by the search engine) by a certain factor. The factor is obtained by parameter search over the 2012 version of the dataset, where the search range was from 1 to 2 with a 0.01 step. Afterwards, we re-rank the articles based on the modified scores and provide the final results. This process is executed for each result list of each query, separately. These two experiments should provide us with an answer to our question. The results from the experiments are shown on Table 1.

TABLE I. RESULTS FROM CASE-BASED RETRIEVAL OVER IMAGECLEF 2013 USING TEXT AND MODALITY DATA

	MAP	Rprec	P10	P20
text	0.2004	0.2101	0.2029	0.1729
text + mc	0.2025	0.2103	0.2086	0.1743

From the results we can note that the there is a slight boost in all metrics in the *text+mc* experiment relative to the *text* experiment. One metric is MAP. The MAP is 0.2004 for the *text* versus 0.2025 for the *text+mc* experiment. This states that our hypothesis is plausible and introducing modality classification of the depicted images in the articles can improve the performance in case-based retrieval. But, there is still room for more substantial performance increases.

VI. CONCLUSION

A new approach to retrieval of medical articles has been presented in this paper. The approach takes into account the modality of the images contained within the articles as well. The results show that such an approach can provide a slight performance boost.

Retrieval of medical cases is an important part of the everyday work of medical professionals. If one can provide more relevant results it can aid the medical process and improve the quality of the work that the medical professionals do. Although, there are systems which offer such services, there is still substantial room for improvement. Our possible future work could go into incorporating different ways of query expansion, so as to define the query more precisely by using different natural language processing techniques and/or by including some on-line medical data stores.

ACKNOWLEDGMENT

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

REFERENCES

[1] I. Ounis, G. Amati, and V. Plachouras, “Terrier information retrieval platform,” Advances in Information Retrieval, pp. 517-519 2005.

[2] J. Kalpathy-Cramer and H. Müller, “Overview of the CLEF 2011 medical image classification and retrieval tasks”, in Proceedings of Cross-Language Evaluation Forum, 2011

[3] A. D. Lindberg, B.L. Humphreys,and A.T. McCray, “The Unified Medical Language System,” Methods of information in medicine, vol.32, no.4, pp. 281, 1993.

[4] [http://pubget.com] last visited: 17.03.2015.

[5] M. Errami, J.D. Wren, J.M. Hicks, and H.R. Garner, “eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications,” Nucleic acids research, vol. 35, no. 2, 2007.

[6] C. Peters, “Cross language evaluation forum,” D-Lib, 2000.

[7] A. H. Vahid, A. Alpkocak, R. G. Hamed, N. M. Ceylan, and O. Ozturkmenoglu, “DEMIR at ImageCLEFMed 2012: Inter-modality and Intra-Modality Integrated Combination Retrieval”, Working Notes of CLEF, 2012.

[8] J. Vanegas, J. Caicedo, J. Camargo, R. Ramos-Pollan, F. Gonzalez, “Bioingenium at ImageCLEF 2012: Textual and Visual Indexing for Medical Images”, Working Notes of CLEF, 2012.

[9] S. Choi, L. Jeongeun, and J. Choi, "SNUMedinfo at ImageCLEF 2013: Medical retrieval task," CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. 2013.

[10] A. Mourao, F. Martins, and J. Magalhaes, "NovaSearch on medical ImageCLEF 2013," CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. 2013.

[11] M. Simpson, et al. "ITI’s participation in the 2013 medical track of ImageCLEF," CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. 2013.

[12] I. Kitanovski, et al. "Medical Image Retrieval Using Multimodal Data." Discovery Science. Springer International Publishing, 2014.

[13] C. Macdonald, P. Vassilis, B. He, C. Lioma, and I. Ounis. "University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming", Accessing Multilingual Information Repositories, pp. 898-907, 2006.

[14] [http://terrier.org], last visited: 17.03.2015.

[15] G. Amati and C.J. Van Rijsbergen, “Probabilistic models of information retrieval based on measuring the divergence from randomness”, ACM Transactions on Information Systems, vol. 20, no. 4, pp. 357-389, 2002

[16] I. Kitanovski, I. Dimitrovski, S. Loskovska, “Fcse at medical tasks of ImageCLEF 2013,” CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. 2013.

[17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision 60 (2), pp. 91-110, 2004

[18] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

[19] C.D. Manning, P. Raghavan and H. Schütze, Introduction to information retrieval (Vol. 1). Cambridge: Cambridge University Press, 2008

Semantic Web Integration with SPARQL Autocomplete

Aleksandar Andreevski, Riste Stojanov, Milos Jovanovik and Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje, Macedonia
andreevski.aleksandar@students.finki.ukim.mk, {riste.stojanov, milos.jovanovik, dimitar.trajanov}@finki.ukim.mk

***Abstract*—This project is intended to ease the writing process of dynamic SPARQL queries for applications. Its goal is to make an autocomplete form that can be reused in different applications and will be up to date with the latest ontologies, thus making the process of using Linked Open Data closer to application developers in general. This is done by having a server with an API that returns a JSONP format for each SPARQL query sent by the user application, i.e. the autocomplete form. The autocomplete feature is implemented with AngularJS and helps the user with writing the SPARQL keywords, the ontology classes and properties.**

***Index Terms*—Semantic Web, SPARQL, AngularJS, Autocomplete**

I. Introduction

The current Web is comprised of text and media files such as pictures, audio and video files. This kind of data representation makes it very easy for humans to understand and absorb the knowledge from that data. However, it is not suitable for the computers, since they see the data structurally without ”understanding” its meaning. Thus, the computers don't understand this data, and they can not fetch, select, filter, combine, aggregate data for us, without someone analyzing that data. This is where the Semantic Web comes in. The Semantic Web represents a system that makes it easier for machines to process and ”understand” web data and also to respond to complex human requests based on their meaning. All that the humans need to provide is relevant information about that data. The Semantic Web is bridging the gap between the human understandable meaning and computer understandable structure with the standards that define representation structure of the data meaning.

The knowledge in the Semantic Web is mostly represented through several formats, such as RDF[1], RDFS[1] and OWL[2]. One can query this knowledge using the SPARQL query language[3]. SPARQL is able to query files containing RDF data regardless if they are exposed on the Web or stored in a local database. Furthermore, SPARQL can query multiple data sources at once and dynamically build a virtual RDF graph from all those sources. It is intended for revealing facts from the semantic data by the applications and expert users that know its syntax.

The base of the SPARQL syntax is similar to the SQL syntax, and their difference is that SPARQL is designed to work with path expressions for graph querying. In the query languages such as SPARQL, the one that writes the query should know the underlying data and its relations. Even though the base of the SPARQL syntax is simple, its main complexity comes from the huge amount of data that is being queried, since each data set can contains millions of triples from different domains. It is hard for a human to remember all the resources from some data set, and it is often practice for the SPARQL users to write several additional queries in order to reveal the underlying data, represented with resources identified by their URIs. In SPARQL queries, the resources must be referenced with their URIs.

In this paper we describe a SPARQL autocomplete system that solves the problem of remembering resource URIs by the users, and significantly speeds up the query writing process. In Section II we describe the SPARQL query building systems with their features. In Section III we describe our SPARQL autocomplete system with the technical and architectural features, and we compare it with similar systems in Section IV. The last section Section V gives the conclusion and describes the future steps for this project.

II. Related Work

There are a few SPARQL query builders that exist. In general they can be split in two groups:

a) Graphical query builders: These query builders lean towards the idea of graphical representation of the resources, variables, type of query. They can be in a form of building a graph [4] where every node is a variable, resource or literal and by choosing the type of query, the system (program) will build the query for you. The other form is building the query based on filling in forms of what the user wants the query to achieve, which would the variables be, selecting the resources from lists, etc. These query builders are wizard-like or have an interface that starts building the query like a tree starting from the query type as a root, and then restricting the possible terms and resources to choose from with each choice made in the branches [5].

One of the main advantages of the graphical query builders is that the person using them does not have to know much about SPARQL, or can just have a basic knowledge of it to construct a query. The user interface is pretty straight-forward and easy to learn for new users. The downside is that in some builders the query takes more time to build due to the conversion from its graphical form to a proper SPARQL query.

[1] http://www.w3.org/RDF/



Fig. 2. *Syntax words autocomplete.*



Fig. 3. *Ontology terms URL autocomplete.*

b) Autocomplete or Textcomplete query builder: These builders are meant for users who have knowledge of SPARQL query writing. They will not create the query for you; instead they will help you writing it. The general idea is to give the user suggestions what to write next or in some builders even check the validity of the current state of the query. Flint[2] and Squebi[3] are text editors for building SPARQL queries that enable this kind of autocomplete and query validation. Although these kind of query builders may not be visually rich as the graphical query builders, they do provide in-depth control of the query and are also faster in terms of query creation time. The query is built by the user and it is validated with every next letter the user types, so the query is ready instantaneously after each new letter. One problem with this kind of builders is that the autocomplete is only for reserved syntax words and maybe some small set of the most popular prefixes, classes and properties. This leaves the user with the problem of finding the URIs (URLs) of the desired resources so he/she can build the desired query. Most of the graphical query builders share this problem too.

Regardless of the type of the query builders, they can be hosted on the Web as part of a web page or they can be a desktop editor or an application. The editor and application types are bound to the user system and the user can use them from there only. The ones hosted on the Web are more flexible in that aspect. They are service-oriented, can be used from everywhere and will always be up-to-date.

III. SPARQL Autocomplete System

The system presented in this paper provides SPARQL query builder that autocompletes the user text with suggestions on what to write next. It also allows query execution against any SPARQL endpoint provided by the user. The SPARQL queries are written in the HTML textarea, which is the main component in the project. The query is automatically completed with SPARQL syntax terms, the classes and properties from the most widely used ontologies, and also supports the most commont ontology prefixes. When a prefix is chosen, it is automatically added in the prefix clause of the query along with its full URI.

[2]http://openuplabs.tso.co.uk/demos/sparqleditor
[3]https://github.com/tkurz/squebi

The SPARQL autocomplete editor is developed as a reusable web component which can be integrated in any web page and can query any SPARQL endpoint. Since there are endpoints that doesn't support the JSONP[4] protocol, and the browsers doesn't support cross site scripting[6], the editor communicates with a dedicated server via the JSONP protocol that provides the autocomplete hints for the users. The architecture of the system is shown in Figure 1.

A. SPARQL Autocomplete Server

The server is responsible for indexing and storing the classes and the properties from the most common SPARQL endpoints. In the current version, the server provides the classes and properties provided by the Linked Open Vocabulary project[5], but it is not limited to them. The server also provides an interface for adding new SPARQL endpoints; when a new endpoint is added, the system queries it and retrieves all the classes and properties available through it. For the prefixes, the server provides the one registered in the Linked Open Vocabulary project, combined with the prefixes defined in the crowd sourced prefix.cc platform[6].

The server has scheduled tasks that re-index the classes, properties and prefixes from the registered sources. In the current version, these scheduled tasks are executed weekly.

The server provides several functonalities. The first one is to serve the script for the autocomplete editor widget. This widget is described in details in Section III-B. The second function is to provide the resources needed for autocompleting the user query. The server supports two modes for communication. In the first mode it transfers JSON objects that are suitable for all clients. However, if the editor is reused in a page with a different domain than the one of the server, this way of communication does not work, since the browsers block the cross side scripting. Because of this, the server also supports communication via JSONP, where the data is sent to a callback provided by the caller.

[4]http://en.wikipedia.org/wiki/JSONP
[5]http://lov.okfn.org
[6]http://prefix.cc/

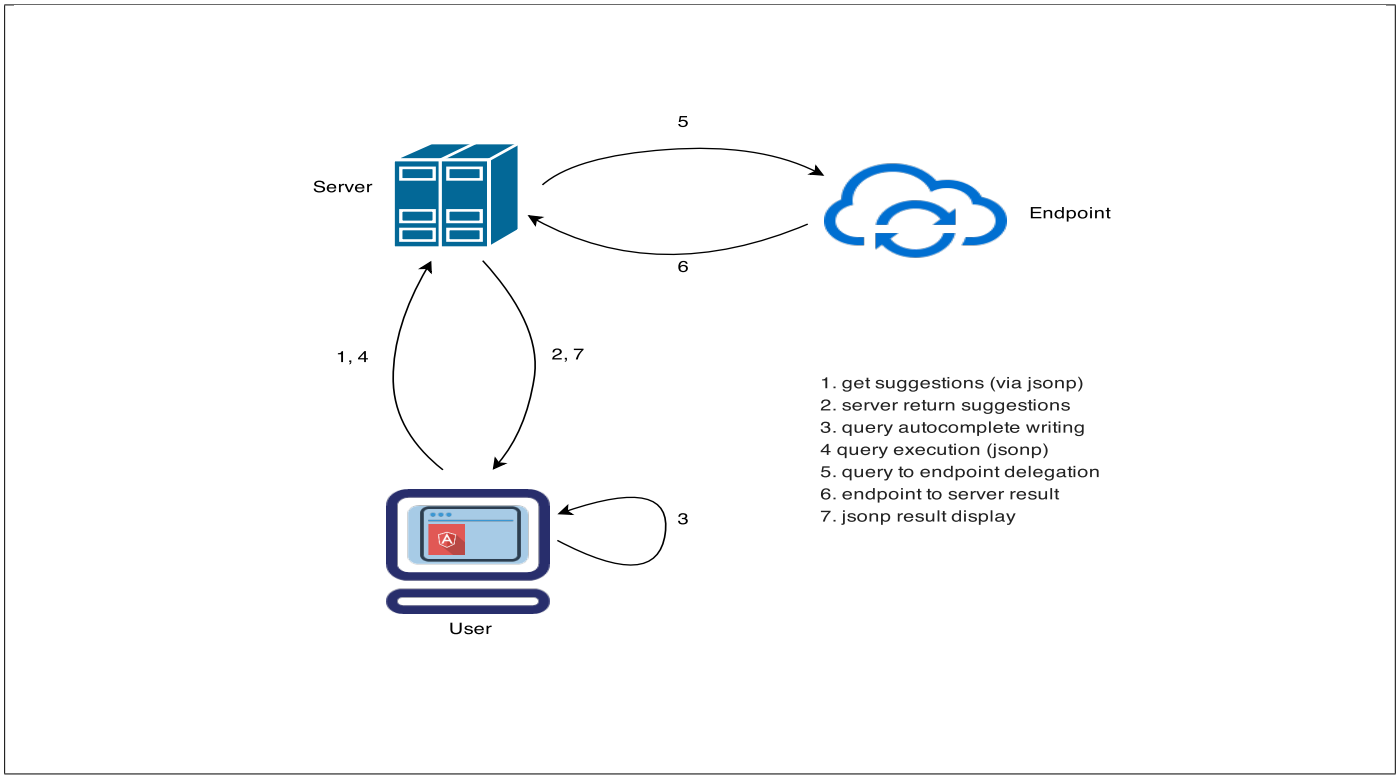


Fig. 1. *System architecture.*

B. SPARQL Autocomplete Editor

The SPARQL autocomplete editor is developed as an Angular JS[7] directive, which can be reused in every web project. The directive provides configuration of the accepted SPARQL endpoints, a default endpoint and the appearance of the component. It can be reused in every web project, and its only dependencies are AngularJS and Text Complete JQuery plugin[7]. This way, each developer can reuse the features of this editor in his/her project. In this mode of functioning, the directive is hosted on a different domain than the autocomplete server, and in order to communicate with eachother, the JSONP protocol is used. This way, the server sends the result as an argument to a callback Javascript function.

Additionally, the system provides an easier way of integration via the widget. This widget is included with an embedded Javascript, hosted by the autocomplete server. This script takes the ID of the target HTML element as an argument, and once it is loaded, it displays the editor at the position of the provided element. The embedded Javascript has all of the dependencies with it, so there is no additional configuration necessary.

The SPARQL autocomplete editor is a custom directive that displays a textarea with auto-completion dropdown list shown while the user types in it. The directive additionally provides list of available endpoints and allows query execution against the selected endpoint.

The autocomplete functionality uses the open source JQuery plugin(Footnote 7) that allows definition of suggested words based on regular expression for matching the text typed in by the user. The component allows definition of a function for narrowing the suggestions, the template for the suggestion list, what to replace after the autocompletion and many more options. There are three matching patterns implemented in the current version:

- SPARQL syntax words
- <resource uri>
- ontology prefix:resource

All of these words are retrieved from the SPARQL autocomplete server and are loaded into the directive. The SPARQL syntax words are hard-coded in an array, because they represent a fixed set of words defined by the SPARQL W3C standard. These words are suggested after a single typed letter, as shown in Figure 2.

The next group of words are shown when the user types the '<' symbol and means that the user can search the ontology terms by their URIs. Once the '<' is typed, a set of URIs is retrieved by the AngularJS service. The service loads these URIs by sending a request to the server who then delegates the response to the AngularJS service. After the response is retrieved by the service, it will be processed and returned as an array to the autocomplete directive. Figure 3 shows the URL autocompletion.

The last word group is the the ontology prefixes group followed by the ontology resources. Autocomplete suggestions are given just after one typed letter. Until the ':' symbol is typed or the user chooses one of the prefixes from the list, the autocomplete shows suggestions only for the ontology prefixes. They are retrieved from http://prefix.cc but this time

[7]http://yuku-t.com/jquery-textcomplete/

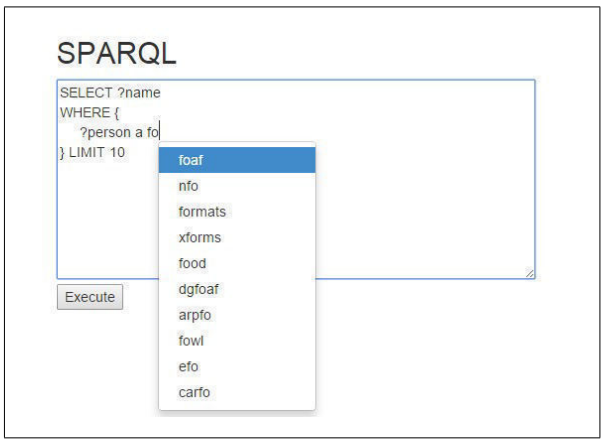


Fig. 4. *Prefixes autocomplete.*



Fig. 5. *Sufixes autocomplete.*

through a direct AJAX[8] call from the AngularJS service itself, not the server. The response contains the most popular ontology prefixes along with their URIs. The prefixes are then sent to the autocomplete script, while the prefix-uri mapping is kept in the service (see Figure 4). After the ':' is being typed, the autocomplete directive tries to match whether the prefix URI is a prefix of some of the URIs we have stored (the URIs retrieved from http://seminant.com/). If there is a match, those matches are shown as suggestions and when the user chooses one of them it is shown in the query as "prefix:resource", while the whole URI is added at the beginning of the query in the prefixe clause (see Figure 5).

IV. Discussion

In the Section III a full description of our autocomplete query builder was given. In this section a comparison is made with the SPARQL query builders and editors from the plain autocomplete category. Let us start with what this editor can not do. Our editor lacks the ability to perform a query validation, while Flint supports this feature by having all the syntax words and the ontology terms categorized so it can apply ordering rules to them. Another drawback is the lack to format the query. This version does not adds tabs or new lines after writing a triple in the query like in Squebi. Coloring of words is also not supported because of the constraints of a HTML textarea and is not intended to support this feature.

Now let us see what is the advantage of our editor over the others. All of the other cited editors have only a few ontologies that they offer as autocomplete suggestions. They are hard-coded in their implementation, while our system fetches them as described previously. The other builders only have autocompletion for the few most common prefixes and some ontology terms URLs. We offer the most used ontologies, about one thousand in size, their prefixes, URLs and terms. That was the general idea for starting this kind of project. This was the common thing all the other SPARQL editors had as a drawback. All of their focus was on either formatting, coloring, visualizing the query or the results. As for our autocomplete system, the goal was to make the autocompletion more powerful, thus easing the query writing process.

V. Conclusion and Future Work

In this paper a review was given of our SPARQL autocomplete system. The idea behind it was described and then elaborated through the project structure and logic. Finally, a comparison was made with the other SPARQL query builders and listed its flaws and advantages over them. After the previous section, it was concluded that our autocomplete system reduces the time to construct a query by not having to search for the ontology terms and resources by ourselves; they are rather provided as suggestions.

In our future work we plan to change the preview of the results. For now we use an AngularJS Table containing links of triplets. We intend to create a separate AngularJS component (directive) to show the result in a graph. The general plan is to have the graph show the resources and to display some metadata about a node on hover. We also plan to support navigation from a node to other nodes (resources) that have an outgoing connection to the one that we navigated from.

Acknowledgement

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje, as part of the research project "Multi-domain Linked Data Integration".

References

[1] D. Brickley and R. V. Guha, "Resource description framework (rdf) schema specification 1.0: W3c candidate recommendation 27 march 2000," 2000.
[2] D. L. McGuinness, F. Van Harmelen *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.
[3] S. Harris and A. Seaborne, "Sparql 1.1 query language," *W3C Recommendation*, vol. 21, 2013.
[4] A. Russell and P. Smart, "Nitelight: A graphical editor for sparql queries," 2008.
[5] O. Ambrus, K. Möller, and S. Handschuh, "Konduit vqb: a visual query builder for sparql on the social semantic desktop," in *Workshop on Visual Interfaces to the Social and Semantic Web*, 2010.
[6] K. Spett, "Cross-site scripting," *SPI Labs*, pp. 1–20, 2005.
[7] P. B. Darwin and P. Kozlowski, *AngularJS web application development*. Packt Publ., 2013.
[8] J. J. Garrett *et al.*, "Ajax: A new approach to web applications," 2005.

Making impact with data in a University setting – the case of SEE-University

Blerta Abazi Chaushi
Faculty of Business and Economics
South East European University
Tetovo, Macedonia
b.abazi@seeu.edu.mk

Agron Chaushi
Faculty of Business and Economics
South East European University
Tetovo, Macedonia
a.caushi@seeu.edu.mk

Zamir Dika
Faculty of Contemporary Sciences
South East European University
Tetovo, Macedonia
z.dika@seeu.edu.mk

***Abstract* — The importance of data collection in higher education institutions is well documented in literature. South East European University (SEEU) gives a high importance to data collection and analytics. This paper uses the Bakers Model as a theoretical background to explain the use of data analytics in the context of SEEU. First information systems that produce data are explained. Second, the data collected from the Quality Assurance Office at SEEU is explained. Third for each of the elements of the Baker's Model examples from SEEU are given. The contribution of this research is to showcase the importance of data for decision making by giving examples from real world and put in a context of a theoretical model.**

Keywords: data management, data mining, quality, higher education information systems

I. INTRODUCTION

Institutions of Higher Education (HE) collect data for many reasons, weather for institutional planning, strategy formation, or just because they are required by the ministry of education in yearly basis. One thing is evident; a lot of data from many resources are collected.

South East European University (SEEU) places a lot of importance to data collection and analysis. Yearly, more than 40 standard reports are produced by the Student Services Department (SSD) sorted in 9 different categories out of which the ones that are most extensively used are the operational reports, faculty operations data, reports and analysis for the executive leadership and decision making, and institutional reports required by law. Moreover, the Office of Quality Assurance and Management (OQAM) at SEEU collects data, conducts analysis, and disseminates information about students' feedback for their experience with the University services, with faculty staff and specific courses with the use of yearly surveys that each student is supposed to fill out for each course. Just during the period of 2009-2012, around 19510 surveys were filled by students, which shows the importance this office and SEEU gives to improvement of services, retention of students, increase of satisfaction, as well as progress of the staff in teaching methods and skills. This office also conducts teaching observations, both formal and informal for each academic staff in yearly basis. It is an important tool for the academic feedback for each faculty member and the continuation of the work in this institutions also relies in the progress a staff member has achieved through the observations as well as the rating he/she gets from the students trough the surveys.

This paper contributes to the literature by showing the use of data in a university setting, the impact it has on many business processes important to an institution of HE, as well as a roadmap that other universities, especially in Macedonia can follow. The limitation of this paper is that only SEEU has been analyzed from the universities in the country, and at this point no data is available in the form that this paper requires for the purpose of comparison and benchmark.

The importance of data collection and analytics at SEEU is taken as a case study to contribute and support other universities to enhance the purpose of data already collected in the systems but never used, and as an idea to progress in terms of increasing and managing the quality of services, staff and student satisfaction in their institutions. This paper is organized in three sections. In section one, we look at the theoretical background on how data can be used in universities for decision making. Particularly, Bakers' model for data analytics is described. In section two, SEEU is given as a case study. It is analyzed from three main perspectives. First, information systems at SEEU are described as the main sources of data. Second, the importance of reports produced from the SSD, OQAM, as well as OLAP tools for many university services, management, student and staff members. Third, Baker's model is used to explain SEEU's use of data analytics.

II. WHAT CAN UNIVERSITIES DO WITH DATA – THEORETICAL BACKGROUND

As depicted earlier, data is collected from many different sources and for many different reasons. In the past, the work of University plotters and decision makers in a lot of countries mainly has been concentrated on issues concerned with

forecasting numbers of students, faculty members, and support staff, as well as predicting the demand for, and location of, the buildings and equipment required by institutions' systems at any one point in time [1] However, as research suggests, many universities do not make decisions about the quality of education, and even more, did not have clear formats of data visualization for various levels at which decisions are made. Empirical studies of data-based decision making have consistently found data-driven decision making is a necessary factor for successful institution. Management in Universities that were able to efficiently use data for analysis and decision making were more devoted to use and built a strong vision for data as an impact factor in a University setting [2]–[8]. A theoretical framework, adapted from the literature, acknowledges that various data categories (input, outcome, process, and satisfaction data) can enlighten the decisions in university scenery, but that the presence of raw data does not ensure its effective use.[9]. Moreover, the study from Marsh et al, implies that data-driven decision making does not guarantee effective decision making since just having the data does not mean that it will be used appropriately or lead to improvements. Suggestion that equal consideration needs to be given to both analysis of the data and taking actions based on the data is implied since there is a need for more research on the outcomes of data-driven decision making on improvement of instruction, student achievement, and other outcomes, such as staff qualifications, trainings needed and academic publication. [10]. Baker in his study acknowledges that data has been used by universities for quite some time now, but he states that educational data mining, which is centered around the development of methods for making discoveries within unique kinds of data, is yet to be developed [11]. As depicted in literature, and explained by Baker, there are five different categories of mining the data in a University setting: prediction, clustering, relationship awareness, discovery, and refinement.

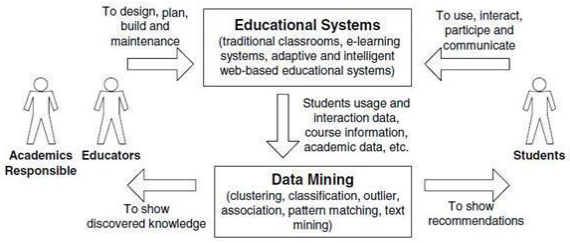


Figure 1: Applying data analysis in educational institutions [12]

As figure 1 depicts, the key people in the process of designing, planning, building and maintaining the educational system are the educator, i.e. professors/teachers and academics responsible, i.e. the management of the University. Students use and interact with the system. So the knowledge that is created and discovered through the five different categories of data mining and analysis can be oriented toward the students, academic staff and management staff.

A. *Data Prediction*

Prediction is the use of data to develop models which can conclude a single aspect of data from combination of other aspects of the data. An example would be predicting student outcomes, enrollment for next academic year, retention rate etc. Students may need help to know which course is best for them based on prediction of how well they will perform in the courses they select. Prediction can be used to enable universities to better allocate resources and staff and proactively manage student outcomes [11], [12]. There are available tools and methods that can be used by academic staff to predict the final grades of the students such as decision tree, induction analysis and genetic algorithms. Ueno's work is taken as an illustration, where he proposes a method of on-line outlier detection by using a Bayesian predictive distribution of learners irregular learning process that can be applied effectively in a e-learning classroom [13]. Other examples of the use of data prediction are the prediction and detection of atypical behavior [14], detection of student misuse of tutoring systems [15], finding sources of error in predicting student's knowledge [16], as well as detection of regularities of deviations in the actions of both the educator and the student in order to improve the learning and teaching process [17].

B. *Data Clustering*

Clustering the data means to group data points that naturally belong together making it useful to create categories of data. Mining student data to cluster for discovery of patterns of user behavior is just one use of this category. The use of clustering model in the case of characterizing groups of online students is conducted by Castro et al by using a topographic mapping model [14]. Data clustering can be very useful in the adaptive and intelligent web based educational systems, content management systems, allocations of resources and staff, management of alumni etc. Luan, in his research proposes the use of both unsupervised and supervised data mining algorithms to do clustering and prediction to enable educational institutions to reach efficacy and effectiveness in university processes [18].

C. *Relationship awareness*

Relationship awareness means discovering the relationship between variables in a data set with a large number of variables. If-then rules, finding linear correlations, temporal association and if one event is the cause of another event are forms that relationship awareness takes. [11]. Ilustration of the use of relationship would be finding a rule that if a student is frustrated with a subject, he will work more and thus ask more questions; determining the path of the students learning that leads to learning event; concluding a casual relationship if an academic event frequently leads to positive learning outcome etc. Romero et al propose a genetic programming with multi-objective optimization techniques to provide a feedback by discovering relationships from students usage information [19].

D. *Data discovery*

Data discovery usually start with a model of a phenomenon that is developed with the use of one or all the

above mentioned three categories: prediction, clustering and/or relationship awareness. A demonstration of data discovery would be the useful knowledge that helps improve the learning process of the student, establishing basis for decision making when designing or modifying a course or teaching approach, as well as manipulating or gaming the system by analyzing complex structures. The last is supported by the study of Walonoski and Heffernan for the detection and analysis of off-task behavior, where students try to manipulate the tutor behavior in order to advance through a curriculum quickly and easily [20].

E. Data Refinement

Data refinement, for the purpose of this paper, is the cleansing of the data for the purpose of human judgment. This is very important, especially in data-driven decision making since a lot of inference relies on the ways that the data are represented. In here, different visualization methods are exploited for the purpose of identification of important data and classification of the same. Mazza and Milani, in their study developed a tool they called GIZMO (graphical interactive student monitoring system) that implement visualization techniques which academic staff finds to be very useful, especially since it allows them to manipulate for the purpose of examining social, cognitive and behavioral aspects of the students in their online learning system [21].

III. THE CASE OF SOUTH EAST EUROPEAN UNIVERSITY

South East European University (SEEU) since its foundation has recognized the importance of information systems and the importance of data collection and its use in decision making. Yearly, more than 40 standard reports are produced by the Student Services Department (SSD). Moreover, the Office of Quality Assurance and Management at SEEU collects data, conducts analysis, and disseminates information about students' feedback for their experience with the University services, with faculty staff and specific courses with the use of yearly surveys that each student is supposed to fill out for each course. To better understand the nature of the data that is collected and analyzed, the case of SEEU is presented at three different levels. First, the types of information systems that SEEU has developed. Second, the types of reports and surveys/questionnaires and trainings it conducts. And last, placement of the data mining in SEEU in the model developed by Baker, which was analyzed in the previous section, is conducted.

A. Information systems in SEE-University

SEEU has invested in different systems such as: Student Information Systems (SIS), Learning Management Systems (LMS), Human Resource Management Systems (HRMS), Accounting Information Systems, Finance System, Library Services System, Research Management System, Electronic Staff Attendance System, Electronic Student Attendance System, as well as various reporting tools for the SEEU middle and senior management. Figure below shows the schematics of an information system, and it can be derived that all the data collected, in fact is utilized for decision making and control.

Next, the most important information systems in use by SEEU will be described as well as how the data collected by these systems are utilized.

1) Student Information Systems (SIS)

The SIS of SEEU is called University Management System (UMS) and its one of its earliest and major investment in information systems. UMS, collects various data regarding to students, courses, facilities, as well as academic staff. For students the data starts to be collected as early as student applications, continuing with data about admissions, registrations, student activities (course enrollments, grades, bursary, curriculum mapping, etc.), graduation, and finishing with alumni and employment data. Regarding courses, the process starts with accreditation of the study programs. Once a study program is accredited, the process of curriculum and course registration takes place in UMS. When a certain curriculum starts to be used, courses become part of the course offerings and they take place in the process of academic planning, later on in in scheduling as well as grading. Data about classrooms is also collected in UMS. These data are used during the process of academic planning, scheduling as well as event management. Data about lecturers is fed from HRMS and this data is used in academic planning, scheduling, and grading.

2) Learning Management Systems (LMS)

SEEU has its own LMS called Libri, which serves the needs of lecturers and students. Libri has the basic capabilities of an LMS with some analytic tools for lecturers as last login for each student, the number of logins for students, download count for the materials, etc. On the administrative side, Libri has some analytics tools for usage of Libri by course instructors and students. The data for lecturers and students is fed from UMS whereas course offerings are manually added to Libri. The same applies for registration of courses in Libri by students.

3) Library Services System

SEEU uses information system for its library as well. The system has data about the books available on library, and data about rental history of its books. The system, through its web interface available to SEEU members ability to search and reserve books, as well as access to online databases such as EBSCO Host. The data collected from this system is used for predicting how many copies of each book the library should buy, which are the most wanted books, which books are not needed. This data is used from the library for predictive analysis of future purchases.

4) Human Resource Management Systems (HRMS), Accounting Information System (AIS), and Finance System (FS)

For HRMS, AIS, and FS, SEEU uses a commercial ERP which is integrated with other systems including UMS. The importance of this ERP, apart from accounting and finance, lies on data about lecturers such as their qualifications, academic titles, remuneration policy, electronic work and class attendance, etc. Through this data HR department as well as

management can track academic performance of its faculty which is one the main resources of the university. Specifically, some examples of utilization of the HRMS are: checking if lecturers start and end the classes on time, working hours of staff members, etc. which combined with the data from Quality Assurance Office, and Research Management Systems give a good feedback on the overall academic performance of faculty members.

5) Research Management System (RMS)

As part of UMS, SEEU has created a RMS called MyResearch. The data in the system is entered from the academic staff itself which is later reviewed and valued by the research committee of SEEU with points. These points are used by the academic staff for academic promotion. The same data is used to measure the overall research progress at university level and as a research database of SEEU staff members.

6) Electronic Student Attendance System

One of the recent systems at SEEU is the Electronic Student Attendance System. The development of this system was partly motivated by the Law for Higher Education in Republic of Macedonia. The system tracks student class attendance for each course in which a student is enrolled. This system is becoming very important for few reasons:

- It enables university to track student attendance at central level,
- It enables university to enforce its attendance policy at system level,
- It enables data analytics regarding student attendance and other variables (time of the day, relationship between attendance and grades, etc.)

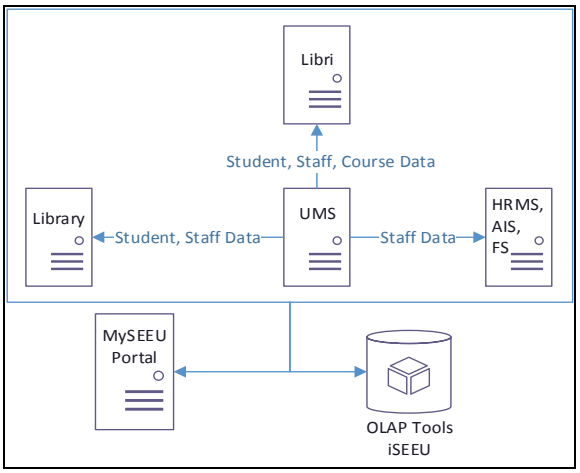


Figure 2: SEEU Information Systems

B. Reports and tools to collect and analyze data in SEEU

Since the data is collected and represented by two main bodies in SEEU, the following section conducst analysis of the Office of QAM, and SSD.

1) Data used in OQAM-SEEU

The Office of Quality Assurance and Management (OQAM) in SEEU use the following data.

a) Teaching Observation

Teaching Observation - reports are fed back individually to teachers, the Faculty has the data and twice a year the Quality Office collates the data, analyses it and sends a report to academic staff and students. This shows comparative data about teaching, learning, management, resources, progress as well as sharing general good practice and points for improvement. OQAD has also regularly used this information for general training sessions. Where data about a staff member were unsatisfactory, this has been followed up with additional support and another observation and has resulted in good improvement measured by the grades. The unannounced observations are seen as being more real and as a way of seeing what faculty staff members are actually doing. However, the data actually shows that the judgments overall are higher than the announced process, plus not everyone is seen and some are seen several times. The individual report must be put in every member of Staff Evaluation report every year up to now and is part of the Promotion in Title. An example of a report from teaching observation in University level is given in the figure 3.

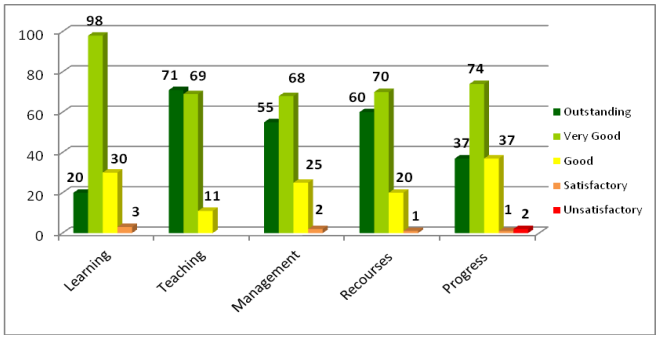


Figure 3: Announced Observation University Profile 2013-14

b) Student Evaluation

The Quality Office coordinates the annual Student Evaluation Surveys in liaison with Student Advisors and Student Services. The Academic survey is paper based, conducted in the second semester and done anonymously by administrators. This is based on every professor having one evaluation and possibly two if also working in different cycles or campuses. The results are analyzed. Every professor receives her/'his own result. The Faculty Deans receive the results and the data is analyzed at university level. The individual data must be considered as part of individual annual Staff Evaluation and the Faculty should consider their own data. There is also a Survey of Admin Departments done electronically. The results are analyzed by every Director/Head and the senior management and they are required to put improvements in their next Action Plan as well as discussing the results and actions with their teams. OQAM did the Teacher Assessment Poll (TAP) in BE at the end of the first semester, which is a course based evaluation using out of faculty facilitators. Students in every course were asked what helped learning, what didn't, quality of assessment, other comments. A report was given anonymously and immediately to individual staff.

The results were collated by the Q Office, given to the Dean's Office and sent by them to all BE staff and students. OQAM plans to conducts TAP in other faculties in future, and thus would give two different methods. Figure below represents a distribution of student evaluation surveys in the different faculties.

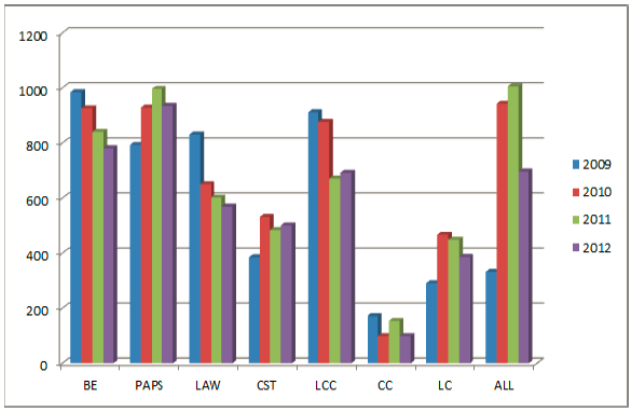


Figure 4: Example of distribution of student evaluation survey in different faculties

c) Training

OQAM sends reports to staff members about the value of the two training sessions for each year. This office also analyzed and reported on the value of Student Induction.

d) Other Surveys

OQAM also conducts other surveys - Client Audit Survey (staff evaluating each other's Office/Dept); Communication Service Standard Survey. The Quality Office and senior management are supporting every department in developing student and client evaluation methods e.g. Helpdesk; e-response, procurement evaluation end user survey, comments box and they follow this up with a selective, departmental audit. OQAM is also contributes to the ISO internationalization process and they are validated externally every year – this is a standard they have retained.

2) Reporting and Data Analytics Tools

There are two sources of data used for creating reports. The first source is the information systems of the university explained earlier. The second source is the data collected from outside of the information systems such as OQAM, also explained earlier. It is important to mention that SEEU does not input the data collected externally in the information systems of SEEU. These data are analyzed with external tools (statistical software). From the data that resides inside the information systems of SEEU, a data warehouse is built for easier data analytics using OLAP tools built in a specialized portal for the university management called iSEEU. The standard reports queried directly from the information systems are mainly produced by the Student Services Department (SSD) in SEEU.

C. SEEU in Bakers' model

1) Data prediction

For data prediction, SEEU uses the SSD data processing reports. These reports are divided into four major categories: 1. Student enrollment progress – for each academic term (fall & spring) and for each of the three cycles (undergraduate, master, and doctoral cycle); 2. Regular exam session progress – for each academic term; 3. Make-up exam session progress – for each of the sessions during an academic year; and 4. Retention progress – for each faculty, in each level of study, in each academic term. These data are used for planning staff and allocating recourses – such as classrooms, equipment needed, optimization of class size etc. Also, these data are used for marketing activities, especially increasing the retention rate amongst students.

2) Data clustering

Data clustering in SEEU, initiated by the SSD, namely the faculty operation data they produce, initially had the purpose of providing information and feedback to Deans' offices in all five faculties. However, now they are extensively used for planning purposes. Example would be reports under the operational reports group, such as student finances where clusters are formed for students with different financial load – that being said, students paying full tuition, students with merit based scholarship, students with scholarships, students with work and study program etc. Moreover, clustering is being applied in one of the faculties of SEEU by applying clustering algorithms with the purpose of generating groups based on some similarity measures. Example is the identification of lecturers with common problems across faculties. If identified, the university devises specialized training sessions that target specific groups osf users across faculties.

3) Relationship awareness

Examples of this type of mining include identifying courses where high failure rate also impacts other courses, identifying relationships between attendance and grades, identifying relationships between high schools GPA and SEEU GPA, examining the impact of scholarship in student performance, identifying courses/lecturers with high grade inflation etc.

4) Data discovery

SEEU tries to perform data discovery with many data they gather. One example here is the discovering of the trends in the use of the LMS and student grade depending on their use of the system and the impact on teaching and learning. The reports generated by e-learning center at SEEU show that there is a positive correlation between the extensive use of LMS and student performance.

5) Data refinement

Although very information rich, SEEU still need to work a lot in defining data analysis techniques for better data refinement. A study by two members of Faculty of Contemporary Sciences in SEEU [22] for improving the quality of service in SEEU through advanced data analysis and visualization, has laid the grounds for better data refinement for the purpose of identifying data hidden in the dataset which might present valuable information for the management.

IV. CONCLUSION

In summary, this study is an attempt to link theory with a real world example in data for decision making. This study

gives a theoretical background about the impact of data in university setting, especially data-driven decision making. It expands the Baker's model for data mining by adding examples from literature, which is not done so far. In this study, a real case is taken, the South East European University, which shows what systems are in place to collect and generate data and how they contribute in decision making. Another contribution of this study is that it shows a real example of how a university is trying to manage the data by analyzing it from the different offices/departments. Additionally, this study may serve to see how much universities, not only in the region, but anywhere in the world, are using analytics and mining by placing their reports and data collection and analysis techniques in the Baker's model, as it is done in this study with SEEU with examples for each of the categories.

ACKNOWLEDGMENT

We would like to express our special thanks of gratitude to the head of OQAM and Executive Advisor to the Rector, Mrs. Heather Henshaw (h.henshaw@seeu.edu.mk) for her genuine support in providing data about the OQAM in SEEU. Also, We would like to thank the Director of Student Services Department, Mr. Abaz Selmani (a.selmani@seeu.edu.mk), for providing us with the data that this department produces, without which this study would be incomplete.

REFERENCES

[1] K. N. Ross, L. Mählck, and Al-nouri Abdel Ghani, *Planning the quality of education: the collection and use of data for informed decision-making*. Unesco, 1990.

[2] M. A. Lachat and S. Smith, "Practices that support data use in urban high schools," *J. Educ. Stud. Placed Risk*, vol. 10, no. 3, pp. 333–349, 2005.

[3] T. Mieles, E. Foley, and others, "From data to decisions: Lessons from school districts using data warehousing," *Provid. RI Annenberg Inst. Sch. Reform Brown Univ.*, 2005.

[4] J. Choppin, "Data use in practice: Examples from the school level," in *annual meeting of the American Educational Research Association, New Orleans, LA*, 2002.

[5] J. Feldman and R. Tung, "Whole School Reform: How Schools Use the Data-Based Inquiry and Decision Making Process.," 2001.

[6] J. Feldman and R. Tung, "Using Data-Based Inquiry and Decision Making To Improve Instruction.," *ERS Spectr.*, vol. 19, no. 3, pp. 10–19, 2001.

[7] J. L. Herman and B. Gribbons, *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles, 2001.

[8] S. Mason, *Turning data into knowledge: Lessons from six Milwaukee public schools*. Madison, WI: Wisconsin Center for Education Research, 2002.

[9] J. A. Marsh, J. F. Pane, and L. S. Hamilton, "Making Sense of Data-Driven Decision Making in Education," 2006. .

[10] K. A. Kerr, J. A. Marsh, G. S. Ikemoto, H. Darilek, and H. Barney, "Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts," *Am. J. Educ.*, vol. 112, no. 4, pp. 496–520, 2006.

[11] R. R. Baker, "Data mining for education," *Int. Encycl. Educ.*, vol. 7, pp. 112–118, 2010.

[12] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[13] M. Ueno, "Online Outlier Detection System for Learning Time Data in E-Learning and Its Evaluation.," in *CATE*, 2004, pp. 248–253.

[14] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of teaching and learning paradigms in intelligent environment*, Springer, 2007, pp. 183–221.

[15] R. S. Baker, A. T. Corbett, and K. R. Koedinger, "Detecting student misuse of intelligent tutoring systems," in *Intelligent tutoring systems*, 2004, pp. 531–540.

[16] M. Feng, N. T. Heffernan, and K. R. Koedinger, "Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required," in *Intelligent Tutoring Systems*, 2006, pp. 31–40.

[17] M. Muehlenbrock, "Automatic action analysis in an interactive learning environment," in *The 12th international conference on artificial intelligence in education, AIED*, 2005, pp. 73–80.

[18] J. Luan, "Data Mining and Knowledge Management in Higher Education-Potential Applications.," 2002.

[19] C. Romero, S. Ventura, and P. De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors," *User Model. User-Adapt. Interact.*, vol. 14, no. 5, pp. 425–464, 2004.

[20] J. A. Walonoski and N. T. Heffernan, "Detection and analysis of off-task gaming behavior in intelligent tutoring systems," in *Intelligent Tutoring Systems*, 2006, pp. 382–391.

[21] R. Mazza and C. Milani, "Exploring usage analysis in learning systems: Gaining insights from visualisations," in *AIED'05 workshop on Usage analysis in learning systems*, 2005, pp. 65–72.

[22] A. Besimi and V. Shehu, "Improving Quality of Service at Seeu Through Advanced Data Analysis and Visualization," *SEEU Rev.*, vol. 9, no. 2, pp. 9–25, Nov. 2014.

New online applications and facilities at IFIN-HH and ELI-NP

Laurentiu Serban and Alexandru I. Nicolin
Department of Computational Physics and Information Technologies,
Horia Hulubei National Institute for Physics and Nuclear Engineering, Reactorului 30, Magurele, Romania
laurentiu@nipne.ro, alexandru.nicolin@nipne.ro

Abstract—**We discuss in this brief report the new version of the Publications Software used by the research personnel of the Horia Hulubei National Institute for Physics and Nuclear Engineering (IFIN-HH) and the Extreme Light Infrastructure – Nuclear Physics (ELI-NP) experimental facility (currently in construction) for the annual evaluation process and also for keeping track of all the scientific output, for institutional reports. Other applications, developed in-house or installed and customized as Open Source software products, are used for File Management, organizing Scientific Events, schedule Meeting Rooms and for Public Acquisitions & Audit.**

Keywords— Assessment of scientific results; Open Source solutions

I. Introduction

The Horia Hulubei National Institute of Physics and Nuclear Engineering (IFIN-HH) is the largest Romanian research institute both in terms of scientific personnel and scientific publications. With a history going back some 65 years, the institute has hosted a series of landmarking achievements of which we mention the development of the first Romanian computer (1955), the first Romanian laser and the Extreme Light Infrastructure – Nuclear Physics (ELI-NP, see http://www.eli-np.ro) which will provide "magnificent new opportunities to study the fundamental processes unfolded during light-matter interaction" using the most intense lasers world-wide and will foster an unprecedented interdisciplinary research plan which addresses "frontier fundamental physics, new nuclear physics and astrophysics as well as applications in nuclear materials, radioactive waste management, material science and life sciences" (see [1] for a detailed discussion). ELI-NP is currently under construction within the premisses of IFIN-HH in Magurele, in the outskirts of Bucharest, and is scheduled to go into operation in 2018 as the nuclear pillar of the larger European research infrastructure ELI (Extreme Light Infrastructure), whose other two pillars are located in Prague, The Czech Republic, namely ELI Beamlines (see http://www.eli-beams.eu), and Szeged, Hungary, namely ELI Attosecond (see http://www.eli-hu.hu).

In this article we discuss the e-services solutions currently in place at IFIN-HH and ELI-NP used for the individual and institutional evaluations, as well as other solutions dedicated to the managament of scientific events, file management, and public acquisitions and audit. The structure of the paper is the following: Section II consists of three subsections dedicated to the *Management of scientific events*, the *Assessment of scientific events*, and *Other services*, while Section III presents our concluding remarks as well as a series of directions of future development.

II. E-Services

A. Management of scientific evens

Indico (see Ref. [2] for the general User Guide) is a well-known web application developed by CERN commonly used for the detailed planning of scientific events, both small and large. Following 10 years of sustained development, the application now offers a complete solution for the management of scientific events, which includes the organizational workflow, an efficient system for room booking, an upload, retrieval, and archival system for all types of documents (and their associated metadata), as well as paper reviewing functionalities. Technical specifications aside, Indico's success in the scientific community is also due to its open source policy (under the terms of the GNU General Public License) which makes it very convenient for all interested organizations to develop personalized solutions tailored to their needs. The Indico implementation currently in place at IFIN-HH and ELI-NP allows the precise management of "Conferences", "General Seminars", "Department Seminars", "Workshops" and "Other events" and has been used for the conferences of the Romanian Tier-2 Federation (which is a member of the Worldwide LHC Computing Grid), those of the Extreme Light Infrastructure – Nuclear Physics (ELI-NP) project, etc. During 2014 the local Indico implementation has been used to organize a total of 20 scientific gatherings which sum up to more than 500 participants.



Fig 1. A typical Indico screen-shot of the General Seminars hosted by IFIN-HH. There is a general web page for all seminars and each seminar has its own individual sub-web page.

B. Assessment of scientific output

The assessment of the scientific output is one of the main recurrent administrative tasks used both for individual and institutional evaluations. Financial elements aside, a correct assessment of the scientific output can uncover emerging and fading research directions, the visibility and impact of the scientists in a given group or department, as well as possible scientific collaborations between the scientists in a given administrative unit, or between scientists from different institutions. In light of its importance, the system developed by IFIN-HH (and now used also by ELI-NP) to keep track of the scientific output, peer recognition and research projects of each scientist has been tailored to allow automatic assessment reports. The first version of the system was implemented in 2005, but the first self-assessment electronic form was introduced in 2013 and functioned in parallel with the classical printed evaluation forms, while the second one was implemented in 2014. Our database covers a wide set of indicators (such as: articles, books and book chapter patents, conference proceedings, invited lectures are national and international conferences, peer-reviews for scientific journals, membership in editorial boards, research projects, PhD coordination, awards, etc.) for which it gathers all relevant metadata. The structure of the database is such that we can generate automatic assessment reports both for the annual internal individual evaluation and for the national funding agencies which are interested in aggregated data which emphasize a particular aspect. The core of the database is represented by the scientific publications for which the most important metadata are the full title of the paper, the names and number of authors (with emphasis on the order of the authors and the name of the corresponding author), the full name of the journal and the associated scientometric factors (say, the impact factor and the eigenfactor) and the numbers of citations (if any). To reduce the workload of the scientists, our e-platform allows for automatic selection of the name of the journal and the associated scientometric factors, but does not import automatically the records from the Web of Science database of Thomson Reuters. Moreover, the platform allows the automatic import of .bib files with all publications relevant for a given scientist, which reduces dramatically the workload. Our experience with the scientific output shows that the name of the institute appears in various erroneous formats, an aspect that precludes the automatic import from Web of Science, and that the projects used to finance the published work are not always acknowledged explicitly. The IFIN-HH database on scientific output can be queried automatically using virtually all recorded metadata as filters and it is used periodically for the global evaluation of the institute. During 2014 the application has been used to organize a total of 3500 scientific outputs (such as articles, books, conference proceedings, patents, etc.) for more than 400 participants. The logs show that the registration of each scientific publication took an average of five minutes (including here the citing papers), and an overall 10 minutes for the other recordings (referee works, memberships in editorial boards, etc.). Among the observed faults of the platforms we note the so-called "rapid writing problem", which is due to the fact that a very rapid fill-in of a specific field on the platform effectively blocks the auto-complete function which does not have enough time to compute the suggestion. This is particularly problematic for the names of the journals where a database of 10000 scientific journals is inspected by the auto-complete function. Another reported problem concerns the errors in adding to the database scientific publications in journals with unusual names. To give just one example the articles published in "The European Physical Journal Plus" were not correctly processed, as "Plus" was also identified with "+". Let us mention that (since 2014) our assessment platform uses the same list of journals as Thomson Reuters, which allows us to efficiently compare the scientific output of IFIN-HH with that of similar universities and research institutes world-wide.

On the technical side, let us mention that the web page through which the information is recorded into the database uses the Ajax (a name derived from (*Asynchronous JavaScript and XML*, see Ref. [3] for a detailed discussion) programming technique that allows for increased interactivity with the web page. Ajax consists of a series of web development techniques used on the client side to create asynchronous web application. The asynchronicity allows the client to send or receive some data from a server without interfering in any way with the page already loaded by the browser of the client. In other words, Ajax effectively avoids the full-page reloads and creates a very pleasant working environment. The drawback of Ajax is two fold: on one hand (which concerns the local development team) the code is quite complex and therefore hard to maintain and debug, while on the other hand (which concerns the users' experience with the web page) the history of an Ajax-based web page does not keeps track of all instances of a given page, which makes it difficult to the user both to return to a specific instance and to bookmark it for future reference. Moreover, most web crawlers do not execute JavaScript code, so without distinct means of accessing the content of an Ajax-based web page it is left largely unindexed by search engines. Please note that the browsers should have the "Javascript enable" option checked in order to have the Ajax scripts running. We also note that the hardware requirements of the platform are minimal and it can run with

difficulty on a standard multi-core computer, while on the software side the main requirement is PhP 5.3.3. We also note that the future versions of the platform are strongly dependent on the assessment criteria (set by the Romanian Ministry of Education and Research) used to evaluate IFIN-HH as well as the scientific criteria used to evaluate PhD students and the minimal criteria for Associate and Full Professorships. In this respect, the development of the platform is motivated by a set of user requirements, which change as often as the national scientific assessment regulations do. As timing is always tight, the system development life cycle is very particular, with some phases (say testing) being only superficially passed-through. Due to this tight timing the current version of the system, for instance, does not import automatically the scientometric factors from Thomson Reuters database which are modified by hand every year. Finally, the monitoring system records a series of information concerning the browser, the operating system and the time each user stays logged on the platform and also includes an email-based online support system through which the users can draw the attention of the system administrator to the faults of the system.

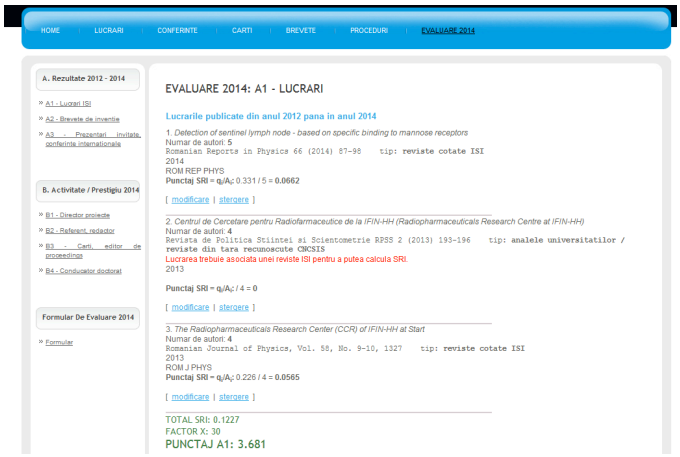


Fig. 2 A typical screen-shot from the IFIN-HH web application which gathers the scientific output of the research personnel.

C. Other services

Other open-source solutions developed at IFIN-HH include the Meeting Room Booking System developed in 2014 which can keep track of the rooms designed for meetings or scientific gatherings situated in distinct buildings of our institute, the Public Acquisition Platform developed in 2009, improved annually with new modules which covers more than 7000 acquisitions, and two ownCloud-based data storage and management solutions which amount tens of TB. The Public Acquisition Platform is a good example of an application which has IP-restricted access (for the Administration Module which is accessible only within the premises of IFIN-HH), while the ownCloud-bases data storage are good examples of institutional cloud systems which allow file versions, various levels of undelete, sharing large amounts of data among different users, etc. Moreover the early versions of ownCloud raised a series of nontrivial questions concerning the security of data which was stored on the disks of the servers in an un-encrypted form. This, in turn, allows the straightforward extraction of the data (files, folders) from the disks and effectively eliminates any type of data privacy.

III. Conclusions and Outlook

In this brief report we have surveyed the new version of the Publications Software used by the research personnel of IFIN-HH used for the annual individual and institutional evaluations, as well as our other open-source used for File Management, organizing Scientific Events, schedule Meeting Rooms. The main directions of future development depend to a large extent on the requirements from the Ministry of Education and Research, but a few directions stand out: the integration of the Audit our current applications and the automatic import of data from the Web of Knowledge database of Thomson Reuters such that the scientists of IFIN-HH only validate (and, where needed, correct) their scientific production without actually spending time on filling-in the online forms.

Acknowledgment

For this work LS was supported through POSCCE ID 1334 (ELI-NP, Code: SMIS-CSNR40741), Contract Nr. 425/12-XII-2012, while AIN was suppoted through PN 09370108/2015.

References

[1] D. Habs, T. Tajima, and N. V. Zamfir, “Extreme Light Infrastructure – Nuclear Physics (ELI-NP): New Horizons for Photon Physics in Europe,” Nuclear Physics News, vol. 21, pp. 23-29, 2011.

[2] See the Indico User Guide at http://indico-sandbox.cern.ch/ihelp/html/UserGuide/index.html

[3] A. T. Holdener III, J. Clerk Maxwell, Ajax: The Definitive Guide, O’Reilly: Beijing, 2008.

Visual Novel – Potentials for Education and Socializing

Katerina Bashova, Veno Pachovski, Adrijan Bozinovski
School of Computer Science and Information technology
University American College Skopje
Skopje, Macedonia
bashova.katerina@gmail.com, pachovski@uacs.edu.mk, bozinovski@uacs.edu.mk

***Abstract*—In this paper, the potentials that the visual novels can offer are explored, especially for education and socialization purposes. In order to present the potentials of the visual novels, the definition of the games, which support branching path storytelling, is presented. The focus of this paper is to show how visual novels can be a powerful tool for creating interactive stories that can educate and in the same time entertain the youth, which is more in demand for games that contain interactivity and multimedia. Within these games, various multimedia elements can be found, such as background sounds, music, and images for each scene, character sprites, character voices, special effects and many more. However, the most important element, which is the core of this game, is the block of text that sequentially appears on the screen. If the story is able to convey its message and is tangible to the player, at the end of the game the player may be encouraged to "read" the visual novel again in order to find out alternative endings. To demonstrate the potentials of the visual novels, several noncommercial demo versions and successful commercial ones are analyzed. Answers are given to the following questions: what kind of stories they have, what kind of messages they send to the player, how many characters does each game have and what kind of multimedia elements they have. The paper will conclude with recommendations for improvements of the aforementioned games.**

Keywords*— *Interactive storytelling; branching path stories; gamebooks; visual novels; visual poetry; educational games; multimedia

I. INTRODUCTION

In the past, computer games were mostly used for entertainment and most of them promoted violence, fear and hate. In recent researches [1], it is concluded that "interactive games can be powerful environments for learning". Many educational games are being developed in order to motivate potential students into a learning activity. Some students, who had been reluctant to follow the traditional didactic type of teaching, have accepted the games with utmost pleasure and have engaged in the productive play [1]. Most of the games have different strategies in order to pass the educational messages to the players, but what is common to all of these games are the multimedia elements like music, sounds, animation, images (two-dimensional or three-dimensional), characters and many other. However, what if the main motive is to encourage the players not just to play the game, but also to read the game story and to provide the player the power to influence the game story? This could be achieved with the use of the games known as visual novels.

II. BRANCHING PATH STORIES

A. *Definition and Brief History*

Branching path stories consist of multiple decision points where they can be found across the game and, as the player progresses in the game, s/he has to make decisions on those points [2]. The decisions can have small-scale impacts on the main plot, but they can also have a vast impact and can change the course of the story completely, by making it branch into an entirely other direction. As a result, the branching path stories have multiple endings, thus allowing the player the freedom and control, by allowing him or her to have the power to influence in the course of the story at the multiple decision points that s/he comes across. The true uniqueness of the branching path stories is the rigid structure of decision points from which it is built and the paths that lead to those points [2].

The earliest concept of non-linear stories had been defined by the Argentinian writer Jorge Luis Borges, who in 1941 had published a short fictional story "An Examination of the Work of Herbert Quain" [3]. The story is in a form of a biographic essay which describes several works by the fictional Irish author Herbert Quain, from which "April March" is a non-linear novel. The novel consists of thirteen chapters, which present nine different story lines. "…The first chapter can lead into any one of three subsequent chapters, and each of those in turn has three possible subsequent chapters." This story can be considers the first example of the elementary structure of the game books [4].

B. *Game books – Definition and Brief History*

A game book can be defined "…as any book in which the reader participates in the story by making choices which affect the course of the narrative" [5]. There are many series and types of game books, from which the most popular series are Choose Your Own Adventure, published from 1979 to 1998,

with 185 known game books, which had been translated into twenty six different languages [6]. Although most of the game books were fictional, from 1958 to 1972 thirty-two educational game books had been published that were known as TutorText. These books covered many subjects such as Mathematics, music, computer programming, physics, economy, chemistry and many others [7]. The books are structured with a section, which explains the reader some subject and at the end of explanation a question with multiple answers is presented; each answer points the reader to a certain page of the book for further explanation, telling the reader whether s/he had got the right answer. The game book "Adventures in Algebra", published 1960, has a partial online implementation and in the near future the rest of the series are planned to be available online [8].

III. VISUAL NOVEL

The visual novel is a popular games genre in Japan, that was mainly released on the PC, but was later also ported on other popular game consoles, such as PlayStation and DS. It is a game that in many ways is an action of reading a book. However, unlike reading an e-book classically, the story is told by blocks of text that appear on the screen [2]. The text blocks can appear in two ways: as an NVL (derived from NoVeL) file, where the text is presented in a frame that fills the whole screen, and an ADV (derived from ADVenture game) file, where the text is presented in a smaller frame that is commonly located on the bottom of the screen. There are exceptions, such as when the frames are in a form of a floating bubbles that appear on the screen or in a rectangular form that covers one vertical half or one vertical third of the screen.

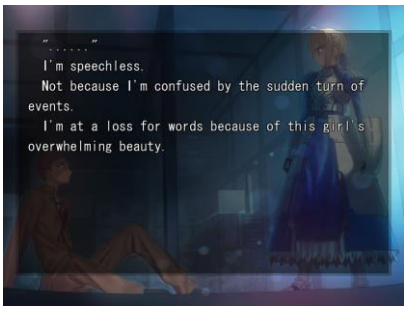


Fig. 1. Fate/stay night - Example of NLV frame



Fig. 2. Quartett! - Example of floating bubble frames



Fig. 3. Ristorante Amore - Example ADV frame

Other characteristics that the e-books don't possess are the multimedia elements like background images, character sprites, background music and sounds effects, image maps, voice acting and so on. The background images are static images that represent the scenes of the story, while the character sprites can be static or animated images that represent the expressions and possessions of the characters that appear in the story. The background music and sounds effects bring up the atmosphere of the story and each character may have its own theme song. The image maps are maps that allow the player to go from one place to another, and can be used in some professional complex visual novels, as well as the use of voice acting, which provides voices for each character.

The visual novels are considered an improvement in the branching path stories that are found in Choose Your Own Adventure books, because some of the decisions points are hidden from the player, and the game, based on the past chosen decisions made by the player, automatically selects a branch. Another pro is the large number of save slots, which enable the player to save the game state at any time s/he wants and later load it and continue to play from that saved point. Additionally, the player can skip through the already-read text until the next decision point or until a certain scene. This enables the player to explore easily the other branches of the story and to find out the rest of the alternative story endings. Lastly, the visual novels, as a digital product, unlike the classical books, do not face the restriction length. An example is the visual novel Fate/Stay Night, which, along with its different branches, has a total word count larger than that of the whole The Lord of the Rings trilogy [2].

Commonly used visual novel (VN) engines in Japan are KiriKiri [9], NScripter [10], and CatSystem2, whereas the VN engines globally used are listed in Table I [11]. Still some developers, who prefer to have more freedom in their coding and/or encryption of their games, or in order to target specific operating systems/consoles, have used alternative tools to develop their visual novels such as YoYoGames' GameMaker Studio [12], ramifygames' jStory [13], RPG Maker [14], Unity Technologies' Unity [15], Adobe Systems' Adobe Flash [16] and many others.

TABLE I. LIST OF VARIOUS VIUSAL NOVEL ENGINES [11]

Name of engine	*Developer*	*Player*	*Description*
Ren'Py	Windows, Macintosh, Linux	Windows, Macintosh, Linux, Android	A visual novel engine based on the Python coding language. Minimal programming experience is needed to create a simple game, though the use of Python allows more complex games to be created. The Lemmasoft Forums contains a board dedicated to the discussion of Ren'Py, both in development and use.
Belle	Windows, Macintosh, Linux	Windows, Macintosh, Linux, Android, Internet Browser	A cross-platform, open source engine based on HTML5 with an interface similar to Novelty. This allows greater versatility for players so they can play on a web browser or offline. Currently in development.
Nscripter	Unknown	Unknown	None
NVList	Windows, Macintosh, Linux	Windows, Macintosh, Linux, Android	NVList is an open source visual novel engine written in Java. All rendering is done through OpenGL, taking advantage of available hardware acceleration and allowing for high-resolution graphics.
Novelty	Windows	Windows	A visual novel engine that requires no technical knowledge to use. Users can drag and drop for certain effects easily.
BASSnovel	Unknown	Internet Browser	A visual novel engine that appears to be browser-based.
NovelStream	Unknown	Internet Browser	A visual novel engine that appears to be browser-based.
WebStory Engine	Windows, Macintosh, Linux	Internet Browser	Browser-based VN engine built using XML.
CatSystem2 (Japanese only)	Unknown	Unknown	None

Most of the visual novels have adult themes that are not suitable for children. Some visual novels like Type-Moon's Fate/stay night [17] and Littlewitch's Quartett! [18] have published 18+ versions at first, but later they have also published 15+ versions, whereas Key's Clannad [19] was published for all ages. Recently educational visual novels have been published as well. An educational visual novel called Amigo for iPad teaches nutrition to young Japanese students with regular eating behaviors. They have managed to teach four out of five topics and managed to motivate half of the students to put the obtained knowledge into real life use [20].



Fig. 4. Extract from visual novel Amigo [20]

Philippine students have developed another educational visual novel as their undergraduate thesis. They have selected a popular epic Ibong Adarna by José de la Cruz, which in the Philippines is a part of required school literature assignment. They have made a testing of Ibong Adarna along with the Japanese visual novel Clannad on high school students and later they have made a survey on them regarding the two visual novels. Their visual novel has received positive feedback from the students and was leveled as good as the Japanese one, and they have managed to make a successful educational visual novel in which young students can enjoy in its story [21].

Another interesting educational visual novel is Hypothesis Testing, which was made by prof. J. David Eisenberg for his class Introduction to Research Methods. What inspired him to make this small visual novel is that he thought that his students were not reading the course book and his lecture did not seem to explain the topic. He had been searching a new way to indulge the students into studying but at the same time to have fun. His solution was a small visual novel that he had made with the visual novel engine Ren'Py (see Table I.) [22]. What is more interesting is that the visual novel is available online so not only his students but also other people interested into learning about hypothesis testing can play it. Unfortunately, there is no record about whether the visual novel has improved the students' studies, since no survey had been made nor feedback was given [23].



Fig. 5. Multiple choice question about Null Hypothesis-extract from Hypothesis Testing

IV. ANALYSIS OF EDUCATIONAL VISUAL NOVELS

Two noncommercial and one commercial visual novel will be analyzed and their characteristics and multimedia elements will be presented in this paper.

A. *Analysis of Cinders*

Cinders is an "atmospheric" visual novel developed by MoaCube. The game was developed with the tool GameMaker from YoYo Games, because, according to the developers, it allowed them more flexibility to write their own VN (Visual Novel) framework, than with the other standard VN tools [24]. The story is based on the Grimm's popular Cinderella, but the retelling is more realistic and has heavy branching through the game. The story has four main endings and each ending had from one to four panels: The Fairytale ending (three panels – Queen type, Love Interest and Advisors), Traveler ending (two panels – Life Style and Company), The Independent Woman (four panels – Carmosa State, Type of Host, Connection with the Sisters and Love Interest) and The Grim Fate ending (one panel – Mourners). Based on how the main character Cinders interacts with all the eight characters in the games, the player can unlock all four endings by replaying. On the first playing, as well on every other replaying, the player obtains eight trophies that will appear on the main game menu.



Fig. 6. Cinders game menu before obtaining the trophies



Fig. 7. Cinders game menu after obtaining the trophies

There are nine characters, which are not represented as typically evil or good, but they all act based on the circumstances and the positions they have in society. The main character is Cinders, which lives with her stepmother and two stepsisters who treat her as a mere servant in her own home. On each decision point, the player shapes Cinders' personality and which can be good, evil, wise and independent woman or traveler, and can have various love interests, advisors and travel companions. In addition, there is The Grim Fate ending were Cinders faces death and if she had made some particular bond with one of the characters, they will mourn for her.

The game consists of various multimedia elements such as lip and eye animation with sound, text animation with sound, eloquent background images, background music and sound. In the game, there is also a map of five location, which based on the time of the game can be accessible for the player to explore.



Fig. 8. Cinders map location

B. *Analysis of The Little Lazy Snail*

The Little Lazy Snail is a noncommercial visual novel developed with the tool Ren'Py (see Table I.) in 2012, as an undergraduate thesis. The game targets preschool and early schoolchildren. The story is an original fable, with the message "the early bird gets the worm", following eleven anthropomorphic bugs in an imaginary forest, where the main character is the little lazy snail Goce who is a typical sleepyhead and doesn't want to go to school. The story covers two days of Goce's life in and out of school, and shows how he interacts with his friends, teacher, family and strangers.

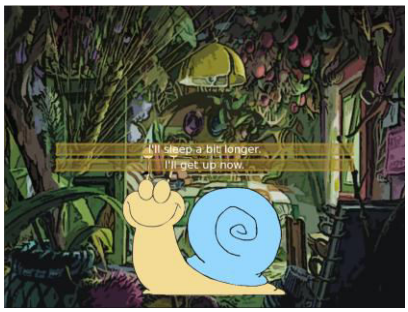


Fig. 9. The little lazy snail's major branching point

The game has only two decision points, from which the last one is a major branching point where the player can get either a happy ending or bad ending. If the player gets the happy ending, s/he unlocks the credits of the games as a video sequence which can be accessible from the main game menu. The game consists of various multimedia elements such as lip and eye animation with voice over only in the Macedonian version, vivid background images, background music and sound, image transformation.

C. *Analysis of The Right Path*

The Right Path is a noncommercial visual novel developed with Ren'Py (see Table I.) in 2013 and it targets teens. The game has an original realistic story about a teenage boy James and his everyday life. He is portrayed as a rebel and lazy

student, and through the game based on how he interacts with his friends and strangers, the player receives points and if s/he collects enough points, s/he gets the good ending. In order to collect the points, James must go to school, offer help to his friends and avoid associating with the school delinquents. If he joins the delinquents, or accepts to assist in pick pocketing, the player loses points and faces the bad ending. As a reward on achieving the good ending, the player unlocks the special CG (Characters Gallery) and BG (Backgrounds Gallery) that become accessible from the main game menu.

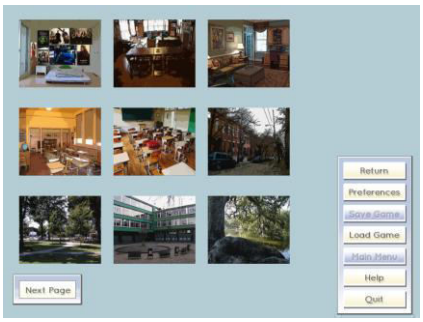


Fig. 10. Background gallery from The Right Path

Besides hidden galleries, the game has static character sprites, carbonized real photos as background images, background music, minimal sounds and snow falling animation. Unlike the previous game, it is only in English and has no voice over.

TABLE II. THE MULTIMEDIA ELEMENTS FOUND IN THE THREE VISUAL NOVELS PRESENTED IN THIS PAPER

Multimedia elements	**Visual novels**		
	Cinders	***The Little Lazy Snail***	***The Right Path***
Backgrounds	Yes	Yes	Yes
Background music	Yes	Yes	Yes
Sound	Yes (text appearance)	Yes	Yes
Character sprites	Yes (lip and eyes animated)	Yes (lip and eyes animated)	Yes (static)
Animation	lip and eyes animation sprites transition	lip and eyes animation sprites transition	Snow animation
Voice over	No	Yes (only in Macedonian)	No
Awards	Trophies and pictures	Video credits	Characters gallery (CG) Backgrounds gallery (BG)
Other	/	Available in two languages	/

V. CONCLUSION

The visual novels, primarily used as entertainment in Japan, have the potential for education and socialization. When the player "reads" the visual novel, s/he has the power to influence the flow of the game story by interaction with the non-playable characters and by making decisions on the decision point within the game.

Although the story is the core of the visual novels, other multimedia elements (background music, background images, sounds and similar) are the ones that can keep the player's attention to continue to follow the story and play the game repeatedly. This could be ideal educational tools for children, which can be easily charmed by vivid characters, pictures, music and animations. However, educational visual novels are not strictly made for children, but also for high school student or college students and can help them improve their studies and use the gained knowledge in daily lives. In fact, a well-designed general educational visual novel can be enjoyable both by youths and adults.

Because the number of Visual Novel engines is increasing every day and because of their user friendliness, even for non-programmers, visual novels can be easily "written". Nevertheless, in order to develop a successful educational visual novel, a strong collaboration of educators and developers must be accomplished, where the educators would define the stories and the developing team/s would "write" the visual novel. In addition, it is important to maintain a balance in the multimedia elements, especially on the story, which is the core of the visual novel. Therefore, it must be carefully constructed in order to avoid unnecessary and problematic branches that could cause a loss of valuable resources and hinder the future maintenance of the game.

REFERENCES

[1] D. A. Lieberman, "What can we learn from playing interactive games?," in *Playing Video Games: Motives, Responses, and Consequences*, Peter Vorderer and Jennings Bryant, Eds., Abingdon, Oxfordshire, Oxford: Routledge, 2006, pp. 379-397.

[2] J. Lebowitz and C. Klug, Interactive Storytelling for Video Games: A Player-Centered Approach to Creating Memorable Characters and Stories, Burlington, MA, USA: Focal Press, 2011.

[3] F. Kistler, D. Sollfrank, N. Bee and E. André, "Full Body Gestures Enhancing a Game Book for Interactive Story Telling," in *Interactive Storytelling 4th International Conference on Interactive Digital Storytelling, ICIDS 2011 Vancouver, Canada, November 28 - December 1, 2011 Proceedings. Lecture Notes in Computer Science / Information Systems and Applications, incl. Internet/W*, vol. 7069, M. Si, D. Thue, E. André, J. Lester, J. Tanenbaum, V. Zammitto, Eds., Heidelberg, Springer Berlin Heidelberg, 2011, pp. 207-218.

[4] J. L. Borges, "Collected Fictions," Trans. A. Hurley, New York, Penguin Books, 1999.

[5] D. Katz, "Frequently Asked Questions - Questions About Gamebooks," 1998-2015. [Online]. Available: http://www.gamebooks.org/show_faqs.php. [Accessed 15 February 2015].

[6] D. Katz, "Choose Your Own Adventure (1979-1998)," 1998-2015. [Online]. Available: http://www.gamebooks.org/show_series.php?id=30. [Accessed 15 February 2015].

[7] D. Katz, "TutorText: Doubleday Series," 1998-2015. [Online]. Available: http://www.gamebooks.org/show_series.php?id=457. [Accessed 15 February 2015].

[8] D. Wuertele, "Tutortext.org," 2011-2015. [Online]. Available: http://tutortext.org/. [Accessed 15 February 2015].

[9] W.Dee, "*KiriKiri Download page (吉里吉里ダウンロードページ)*",2013-2015. [Online]. Available: http://kikyou.info/tvp/. [Accessed 16 April 2015].

[10] N. Takahashi, "nscripter.com". 2015. [Online]. Available: http://www.nscripter.com/. [Accessed 16 April 2015].

[11] OELVN Community Wiki, "Game Engines - OELVN Community Wiki", 2013. [Online]. Avaiable: http://wiki.renai.us/wiki/Game_Engines [Accessed: 01 March 2015].

[12] YoYo Games, "GameMaker: Studio | YoYo Games", 2015. [Online]. Available: https://www.yoyogames.com/. [Accessed 1 March 2015].

[13] P. Moore, "jStory | Make your own adventure", 2015. [Online]. Available: http://ramifygames.com/. [Accessed 1 March 2015].

[14] Enterbrain and Degica, "Make Your Own Game with RPG Maker", 2015. [Online]. Available: http://www.rpgmakerweb.com/. [Accessed 1 March 2015].

[15] Unity Technologies, "Unity - Game Engine," 2015. [Online]. Avaiable: http://unity3d.com/. [Accessed: 03 March 2015].

[16] Adobe Systems Incorporated, "Adobe Flash runtimes | Adobe.com," 2015. [Online]. Avaiable: https://www.adobe.com/products/flashruntimes.html. [Accessed: 03 March 2015].

[17] Type-Moon, *Fate/stay night,* Tokyo: Type-Moon, 2004.

[18] Littlewitch, *Quartett!,* Tokyo: Littlewitch, 2004.

[19] Key, *Clannad,* Osaka: Key, 2004.

[20] F. L. Salazar, T. Nakajima and T. Alexandrova, "Visual Novels: An Methodology Guideline for Pervasive," in *Grid and Pervasive Computing: 8th International Conference, GPC 2013, and Colocated Workshops, Seoul, Korea, May 9-11, 2013, Proceedings. Lecture Notes in Computer Science / Theoretical Computer Science and General Issues (Book 7861).*, vol. 7861, J. J. H. Park, H. R. Arabnia, C. Kim, W. Shi and J. Gil, Eds., Heidelberg, Springer Berlin Heidelberg, 2013, pp. 234-243.

[21] C. T. J. Agos, M. L. C. Bal-Ut, L. G. Calam, R. N. Cantuba, M. H. M. Vallo and M. R. Apsay, "Game Development Of Ibong Adarna Visual Novel," *International Journal of Scientific & Technology Research,* vol. II, no. 9, pp. 259-266, 2013.

[22] jdeisenberg, "Lemma Soft Forums - Educational Visual Novel," 13 March 2010. [Online]. Available: http://lemmasoft.renai.us/forums/viewtopic.php?f=16&t=6809&sid=fb645d3f8d6131da4755e16747ae6a96. [Accessed 26 February 2015].

[23] J. David Eisenberg, "Inquiry for educational visual novel," [E-mail] Message to: Katerina Bashova, 01 February 2015.

[24] TeeGee, "Cinders is code complete!| MoaCube," MoaCube, 28 February 2015. [Online]. Available: http://moacube.com/news/cinders-is-code-complete/. [Accessed 28 February 2015].

Detecting protein binding sites by using the fuzzy nearest neighbours clasifier

Georgina Mirceva, Andrea Kulakov
Department of intelligent systems
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
georgina.mirceva@finki.ukim.mk, andrea.kulakov@finki.ukim.mk

***Abstract*—The number of determined protein structures rises with the innovation of new technologies. However, this information is not efficiently used if these structures are not functionally annotated. Namely, the knowledge about the functions of the protein molecules is of highest importance and could be used for designing new drugs that can help in improving humans' lives. There are various experimental methods for determining protein functions, but they are labor intensive and time consuming, so cannot provide fast annotation of all proteins whose structures are determined. This inspired many researchers to investigate new computational methods for functionally annotating the protein structures. One group of methods makes decisions about the protein functions based on the characteristics of the binding sites where the inspected protein interacts with another one. In that context, in this paper we introduce a novel approach for detecting the protein binding sites by using a fuzzy nearest neighbours classifier. We present some experimental results from the evaluation of the proposed approach.**

Keywords—protein function; protein binding site; fuzzy nearest neighbours classifier

I. INTRODUCTION

Protein molecules and their understanding is very important because they participate and influence many processes in the living organisms. Based on the knowledge about the interactions in which they are involved and the functions they have in these interaction, we may control the processes in the living organisms, in terms of stimulating or deactivating some interactions. Therefore, the knowledge about the functions of the proteins could be used for designing new drugs. There are many methods for determining protein functions, but with manual inspection there is an enormous increase of the number of protein molecules that are not yet investigated in terms of their functions. Moreover, these methods are very expensive. As a result of this, functional annotation of protein structures is one of the hottest topics in bioinformatics, which stimulates many research groups to investigate and develop various computational methods in order to provide fast and accurate determination of the protein functions.

The literature provides a wide palette of methods for annotating protein structures. The homology between proteins could be used as indication that they share common functions because they have common predecessors. Therefore, some methods annotates protein structures by identifying the homologous of the inspected protein structure [1]. Due to the fact that proteins that are homologous are very likely to have similar functions, it is presumed that the parts of the structures that remain the same during evolution are those who determines the functions. Therefore, a second group of methods try to identify and analyze the conserved parts of the protein sequence and structure [2]. Based on the knowledge about the interacting pairs of structures, the protein-protein interaction networks are obtained, and afterwards, various methods for analyzing graph are used in order to predict the functions of the examined protein [3]. Another very popular methods are those who determine protein functions by identifying and analyzing the binding sites where an interaction between proteins occurs. In [4], a wide overview of the tools for identifying protein binding sites is provided. In this research, we focus on the last group of methods.

The first step in identifying binding sites is to determine the characteristics of it amino acid residues. There are various characteristics that are considered in the methods provided in the current literature. In this research we take into consideration the most common features of the residues that are used for this purpose, which are the Accessible Surface Area (ASA) [5], depth index (DPX) [6], protrusion index (CX) [7] and hydrophobicity [8].

After extracting the features of the residues, then by using some method for classification we should estimate whether a given residue is part of binding site or not. For this aim, various methods could be used. In our previous research [9], we have used several classification methods that are based on the classical set of theory. However, in the context of annotation of protein structures, the fact that some mutations occur during protein evaluation should be considered. This means that we should pay attention on the fact that the small changes in the features obtained during evaluation should not have high influence on the final decisions. In order to provide this, in our previous research we applied several methods that are based on fuzzy set theory in order to generate models for protein binding sites prediction. We used methods that induce fuzzy decision trees (FDTs) [10] and fuzzy pattern trees (FPTs) [11]. In [12] we used the FPT for predicting protein binding sites, while in [9] we made additional improvements by selecting the most relevant features of the amino acid residues. There are also

other classification methods based on the fuzzy set theory that could be used. In this research we introduce the fuzzy nearest neighbours classifier [13] for protein binding sites prediction.

The rest of the paper is organized as follows. Section 2 describes the process of extracting the amino acid residues features and the process of predicting the protein binding sites by using the fuzzy nearest neighbours classifier. Some experimental results of the evaluation of the proposed approach are given in section 3, while section 4 gives conclusion and directions for further improvements.

II. Our Approach

In our approach, the detection of the protein binding sites is performed in two steps. In the first step, the characteristics of the amino acid residues are extracted. Then, in the second step the protein binding sites are identified by using the fuzzy nearest neighbours classifier [13].

A. Extraction of the Characteristics

We extract the following characteristics of the residues that constitute the inspected protein structures: Accessible Surface Area (ASA) [5], depth index (DPX) [6], protrusion index (CX) [7], and hydrophobicity [8].

The first characteristic, Accessible Surface Area (ASA), is the most common feature used for binding sites prediction. It provides information about the surrounding volume around the inspected residue that could be used in the decision whether the inspected residue would be part of some binding site or not. The value of the ASA characteristic is estimated by using the rolling ball algorithm [5]. This is done by rolling a sphere with a radius of 1.4 Å around the surface of the protein, and for each residue we estimate its area that is touched by this sphere. In order to reduce the dimensionality of the problem in the second step, we filter only the residue that are located at the protein surface. This filtering is done by considering only those residues for which at least 5% of their surface area could be touched by the rolling sphere, as it is suggested in [14].

The second characteristic that is considered in this research is the depth index (DPX) [6]. This characteristic for a given atom is equal to the Euclidean distance between the inspected atom and its nearest atom among the atoms that are touched by the rolling sphere used for calculating the ASA characteristic. Because each amino acid residue is constituted from several atoms, therefore we aggregate their DPXs by averaging the values.

The third characteristic is the protrusion index (CX) [7], and it is related to the density of the surrounding volume around a given atom. The protrusion index is calculated as a ratio between the non-occupied (empty) part of the volume around the inspected atom and the occupied volume. The occupied volume is estimated by multiplying the number of non-hydrogen atoms in the inspected volume and the mean volume of the atoms. Similarly as for the depth index, first we calculate the protrusion indices for all atoms that constitutes the inspected residue, and then we aggregate them by averaging.

The fourth characteristic is the hydrophobicity. This characteristic is also very widely used for binding sites prediction because it indicates the preferences a given type of amino acid to be located near the protein surface or in its deep interior. In this paper we use the scale for hydrophobicity that is introduced in [8].

B. Fuzzy nearest neighbours classifier

After extracting the characteristics of the residues, next, in the second step we identify the residues that constitute the binding sites by using the fuzzy nearest neighbours (FNN) classifier [13]. The FNN classifier is inspired from the well-known *k*-nearest neighbours (KNN) classifier. In KNN, if all neighbours have equal weights during voting, then the majority class will be chosen although it could be related with the more distant neighbours. This could be avoided by using distance weighted voting. However, the choice of the number of nearest neighbours that are considered is of high importance and has high influence on the final decisions. This could be solved by using the FNN classifier that is introduced in [13], which is based on the fuzzy set theory. With the FNN algorithm, first the *k* nearest neighbours are determined thus obtaining the set of nearest neighbours denoted as *NN*, and then the inspected test sample *t* is classified in a class *c* by maximizing

$$\frac{\sum_{x \in NN} S(x,t)C(x)}{\sum_{x \in NN} S(x,t)}, \tag{1}$$

where $C(x)$ is the class membership function for the inspected class, while $S(x,t)$ is the similarity between the inspected sample t and its neighbour x and it is calculated as

$$S(x,t) = \frac{1}{D(x,t)^{\frac{2}{m-1}}}. \tag{2}$$

In equation (2), $D(x,t)$ denotes the distance between the samples x and t, while with m we define the extent of the distances in the distance weighted voting approach. In this research we use m=2, thus the similarity between two samples is $S(x,t) = 1 / D(x,t)^2$.

For the class membership function, we may use the crisp membership function or the gradual function proposed in [13]. With the crisp function, $C(x)$ equals 1 if sample x belongs to the given class c and zero otherwise. With the gradual approach [13], the class membership function is defined as

$$C(x) = \begin{cases} 0.51 + 0.49\frac{n_C}{K}, & x \in C \\ 0.49\frac{n_C}{K}, & x \notin C \end{cases}, \tag{3}$$

where C denotes the set of samples in the examined class, n_C is the number of samples in C and K is the number of nearest neighbours that are considered for making the decision. In this research we use the gradual approach for the class membership function $C(x)$.

III. Experimental Results

The dataset used for evaluation of the proposed approach contains the four features that were described in section 2, and the class attribute, which is a binary attribute who indicates whether the residue belongs to a binding site or not. The value for the class attribute is obtained based on the knowledge stored in the BIND database [15] that contains information about the protein binding sites that are determined in experimental way.

In order to make evaluation by using the most representative protein chains, we filter the database by considering the chains that have less than 20% similarity in their sequences based on the criteria proposed in [16]. Then, using the same criteria, the chains that have less than 10% similarity in the sequences are considered as test samples, while the remaining chains are used as training samples. In order to reduce the dimensionality of the problem, we eliminate the residues of these chains that are not at the surface, as it was described in the previous section. Regarding the test dataset, we additionally filter the residues from the 1549 test chains contained in the B1549 test dataset that was used in [9]. The samples correspond to the residues of these chains, so in this way the number of test samples is 277735. From these test samples only 16.42% belong to the binding sites class according to the BIND database [15]. Similarly, from the 115579 residues from the training chains, 15696 (13.58%) are residues from the binding sites stored in the BIND database. From this, it is evident that the class that corresponds to the residues that are not part of binding sites is dominant, therefore we perform balancing of the training dataset in order to give equal treatment to each class (to avoid favoring of the dominant class). The balancing is done by down sampling the training set until 27% of its original size is achieved by using sampling without replacement (meaning that each sample could be considered only once in the newly obtained set).

We use the implementation of the fuzzy nearest neighbours method [13] provided in the Fuzzy Weka software. In the experiments, we set up the parameters to their default values. The number of nearest neighbours is set to $K = 10$, while the parameter m, which defines the degree of a given distance, is set to $m = 2$. We use the Euclidean distance for measuring the distance between the inspected test sample and its neighbours.

The proposed approach achieves AUC-ROC = 0.554, Precision = 0.754, Recall = 0.565 and F_1 measure = 0.622. The testing time for making the final decisions for all 277735 test samples is 5762 seconds (1.6 hours). Since we are using a lazy approach where instance based learning is performed, therefore the testing time is significantly higher than by using some other classification method where prediction model is generated.

IV. Conclusion and Future Work

In this paper we proposed a novel approach for protein binding sites prediction. The decisions about whether a given amino acid residue is one of the constituents of some of the binding sites of the protein is made in two steps. In the first step, four characteristics of the residues are extracted, and then in the second step the inspected residue is classified by using the fuzzy nearest neighbours classifier.

The proposed approach achieves AUC-ROC = 0.554, Precision = 0.754, Recall = 0.565 and F_1 measure = 0.622 by considering the 10 nearest neighbours of each test sample.

Further, we may make additional analysis by using various settings in the method. Namely different number of nearest neighbours could be used. Also, some other measure could be used for measuring the distance between the samples, and by the parameter m we may give different degrees to this distances.

Besides the four characteristics used in this paper, also some other features of the amino acid residues could be included in order to provide better distinction between the binding and non-binding residues. Regarding the classification method, in future we will continue our search for the most appropriate classifier that could lead to better predictions.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, Macedonia.

References

[1] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Evolution of function in protein superfamilies, from a structural perspective," J. Mol. Biol., vol. 307, no. 4, pp. 1113–1143, 2001.

[2] A. R. Panchenko, F. Kondrashov, and S. Bryant, "Prediction of functional sites by analysis of sequence and structure conservation," Protein Science, vol. 13, no. 4, pp. 884–892, 2004.

[3] M. Kirac, G. Ozsoyoglul, and J. Yang, "Annotating proteins by mining protein interaction networks," Bioinformatics, vol. 22, no. 14, pp. e260–e270, 2006.

[4] N. Tuncbag, G. Kar, O. Keskin, A. Gursoy, and R. Nussinov, "A survey of available tools and web servers for analysis of protein-protein interactions and interfaces," Briefings in Bioinformatics, vol. 10, no. 3, pp. 217–232, 2009.

[5] A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms," Lysozyme and insulin, J. Mol. Biol., vol. 79, no. 2, pp. 351–371, 1973.

[6] A. Pintar, O. Carugo, and S. Pongor, "DPX: for the analysis of the protein core," Bioinformatics, vol. 19, no. 2, pp. 313–314, 2003.

[7] A. Pintar, O. Carugo, and S. Pongor, "CX, an algorithm that identifies protruding atoms in proteins," Bioinformatics, vol. 18, no. 7, pp. 980–984, 2002.

[8] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," J. Mol. Biol., vol. 157, no. 1, pp. 105–132, 1982.

[9] G. Mirceva and A. Kulakov, "Improvement of protein binding sites prediction by selecting amino acid residues' features," J. Struct. Biol., vol. 189, no. 1, pp. 9-19, 2015.

[10] C. Z. Janikow, "Fuzzy decision trees: issues and methods," IEEE Transactions on Systems, Man, and Cybernetics, vol. 28, no. 1, pp. 1–14, 1998.

[11] Z. H. Huang, T. D. Gedeon, and M. Nikravesh, "Pattern trees induction: a new machine learning method," IEEE Transaction on Fuzzy Systems, vol. 16, no. 3, pp. 958–970, 2008.

[12] G. Mirceva and A. Kulakov, "Fuzzy pattern trees for predicting the protein binding sites," 9th Conference for Informatics and Information Technology (CIIT 2012), Bitola, Macedonia, 2012.

[13] J. M. Keller, M. R. Gray, and J. R. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. 15, no. 4, pp. 580–585, 1985.

[14] C. Chothia, "The Nature of the Accessible and Buried Surfaces in Proteins," J. Mol. Biol., vol. 105, no. 1, pp. 1–12, 1976.

[15] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," Nucleic Acids Res., vol. 29, no. 1, pp. 242–245, 2001.

[16] J. –M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner, "The ASTRAL Compendium in 2004," Nucleic Acids Res., vol. 32, pp. D189–D192, 2004.

Autonomious robot for temperature monitoring in the datacenter

Vladislav Bidikov
Faculty of Computer Science and Engineering
Skopje, Macedonia
vladislav.bidikov@finki.ukim.mk

Nevena Ackovska Ph.D.
Faculty of Computer Science and Engineering
Skopje, Macedonia
nevena.ackovska@finki.ukim.mk

Monitoring the data center is always a specific task especially since most data center only has temperature sensors on specific location. These sensors not always show the real situation. We like to create a concept which will allow a small vehicle to check specified locations in the data center.

Keywords: robotics, temperature, cloud, datacenter

I. INTRODUCTION

With all the latest advancements in the cloud and virtualization markets, who needs a power-draining, expensive data center? Unfortunately, more and more organizations are still continuing to run with this overhead. In new data center designs, capacity provisioning for ever-higher power requirements has always been an area of concern, with a problematic question on whether conventional room-conditioning systems can manage future information technology loads. Within existing data centers, computing capacity typically increases over time as IT requirements increase, resulting in increased power and cooling requirements. Data center operators are challenged to provide adequate support infrastructure that is provisioned, or adapt accordingly, to achieve future IT mission requirements while minimizing energy use.

With increased reliance on data centers, full software and hardware robotics automation is no longer a question of "if," but a matter of "when." Simple RFID tags, laser and barcode identifiers can create true data center automation. Data center, automation and robotics technologies have come a very long way over the past decade. From the warehousing or inventory perspective, robots are equipped to sense location, put the assets or inventory in order, and can directly interact with the human-created automation scenarios.

We describe the concept for an inexpensive robot that serves as a physical autonomic element, capable of navigating, mapping and monitoring data centers with little or no human involvement. The concept will be based on standard open hardware platforms which can be integrated with existing power efficiency and building management solutions.

II. ANALISYS OF THE ROBOT PLATFORM

As the demand for IT computing and network devices continues to increase, there is a growing need to remove heat generated by the equipment, effectively and efficiently. To improve efficiency, it makes sense to locate cooling (such as computer room air conditioners or computer room air handlers) close to the heat source, or another approach would be to have a removal device, such as a hot air chimney, close to the heat source.

Now, with techniques in place, are we really sure that the amount of cooling given per rack or equipment is just right? A typical data center should run anywhere between 18 degrees C to 27 degrees C as per ASHRAE (American Society of Heating, Refrigerating, and Air-Conditioning Engineers). The lower the temperature, higher will be the energy consumption and higher the temperature resulting in a possible problem with equipment availability. This raises the risk of temperature fluctuation inducing equipment failures. Thus, an accurate responsive data center temperature monitoring and control has become increasingly important. Running the data center at optimum temperature is thus a challenging technical effort.



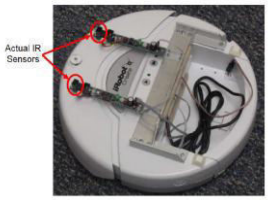
Some of the earliest work on data center energy monitoring and management was based on sensors installed at fixed locations. Alternatively, another method used was a hand-pushed mobile sensing station that accepts human input to record data center layout while automatically recording a set of temperature and humidity readings. Both concepts have their advantages but their main problematic area is the fact that their limited mobility can create so called cold/warm areas which in general will be a major consumer of power.

The quest for an alternative to these approaches led to the development of an inexpensive robot that automatically maps and navigates a data center, collects temperature, humidity and other data, and feeds such data into a software tool for efficiency analysis and data center energy management. This idea is greatly researched by a set of engineers from EMC (www.emc.com), brainstormed and came up with the idea to build a low-cost robotic platform to monitor environmental parameters in a data center. The

prototype then designed of the "DC Robot" as they call it consists of three sensors mounted to a vertical tube as per ASHRAE recommendations on measurement points, and is also equipped with remote navigation capabilities through cameras and remote control software. This DC Robot collects temperature data using three digital sensors and relays it through a Wi-Fi access point for post-processing. An algorithm converts the temperature data into a thermal map. The thermal map can then be used to easily identify the hot and cold spots of the data center aisles with location information. After further analysis of this concept we found that this idea works very well for large data centers since the algorithms behind the platform can easily learn the structure and form of the data center. When we try this concept in a smaller data center (like the one in the Faculty for Computer Science and Engineering) the limited paths which are available make this robot platform limited since there will be places where the robot will not be able to reach.

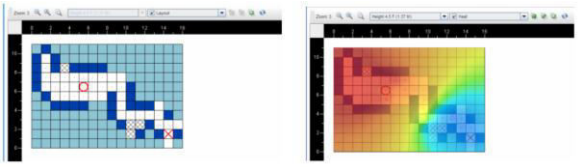
This problem can be solved in two ways. The first way is to make the robot smaller in its psychical dimensions. That way, the robot can easily go in tighter spots so I can walk in the small corridors of the data center. Unfortunately, this can cause stability problems since the robot has to carry the boom with sensors in a vertical position. The same problem was investigated by IBM researchers and they found that the minimal psychical size of the robot should be as minimal as the iRobot platform.

The second way of solving this problem will be possible if we can do some infrastructural changes in the datacenter itself so there is a circular path for which the robot can follow. The idea of making the robot "fly" in the data center was abandoned since the air flow of the cooling systems in the datacenter can create unforeseen reactions from the flying robot.

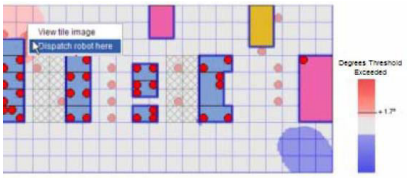


Analysis of the movement of the robot and its autonomous movement can be achieved by using basic Depth First Search (DFS) or an optimized version so the robot can map the datacenter in a more robot like way. The accuracy of the algorithm will be improved if the robot has more sensors: bump sensor, IR sensors, line tracing sensors etc. Usage of ultra-sonic sensors should be avoided since their usage can create harmful interference for other systems in the data center.

Integration of data coming from the robot is another challenge since the data from the temperature sensors has to be processed in order to make sense. Some concepts of the data center can also influence these measurements especially if some racks are filled with different equipment as others (for example storage appliances and servers have different working temperature).



Usage of special integration and building management software platforms can help create so called heat maps of the data center. Further analysis of these heat maps can help diagnose problems or even allow the data center operator to dispatch more robots to further assess the situation. Additionally, if the robot has capabilities for sending live video feed to the data center operator, than this platform can be a valuable tool for making diagnostics of problems in the data center even easier.



III. CONCLUSION

Fixing a simple cooling leak saves a lot of energy. Small adjustments in the data center's cooling temperature can also provide large energy savings. Poor control and monitoring of the conditions in a data center may shorten the life of the equipment, overheating can cause intermittent faults, and in extreme cases cause equipment to fail catastrophically. The cost in time, money and lost business productivity can be considerable.

Having an independent system like the "DC Robot" check on the air conditioning unit and monitoring devices is a wise idea. It can also help give a clear picture of the conditions in different parts of the data center and coupled with alerts such a system can give ample warning before conditions become critical or out of control. Robots become the extension of data center staff and managers and assist in keeping the data center environment running smoothly. Future platforms based on open hardware can further increase the possibilities in making monitoring of the data center even easier. Further development of these platforms can even provide some repairing tasks for which valuable engineer time can be saved.

REFERENCES

[1] Lenchner, J., Isci, C., Kephart, J. O., Mansley, C., Connell, J., & McIntosh, S. (2011, June). Towards data center self-diagnosis using a mobile robot. In Proceedings of the 8th ACM international conference on Autonomic computing (pp. 81-90). ACM.

[2] J. Borenstein, H. R. Everett, L. Feng, and D. Wehe. Mobile robot positioning: Sensors and techniques, 1997.

[3] R. Sharma, C. Bash, C. Patel, R. Friedrich, and J. Chase. Balance of power: Dynamic thermal management for internet data centers. IEEE Internet Computing, 9(1):42–49, January 2005.

New Method for Protein Function Prediction

Kire Trivodaliev, Ilinka Ivanoska, Slobodan Kalajdziski, Biljana Risteska Stojkoska
Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, Macedonia
{kire.trivodaliev, ilinka.ivanoska, slobodan.kalajdziski, biljana.stojkoska}@finki.ukim.mk

Abstract — **Clustering algorithms play an important role in the analysis of biological networks, and can be used to uncover functional modules and obtain hints about cellular organization. High-throughput experimental technologies, along with computational predictions, have resulted in large-scale protein interaction networks (PINs) for numerous organisms. A significant number of proteins in a PIN remain uncharacterized and predicting their function remains a major challenge. We propose a function prediction approach based on homogeneity optimization of the PIN employing a clustering algorithm which optimizes the ratio of between cluster variety and within cluster variety having the measures defined as distances using semantics. Validation of the new method is done using a purified and reliable Saccharomyces cerevisiae PIN. The results are comparable with state-of-the-art algorithms. As a conclusion new progress and future research directions are discussed.**

Keywords— protein interaction networks; semantics; network clustering; homogeneity; protein function prediction

I. INTRODUCTION

High-throughput experimental technologies, along with computational predictions, have resulted in large-scale biological networks for numerous organisms. In recent years, much research effort has focused on analyzing these biological networks in order to obtain hints about cellular organization and functioning. Protein-protein interaction (PPI) data produced by high-throughput techniques, such as microarray co-expression analysis and yeast2hybrid experiments, is one of the most commonly used sources for network representations of the biological data for an organism. PPI data has the nature of networks. There is more information in a protein interaction network (PIN) compared to sequence or structure alone. A protein in a PIN is annotated with one or more functional terms. Gene Ontology (GO) [1] is a structured and controlled vocabulary for describing gene and protein products and as such is an important step in the computational protein function prediction. GO term sets associated with proteins defined as interacting in a PIN can be used to compute the so called semantic similarity of the interacting proteins. Several metrics have been proposed to calculate protein semantic similarity in the context of the GO [2]. We use the concept of semantic similarity to map the PIN in a semantic metric space.

The protein function prediction can be seen as the process of understanding the proteins context in the PIN. The simplest view of a PIN is an unweighted, undirected graph in which every protein is represented with a single node and interactions between two proteins are represented with edges between the corresponding pair of nodes. This representation can be augmented by taking into account the content and/or topology of the graph and thus perform a more informed decision on the potential function(s) of an unknown protein. Clustering is perhaps the most common approach for global network analysis, and is frequently applied to uncover functional modules and protein complexes, and to infer protein function.

Traditional graph-based clustering methods use different metrics to partition an unweighted PIN and usually produce few giant core clusters with many tiny ones [3]. First step towards improvement has been made by weighting the PIN based on its topological properties [4-7]. Integration of additional information for PPI data within the PIN in order to produce weights is used in [8]. More sophisticated weighting schemes combined with optimization techniques are used in [9-11]. [9] uses neighborhood density to assign weights to nodes and then finds densely connected regions by outward traversal starting from a high weight protein. The numbers of inter- and intra-cluster edges are used in [10] to compute a cost function optimized by a local search algorithm. [11] combines random walks and Markov chains to simulate a flow on the graph during which currents in edges are strengthened or reduced which leads to partitioning in the graph.

More recent approaches exploit semantic similarity measures based on GO between pairs of proteins within the PIN. PROCOMOSS [12] uses a multi-objective evolutionary approach in which graphical properties as well as biological properties based on GO semantic similarity measure are considered as objective functions for detecting protein complexes in a PIN. CSO [13] performs clustering based on network structure and ontology attributes similarity on GO attributed PINs. Modularity optimization algorithms are employed on a semantically enriched PIN in [14]. Results presented in these papers are proof of the significance of the PIN content in clustering performance and function prediction.

In this paper we present a novel approach for protein function prediction based on homogeneity optimization in the PIN. The first step is mapping the PIN graph to a new semantic metric space, using the network content. The partitions of the graph are produced by iteratively increasing the ratio between the average distances between clusters and average distances within clusters. Once an optimal solution is reached the clusters are used to perform function prediction using cluster based function enrichment.

The rest of this paper is organized as follows. Section 2 gives the overview of our approach. Section 3 describes in detail each step of our method. The results and discussions are presented in Section 4, and finally this paper is concluded in Section 5.

II. Approach Overview

Fig. 1 gives a schematic overview of the proposed approach to protein function prediction. As can be seen the first intermediate result is the weighted PIN graph. In order to accomplish this we first need to define the PIN dataset which will be used in the function prediction process.

We conduct our experiments on Saccharomyces cerevisiae PPI data which are compiled from a number of established datasets used in previous research on PPI. In the first step we merge the PPI datasets of Uetz [15], Ito [16], Ho [17], Krogan [18], and Gavin [19]. The merging and additional processing of the data is done as proposed in [14]. The final dataset is highly reliable and consists of 2502 proteins with 6354 interactions between them and has a total of 888 functional terms.

We first define the simplest PIN graph representation i.e. the *unweighted PIN graph* with $G_1=(V,E)$ where nodes $i,j \in V$ correspond to proteins, and edges $(i,j) \in E$ correspond to interaction between "proteins" i and j. For our data we have $|V|=2502$ and $|E|=6354$.

The simplest way to augment the unweighted graph is to add weights to each edge. We choose to weight each edge in accordance to the similarity of the proteins which the edge connects. Each of the proteins is associated with a set of terms and the similarity of a pair of proteins is a function of the similarity of their corresponding term sets. We use the Gene Ontology and we quantify the similarity as semantic similarity derived on the basis of the directed acyclic graph structure of GO. Edge-based semantic similarity is determined by measuring the length of the path between two terms in the ontology. Node-based semantic similarity compares the terms, by comparing the terms themselves, their ancestors, or their descendants. Information content (IC) is defined as a measure of how specific and informative a term is. The IC of term t is defined as negative log likelihood $-log\ p(t)$, where $p(t)$ is the probability of occurrence of t in the dataset we are using. The third type of similarity measures are the hybrid measures which combine the previous two by giving weights to the GO nodes or edges according to their type.

Once a similarity measure has been chosen, we define a new graph representation $G_2(V,E,W)$, where W is a matrix whose elements w_{ij} are the weights of the edge $(i,j) \in E$. We compute the weight matrix as a function of the similarity measure.

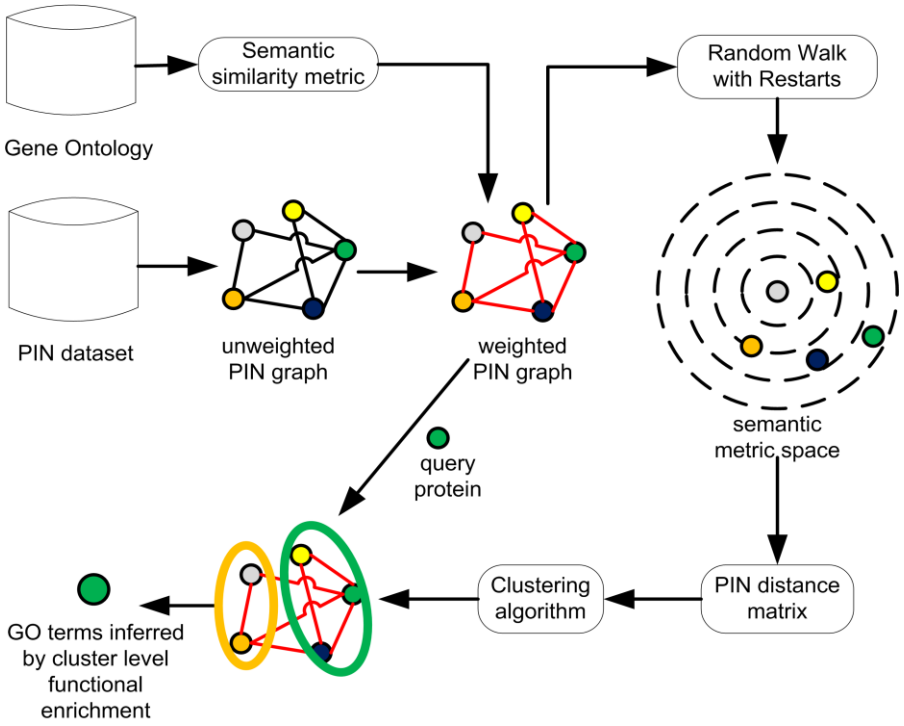


Fig. 1. Schematic overview of function prediction using our approach

In the next step we employ a Random Walk with Restarts in order to define the semantic metric space i.e. define the distance between each pair of proteins in the PIN graph. We use this metric space in the clustering algorithm in order to calculate the ratio between the semantic dissimilarity and similarity of the produced clusters. Once the partitioning of the PIN graph converges we can use the produced clusters in the function prediction of a query protein. For a chosen query protein we find its corresponding cluster and perform function enrichment within that cluster in order to assign the most likely functions to the protein.

The next section gives the implementation details of each step in our approach.

III. IMPLEMENTATION DETAILS

A. Semantic similarity measure

In a previous research [20] we conducted experiments with the same dataset comparing different semantic similarity measures and the performance they yield when used in protein function prediction. Resnik's [21] semantic similarity metric, originally developed for WordNet and later applied to GO, gives the best results. According to Resnik the similarity between two terms is the IC of their most informative common ancestor (MICA):

$$sim_{Resnik}(t_1,t_2) = max_{t\in A(t_1,t_2)}(-log\ p(t)) \quad (1)$$

with $A(t_1,t_2)$ being common ancestors set for terms t_1 and t_2. We define a similarity matrix for proteins p_1 and p_2 as $SIM = [sim_{ij} = sim_{Resnik}(t_{1i},t_{2j})]_{m\times n}$ where the two proteins are assigned the term sets $T_1 = \{t_{11},t_{12},...,t_{1m}\}$ and $T_2 = \{t_{21},t_{22},...,t_{2n}\}$. The similarity between p_1 and p_2 is:

$$sim(p_1,p_2) = max\left(\frac{1}{m}\sum_{i=1}^{m} max_{1\le j\le n}\ sim_{ij},\ \frac{1}{n}\sum_{j=1}^{n} max_{1\le i\le m}\ sim_{ij}\right) \quad (2)$$

B. Weighting the PIN graph

The weight of a protein edge has two parts: content and structure part.

The content part w_{ij}^1 of edge (*i, j*) is calculated by taking into account only the terms ("contents") assigned to nodes *i* and *j*. We compute w_{ij}^1 as follows:

$$w_{ij}^1 = \frac{max([sim(i,j)]) - sim(i,j)}{max([sim(i,j)]) - min([sim(i,j)])} \quad (3)$$

where *sim*(*i*,*j*) is calculated according to (2); *max* and *min* are the maximum and minimum value over all possible protein pairs similarities.

The structure part of the weight takes only the topology context of the nodes *i* and *j* into account when calculating weight w_{ij}^2 for edge (*i, j*). The topology context is naturally encoded in the adjacency matrix $A=[a_{ij}]$ so we define the structure part of the weight matrix $W^2 = [w_{ij}^2]$ as follows:

$$W^2 = W^1 \times A + A \times W^1 \quad (4)$$

where $W^1 = [w_{ij}^1]$ is the content part of the weight matrix. Since $a_{ij} = 0$, $\forall(i,j)\notin E$, for each w_{ij}^2 the first part of (4) is the sum of content parts of the weights between node *i* and all neighbors of *j*, and the second part is the sum of content parts of the weights between node *j* and all neighbors of *i*. In order to avoid the bias (4) has for hub proteins, it is normalized and averaged and we get:

$$W^2 = \left(W^1 \times A^1 + A^2 \times W^1\right)/2 \quad (5)$$

with $A^1 = [a_{ij} / (\sum_{n=1}^{N} a_{nj})]$, $A^2 = [a_{ij} / (\sum_{n=1}^{N} a_{in})]$, and $N=|V|$.

The final weight of an edge is defined as an average of its content and structure parts:

$$W = \left(W^1 + W^2\right)/2 \quad (6)$$

The two parts of the weight calculation are essential in the validation phase. When we choose a protein to be a query we take that its term set is empty. In order to have a senseful function prediction each edge attached to the query protein needs to be weighted which would be impossible without the structure part of the weight as defined in (5).

C. Mapping to a semantic metric space

The mapping of the weighted PIN graph into a semantic metric space is done by using the distance metric defined in [22]. We first employ a Random Walk with Restarts (RWR) on the graph. A RWR starting at node *i* is defined with:

$$P_t^i = (1-c)W^T P_{t-1} + cr \quad (7)$$

where $P_t^i(j)$ is the probability that at time *t* the random walker would end up at node *j*, *c* is the restart (return to start node) probability, and *r* is a column vector for which only the *i*-th element equals 1 and all other are 0. The restart probability defines the diameter around the start node which the walker would traverse before returning to the start node, with *c*=1 traversing only the immediate neighborhood, and *c*=0 the whole graph. In our experiments we got best results using *c*=0.5 and that is the default setting in the results section. The steady state distribution, i.e. $P_{t+1}^i = P_t^i$, is a measure of affinity or closeness of node *i* to other nodes in the network. We can now define an intermediate distance from node *i* to node *j* as:

$$d'(i,j) = 1 - P_t^i(j) \quad (8)$$

Note that since the PIN graph is irregular we have $d'(i,j) \neq d'(j,i)$. The distance between nodes *i* and *j*, based on (8), is now defined as:

$$D(i,j) = \left[d'(i,j) + d'(j,i)\right]/2 \quad (9)$$

D. Clustering by semantic homogeneity optimization

We define the clustering as the process of finding the partitioning of the graph which maximizes the quality function defined as:

$$Q = BCV / WCV \quad (10)$$

where *BCV* quantifies the between cluster variety, and *WCV* quantifies the within cluster variety. In order to mathematically define these values we first define the distance between clusters *A* and *B* using (9) as:

$$D(A,B) = \frac{1}{|A||B|}\sum_{i\in A}\sum_{j\in B} d(i,j) \tag{11}$$

We can interpret (9) and (10) as a "repulsiveness" measure for a pair of nodes and clusters, respectively. This interpretation comes from the fact that the steady state distribution of a random walk, which is in the core of the mapping to the new metric space, represents the mutual affinity of nodes. We define

$$BCV = \frac{2}{N_c(N_c-1)}\sum_{i=1}^{N_C-1}\sum_{j=i+1}^{N_C} D(C_i,C_j) \tag{12}$$

where N_C is the total number of clusters. This definition of the between cluster variety yields a maximum value of 1 when each pairing of clusters has distance 1 (in theory that can happen only if the two clusters are in complete contrast, e.g. all nodes in *A* are white, while all nodes in *B* are black). The within cluster variety (or single cluster homogeneity) for a cluster C_i is defined with

$$WCV_{C_i} = \frac{2}{|C_i|\cdot(|C_i|+1)}\sum_{v,w\in C_i} d(v,w) \tag{13}$$

The iterative optimization process is defined as follows. We initialize the partitioning with all nodes in separate clusters. We iterate through all nodes and for each check whether its joining to a neighboring cluster (a cluster where a neighbor node belongs) will improve the *BCV/WCV* ratio. When a node joins a cluster it automatically leaves its original cluster, and if the original cluster had a single node we have cluster merging. In order to speed up the algorithm convergence we iterate through the list of PIN graph nodes sorted according to their degree (cumulative weight of edges attached to the node) in a descending order. Once all nodes from the list have been traversed we start from the beginning i.e. the node with the highest degree. The algorithm converges if the whole node list is traversed and the BCV/WCV ratio does not change. We can produce hierarchical clustering (modular structure) if we use the approach of [23] where the algorithm is run multiple times, using the steady state partitioning of a previous run as an input to the next, having each cluster from one run replaced with a single node in the next run. In order to be able to perform the next run we need to define the weights of the graph produced by replacing clusters with single nodes. We use the following:

$$WWC_k = \sum_{i,j\in C_k} a_{ij} \tag{14}$$

$$WBC_{kl} = \sum_{i\in C_k}\sum_{j\in C_l} a_{ij} \tag{15}$$

where WWC_k is the sum of all weights of edges in cluster C_k, WBC_{kl} is the sum of all weights of edges between clusters C_l and C_k, and a_{ij} is the adjacency matrix element, which for the weighted PIN graph is the same as the graph weight matrix i.e. *A*=*W*. The new adjacency matrix is defined as:

$$b_{ij} = \begin{cases} WWC_i, i = j \text{ (loop on new node)} \\ WBC_{ij}, i \neq j \text{ (edge between new nodes)} \end{cases} \tag{16}$$

This new matrix is now used in the semantic homogeneity optimization process in the same way as explained before. This process can continue until we reach a single cluster, or a predefined threshold for number of clusters. The automatic stopping criterion is part of our future research.

E. Cluster based functional enrichment

Once we obtain the clusters we can predict the functions of a query protein by scoring the functions within its cluster, *K*. Each term is ranked by its frequency of appearance as a term assigned to nodes within the cluster:

$$s(j)_{j\in T_K} = \sum_{i\in K} z_{ij} \tag{17}$$

where T_K is the set of terms present in the cluster *K*, and

$$z_{ij} = \begin{cases} 1, \text{ if i-th node from K has j-th term from } T_K \\ 0, \text{ otherwise} \end{cases} \tag{18}$$

All ranks are then normalized to [0,1]. The query protein is annotated with all functions with rank above a previously determined threshold ω.

IV. RESULTS AND DISCUSSION

The effective evaluation of protein functional annotation is challenging. The lack of agreed measures and benchmarks used for assessment of the methods performance makes this task difficult. We use the leave-one-out approach when only one protein at time plays the role of a query protein and is considered as unannotated. We rank the functions in the query protein cluster as explained in the previous section. We change the threshold in the [0,1] range and compute the numbers for the four possible different classes which can occur during the assignment process: True Positive (TP) - when annotation is assigned and is part of the true annotation set, True Negative (TN) - when annotation is not assigned to the protein and is not part of the true annotation set, False Positive (FP) - when annotation is assigned but is not part of the true annotation set, False Negative (FN) - when annotation is not assigned but is part of the true annotation set.

Each annotation is assigned to one of the four classes. Using the number of annotations in each class we can calculate the following statistical measures:

TABLE I. FUNCTION PREDICTION RESULTS USING OUR APPROACH

PIN grpah	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.8174	0.6411	0.5046	0.4138	0.3344	0.2637	0.2165	0.1574	0.1281	0.0849	0.8277
	fpr	0.2831	0.1112	0.0583	0.0340	0.0237	0.0117	0.0086	0.0044	0.0030	0.0016	
Weighted Content	Sens.	0.8707	0.6810	0.5391	0.4479	0.3663	0.3153	0.2693	0.2184	0.1821	0.1401	0.8724
	fpr	0.2241	0.0905	0.0563	0.0315	0.0229	0.013	0.0076	0.0040	0.0025	0.0015	
Weighted Structure	Sens.	0.8942	0.7162	0.5517	0.4807	0.3812	0.3327	0.2986	0.2340	0.1926	0.1498	0.8903
	fpr	0.2103	0.0942	0.0513	0.0246	0.0207	0.0096	0.0069	0.0035	0.0023	0.0015	
Weighted Hybrid	Sens.	0.9013	0.7250	0.5585	0.4948	0.3891	0.3375	0.3008	0.2358	0.1944	0.1517	0.8991
	fpr	0.2009	0.0846	0.0419	0.0229	0.0192	0.0088	0.0071	0.0036	0.0023	0.0015	

$$Sensitivity\ (TruePositiveRate) = \frac{TP}{TP + FN} \tag{19}$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{20}$$

Graphed as coordinate pairs, the Sensitivity and the FalsePositiveRate form the receiver operating characteristic (ROC) curve. The Area Under Curve (AUC) of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

Table I shows the results we get when using our approach from aspect of different PIN graph representations. The simple PIN graph is the unweighted graph, while weighted content and weighted structure refer to weighted PIN graph representation which use only the content part and only the structure part of the total edge weight, respectfully. The weighted hybrid PIN graph is the one proposed and defined in this paper. The results show that the mapping to a semantic metric space significantly improves the function prediction performance.

We compared our approach with the function prediction via clustering approaches presented in [14], i.e. BGLL [23], FC [24], TimeBGLL [25], EdgeCluster [26], and Infomap [27]. Table II gives a comparison of the AUC values of different approaches using different graph representations. Table III gives the comparison of the FalsePositiveRates of different approaches when the threshold ω is zero, while Table IV gives the sensitivity values at that same threshold of ω=0.

TABLE II. COMPARISON OF AUC VALUES OF OUR METHOD WITH OTHER LEADING CLUSTER BASED FUNCTION PREDICTION APPROACHES

Representation / Algorithm	simple graph	weighted graph (content)	weighted graph (structure)	weighted graph (hybrid)
FC	0.8432	0.8871	0.9025	0.9079
BGLL	0.8446	0.8877	0.9026	0.9079
TimeBGLL	0.8653	0.8892	0.9077	0.9126
Infomap	0.8589	0.8882	0.9056	0.9086
EdgeCluster	0.8789	0.8956	0.9111	0.9189
Our approach	0.8277	0.8724	0.8903	0.8991

TABLE III. COMPARISON OF FALSEPOSITIVERATES FOR DIFFERENT APPROACHES WHEN ω=0

Representation / Algorithm	simple graph	weighted graph (content)	weighted graph (structure)	weighted graph (hybrid)
FC	0.2995	0.1929	0.1825	0.1779
BGLL	0.0447	0.1899	0.1794	0.1753
TimeBGLL	0.1724	0.1482	0.1337	0.1356
Infomap	0.0614	0.0502	0.0499	0.0499
EdgeCluster	0.1403	0.2950	0.2818	0.2731
Our approach	0.2831	0.2241	0.2103	0.2009

TABLE IV. COMPARISON OF SENSITIVITY VALUES FOR DIFFERENT APPROACHES WHEN ω=0

Representation / Algorithm	simple graph	weighted graph (content)	weighted graph (structure)	weighted graph (hybrid)
FC	0.8343	0.8693	0.8847	0.8919
BGLL	0.7166	0.8681	0.886	0.8929
TimeBGLL	0.8185	0.852	0.8762	0.8853
Infomap	0.7523	0.8053	0.8368	0.8417
EdgeCluster	0.8184	0.8999	0.9207	0.9261
Our approach	0.8174	0.8707	0.8942	0.9013

As can be seen from the results our approach is better or comparable with these state-of-the-art clustering algorithms. The AUC value as an overall performance indicator is close to the best performances. However the results in Table III and Table IV give a more specific view in the whole process. Table III shows that at ω=0 our approach is one of the most sensitive, but also one that has a high FalsePositiveRate, and we have the problem of noisy prediction. This is due to the fact that clusters generated during our approach have big coverage, The coverage of a cluster here is defined as the ratio between the number of terms present in the cluster and the number of terms present in the whole network. The lower coverage clusters lead to fewer mistakes being made during the term

assignment process, but on the downside these clusters may lack the necessary terms needed for correct and complete annotation of a query protein. This yields that when the priority is the completeness of the function prediction our approach is the same in performance as the best, while it's lacking in performance when more stringent clustering is needed i.e. the priority is getting a narrow (although incomplete) set of accurate functions.

V. CONCLUSION

In this paper we presented a novel approach for protein function prediction using homogeneity optimization in protein interaction networks. We give a general architecture of our approach which can be used with different combinations of the specific steps included in the process. We showed the process of transforming the PPI data to a weighted graph representation for the PIN exploiting both the content of the network and its topology through GO based semantic similarity measures. We then mapped this graph in a semantic metric space where we use a distance measure which quantifies the "repulsiveness" of node and/or clusters. We use this distance to iteratively optimize the semantic homogeneity of the partitioning of the graph by maximizing a quality function defined as the ratio of between and within cluster semantic variation. Our approach can work in a hierarchical manner thus discovering a modular structure within the graph. The validation of our approach and the comparison with state-of-the-art algorithms show that our approach can compete in terms of function prediction performance. It has top properties when we aim to perform function prediction which is complete i.e. captures all possible functional terms, because the clusters generated during the semantic homogeneity optimization have good coverage. More research needs to be done in terms of reducing the FalsePositiveRate and making the approach more accurate in border line conditions.

ACKNOWLEDGMENT

This work is supported by the project "Computational analysis of human brain data", financed by the Faculty of Computer Science and Engineering University "Ss. Cyril and Methodius" in Skopje, Macedonia.

REFERENCES

[1] Ashburner, M., et al (2000). Gene Ontology: tool for the unification of biology.*Nature genetics*, *25*(1), 25-29.

[2] Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., & Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, *9*(Suppl 5), S4.

[3] Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat Rev Genet 5, 101--113 (2004)

[4] Arnau, V., Mars, S., Marin, I.: Iterative cluster analysis of protein interaction data. Bioinformatics 21, 364--378 (2005)

[5] Rives, A.W., Galitski, T.: Modular organization of cellular networks. PNAS 100, 1128--1133 (2003)

[6] Friedel, C.C., Zimmer, R.: Inferring topology from clustering coefficients in protein-protein interaction networks. BMC Bioinformatics 7, 519 (2006)

[7] Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. Proteins 54, 49--57 (2004)

[8] Luo, F., Yang, Y., Chen, C.F., Chang, R., Zhou, J., et al.: Modular organization of protein interaction networks. Bioinformatics 23, 207–214 (2007)

[9] Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2 (2003)

[10] King, A.D., Przulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics 20, 3013–3020 (2004)

[11] Enright, A.J., Dongen, S.V., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 30(7):1575-84 (2002)

[12] Mukhopadhyay, A., Ray, S., De, M.: Detecting Protein Complexes in PPI Network: A Gene Ontology-based Multiobjective Evolutionary Approach. Molecular BioSystems, 8(11): 3036-3048 (2012)

[13] Zhang, Y., Lin, H., Yang, Z., Wang, J., Li, Y., Xu, B.: Protein Complex Prediction in Large Ontology Attributed Protein-Protein Interaction Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(3): 729-741 (2013)

[14] Trivodaliev, K., Bogojeska, A. and Kocarev, L. "Exploring Function Prediction in Protein Interaction Networks via Clustering Methods."*PloS one* 9.6 (2014): e99755

[15] Uetz, P., et al. "A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae." *Nature* 403.6770 (2000): 623-627.

[16] Ito, T., et al. "A comprehensive two-hybrid analysis to explore the yeast protein interactome." *Proceedings of the National Academy of Sciences* 98.8 (2001): 4569-4574.

[17] Ho, Y., et al. "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." *Nature* 415.6868 (2002): 180-183.

[18] Krogan, N. J., et al. "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." *Nature* 440.7084 (2006): 637-643.

[19] Gavin, A.C., et al. "Proteome survey reveals modularity of the yeast cell machinery." *Nature* 440.7084 (2006): 631-636.

[20] Ivanoska, I., Trivodaliev, K., & Kalajdziski, S. (2014). Protein Function Prediction Using Semantic Driven K-Medoids Clustering Algorithm. *International Journal of Machine Learning and Computing*, *4*(1), 52-56.

[21] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *Proc. 14th Int',l Joint Conf. Artificial Intelligence,* 1995.

[22] Trivodaliev, K., Ivanoska, I., Kalajdziski, S., & Kocarev, L. (2014). Novel Gene Ontology Based Distance Metric for Function Prediction via Clustering in Protein Interaction Networks. *ICT Innovations 2014: World of Data*, Advances in Intelligent Systems and Computing Volume 311, Springer International Publishing, 2015, 167-176.

[23] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

[24] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

[25] Lambiotte, R., Delvenne, J. C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.

[26] Evans, T. S., & Lambiotte, R. (2010). Line graphs of weighted networks for overlapping communities. *The European Physical Journal B-Condensed Matter and Complex Systems*, *77*(2), 265-272.

[27] Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118-1123.

The challenges of indoor movement analysis

Ivana Nižetić Kosović
Department of Applied Computing
Faculty of Electrical Engineering and Computing
Zagreb, Croatia

Abstract— **Increasing need for development of location-based services is already pushing its limits. Besides providing current object's position, there exists a demand for the analysis of previous movements in order to extract more meaningful information and knowledge about objects behaviour. The most of the studies considering movement analysis are focused on outdoor environments relying on frequent GPS data recordings. Limitation of current indoor positioning techniques leads to the lack of reliable indoor location-based services and also to the lack of research on analysis and discovery of indoor movement patterns considering semantics of the movement. In this paper, the process of indoor movement analysis is proposed and is given an overview of existing methods, trends and challenges for each step of the process. Several emerging and challenging scientific and technological topics are addressed: (1) ubiquitous positioning especially considering indoor environments; (2) storing and representing tracking data and heterogeneous environmental data (such as contextual information, environmental setup and time attributes); (3) enriching trajectories with semantic and performing exhaustive real-time queries and algorithms for analysis.**

Keywords—indoor positioning; movement analysis; semantic annotation;

I. INTRODUCTION

In recent years, research on moving objects has become more accessible and more popular due to the development of technology that allows easy tracking of moving objects, storing tracking data and performing demanding algorithms [1, 2]. Almost every person owns a mobile device that has access to wireless network and is equipped with GPS device. Still, ubiquitous tracking is often not omnipresent as expected, especially when it comes to indoor environments where signal is obstructed by obstacles, equipment properties and environmental changes and desired accuracy and precision is not reached [3].

Trends in positioning technology will push the development of more location-based services to facilitate human mobility and will enable collecting of a large amount of tracking data which can be analyzed in order to get better insight into human behaviour. Although there is a large body of literature considering movement analysis in outdoor environments, analysis of indoor movement is studied mainly in the field of robotics in which often only vision-based positioning techniques are applied.

There are numerous of applications in which reliable indoor positioning technique and adequate analysis of tracking data can be applied: commercial (service industry, housekeeping), environmental (wildlife tracking, pollution monitoring), civil or military applications [4]. Some of examples are: at the airports where passengers need assistance to find the gate and where real-time analysis of trajectories could resolve jams during the check-in and redirect passengers; in parking garages to navigate driver to a free space or to indicate parking places which are rarely used; in the museum during the exhibition to show additional information about the certain (nearest) exhibit or to show fluctuation of the people to determine the future organization of the exhibit setup.

A detailed analysis of trajectories gains insight into the behaviour of the observed object or group of objects, as well as insights into their interaction [5]. Lately, the analysis of the movement has generally shifted from raw movement data analysis to analysis of to semantically rich trajectories, which is necessary for a complete understanding of movement. Mobility analysis, both indoor and outdoor, remains one of challenging research topics considering concept of future smart cities [6].

The rest of the paper is organized as follows. The next section illustrates the process of movement analysis consisting of several steps. In the third section the overview of existing methods and future trends in indoor positioning is given. The fourth section covers the challenging topics in research on storing movement data and movement analysis. Finally, conclusion is given in section five.

II. THE PROCESS OF MOVEMENT ANALYSIS

The process of movement analysis in general consists of the following phases illustrated in Fig. **1**: starting from positioning, to tracking, semantic enrichment, analysis and modelling and finally learning about the behaviour of moving objects. In order to be able to perform efficient analysis of indoor movements all the steps need to be satisfactory and efficiently performed. Each step has both scientific and technological challenges which need to be considered in order to achieve complete and accurate process of indoor movement analysis.

Positioning is the first and very important step of the process. Inaccuracy present in this phase affects all following phases of the process. Second phase Tracking is more than collecting sequential positions - it is the process itself which tackles challenges such as storing spatio-temporal data, simultaneously tracking of multiple objects, organizing and manipulating large amount of data.

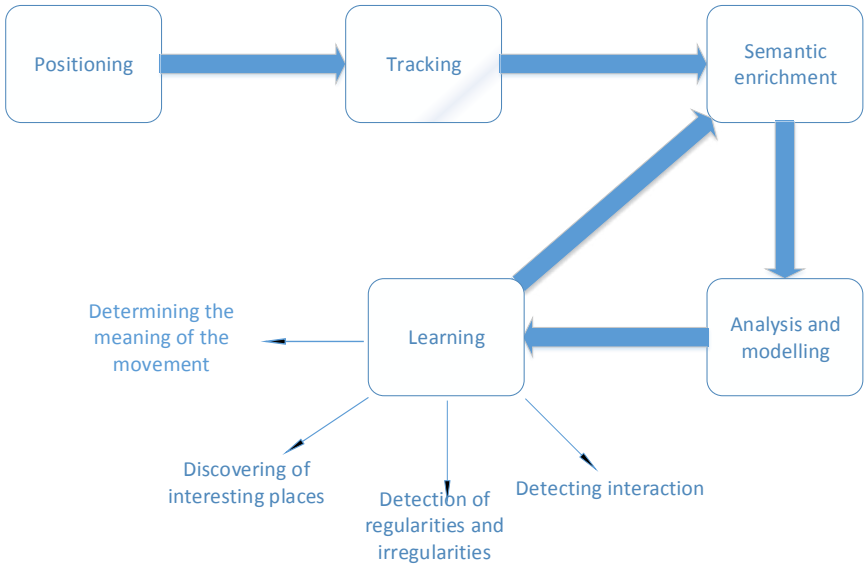


Fig. 1. Schema of a process for movement analysis

Analysis of tracking data starts with the step of semantic enrichment of data - mapping tracking data with the context such as environment, object properties or other to obtain meaningful behavioural patterns. Analysis and modelling phase consists of performance of algorithms and methods for pattern extraction, clustering interesting locations and so on, leading to the step of learning about object's behaviour (determining the meaning of the movement, discovering interesting places, detecting regularities and irregularities, and detecting interaction if considering multiple object tracking). Conclusions gained in this step can be used to enhance the process of the semantic enrichment.

III. POSITIONING AND TRACKING

As illustrated in Fig. **1**, positioning is the first step in the process of feasible movement analysis. The most of the studies considering movement analysis are focused on outdoor environments relying on frequent GPS data recordings. GPS, which is the most common outdoor positioning technology, does not work well in indoor environment due to the presence of obstacles between satellites and the receiver, such as building walls. Therefore, as a consequence of unreliable indoor positioning techniques, indoor environments are rarely studied [7].

Without having further movement analysis in mind, there are many successful techniques and technologies used for indoor positioning. The common categorization of indoor positioning techniques consists of three general categories [8]: proximity-based, triangulation and vision-based techniques. In [9] authors listed one more category: fingerprinting.

A. *Proximity-based techniques*

Proximity-based localization techniques consist of tag scanners (such as RFID tags, bar codes or similar) put at certain (known) positions. When object passes by, her device measures nearness to these positions. Such systems have in general complex setup and only local coverage. They are less accurate than other techniques. The accuracy could be increased if using multiple readings or multiple sensors to detect someone's position.

One of the most popular and one of the first indoor personal location systems using infrared technology is the Active badge [10]. The comparison of technologies for proximity-based techniques is given in [8].

B. *Triangle-based (Triangulation)*[1]

The triangle-based techniques use the geometric properties of triangles to compute location of an object. This category is further divided into the distance-based (lateration, centroid-based) and angle-based techniques (angulation) [8] [3]. Further subcategories are described in [3] [1].

Various technologies are used such as infrared radiation, radio frequency (RF) and ultrasound system technology [3]. Positioning based on Wi-Fi has some comparative advantages over competing technologies [9]. Since radio waves are able to penetrate through the walls and human bodies easily, it is applicable to almost every indoor

[1] I prefer the term "triangle-based" rather than commonly used "triagulation". "Triangulation" is confused with positioning method based on information about three angles from object to known nodes.

environment, has a larger coverage area and needs less hardware comparing to other systems. According to the Wi-Fi properties, positioning accuracy is expected to be between 1 and 5 meters [3]. In practice, achieved accuracy is up to 3m [11] [12] which is in many applications unsatisfactory.

C. Fingerprinting

Fingerprinting technique is based on creation of a so-called radiomap, a collection of pre-measured signal strengths for all visible access points in a particular location [13]. Although this method is rather simple, and does not require any specialized hardware, it has some drawbacks. The main drawback of such methods is computation of path loss exponent which is extensive and error-prone. Further on, the process of creating a radiomap is time-consuming, and since radiomaps are "static" and cannot adapt to environment changes, calibration should be repeated every time a significant change in environment occurs [3]. However, this technique is widely used and represents good solution to many applications. The most popular system based on fingerprinting technique is RADAR [14].

D. Vision-based techniques

The vision-based (or scene analysis) techniques use features of a scene observed from object's position to determine her position or position of the objects in the scene. There is plenty of literature in the field of robotics considering robot navigation in indoor environments [15, 16]. In self-localization step of robot navigation, robot learns its position mostly by matching video images from attached camera with landmark information [16]. Recognition of landmark from the visual systems still suffer from low image resolution, objects deformation, noise and variation of scales [17]. Artificial landmarks can ease the process of recognition but it is often inconvenient to put such landmark in for example offices or some public spaces.

By vision-based localization techniques is usually meant determination of location of the object carrying camera. There is another approach which can be used to track moving objects - by analyzing and detecting positions of the objects from the scene [18]. That is even more complex problem since people tend to move unpredictably and multiple people interact with each other and to the environment which can complicate their recognition. Nowadays, vision-based techniques are broadly used for UAVs flight.

E. Trends and challenges

Although many projects and systems that enable localization exist nowadays, there is still a lack of mobile applications for indoor positioning with simple usage, without additional devices needed, minimizing additional user involvement for its operation [19].

Each of aforementioned techniques has both advantages and disadvantages. To take advantages of various positioning techniques and combine them to determine location of the moving object, concepts of sensor data fusion and hybrid methods for positioning are lately introduced [20, 21].

Considering triangle-based techniques, the results of our studies showed that accuracy of indoor positioning can be increased by using custom set-up of positioning devices (such as access points in the room in Fig. **2**) and by considering orientation of the person using inside the room [19]. We achieve accuracy of about 1m. Further improvement of the quality and accuracy of the localization system could be achieved considering the context, i.e. taking into account a range of other factors from the environment, like existing obstacles (e.g. furniture), time periods between localizations or the other nearby users.

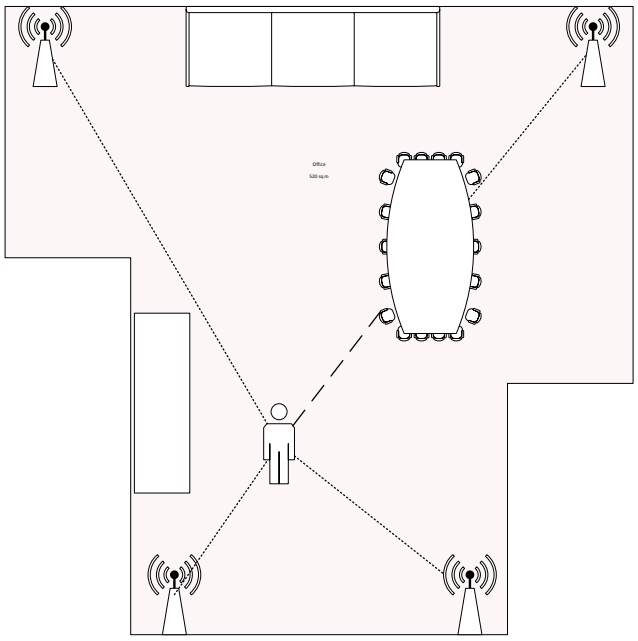


Fig. 2. The example of indoor setup

F. Tracking

Observing the sequence of positions or locations (trajectory) of certain object is tracking [22]. Tracking is the next step towards movement analysis (Fig. **1**) and represents more complex problem than positioning itself. Since positioning error could be larger that distance between two sequential positions of an object in indoor environments, obtaining meaningful trajectory could be non-trivial task. In Fig. **3** is given an example of real movement (filled line) and estimation of that movement with positioning errors (dashed line). Relatively small positioning error yields to quite unsatisfactory movement representation.

Additional information such as orientation or altitude, or previous object's position could enhance accuracy of indoor tracking. The further challenging problem is tracking of multiple objects. Multiple objects affect signal measurements and produces positioning errors. Ideas on of

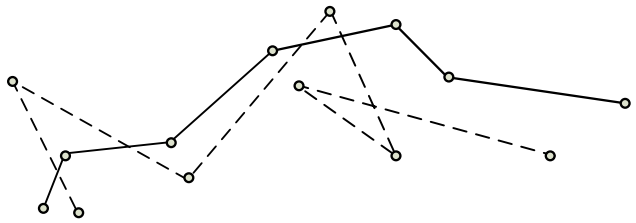


Fig. 3. An example of unsatisfactory movement estimation

localization of nodes (objects) in GPS-less crowded environments (such as shopping malls) is presented in [23].

IV. MOVEMENT ANALYSIS

Semantic enrichment, movement analysis and finally learning about the behaviour of moving objects are the final steps in the process of movement analysis (Fig. **1**).

The most of the studies considering movement analysis are focused on outdoor environments. Analysis of indoor movement is useful for many applications and research of moving objects is recently moving in that direction which will reveal unsolved problems and many challenges of indoor movements analysis.

Best practices from the research on movement analysis in outdoor environments or from research independent on application domain should be considered for indoor analysis. Still, they need to be adapted to the specifics of indoor environments.

A. *Movement patterns*

Behaviour of a moving object manifests in her movement [24]. Generally, movement patterns include any recognizable spatial and temporal regularity or any interesting relationship in a set of movement data, whereas the proper definition of "pattern interestingness" depends on the application domain [25].

B. *Novel algorithms*

Increasing number of analysis and prediction algorithms are developing: data mining techniques to extract behavioural patterns from moving objects data, clustering algorithms to detect important places, prediction techniques to model and predict moving object's future location, such as neural networks, Markov models, and specific types of dynamic Bayesian networks like hidden Markov models or Kalman filter [2] [26].

However, most of them takes only raw trajectories into analysis. By analyzing raw trajectories, only generic patterns can be extracted e.g. patterns that may be found in any form of behaviour that builds on movement, for example periodic pattern (3) or dense regions (1) and (4) from Fig. **4**. These patterns are usually insufficient to explain specific behaviour of certain moving objects [25]. On the other hand, behavioural patterns are higher-level movement patterns that correspond to a particular behaviour typical of a certain object, for example stopping at the cloakroom (1), observing the exhibit (4) or climbing the stairs (3) from Fig. **4**. To extract behavioural patterns, it is necessary to know what exactly the parameters are that define movement and what external factors influence movement. In [25] authors distinguish four groups of influencing factors which should be taken into account when analyzing the movement: intrinsic properties of the moving object, spatial constraints, environment against which the movement takes place and influences of other moving objects.

C. *Semantic enrichment of trajectories*

Over the last years, approach on moving objects moved to semantically rich trajectories [27]. The authors [28-30] emphasize the pursuit of new ways of designing prediction problems, including the enrichment of data on the movement of geographic and semantic information. There are two main approaches to identify semantic patterns of trajectories: episode view of trajectory or stop and moves view. In the first approach, authors denote an episode as a discreet time period for which the object's spatio-temporal behaviour was relatively homogeneous [30]. The trajectory is presented as set of episodes, with corresponding characteristics such as meaning, description, purpose etc. In the other approach trajectory is viewed as series of stops and moves [31] [28]. Stops are the important places of the trajectory where the object has stayed for a minimal amount of time. The remaining parts of the trajectory form the moves. Stops are associated to known geographical information. Both approaches open up new view on trajectories - semantic trajectories [27]. The inclusion of contextual (semantic) information is necessary for a complete understanding of movement [27] [32, 33].

Considering semantics, the results of our previous studies demonstrated that the inclusion of additional information about the environment in which they move greatly contributes to the understanding of their behaviour [34] [35]. We expect additional information to contribute to indoor movement analysis as well.

D. *Modelling movement data*

Advancements in positioning and tracking technology enable collection of a large amount of data. Together with the need of semantic and contextual enrichment of movement information, it opens up problem of storing and querying heterogeneous data on moving objects. Storing moving objects data with both spatial and temporal aspects is a complex problem [36]. Some notable ideas of centralized solutions for managing moving objects' data and performing algorithms were proposed. Considering wild life research the vision of centralized solution for wildlife data management is presented in [37]. In [38] authors propose a generic data model for moving objects that can apply in more than one environment and applied it to transportation model. Conceptual data model for representing semantic trajectories applied to tourism and animal movement is shown in [32] giving the baseline for future research on semantic trajectory.

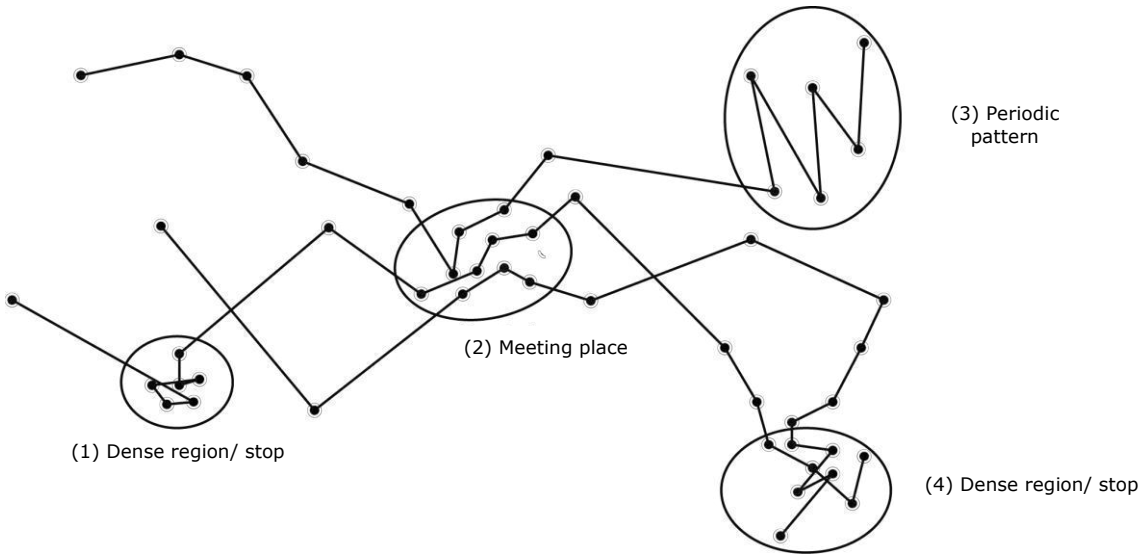


Fig. 4. Examples of movement patterns

We have also presented generic model and proposed the conceptual data model for analysis and movement prediction independent of application area and moving object type [39]. Model of moving objects and contextual information should be as general as possible and adaptive to application domain (meta-modelled). Mechanisms for semantic annotation of trajectories as well as movement patterns analysis should be independent on the application domain.

V. CONCLUSION

There is an increasing need in development of location-based services to facilitate human mobility, consisting of not only providing current position of the object but of analysis of previous object's movements and learn about objects behaviour.

Indoor movement analysis is the process with several steps where each step still has limitations and problems to be solved in order to be able to perform complete and accurate process of movement analysis. The step of indoor positioning is the weakest link in the process of analysis. Map matching and semantic enrichment of indoor movements provide basis for meaningful analysis and are also facing issues specific to indoor environments.

The following challenges summarize future research and technological topics considering indoor movement analysis:

- Development of more efficient solution to ubiquitous positioning considering indoor environments.
- Modelling of moving objects and contextual information for indoor environments - meta-model, adaptive to any domain.
- Development of the mechanism for enriching objects' movement with contextual information (with the emphasis on indoor environments).
- Development of efficient solution for indoor movement pattern recognition considering semantics of the movement.

REFERENCES

[1] Y. Liu and Z. Yang, Location, Localization, and Localizability: Location-awareness Technology for Wireless Networks: Springer, 2010.

[2] Y. Ishikawa, Data Mining for Moving Object Databases. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010.

[3] Z. Farid, et al., "Recent Advances in Wireless Indoor Localization Techniques and System," Journal of Computer Networks and Communications, vol. 2013, p. 12, 2013.

[4] I. Amundson and X. D. Koutsoukos, "A survey on localization for mobile wireless sensor networks," presented at the Proceedings of the 2nd international conference on Mobile entity localization and tracking in GPS-less environments, Orlando, FL, USA, 2009.

[5] C. Renso, et al., "How you move reveals who you are: understanding human behavior by analyzing trajectory data," Knowledge and Information Systems, vol. 37, pp. 331-362, 2013/11/01 2013.

[6] M. Batty, et al., "Smart cities of the future," The European Physical Journal Special Topics, vol. 214, pp. 481-518, 2012/11/01 2012.

[7] T. Liebig, et al., "Methods for Analysis of Spatio-Temporal Bluetooth Tracking Data," Journal of Urban Technology, vol. 21, pp. 27-37, 2014/04/03 2014.

[8] J. Hightower and G. Borriello, "Location Systems for Ubiquitous Computing," Computer, vol. 34, pp. 57-66, 2001.

[9] Y. Gu, et al., "A Survey of Indoor Positioning Systems for Wireless Personal Networks," IEEE COMMUNICATIONS SURVEYS & TUTORIALS, vol. 11, 2009.

[10] R. Want, et al., "The active badge location system," ACM Trans. Inf. Syst., vol. 10, pp. 91-102, 1992.

[11] K. Chintalapudi, et al., "Indoor localization without the pain," presented at the Proceedings of the sixteenth annual international conference on Mobile computing and networking, Chicago, Illinois, USA, 2010.

[12] Q. Dong and X. Xu, "A Novel Weighted Centroid Localization Algorithm Based on RSSI for an Outdoor Environment," Journal of Communications, vol. 9, pp. 279-285, 2014.

[13] K. Kaemarungsi and P. Krishnamurthy, "Properties of indoor received signal strength for WLAN location fingerprinting," in Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004. The First Annual International Conference on, 2004, pp. 14-23.

[14] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," presented at the INFOCOM, 2000.

[15] B. Kuipers and Y.-T. Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," in Toward learning robots, V. Walter Van de, Ed., ed: MIT Press, 1993, pp. 47-63.

[16] A. Hornung, et al., "Efficient Vision-based Navigation - Learning about the Influence of Motion Blur," Autonomous Robots, vol. 29, pp. 137-149, 2010.

[17] N. Nithya and D. Tamil Selvi, "Human Cognition and Vision Based Earlier Path Determination System for Indoor Mobile Robot Path Planning," in Progress in Systems Engineering. vol. 330, H. Selvaraj, et al., Eds., ed: Springer International Publishing, 2015, pp. 727-732.

[18] C. Wongun, et al., "A General Framework for Tracking Multiple People from a Moving Camera," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, pp. 1577-1591, 2013.

[19] I. Nižetić Kosović and T. Jagušt, "Enhanced Weighted Centroid Localization Algorithm for Indoor Environments," presented at the International Conference on Communication Networks and Applications, London, 2014.

[20] A. Campbell and T. Choudhury, "From Smart to Cognitive Phones," Pervasive Computing, IEEE, vol. 11, pp. 7-11, 2012.

[21] C. Lyu-Han, et al., "Intelligent Fusion of Wi-Fi and Inertial Sensor-Based Positioning Systems for Indoor Pedestrian Navigation," Sensors Journal, IEEE, vol. 14, pp. 4034-4042, 2014.

[22] S. Alamri, et al., "A connectivity index for moving objects in an indoor cellular space," Personal Ubiquitous Comput., vol. 18, pp. 287-301, 2014.

[23] B. Risteska Stojkoska and D. Davcev, "MDS-based Algorithm for Nodes Localization in 3D Surface Sensor Networks," presented at the The Seventh International Conference on Emerging Security Information, Systems and Technologies, Barcelona, Spain, 2013.

[24] S. Shekhar and H. Xiong, Eds., Encyclopedia of GIS. New York: Springer, 2008, p.^pp. Pages.

[25] S. Dodge, et al., "Towards a taxonomy of movement patterns," Information Visualization, vol. 7, pp. 240-252, 2008.

[26] R. Joo, et al., "Hidden Markov Models: The Best Models for Forager Movements?," PLoS ONE, vol. 8, p. e71246, 2013.

[27] C. Parent, et al., "Semantic trajectories modeling and analysis," ACM Comput. Surv., vol. 45, pp. 1-32, 2013.

[28] L. O. Alvares, et al., "A model for enriching trajectories with semantic geographical information," presented at the Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, Seattle, Washington, 2007.

[29] N. Andrienko, et al., "Basic Concepts of Movement Data," in Mobility, Data Mining and Privacy, F. Giannotti and D. Pedreschi, Eds., ed Berlin Heidelberg: Springer-Verlag, 2008, pp. 15-38.

[30] B. Guc, et al., "Semantic Annotation of GPS Trajectories " presented at the 11th AGILE International Conference on Geographic Information Science 2008, University of Girona, Spain, 2008.

[31] S. Spaccapietra, et al., "A conceptual view on trajectories," Including Special Section: Privacy Aspects of Data Mining Workshop (2006) - Five invited and extended papers, vol. 65, pp. 126-146, 2008.

[32] V. Bogorny, et al., "CONSTAnT – A Conceptual Data Model for Semantic Trajectories of Moving Objects," Transactions in GIS, vol. 18, pp. 66-88, 2014.

[33] M. Buchin, et al., "Context-Aware Similarity of Trajectories," in Geographic Information Science. vol. 7478, N. Xiao, et al., Eds., ed: Springer Berlin Heidelberg, 2012, pp. 43-56.

[34] I. Nižetić Kosović, et al., "An expert system for discovering biogeographically interesting locations from animal movement data," presented at the European Computing Conference, Paris, France, 2011.

[35] I. Nižetić Kosović and K. Fertalj, "The method to predict the animal movement considering context and expert knowledge," unpublished, 2015.

[36] R. H. Güting and M. Schneider, Moving Objects Databases. San Francisco: Morgan Kaufmann Publishers, 2005.

[37] F. Urbano, et al., "Wildlife tracking data management: a new vision," Philosophical Transactions of the Royal Society B-Biological Sciences, vol. 365, pp. 2177-2185, Jul 27 2010.

[38] J. Xu and R. H. Guting, "A generic data model for moving objects," Geoinformatica, vol. 17, pp. 125-172, 2013.

[39] I. Nizetic and K. Fertalj, "Automation of the Moving Objects Movement Prediction Process Independent of the Application Area," Computer Science and Information Systems, vol. 7, pp. 931-945, Dec 2010.

Object tracking using a mobile camera as a sensor on a hexapod robot

Dragan Nakevski, Kosta Cincaroski, Ivan Ilioski, Nevena Ackovska, Vesna Kirandziska
Institute of intelligent systems
Faculty of Computer Science and Engineering
Skopje, Macedonia
dragannakevski@gmail.com, kostacincaroski@yahoo.com, ivan.cfc@hotmail.com, nevena.ackovska@finki.ukim.mk, vesna.kirandziska@finki.ukim.mk

***Abstract*— Hexapod robots are mobile robots that are widely used for research purposes. Their construction enables them to move easily in a ground environment. As a consequence, applications where an environment needs to be explored are common for these robots. This paper presents an implementation of a custom version of object tracking algorithm on a robotic system that uses a hexapod robot. The hardware components of the robotic system include two more elements: a camera from a mobile phone and a computer. The camera is used as a sensor, while the computer is used as a mediator between the robot and the mobile phone. The presented robotic system enables the hexapod robot to "see" its environment, even though the hexapod robot itself has no sensors included. In this paper the whole robotic system built is presented. The results about the performance of the hexapod robot are presented and a discussion about the positive and negative features of the robotic application is given.**

Keywords—robotics; mobile application; object tracking; hexapod

I. INTRODUCTION

Robotics is an ongoing research area that composes mechanical engineering, electrical engineering and computer science. Computer engineers program robots to perform given tasks. Robots can be programmed to move in their environment. Different robots vary in the way they move in their environment. There are wheeled robots, flying robots, swimming robots, humanoids, hexapods and other types of moving robots.

Hexapods can be programmed to move on a flat terrain, to walk on stairs and walk on uneven field. Their construction enables them to stand stable on the ground. But the success of the hexapod's walking action strongly depends on the terrain of the environment. Some researchers are concerned with the design of a hexapod robot that can adapt to a variety of terrain [6]. Others design hexapod robots with other specific valuable features [11]. In this paper the robotic design is not investigated and for simplicity the robot terrain is considered to be flat.

In the process of interacting with the environment, the robotic system should have the ability to acquire information of its surrounding environment. That is why it has to be equipped with sensors. Programming robots to use their sensors as well, can work miracles. Robots could move in an environment with obstacles, they can avoid obstacles, track objects and follow them. Hexapod robots could also be used for environment explorations, path tracking [7], object transportation, human assistance, terrain classification [9], [10] etc. In this paper a hexapod is used for object tracking.

For the research presented here the PhantomX [1] hexapod robot is used. It is a fairly complex robot. Wireless Xbee control via computer or handheld device enables easy control of the hexapod robot. This robot is mostly used by hobbyists, but it is also used in education and research. As a hobby of many, PhantomX has been used in Mech Warfare robotic competitions [13]. On the other hand, its great features motivated many researchers to use PhantomX for scientific purposes. Some of the great features of this robot are presented in the sequel. Firstly, it is fully programmable. Next, all of the software used by PhantomX is developed by Vanadium Labs [12] and it is open source by nature. This fact has the intention of encouraging people to modify, hack and change the robot's source code any way they see fit.

More input/output pins are available for the robot's communication with external devices. As a result arms, grippers, cameras, tilts and even computers can be added as extension to this robot. This would enable many additional features for the robot. In [8] the PhantomX was used with a Microsoft Kinect sensor, board camera and 2D laser range finder. In the approach presented here a camera from a mobile phone was used as a sensor for the hexapod robot.

In the paper presented here a robotic system used for object tracking is described. In the next two chapters the hardware and software components of the robotic system are explained. Next, the results are shown and a discussion about the challenges of the system is made. In the end a conclusion is given.

II. HARDWERE COMPONENTS

A. *Phantom hexapod overview*

As mentioned, the PhantomX hexapod robot is used in this research. It has 6 legs that enable it move. The body of the hexapod is a circle metal construction that gives the robot stability of its construction.

Advanced dynamixel AX Series robot servo motors are used for movement. Each leg has 3 degrees of freedom. Each

leg has 3 servos that make a total of 18 AX servo motors. These servos have integrated sensors built in, that return the current position, voltage and torque of the motor. A high quality LiPo battery is used to power the servos.

B. Arbotix controller introduction

PhantomX is controlled with the ArbotiX controller. The ArbotiX robocontroller is an Arduino compatible microcontroller board, designed to control small-to-medium sized robots, especially walkers that use Bioloid AX-12 and Bioloid AX-18 servos. In the hexapod scenario, AX-18 servos are used.

This microcontroller has an advanced inverse kinematics driven gait engine (NUKE, Nearly Universal Kinematics Engine). The controller runs an inverse kinematics that tells the dynamixel network where, when, and how to position the hexapod legs.

The ArbotiX has a 16Mhz ATMEGA644P microprocessor with 64KB flash ROM and 4KB RAM. It has 2 serial ports, 1 dedicated to the Bioloid servo controller [4] and the other to the XBEE radio. ArbotiX is 2.8"x2.8" board with mounting holes designed to match many Bioloid configurations.

The robot's navigation goes through the controller. The ArbotiX accepts navigation commands using the 'Commander Protocol'. It is a simple serial protocol that allows control of the robot's movement in its environment. The ArbotiX Commander is a ready-to-run handheld controller that outputs this protocol. Using a pair of Xbee wireless modules, the robocontroller and the commander can communicate. This same protocol can also be used to control the hexapod robot from a computer using USB Xbee interface. This research supports the aforementioned approach to control the hexapod from a computer application.

A total of 32 input/ouput pins can be found on ArbotiX robocontroller. Eight of these can function as analog inputs and eight of which can function as digital inputs/outputs. As a consequence, there is an opportunity to connect a camera to the ArbotiX controller. But unfortunately, directly connecting a camera to the ArbotiX robocontroller performs poorly due to the lack of memory on the controller. The robocontroller's memory storage is not enough for image processing.

In order to overcome this problem in our task to implement object tracking on PhanotmX, a camera sensor that is connected to the PhantomX through another device was used. More precise, a camera sensor form a mobile phone is used by the hexapod robot via a computer. In the next section this is discussed in more detail.

C. Completing the system

A schematic view of the connections between the hardware components of our system built is shown on Fig. 1. The main component shown on Fig. 1 is the hexapod. Through a wired connection the hexapod robot is connected to a computer. The connection is one-sided. The information flow comes from the computer to the hexapod robot. Using serial communication the computer sends signals to the robot. The robot's actuators receive these signals and the robot performs some action accordingly.

On one side, the computer communicates with the hexapod robot and on the other side, the computer communicates with a mobile phone (Fig. 1). Information from the mobile phone is sent to the computer via Bluetooth. The reason why the computer is used as mediator between the mobile phone and the robot is because of the lack of hardware support on the mobile phone to communicate with the robot directly using serial communication. This especially comes as a shortage if we want to use the benefits of the wireless communication using Xbee devices that the robot actually supports.

The mobile phone is used not only to capture images from the camera in real time, but also to perform the image processing itself. Machine learning algorithms for object tracking were implemented on the mobile phone application created to perform the presented tasks. The output of the mobile application is sent to the computer via Bluetooth. The software implementation of the mobile application is explained in more detail in the next section.

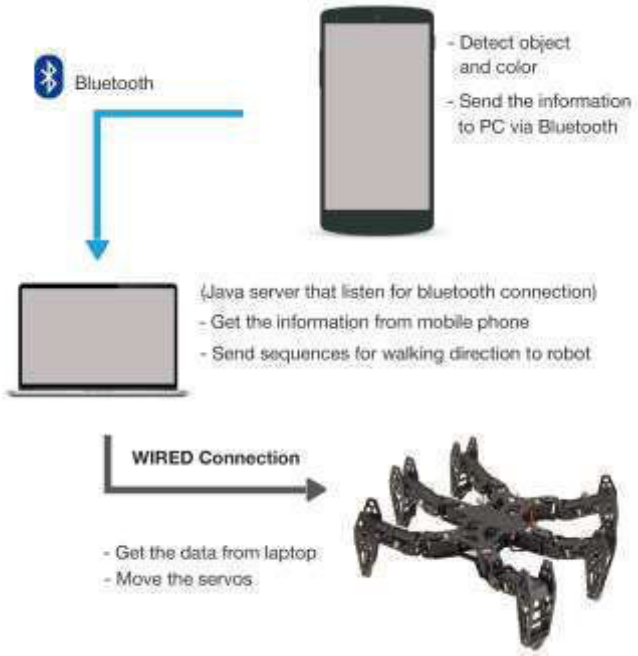


Fig. 1. Schematic view of the connections between the hardware componetns.

To resume, the information from the camera is processed by a software application. The computer receives this information by a Java server installed that listens for the Bluetooth connection information from the mobile phone. From this information a computer application sends signals via serial wired communication to the hexapod robot. These signals control the robot's movement and trigger the robot's actuators to perform as programmed.

These three components (hexapod robot, computer, mobile phone) compose the robotic system that is used for object tracking.

III. SOFTWERE IMPLEMENTATION

As mentioned in the previous sections, in the presented robotic system two software applications were created to complete the task of object tracking. One is an Android mobile application and the other is a Java based computer application used for intercommunication between the mobile phone and the hexapod robot (Fig. 1).

A. Mobile application

A mobile application is built for colored object detection. More specific, the application is used to detect circled objects that are colored in red. This mobile application is an Android application, written in native Java code using Android libraries. Three different phases are executed sequentially with the application. In the first phase an image from the camera of the mobile phone is captured. Next, image processing on the image is performed. In the final phase the output of the application is generated and sent via Bluetooth connection (Fig. 1) to the computer. The output of the mobile application is the position of the detected object. After the third phase, the first phase is executed again.

In the second phase, where the processing of the image is done, algorithms for image processing were used. Image processing is performed using two distinct modules. One for color detection and the other for object detection. Both modules are combined to detect a circular object that has one specific color. The red color is chosen for the experiment.

The image processing was built to meet the requirements for our problem of object tracking. On the user phone the OpenCV Manager application, an Android service targeted to manage OpenCV library binaries on end user's devices, needs to be installed. OpenCV Android SDK is also used as a wrapper for OpenCV Java API. For color detection and object detection, the application uses native java classes and OpenCV methods. Because there are a lot of Android devices with different cameras, we must use a range of red colors for better detection. The HSV thresholding is used for segmentation of the red color from an image. It converts an input image from one color space to another. We choose the HSV transformation, because there were a lot of problems with color detection with different light source, and HSV shows as best. For circle detection the application also uses the OpenCV library and its function HoughCircles. The HoughCircles function uses the HoughTransform technique. The purpose of the technique is to find imperfect instances of objects within a certain class of shapes by a voting procedure. [3].

B. Computer application

The communication between the mobile phone and the robot is enabled via a computer. A computer application is created to establish the connection between this two in the direction from the mobile phone to the robot.

A Java application is created and programmed on the computer. It listens for Bluetooth connections and accepts data with the connected devices. The application uses the BlueCove library [2] for setting Bluetooth connection with a device. After the computer is paired with the device, the device, and in this case the mobile phone, sends data to the computer. To establish the connection with the robot, Java Communications API is used to send information via COMM ports to the robot. Based on the information gathered from the phone, the application is sending information to the robot so it can move to the desired pose and position. In the next section more about the robot control is described.

C. Robot control

A number of software programs have been designed to work with ArbotiX and help to control all servo motors of PhantomX. The main tool that is used for robot control is PyPose[5]. This tool is used to produce sequence of poses. It enables defining sequences of poses for each robotic movement quite easily.

PyPose is written in Python and uses pySerial to communicate with an ArbotiX board. PyPose is a "pose and capture" software. This means that the robot can make any pose by controlling each servo of the hexapod robot. Also PyPose can register and save any pose done by the robot. Using PyPose, for each movement or action, one can make a PyPose sketch that represents a sequence of poses. The PyPose sketch contains information about the position of each robot's servo in each pose and also the time frame for which the sequence should be performed. This information can be automatically saved in a C header file. This file is the output of PyPose and this file can be used by the ArbotiX robocontroller to control the robot so it performs a wanted movement.

However, in the application presented here generated sequences are not used. Instead, a protocol that is used in the wireless Xbee control of the robot is used in the wired communication between the computer and the robot. Messages contain information about the control of the robot, so a specific movement is performed. These messages are sent through a specially designed protocol that uses 8 byte data chunks. The robocontroller is programmed to receive this message. The Commander library translates the received message into a format that uses variables and that can be read by NUKE (The Nearly Universal Kinematics Engine). The Commander library actually represents a middleware library between the data received and the code generated for the robot's movements.

NUKE is used by the robocontroller to calculate the robot's internal states that should be used to control the robot so that the robot is controlled according to the data received by the computer. Between each two consecutive poses of the robot's movement, using kinematics theories, NUKE calculates the internal state of each servo for each moment. These calculations are used to control the robot to perform a given action.

Using the Bioloid library the actual movement of the servo motors is done based on the calculations NUKE has performed. BioloidController is a Bioloid bus library that has a full interaction with the AX servos of PhantomX. Using Bioloit library the hexapod robot makes the actual movements as defined in the message sent by the computer. The hexapod robot can move forward, move backward, move left or right for a given angle and even rotate for a specific angle.

IV. PRELIMINARY RESULTS

In this research a robotic system for object tracking is done using PhantomX hexapod robot. The hexapod itself does not have a sensor of any kind. In order to make the robot act to track objects, sensors with which the robot could perceive the environment should be installed. Due to the lack of sensors on the robot, the robot is extended with a camera from a mobile phone. This mobile camera, which is positioned on the hexapod itself, enables the robot to sense the robot's environment in real time.

The communication between the mobile camera and the robot is enabled by a mediator – a computer. For that purpose, a custom built protocol is created. This protocol is used for sending information from the mobile phone to the computer stating the position of the object that is of interest.

The proposed construction of the robotic system (Fig. 1) enables a stable flow of the signals from the mobile phone camera to the robocontroller and the robot itself. A mobile application, a computer application and a robocontroller program is done to enable this communication in a software level. On Fig. 2 the robotic system is shown.

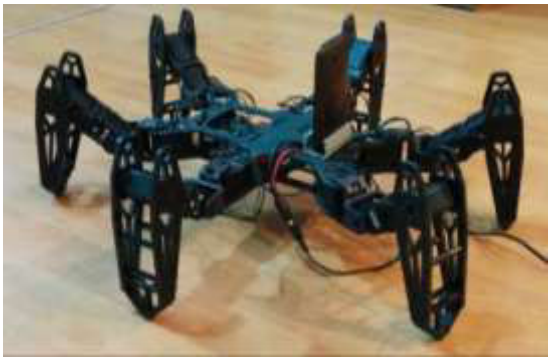


Fig. 2. A picture of the robot with attached mobile device.

The preliminary experimental results showed that the mobile application detects a red circular object. This may also vary from the surface and the image quality taken from phone's camera. If the surface is slippery, the robot won't be stable and the detection won't be good enough. The information from the object detection is processed by the computer application and provides good controls for the robot's next movement.

There were several challenges in the construction of our robotic system used for object tracking. Setting up the components of the system and enabling the exchange of information between them was the biggest challenge. However, at the end a successful communication between the PhantomX, a mobile phone and a computer was established.

Other challenge was to determine the correct position of the object. Several problems arise in the image processing on a picture from the camera. The quality of the camera and the noise in the picture can bring big uncertainty in this system. This should be taken care in the next step of the evolution of this robotics system.

V. CONCLUSION

In this paper the hardware and software implementation of a robotic application for object tracking are described. It is a complex robotic system consisted of more components.

An important robotic feature is added to PhantomX. Indeed, the intercommunication between this robot, a computer and a mobile phone enables the robot to visually sense the environment. One application that uses the visual information is implemented in this research. The goal of the hexapod robot is to track an object in one color. This goal was accomplished with the robotic system built (Fig. 2).

Many other applications can be created using the visual signal from the mobile camera. Examples include environment exploration, robot localization, obstacle avoidance etc. Some can be implemented in the future work.

Next, feedback information from the robot servo motors can be used to improve these applications. The robot could be programmed to move and adjust to different environments where it may be placed.

Furthermore, another form of communication between the mobile phone and the PhantomX could be used in order to exclude the wired communication between the robot and the computer that disables the robot's mobility.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University.

REFERENCES

[1] Trossen Robotics. *PhantomX AX Hexapod Kit*. [Online]. Avalilable: http://www.trossenrobotics.com/phantomx-ax-hexapod-mk1.aspx

[2] BlueCove (2008, December 25). *BlueCove Library*. [Online]. Available: http://www.bluecove.org/

[3] OpenCV (2015, February 25). *OpenCv Library*. [Online]. Available: http://docs.opencv.org/modules/core/doc/intro.html

[4] Google Project - Wiki (2010, Febryary 19). *BioloidController*. [Online]. Available: https://code.google.com/p/arbotix/wiki/BioloidController

[5] Google Project - Wiki (2010, May 4). *PyPose*. [Online]. Available: https://code.google.com/p/arbotix/wiki/PyPose

[6] Li. Junke, Y. Wand and T. Wan, "Design of a hexapod robot", Proc. Of CECNet, pp. 1768-1771, 21-23 April 2012.

[7] W. Jeong, H. Kim, S. Kim and B. Jun, "Path tracking controller design of hexapod robot for omni-directional gaits", Proc. Of Control Conference, Istanbul, pp. 1-6, 23-26 June 2013.

[8] L. Weydert, "Scurrying Robot", Bachelor thesis, Dept. Auto. Systems Lab, Swiss Federal Institute of Techonology, Zurich, Switzerland, 2013.

[9] G. Best, P. Moghadam, N. Kottege and L. Kleeman, "Terrain classification using a hexapod robot", Proc. of Australasian Conference on Robotics and Automation, Sydney Australia, 2-4 Dec 2013.

[10] A. Schmidt and K. Walas, "The classification of the terrain by a hexapod robot", Proc. Of CORES, Yichang, vol. 226, pp. 825-833, 2013.

[11] W. A. Lewinger, "Insect-inspired, actively compient robotic hexapod ", Dept. El. Engineering and Comp. Science, Western Reserve University, May 2005.

[12] Vanidium Labs. [Online]. Available: http://www.vanadiumlabs.com

[13] Mech Warfare Robot Competition. [Online]. Available: http://mech-warfare.com

Complexity of software navigation and implementation of small UAVs

Igor Sekuloski
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

Nevena Ackovska
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

Abstract—**The small Unmanned Aerial Vehicles (UAVs) are a viable alternative to manned aircraft for a variety of applications, including environmental monitoring, photo and video production, agriculture, and surveying. The challenge mainly lies in the development of intelligent UAV systems that will have appropriate attitude and navigate in unpredictable situations, like different weather conditions. Those factors tend to affect, more or less, the entire system's reliability, flight time and range, stability of flight and maneuverability. This paper gives a short overview of small multirotor UAVs, principles of their hardware construction, software integration, navigation and controlling. The practical and experimental part consists of custom built of an experimental quadcopter based on the open source platform ArduCopter, as a part and first phase of an ongoing work in progress. This ongoing research includes theoretical and practical approach through building and use of experimental quadcopter, testing and simulation methods under different weather conditions. The paper presented here intends to focus on building an experimental quadrotor, its evolution, challenges the authors have faced during that process until reaching fully functional UAV. The paper will also try to explain the main engineering and scientific aspects of small UAVs, entering only basically into complexity of influence of weather conditions on quadrotor flying performances.**

Keywords—UAV; multirotor; software; navigation; flying; custom building; quadcopter; weather influence; calibration; VTOL

I. INTRODUCTION

During the past years, research and development of small unmanned aerial vehicles (UAV) got the attention of many academic institutions and military organization across the world [1], mainly because of its potential application for different civil, commercial and military purposes. The UAVs are equipped with necessary data processing units, sensors, automated control and communication systems, capable of performing controlled and autonomous flights.

The challenge mainly lies in the development of intelligent UAV systems that will have appropriate attitude and navigate in unpredictable situations, like different weather conditions, unexpected obstacles, enemy attacks or device failures. In general, the problem is very complex and many factors have to be planed and considered. The UAV's control systems need integration of not only input/output data, but also high level functionality for planning, making decision and automation of particular tasks and missions.

Boosted by technological development and miniaturization, high degree of commercially available necessary hardware components, development of different software platforms including open source control software [2], the small multirotor UAVs became popular among hobbyists and also business community, mainly due to relatively low price, their agility, hovering capabilities, ease of control, wide spectrum of potential application in areas of photo/video production, agriculture, mining and oil industry, archeology etc.

Small multirotor UAV flying, its stability and attitude in the air may be extremely dependant on weather conditions. The main weather factors that influence the flight of small UAVs are wind and turbulence, air temperature and density, precipitation and icing. Those factors tend to affect, more or less, the entire system's reliability, flight time and range, stability of flight and maneuverability.

The next section gives an insight of the problem.

II. FLYING UNDER DIFFERENT WEATHER CONDITIONS

The effect and negative impacts of some weather factors are greater to small UAVs compared to other types of UAVs. By nature of their generally smaller size and weight, small UAVs are adversely impacted by a wider range of raindrop size. The only current solution to precipitation is procedural: avoid flying in moderate or heavy precipitation, while in light precipitation use waterproofed frame and motors, and do not use wooden propellers [3].

In comparison to manned aircraft, the impact of wind also tends to pose a greater challenge to UAVs in general, and in particular to smaller members of the UAV family [4]. Some of the wind conditions may even hinder the flight of small UAVs. High wind speeds play a role in all phases during flight, take-off and landing (crosswind), but also inflight in the form of turbulence. The smaller the UAV, the more it inherently suffers from wind and turbulence. In practice, there is evidence on this susceptibility resulting in jumpy video (if the sensor is unstabilized), erratic flight (heard as sudden changes in propeller sound), possible loss of link (signal

dropout), loss of control and potential crash when an upset exceeds the pilot's or autopilot's ability to recover. Jumpy video and erratic flight are often mitigated by designing small UAV by mounting anti-vibration bumpers and mounting imaging sensors in stabilized gimbals and replacing fixed with variable pitch propellers, respectively [4].

The influence of weather conditions on PID tuning is remarkable. For example, personal experiments with small multirotors has shown that PID values tuned in calm weather conditions are not adequate for stable flight in windy conditions and some additional fine PIDs tuning is necessary.

This research paper gives a short overview of small multirotor UAVs, principles of their hardware construction, software integration, navigation and controlling. The practical and experimental part consists of building quadcopter based on the open source platform ArduCopter.
The research work includes theoretical and practical approach through building and use of experimental quadcopter using different research and simulation methods for inspecting and changing the algorithms, trying to contribute the worldwide research of software solutions for UAV's navigation and validation of the influence of simulated weather conditions and very often unpredictable weather conditions of the real environment.

III. CUSTOM BUILDING SMALL UAVS

For this specific research several multirotors have been custom designed. The first practical experience and work in building multirotors began in 2009 and the challenge from the very first moment was to build a quadcopter from scratch.
It was a challenge to make good frame, geometrically accurate, appropriate lift-to-weight ratio, good positioning of the motors, strong, but also lightweight. The first platform (Fig. 1) used was Aeroquad [5], the controller was handmade with Atmel processor, additional components and PCB designing. Building the first quadcopter has started with many trials and errors. The quality and precision of the sensors advanced over time, resulting in upgrading the platform and even changing the platform to another like Wiicopter [6], and finally Ardupilot [7]. The result of this progress was building several prototypes of multicopter models, as the last one is octocopter which is capable of carrying payload up to 8 kg.



Fig. 1. First version of the experimental quadrotor (2010)

Compared with today's resources found online, 5 years ago, there were limited resources, both theoretical and practical, since this technology was in its infancy, which lead to core difficulties of building multicopter, issues like designing and building frame of appropriate quality, dealing with strange vibrations from unknown sources, fine Proportional Integral Derivative (PID) controller tuning, calibrations of the sensors, motors and propellers.

IV. THE RESULTING COPTER

The practice showed that it is a big challenge to implement small multirotor from the ground up, not only because of its complexity, but also due to the poor availability of the required parts for its construction, relatively high prices of the parts and the time needed for delivery. The complexity is also in planning for the proper hardware parts, many of which are critical elements, like choosing the right materials for construction, motors, Electronic Speed Control (ESC) modules, propellers, etc. Also additional work is needed for balancing the propellers and motors, calibrating the ESCs.
The software platform which the multicopter is based on, is solely a constructor's choice, especially today, when there are plenty of open source platforms for multirotors. The selection is based on the needs of the multicopter, which will offer the features needed for its purpose and also for the further upgrades. But firstly and most important is the reliability of the software that will provide stable and safe flights and fail-safe features. Stable and safe flying is needed to fine tune lots of parameters, sometimes with experimentation. The practical experience and the experimentation showed that the weather conditions influence the flight characteristics significantly. Some of the multicopter characteristics are shown on Table 1.

TABLE I. MULTICOPTER CHARACTERISTICS

Parameter	Value	Measurement unit
All-up Weight:	**2100**	***g***
Additional. Payload	1542	g
Maximum tilt	2100	°
Maximum speed	40	km/h
Voltage	13.48	V
Capacity	5000	mAh
Energy	74	Wh
Flight time	3.8	minutes
Mixed Flight Time	9.1	minutes
Hover Flight Time	12.5	minutes

V. COMPLEXITY OF SOFTWARE CONTROL AND NAVIGATION

The UAVs have to somehow cope with the problem of navigation and control. In small flying robots it becomes particularly acute because of static instability and, with some morphology, the necessity of motion for remaining aloft. A term controller designates the overall hardware and software aggregate which generates real-time behavior of the UAV. A good part of this aggregate is used up by sensor acquisition and preprocessing to realize an onboard state estimation

capabilities with an adequate degree of correlation to environmental conditions.

A quadrotor is a highly operative flying vehicle (Fig. 2). However, a quadrotor is an unstable platform and impossible to fly in a full open loop system. Therefore, advanced embedded systems are widely used for controlling these platforms. The dynamics of a flying vehicle are very complex, and controlling nonlinear plant is a problem in both practical and theoretical terms. The attitude controller is an important feature since it allows the vehicle to maintain a desired orientation, and hence prevents the quadrotor from crashing when the pilot performs the desired maneuvers. Improved performance in the new generation of Vertical take-off and landing (VTOL) vehicles is possible through derivation and implementation of specific control techniques incorporating limitations related to sensors and actuators. The most well - known approach to the decoupling problem is based on the Nonlinear Inverse Dynamics (NID) method, which can be used if the parameters of the plant model and external disturbances are exactly known [9]. Altitude here is used as some sort of measure for the distance of the UAV from the ground. For vehicles that are able to hover, this is the most important variable to be controlled. It enables the vehicle to take of from the ground and either maintains a certain distance to the ground or to maintain a certain rate of ascent and, more generally, keep clear of a large object or bounding surface. In an environment where the ground is not a flat surface, the question does become very intricate. Several factors are interacting, such as overall type of desired behavior, desired speed relative to the environment to objects in the environment and the spatiotemporal dynamics of sensor responses. Several different behaviors may be recognized: terrain following, absolute altitude stabilization, and vertical rate control. The former two emerge from the sensor response dynamics. Depending on whether the UAV is within a range of distance from an object or surface, a valid range measurement overrides and resets the absolute altitude information. Vertical rate on the other hand can be directly extracted from an onboard camera using optic flow estimation.

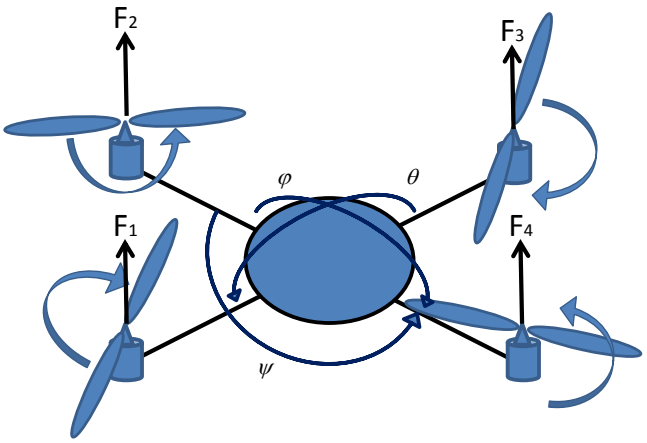


Fig. 2. Quad-thrust multirotor wit

To illustrate complexity of controlling an UAV and difficulties in enabling a stable flight, a brief mathematical description of the quadrotor nonlinear model will be given in the following sub – section.

A. *Quadrotor Rigid Body Model and Attitude estimation*

The aerial vehicle consists of a rigid cross frame equipped with four rotors as shown in Fig. 2. The two pairs of propellers (1,3) and (2,4) turn in opposite directions. By varying the rotor speed, the lift force can be changed and motion created. Thus, increasing or decreasing the four propellers' speeds together generates vertical motion.

Changing the 2 and 4 propellers' speed conversely produces roll rotation coupled with lateral motion. Pitch rotation and the longitudinal motion result from 1 and 3 propellers' speed being conversely modified. Yaw rotation is a result of the difference in the counter - torque between each pair of propellers [9].

The quadrotor is a six degrees of freedom system defined with 12 state and control vectors. Six out of 12 states govern the attitude of the system. These include the Euler angles (Φ, Θ, Ψ) and angular rates around the three orthogonal body axes. The other six states determine the position (x, y, z) and linear velocities of the centre of quadrotor mass with respect to a fixed reference frame.

The dynamic model is derived using Euler - Lagrange formalism under the following assumptions:

- the frame structure is rigid,
- the structure is symmetrical,
- the centre of gravity (CoG) and the body fixed frame origin are assumed to coincide,
- the propellers are rigid,
- thrust and drag are proportional to the square of the propeller speed.

The main role of the inertial sensors is to provide attitude and velocity estimates to be used in the closed-loop stabilizing control algorithm. Inertial navigation is performed using the following principles and facts [10].

- The accelerometers and gyroscopes can be used together as inputs in the motion dynamics, which, once integrated, give the estimates of attitude angles and velocities.
- Noting V = [u v w] ^{T} the vector velocity of the center of gravity of the Inertial Measurement Unit (IMU) in the body frame, Q=[Φ Θ, Ψ]^{T} the Euler angles (roll-pitch-yaw), i.e. the angles between the inertial frame and the body, and Ω = [p q r]^{T} the angular rate of turn in the body frame, and F the external forces, the governing equation is [10]:

$$\dot{V} = -\Omega \times V + F \tag{1}$$

$$\dot{Q} = G(\Omega, Q) \tag{2}$$

$$G(\Omega, Q) = \begin{bmatrix} p + (q\sin\phi + r\cos\phi)\tan\theta \\ q\cos\phi - r\sin\phi \\ (q\sin\phi + r\cos\phi)\sec\theta \end{bmatrix} \tag{3}$$

The inertial sensors suffer from the following flaws.

- the accelerometers, which do not measure the body accelerations but its acceleration minus the gravity, expressed in the body frame, are biased and misaligned. Classically, we consider that the accelerometer signal Y_V has a bias B_V

(independently on each axis) and suffers from additive white noise μ_V. The measurement equations are [10]:

$$Y_V = F - Rg + B_V + \mu_V$$

where g stands for the gravity acceleration and R is the rotation matrix from the inertial frame to the body frame [10]:

$$R = \begin{bmatrix} \cos\psi\cos\theta & \sin\psi\cos\theta & -\sin\theta \\ \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi & \sin\phi\sin\psi\sin\theta + \cos\phi\cos\psi & \cos\theta\sin\phi \\ \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi & \cos\phi\sin\psi\sin\theta - \sin\phi\cos\psi & \cos\theta\cos\phi \end{bmatrix} \quad (4)$$

- the gyroscopes are biased and noisy. The measurement equation is [10]:

$$Y_\Omega = \Omega + B_\Omega + \mu_\Omega \quad (5)$$

where μ_Ω is a noise vector, and B_Ω is a bias vector.
To overcome the preceding fundamental issues, various algorithms for attitude estimation can been considered: Kalman filter, extended Kalman filter or complementary filter in both linear and nonlinear implementations.

VI. Implementation of Small UAV – Experimental Quadrotor

The following subsections describe the specific sub-systems and their implementation in order to achieve useful and manageable small UAV. The subsection A describes the airframe and the hardware components, the B subsection describes the electronic sub – system, the sub-section C elaborates on the challenges in the navigation and control, and the last subsection focuses on the sensing related to vibrations.

A. *Airframe and other hardware components*



Fig. 3. The resulting experimental quadrotor

The experimental quadrotor's airframe satisfies a challenging range of requirements. Foremost, the aircraft has a strong and lightweight airframe, thus maximizing the potential payload. The central plates are double carbon fiber (CF) plates with four booms made from aluminium. A mount was included for small video camera, in addition to the control electronics and batteries used to power the UAV. The quadrotor uses "slow fly" propellers design for better thrust performances and more stable flights. The GoPro camera has stabilized gimbal mount. The complete experimental quadrotor completed as part of this research is pictured in Fig. 3.

B. *Electronics*

A 6-channel, 2.4 GHz RC radio system is used for manual ground control, and Turnigy components were used in the power system with four 2217 brushless electric motors and 30A ESCs.

The ArduPilot 2.5 controller board was used as the main platform for the flight control system due to the functionality and availability of open-source support. The ATmega2560 is a low-power CMOS 8-bit microcontroller based on the AVR enhanced RISC architecture. By executing powerful instructions in a single clock cycle, the ATmega2560 achieves throughputs approaching 1 MIPS per MHz allowing the system designer to optimize power consumption versus processing speed. Based on the Arduino open-source hardware platform, ArduPilot is a full-featured autopilot based on ATmega2560 microprocessor, using an Inertial Measurement Unit (IMU) for stabilization and GPS for navigation. The APM hardware is seen in Fig. 4.

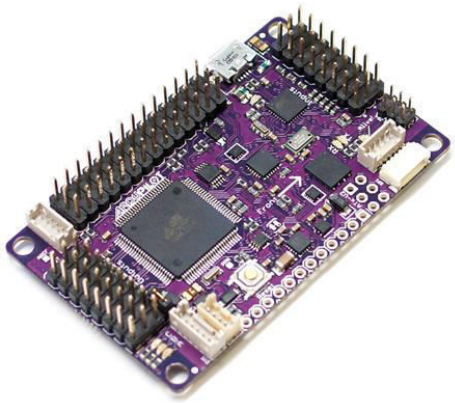


Fig. 4. APM 2.5 board

The ArduPilot features include: use for an autonomous vehicle, built-in hardware failsafe that uses a multiplexer chip and ATTiny processor to toggle control from the RC system to the autopilot, ability to reboot the main processor in mid-flight, multiple 3D waypoints (limited only by memory), altitude controlled with the elevator and throttle, a 6-pin GPS connector for the 4Hz uBlox5 GPS module, six spare analog inputs (with ADC on each) and six spare digital input/outputs for additional sensors.
Furthermore, the board supported the addition of wireless modules for real-time telemetry and was based on a 16MHz Atmega2560 processor. Eight RC-in and –out channels could be processed by the autopilot, in addition to the autopilot on/off channel. The board offered LEDs for power, failsafe (on/off), status and GPS satellite lock.
Programmability and use required the free Arduino IDE to edit and upload the code to the board. The Arduino environment

simplified the code production, modification, and upload procedures for the ArduPilot. Capable of running on Windows, Mac OS X, and Linux, the environment was composed in Java and based on Processing, avr-gcc, and other open source software. Simplified flowchart of complete system is given in Fig. 5.

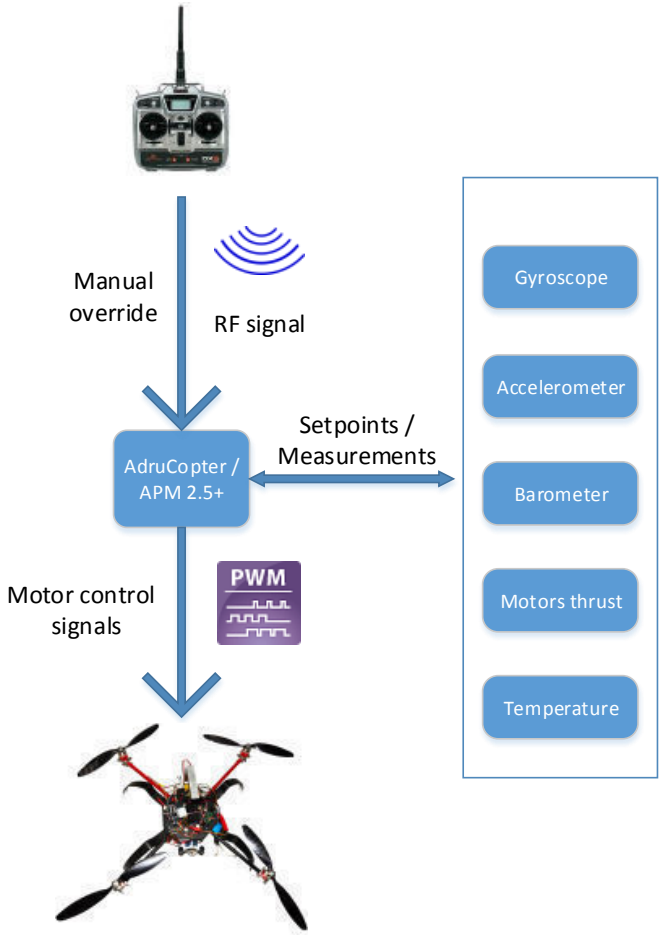


Fig. 5. Simplified flowchart of complete system

C. *Navigation and control system*

The navigation system consists of one microcontroller and several sensors. The first sensor required for flight navigation is 3-axis gyroscope with which the navigation system will have 3 degrees of freedom (DOF) [8]. With this sensor the multirotor is similar to the traditional RC helicopters. The gyroscope detects any rotation on all axes (changes in torque on different motors which results in yawing, tilting and wind gusts) which they disbalance the flight of the multirotor and those errors are compensated and corrected by the software. The second sensor is the 3-axis accelerometer with which the navigation system will have 6 degrees of freedom (DOF). The main purpose of this sensor is the multirotor to keep leveling without any pilot input, and the other use is for correcting the unwanted accelerations on the 3 axes, but mainly for the Z axis (altitude hold mode). The third sensor is the 3-axis magnetometer for measuring the earth's magnetic field to determine the orientation of the multicopter according to the poles and for keeping that direction. This sensor is mandatory for flying autonomously with preprogrammed GPS coordinates and for some other flight modes.

There are other additional sensors that may be applied to the quadcopter: Barometer, Gps, Optical flow measurement.

There are many factors that affect the characteristics of the flight like: Air turbulence, Propeller Acoustic Noise, Grounding and Power Conducted Electrical Noise, Radiated Electrical Noise, Frame Vibrations, Motor vibration, Propeller vibrations. These factors imply need for software and hardware tuning.

D. *Sensing Challenges related to vibrations*

Keeping a small aircraft in the air requires information regarding the attitude of the plane that is constantly updated and, generally speaking, accurate. To maintain level flight, the quadcopter uses accelerometers and gyroscopes, the combination of which is referred to as an inertial measurement unit, or IMU. Accelerometers use the acceleration due to gravity to help us calculate the direction of "down," and gyroscopes measure the rotational acceleration around the x-, y-, and z-axes. By combining these two sets of data into one angle measurement per axis, we can accurately calculate the plane's attitude with respect to the horizon.

A major challenge with the sensors on an IMU was that they are exposed to a great deal of vibration as a matter of course. The motor, which is directly attached to the airframe, is one significant source of vibration; and external conditions like gusts of wind and turbulence are another. Both the accelerometers and gyroscopes in the IMU are negatively impacted by the environment in differing ways. Vibration, external forces (such as from turns), and time all play a role in decreasing the accuracy of the sensor data being reported.

The main source of vibration on an aircraft is from the power system. Even a well-isolated and well-balanced electric motor is a significant source of airframe vibration. Vibrations will always occur due to efficiency losses in the propeller, as well as issues surrounding balance. These vibrations affect both accelerometers and gyroscopes, and cause variations in sensor data that must be managed by our controller.

Ideally, our accelerometer would measure only the force of gravity, or "down." Unfortunately, this is not how an accelerometer works in flight: centripetal forces (generated when the aircraft is moving in a circular path, or turning) cause accelerometers to read inaccurately because they are no longer measuring just gravity. With this extra force, it is not possible to determine our orientation in relation to the ground solely based on the accelerometer. Gyroscopes are not typically affected by external forces (but are affected by vibration) as they only measure instantaneous acceleration. For this reason, we use a combination of sensors to augment and correct one another's observation of the aircraft's orientation.

VII. FURTHER WORK

The successful building and testing of an experimental quadcopter means accomplishment of a working solution as a first phase in an ongoing Master thesis work. The further research needs to be focused on software navigation under influence of different weather conditions, emphasizing influence of air temperature and air density on the flight planning, range of operation, navigation precision and maneuverability.

The next step will be conducting experimental flights and measurements for the calculations and evaluations together with software changes in order to achieve better UAV

performances. The experimental flights will be done under calm and difficult weather conditions to gather more data about the weather influence on the flight platform. The PID algorithm and configuration will be evaluated for getting optimal control and better performances of the quadrotor in hover mode. The feasible realization of the flying mission, flight endurance and the attitude stabilization system is important in the development process of more advanced UAV functionality, therefore many complex factors including weather factors will be taken into consideration and examined. It is expected to result in additional failsafe features on pre-flight and in-flight calculations for flight mission planning and evaluation.

VIII. CONCLUSION

The work presented in this paper describes the complexity of designing, building and tuning multirotor UAV. The result is a multirotor model intended to be used further as an experimental platform in examining mission planning and flight endurance in relation to weather conditions as a part of a broader work for Master thesis preparation.

Throughout the setting and tuning of the multirotor for achieving stable flight, the weather conditions affecting the flight performance was significant. The authors have succeeded good flying performances, maneuverability and controllability of the experimental quadrotor in air temperature range from 0°C to 40°C and during cloudy and windy conditions with wind speed up to 10 knots. The initial examinations of the flight endurance showed that it is dependent not only on the configuration of the platform, i.e. characteristics of the used battery, but also from the weather conditions, especially wind speed and air temperature and density. The ArduCopter software is open source and as such it can be modified according to the needs for the further planned work under the ongoing Master thesis preparation.

IX. ACKNOLEDGEMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, R. Macedonia.

REFERENCES

[1] Guowei CAI, Ben M. CHEN, Tong H. LEE, "An overview on development of miniature unmanned rotorcraft systems", Unmanned Systems, Vol. 2, No. 2 (2014) 1–25, ©World Scientific Publishing Company, DOI: 10.1142/S2301385014300017

[2] Guowei CAI, Ben M. CHEN, Tong H. LEE, "An overview on development of miniature unmanned rotorcraft systems", © Higher Education Press and Springer-Verlag 2009Front. Electr. Electron. Eng. China 2010, 5(1): 1–14, DOI 10.1007/s11460-009-0065-3,

[3] Advanced MultiCopter Design, [Online], Available: http://copter.ardupilot.com/wiki/advanced-multicopter-design/

[4] US Office of the Secretary of Defence "Unmanned Aerial Vehicle Reliability Study", p.35, 36, 2003, [Online]. Available: http://www.uadrones.net/military/research/acrobat/0302.pdf

[5] http://www.aeroquad.com/

[6] http://www.multiwiicopter.com/

[7] http://www.copter.ardupilot.com/

[8] http://en.wikipedia.org/wiki/Degrees_of_freedom_(mechanics)

[9] Roman Czyba1, and Grzegorz Szafranski1, "Control Structure Impact on the Flying Performance of the Multi-Rotor VTOL Platform - Design, Analysis and Experimental Validation", International Journal of Advanced Robotic Systems, 2012, [Online]. Available: http://cdn.intechopen.com/pdfs-wm/42226.pdf

[10] Pierre-Jean Bristeau, François Callou, and David Vissière, Nicolas Petit, "The Navigation and Control technology inside the AR.Drone micro UAV", Preprints of the 18th IFAC World Congress Milano (Italy) August 28 - September 2, 2011

Noise-pollution mapping, measurement and analysis system using volunteered geographic information

Natalija Najdova
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, R. of Macedonia
natalija.najdova@sih.finki.ukim.mk

Ivan Chorbev
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, R. of Macedonia
ivan.chorbev@finki.ukim.mk

Abstract— **Noise pollution is defined as unwanted or excessive human-created noise that can have harmful effects on human health or welfare. Therefore, noise pollution control, especially outdoor noise control, is highly important for the community. By raising awareness, more people will participate in reducing the noise-pollution and spreading the knowledge of possible solutions. This paper presents a measurement and analysis system using VGI to produce a mapping tool for noise-pollution in Skopje. With the help of a mobile application the users are able to collect noise level data, along with a time-stamp and a geo location of the entry. Eventually, by storing such data, we are actually providing a dataset available for further data mining.**

Keywords— Noise-pollution, VGI, Citizen Sensor, Mobile, Noise Mapping, Maps, Geographical Analysis, Maps Visualization, Geographical Data mining

I. INTRODUCTION

Today, according to a United Nations 2014 Report, 54 % of the world's population lives in urban areas with a tendency to increase to 66% by 2050. Projections show that urbanization combined with the overall growth of the world's population could add another 2.5 billion people to urban populations by 2050.[1] In some cities, population growth and unplanned or inadequately managed urban expansion leads to rapid increase of noise levels. The growth in noise pollution has direct impact on the nature and the environment and may cause adverse effects on human health such as noise-induced hearing loss, sleep disturbance effects, cardiovascular and psychophysiological effects, performance effects and effects on social behavior.[2][5]

Accordingly, the European Community passed the directive 2002/49/EC[4] which states that noise protection is essential to ensure health and environmental conservation beneficial to everyone. The directive established a common framework for the assessment and management of exposure to environmental noise. It includes mapping of noise in larger cities via noise maps, as one of the planned actions. The produced maps should be a starting point for further plans in countering with noise pollution [4]. Currently Skopje lacks publicly available data on noise exposure. On the other hand, the rapid urbanization has given rise to a constant measurement demand on the changing noise pollution. An effective way of monitoring our fast changing environment is through crowdsourcing.[6] It enables people with smart phones to produce geographic information at a very low cost and no extra effort. In this sense, multiple citizens might potentially produce data-sets large enough for a reliable data analysis. This approach has already proven accurate enough in measuring noise-pollution levels.[7]

Therefore, the work presented in this paper tries to provide a general overview of the potential of a volunteer mobile noise mapping tool. The overall objective of this project is to provide the missing knowledge of noise pollution in every city and consolidate it with the scientific knowledge on the health impacts of outdoor noise. The results are expected to serve as guidance to environmental health authorities and professionals trying to protect people from the harmful effects of noise. Additionally, the tool can be used by the community when deciding on real state purchases or renting apartments.

In Section 2 "Noise Mapping today" we describe the present examples of the most related work to our system. Section 3 "Current limitations and challenges" presents the reasons and barriers why the previous research attempts failed to reach more cities. The Section 4 "System Architecture" contains the architecture design of our noise mapping crowdsourcing platform. Finally, Section 5 "Conclusion and future work" summarizes the finding of our work and presents directions for future work.

II. NOISE MAPPING TODAY

This paper is not the first, nor the last, to make an attempt to use mobile sensors to measure outdoor noise. The result of researches about projects related to our work showed hundreds of sound measurement apps available.[16] Most of the available applications are free, either ad-supported or with limited features, with a fee to upgrade to full features. However, the majority of these apps are made with little regard for accuracy and they are not widely used in most countries. They lack information provided prior to the purchase or the download is often limited, incorrect, contradictory or poorly formulated. Some of the apps have complicated UI design and some of them don't show the progress on the map in real-time.

Therefore, we present examples of related work to our system. We looked at Android based mobile applications that are to some extent similar to our platform in purpose, free to use and showed more accurate and reliable measurements.

A. NoiseTube[11]

NoiseTube is a research project, started in 2008 at the Sony Computer Science Lab in Paris and currently hosted by the BrusSense Team at the Vrije Universiteit Brussel, which proposes a participative approach for monitoring noise pollution by crowdsourcing. This app was the first to introduce social tagging, remains one of the few to support A-weighting and the only one that can be calibrated remotely via downloadable settings. Perhaps most importantly, they were the first to open their solution to the public. It contains very few measurements in most countries.

B. *WideNoise[15]*

This an Android application with the source code of v3.0 released under an open source license. It applies a linear interpolation algorithm to account for microphone sensitivity. However, the apps "correct" measurements using the same, hard-coded array of calibration points, no matter on which device they run and doesn't support A-weighting. Also the Android application contains very few recent measurement records.

C. *Noise Droid[13]*

This is an Android application developed by the Institute for Geoinformatics (IFGI) at the University of Muenster, Germany as part of the Open Noise Map research started at 2010. The creators provide no information on the way SPL measuring is implemented, or about frequency weighting, calibration or accuracy. Also, this application doesn't contain noise measurements data from a lot countries and it is not released under an open source license.

D. *NoiseMap[14]*

The Darmstadt Sensor Netzwerk, shorted to da_sense, is an urban sensing platform created by Schweizer et al. of the Technische Universität Darmstadt as part of a research started 2011. The platform is currently focused exclusively on the city of Darmstadt and its code is not open source. They created an Android application which is currently offered for download but seem to have calibration currently available only for limited number of phone models(6).

E. *NoiseWatch[10]*

NoiseWatch app allows users to measure SPL over an interval of 10s. Users have to manually initiate each measurement and after capturing a measurement it can be submitted, along with GPS coordinates and a timestamp, to the Eye on Earth platform. No details on the SPL measuring algorithm, calibration or accuracy are provided for the current version of the application. Another drawback is the users have to go to the website platform to see all the measurements made and its code is not open source. It has very few recent measurement records made.

F. *AirCasting[12]*

AirCasting is an initiative of an NGO based in NewYork, launched in December 2011. It is a platform for recording, mapping, and sharing health and environmental data using a smartphone. Both the mobile app and the Web platform are open source. So far there are no records for most countries apart from USA.

III. CURRENT LIMITATIONS AND CHALLENGES

In the previous section we mentioned 6 projects that are to some extent similar to our platform in purpose, but what is common for all of them is the lack of noise pollution data for a lot of countries. The overall objective of the paper is to provide the lacking knowledge of noise pollution of (but not only) our city. Accordingly, we need to identify the reasons and barriers why the previous research attempts failed to reach every city in the world and more broadly, every country.

A. *Duty to collect pollution data*

Possible barriers to environmental and climate friendly behavior include that most of the population consider that it is not their duty, rather the obligation of the Government, companies and industries to address it.[19][20]

B. *Localization*

In case of "environmentally-oriented" systems, we are convinced that concerns about the (local) environment can be a motivating factor. We think that personalized environmental information and tools may have a bigger impact on public awareness and behavior than the global environmental statistics currently provided by governmental agencies.

C. *Recruitment & retaining of contributors*

Finding, motivating and retaining participants, is always a major issue for any system and requires more effort than purely technological solutions. Additionally these systems face the standard crowdsourcing issues like how to recruit contributors, how to manage abuse and how to balance openness with quality.

D. *Privacy concerns*

User privacy is a major concern for any software system that uses the internet. The issue with a mobile sensing system which uses the phone's microphone is capturing of noises automatically presumably not someone's conversations and not without user's approval.[23] Additionally, being a developing country we can classify our population as late adopters of technological solutions and scared to adopt new trends.[22]

E. *Calibration & Data quality*

Without proper calibration, sensor devices produce data that may not be representative or can even be misleading. Most applications on the market showed inconsistence across different devices since android devices come in many shapes and sizes. With a wide range of device types, in order to be as successful as possible on Android, the app needs to adapt to various device configurations and still remain free or low priced.

F. *Human-centered mobile sensing application*

Creating a system that enables a mobile phone to gather geo-localized noise measurements, annotate them and send them automatically to map local noise pollution, providing helpful data for public institutions to support decision making on local issues without waiting for officials (environmental agencies or government funding for expensive measuring campaigns) to turn their attention to our neighborhood, is a delicate system that needs human-centered design approach. The existing applications showed no data of the country which

implies the design was made to serve the purpose of another local environment and the design thinking methodology wasn't applied for the citizens of our (and every) country.

IV. System Architecture

A. General architecture

This section of the paper describes the conceptual architecture we used. For this project, we followed a multi-layer architecture that could be easily implemented, adopted and extended by other applications with a similar purpose.

Figure 1 depicts the general architecture of the project. As seen on the figure, the modules are divided in three parts: Presentation layer (Mobile Client-side android application), Business (Service) layer and Persistent (Data) layer.

In the Presentation layer, a mobile application has been developed, for the Android operating system. This layer is responsible of processing the user's measurements input and returning the correct response back to the user.

The Business layer acts as a transaction boundary and contains both application and infrastructure services. The infrastructure services offer data access that goes beyond the need of the VGI application itself, organizing access to the underlying data tables.

The Persistent layer is the lowest layer of the application, responsible of communicating with the used data storage. All the layers are in details described below.



Fig. 1. Architecture design

B. Presentation Layer – Android Mobile Application

The Mobile application provides the front-end that allows user to capture measurements, submit them to our server and check the general progress of the noise mapping.

For the User Interface we used Android SDK because it is a free and growing development platform and provides a Google Maps Android Heatmap Utility, the layer of Google Maps we used for our Mapping module. Another reason why the Android platform was selected is due to its popularity in our country (almost 80% of all smart phone users).[9]

For the Measurement Module we used part of the noise pollution measurement application called NoiseTube. It is an open-source mobile application that allows the user taking noise samples. [8] Due to the open-source condition of NoiseTube, we reused source code related to the collection and assessment of the noise samples, and include it in our Measurement Module.

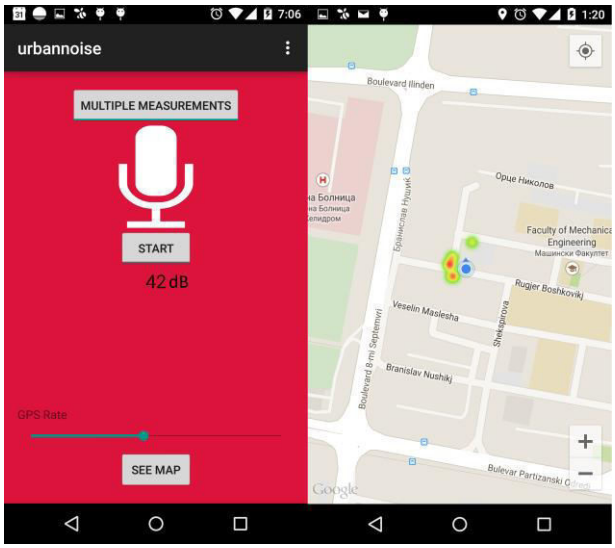


Fig. 2. (A: Measurement View, B: Map View)

. The Measurement Module has custom designed User Interface (as shown on the Figure 2A), provided to ease the process of capturing noise measurements and sending the data to the server. It contains only the basic and necessary components like, toggle button to choose between single and multiple measurements, toggle button to start and stop the measurements (in case of multiple measurements) or measure (in case of single measurement), seek bar to adjust the GPS rate when making multiple measurements and button to start the map view (as shown on the Figure 2B). The map shows noise measurements made in different days but in the same hour as the hour of the system time so the user can see the average noise in the current hour.

C. Service Layer – Java Spring web application

The business logic of our model lies in the service layer which manages the data of the domain objects. For the implementation of the Service layer, we used the Java Play framework to support the creation of a REST Web services running under Java. These services run from the same server that hosts the database server, but it can be directed to a separate machine, depending on the expected volume of user requests.

According to the framework's architecture, HTTP is used as the message transmission protocol, and messages are encoded using JSON. The service offers mediation in the insertion of new measurement contributions, as well as for data recovery on the uploaded noise measurements is presented over a base map in the Android app. For the access drivers, we used JDBC drivers to support connection to PostgreSQL.

D. Persistence Layer – PostgreSQL with JDBC drivers

Finally, the persistence layer involves the necessary components to store the data related with the measurements taken by the users.

For the data server we decided to use PostgreSQL as a powerful, open source object-relational database system. It uses a different approach that supports the same modeling capabilities and has an implementation that may have better

performance. PostgreSQL JDBC driver allows Java programs to connect to a PostgreSQL database using standard, database independent Java code. The JPA repository allows storing and retrieving data in our database.

V. CONCLUSION AND FUTURE WORK

This paper presented a framework for the creation of a VGI application, capable of capturing noise contributions from an Android mobile application which can be used for the citizens of this country. Also we provided an overview of previous efforts that have done similar work covered in this paper, as well as overview of the possible barriers for lacking data of our (and every) country.

Clearly, launching new Android mobile application to gather the missing data takes more than a good idea and a technically sound implementation. The main challenge is an approach towards this idea which should be done using the design thinking methodology (human-centered design) which suggests starting the design process with gathering human opinions on this topic. The approach should cover all the barriers explained in the previous section and find a strategy on how to overcome them, but with the accent of keeping the simple UI of the app. Therefore, we see our project as the starting solution of the current noise mapping limitations and challenges.

Firstly, by providing immediate noise mapping after a measurement is made, we create a role shift of the issue. With the application installed on the devices we make sure the noise mapping of the city is rather seen as a necessary citizens' contribution than as officials' duty.

Secondly, the application provides a very good understanding of what data is shared with the system by presenting the privacy policy at the first use of the application. Additionally by not creating user profiles and opening our source code, we ensure anonymity to our users.

By centering the map view with the user's location, we ensure solving the lack of localization of the noise mapping tools.

Lastly, our system is a fast and robust solution to encourage citizens' participation, increase their motivation and support retaining the participants. It is very simple solution that we believe should be adopted by the users first and then be upgraded to more complex and human-centered application. We believe this should be our first and main focus in our future work.

In future work, we also want to give attention on the problem of calibrating mobile devices, interpolating missing values, implementation of mechanisms for filtering and correcting volunteer data, opening and linking our data with a suitable ontology and increasing the general quality of the data acquired.

Another possible avenue for future efforts to tackle this challenge is the introduction of elements of competition and a gaming system in a joint pollution platform for our country. It means extending the framework by integrating with a platform that include more sensors in order to capture more health factors about the user and to provide a more complete monitoring on his health status and the detect critical events.

REFERENCES

[1] United Nations, Department of Economic and Social Affairs, Population Division (2014). World Urbanization Prospects: The 2014 Revision, Highlights (ST/ESA/SER.A/352).

[2] Basner, Mathias, Wolfgang Babisch, Adrian Davis, Mark Brink, Charlotte Clark, Sabine Janssen, and Stephen Stansfeld. "Auditory and non-auditory effects of noise on health." The Lancet 383, no. 9925 (2014): 1325-1332.

[3] European Union. Future noise policy. com (96) 540 final. European Commission Green Paper, 1996.

[4] The European Parliament. Directive 2002/49/ec of the european parliament and of the council of 25 june 2002 relating to the assessment and management of environmental noise. Official Journal of the European Communities, L189:12–25, 2002.

[5] Berglund, Birgitta, and Thomas Lindvall, eds. Community noise. Stockholm: Center for Sensory Research, Stockholm University and Karolinska Institute, 1995.

[6] M. Haklay and P. Weber P. OpenStreetMap: Usergenerated street maps. IEEE Pervasive Computing 7(4):12-18 2008.

[7] Garcia-Martí, Irene, Joaquín Torres-Sospedra, Luis E. Rodríguez-Pupo, and Joaquín Huerta Guijarro. "A comparative study on VGI and professional noise data." (2014).

[8] Maisonneuve, Nicolas, Matthias Stevens, and Luc Steels. "Measure and map noise pollution with your mobile phone." In Proceedings of the 27th annual CHI conference on human factors in computing systems, pp. 78-82. 2009.

[9] StatCounter Global Stats. Accessed March 6, 2015. http://gs.statcounter.com/#mobile_os-MK-monthly-201501-201503-bar.

[10] European Environment Agency. July 23, 2014. Accessed March 6, 2015. http://www.eea.europa.eu/mobile/apps#noisewatch.

[11] Steels, Luc. "Community memories for sustainable societies." (2008).

[12] AirCasting. Accessed March 6, 2015. http://aircasting.org/about.

[13] Foerster, Theodor, Laura Díaz, and Arne Bröring. "Integrating Volunteered Human Sensor Data into Crowd-sourced Platforms: A Use Case on Noise Pollution Monitoring and OpenStreetMap." (2011).

[14] Schweizer, Immanuel, Roman Bärtl, Axel Schulz, Florian Probst, and Max Mühläuser. "NoiseMap-real-time participatory noise maps." In Proc. 2nd Int'l Workshop on Sensing Applications on Mobile Phones (PhoneSense'11), pp. 1-5. 2011.

[15] WideNoise. Accessed March 6, 2015. http://www.widetag.com/widenoise/.

[16] Kardous, Chucri A., and Peter B. Shaw. "Evaluation of smartphone sound measurement applicationsa)." The Journal of the Acoustical Society of America 135.4 (2014): EL186-EL192.

[17] Haklay, Muki. "Geographical citizen science–Clash of cultures and new opportunities." (2010).

[18] Goodchild, Michael F. "Citizens as sensors: the world of volunteered geography." GeoJournal 69, no. 4 (2007): 211-221.

[19] First Biennial Report on Climate Change. March 2, 2015. Accessed March 5, 2015. http://unfccc.int/resource/docs/natc/macbur1eng.pdf.

[20] Bojovikj, Ph.D Dragana. Климатски Промени. December 1, 2014. Accessed March 5, 2015. http://unfccc.org.mk/content/FBUR/Climate change survey FBUR_MK.pdf.

[21] Abras, Chadia, Diane Maloney-Krichmar, and Jenny Preece. "User-centered design." Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications 37, no. 4 (2004): 445-456.

[22] Janeska-Sarkanjac, Smilka, and Ss Cyril. "Comparative Analysis of the Government ICT Projects in Macedonia, Estonia and Slovenia." (2012).

[23] Lane, Nicholas D., Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. "A survey of mobile phone sensing." Communications Magazine, IEEE 48, no. 9 (2010): 140-150.

SDN Controller: Where to Drop the Pin

Mile Piperkoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
mile.piperkoski@hotmail.com

Igor Mishkovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
igor.mishkovski@finki.ukim.mk

***Abstract*—The main goal of Software Defined Networking (SDN) is moving the control logic from the network devices to external, logically centralized controllers. Doing so enables faster innovation, easier configuration and simpler management. This paper focuses on picking the right node in a network, in regards to network resilience and performance. Various authors propose propagation latency as the main parameter for choosing the best node as a controller. However, upon analyzing several networks, that placement is not the best regarding network resilience. We will show that the betweenness value is also important as a placement metric and show that for this topology, it is best to consider it as an equal factor in picking the best controller node, along with propagation latency.**

Keywords—SDN, controller placement, network resilience

I. INTRODUCTION

Traditional network devices tightly couple control and data planes. In these networks the decisions where and how to forward packets are performed on the same device as the actual packet forwarding. However, recent trends decouple the control and data planes. Network devices are merely used as simple forwarding elements, while the control logic is logically centralized.

The proponents of SDN claim that these networks offer easier configuration, simpler management while reducing CapEx and OpEx, respectively [1].

While the benefits are clear and stated by companies that connect a number of data centers around the globe [2], issues like reliability, scalability, security and controller placement persist, which is the main theme of this paper, are sparking research.

We narrow our focus on one question:

How is network resiliency affected if a controller fails?

This is regarded as a huge problem in the network designing process. While traditional networks were purposely built with vertically integrated devices, SDN networks are quite the opposite. The controller placement choice directly influences whether, in the event of a disaster, the network would stay connected or collapse in several, isolated network islands.

Our goal is not to show that there is only way to solve the controller placement problem. There are already quite a number of different approaches [3]. The goal is to show that when tackling the problem, one must take in account various parameters and make the choice considering the objectives of the task at hand [4].

This work is divided in several parts. Section II is focused on the motivation for developing the SDN concept, whereas Section III states the problem. Section IV defines the metrics used for the research. Section V is reserved for analyzing the OS3E network and Section VI defines a new algorithm for controller placement. Section VII provides further discussion into the matter and future research. Lastly, Section VIII summarizes the results in one conclusion.

II. WHY DECENTRALIZE THE CONTROL PLANE?

Traditional networks rely on expensive and proprietary network devices to maintain connectivity. Each of these devices come with a vendor specific operating system and fully integrated control and data planes. While this adds a certain degree of network resilience, which was one of the main purposes of developing ARPANET, this concept has several drawbacks.

Every device has to be configured individually. In big networks, this is a major problem and a cause for a lot of headaches. Configuration mistakes as a result of human error are not that uncommon. To make matters worse, when a new device is added to the network, certain changes have to be made to the existing network and devices. An error in configuration would most likely lead to several hours, in certain circumstances days, of troubleshooting. Add monitoring and managing and this concept becomes a challenge not everyone is prepared to confront.

Another drawback is the vendor lock-in, which inevitably happens. Once a certain vendor-specific implementation is used it is very difficult to shift to another one. It is an expensive and time consuming feat, involving additional training and a lot of overtime.

Motivated by these drawbacks, SDN as a concept is totally the opposite. Its goal is a clear separation of the control and data plane. The control plane is designed to work as a logically centralized entity, running as a software on x86 server architecture, thus the term software-defined networks.

The network devices are only seen as simple forwarding devices, making their decisions via the SDN controller, the network brain.

This new concept addresses the drawbacks of the traditional networks. Configuration is done at one central location, reducing the possibility of human error and making the network easier to scale, while management is eased by the same reason. Regarding the vendor lock-in problem, SDN introduces unrivaled freedom in choosing network hardware and software. Because one of the goals of SDN is network openness, network hardware from different vendor would have no conflict in functioning in one environment. Moreover, even the SDN controller, which essentially is software, can be substituted for a controller from a different vendor, according to specific needs.

In terms of architecture, a parallel can be drawn with the peer to peer and client-server application architectures. While traditional network devices can be observed as peers in a network, the SDN concept provides us with clients i.e. the forwarding devices and a server i.e. the SDN controller, which offers better control over the entire network [5].

III. Problem Statement

SDN controller placement should not be taken lightly. As previously stated, the controllers are the brains of SDN networks. Without a control logic supporting the underlying infrastructure, the network does not function properly and does not produce the same level of productivity.

As previously stated in the introduction, addressing only one parameter does not successfully solve the controller placement problem. Furthermore, it raises issues regarding network resiliency.

This brings us to the main problem that we address in this paper – a controller placement algorithm taking resiliency into account.

IV. Research and Metrics

Our research is based on propagation latency, approximated through distance in miles, rather than seconds and betweenness centrality i.e. the number of shortest paths from all vertices to all others that pass through that node.

At first, we will analyze algorithm 1, which takes into account the propagation latency parameter.

For a network graph *G(V, E)* where edge weights represent propagation latencies, where *d(v, s)* is the shortest path from node $v \in V$ to $s \in V$ and the number of nodes is $n = |V|$, the average propagation latency (APL) for each node is [6]:

$$\mathrm{APL} = \frac{1}{n} \sum_{v \in V} \min d(v, s) \qquad (1)$$

Since in most topologies, one controller is sufficient, we base our choice for the location of the controller from (1) and the betweenness centrality value. These value are often in correlation i.e. the node with the lowest average propagation latency has the largest betweenness centrality value, so the controller elect is the controller with the lowest average propagation latency and highest betweenness centrality value. How this is affecting network resilience will be analyzed in the following section.

By analyzing Fig. 1, one can predict that the controller should be placed in one of the nodes located in central USA.

V. OS3E Network Analysis

Having defined our metrics, we will now analyze the OS3E network. This network consists of 34 nodes. It is densely populated on the east coast of the USA, while the west coast consists of a lesser number of nodes, increasing the latencies between them.



Fig. 1. The OS3E Network

Utilizing (1) and the betweenness centrality value for each node, the controller should be placed in Chicago, as seen in Fig. 2.

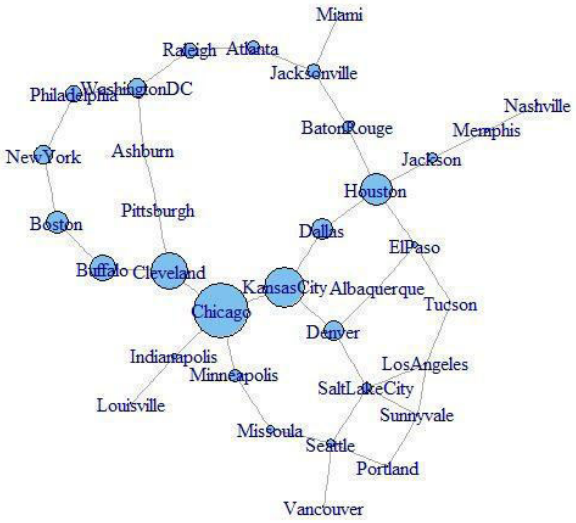


Fig. 2. Controller placement

The size of each node is dependent on its corresponding betweenness centrality value.

Network performance is not the main goal of this paper. As stated in the introduction, the goal is to answer what happens if a controller fails, while various parameters are taken into account.

From a functional perspective, one controller is sufficient. However, from a security perspective this statement is debatable. Since SDN networks rely on a few controllers to make forwarding decisions, they would be the primary target of any attacker looking to disrupt normal work [7]. Since IT security as a whole relies on "prepare for the worst, hope for the best", we will now analyze what happens when

the chosen controller fails or is compromised. In the next figure, Chicago is removed from the network graph.

As seen in Fig. 3, when the Chicago controller experiences a disaster, the network is divided in two clusters, with no connectivity between them. However, since only two nodes belong in the isolated cluster, we can state that the network is not experiencing a great connectivity problem. So, we will remove the next controller elect – Houston.

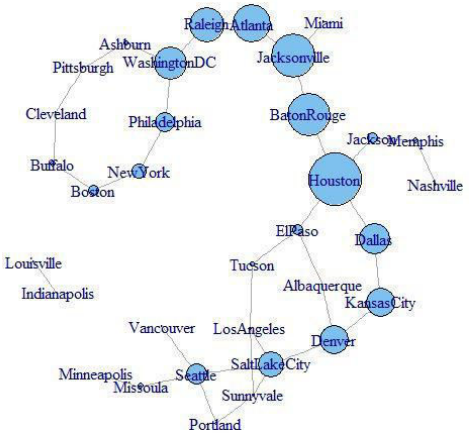


Fig. 3. OS3E network topology after Chicago failure

As we can see in Fig. 4, after just only two controller failures, network connectivity collapses. We witness the forming of four, unconnected clusters, which in essence makes the network nonfunctional.

Since the results of this placement algorithm are not very satisfying in terms of network resilience, a different algorithm is needed.

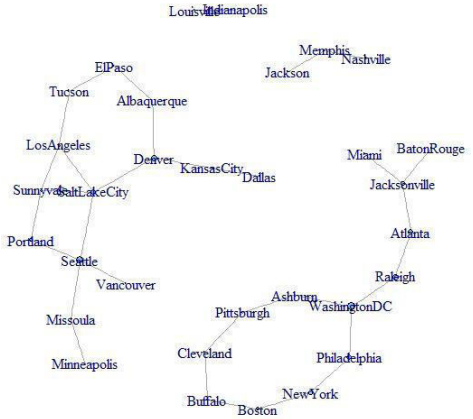


Fig. 4. OS3E network topology after Chicago and Houston failures

VI. NEW CONTROLLER PLACEMENT ALGORITHM

Advances in network bandwidth speed are made daily and the trend is that prices for link speed are dropping. Bearing that in mind, we will use the same parameters to propose a slightly new algorithm.

Using (1), we will calculate the average propagation latency - APL. However, the choice for a network controller will be based not on the minimum average propagation latency value, but rather on the node with small average propagation latency and a small betweenness centrality value. The values for average propagation latency and betweenness centrality for each node are normalized and have a value between 0 and 1. For controller placement, we will use the following equation:

$$C = \max \sum_{i=1}^{n} (\alpha * BC(i) + (1 - \alpha) * APL(i)) \quad (2)$$

where C is the controller node, n is the number of nodes in the network, $\alpha \in [0,1]$, BC is the betweenness centrality of each node and APL is the average propagation latency (1).

The new algorithm leaves a choice of two preferences – network performance or network resiliency. As stated in part III, the new algorithm should prioritize network resiliency over network performance. This is achieved by choosing an appropriate value for α. A bigger value for α means that network performance is preferred, while a smaller value means that network resiliency is preferred. The reasoning of this is clear – the bigger the average propagation delay, the smaller the betweenness centrality value of a given node i.e. the network stays connected even when the controller fails and another one is taking its place. Using (2) the node which satisfies the calculation is the controller C in the network.

We will choose a value for α = 0.1 and prefer network resiliency, or 90% network resiliency and 10% performance. After running our algorithm, the results show that Los Angeles is mostly suited to be chosen as a controller.

When Los Angeles suffers a failure (Fig. 5), the network graph stays connected and the SDN network remains functional.

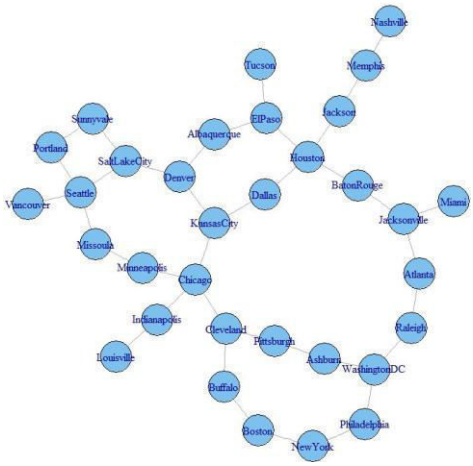


Fig 5. OS3E network topology after Los Angeles failure

Now, if another failure occurs at Tucson (Fig. 6), which was chosen as a controller by the same algorithm and same value for α, again, the network graph stays connected and the network remains functional. As in the previous figure, we can say that the network graph stays connected and the network remains functional.

VII. DISCUSSION

Are similar results to be expected on other topologies? Is this network placement algorithm valid for other topologies?

Based on the analysis on the OS3E network, we analyzed four more topologies available on The Internet Topology Zoo. Using the first placement algorithm and subsequent disaster scenario, we can conclude that selecting the node with the smallest value for average propagation delay and biggest betweenness centrality value, it is expected that in the event of a controller failure, the network will be divided in isolated clusters. The number of isolated clusters and their size varies from topology to topology. If we have a scenario similar to the OS3E network and presume that there are two clusters, of which one is with only one member is acceptable, a second controller failure provides more serious consequence.

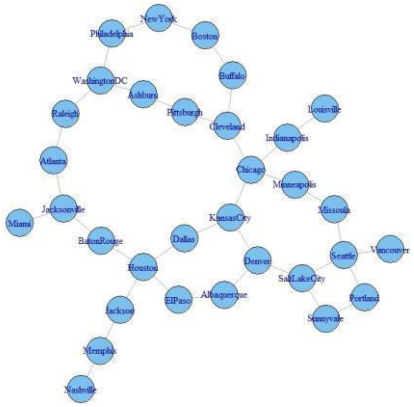


Fig. 6. OS3E network topology after Los Angeles and Tucson failures

That being said, if we again use the new algorithm, the results are the same as for the OS3E network. There are no isolated clusters, the network functions as one whole and resiliency is maintained, since the new algorithm takes in account two parameters.

For example, let's briefly analyze the AGIS network. On the left side of Fig. 7 is the full network topology. The chosen controller is Minneapolis, using (2), for $\alpha = 0.1$. After experiencing a disaster, a new controller (Chicago) is elected, using the same value for α. Even in the event Chicago experiences a disaster and another node takes the role of a controller, the network maintains connectivity.

However, drawing such a conclusion, at the moment, may seem a little far-fetched. To further develop this algorithm and make it applicable in SDN networks, more analysis must be concluded in regards to various network applications and services, thus providing satisfying quality of service and user satisfaction.

VIII. Conclusion

In this paper, we introduced the concept of SDN and a brief comparison with the traditional network architectures. Based on this comparison, we outlined that there is a shift in our understanding in networking from a peer to peer model, to a client-server model. Since the clients without the server in that model cannot function as a whole without server control, security is a must. Because we always have to prepare for the worst, it is imperative that we add controller redundancy in an SDN network.

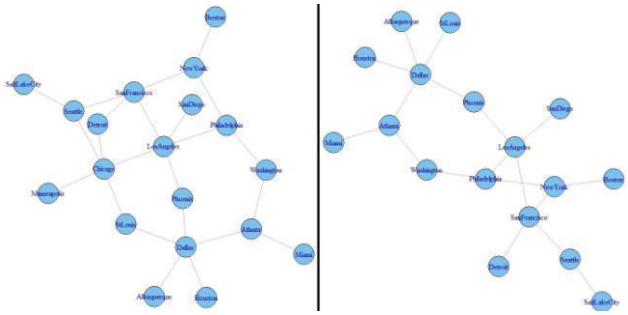


Fig 7. AGIS network topology comparison - Full network vs. Two failed controllers

Based on that experience, one of the most important factors in designing a network is its resiliency. The simple answer is that by adding more than one controller will make the network more resilient and robust. However, we showed that more than one metric needs to be taken into account when placing the controller. Moreover, we showed that it is not always the best metric results associated with a certain node are clear picks.

Finally, as stated in the introduction, the most important factor in picking a placement algorithm is the goals that need to be achieved. If we took in account only the average propagation latency and discard the factor of potential failures, which is a little utopist, one only needs one controller for optimal performance and minimal expenses. However, in a realistic scenario, one must make tradeoffs between different metrics and based upon them place at least 2, preferably more controllers in one SDN network.

Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at the University "Ss. Cyril and Methodius".

References

[1] Wolfgang Braun, Michael Menth, "Software-Defined Networking Using OpenFlow: Protocols, Applications and Architectural Design Choices", Future Internet 2014, 6, pages 302-336, May 2014

[2] Sushant Jain, Alok Kumar, Subhasree Mandal et al., "B4:Experience with a Globally-Deployed Software Defined WAN" SIGCOMM '13, August 12-16, 2013, Hong Kong, China.

[3] Hu Yannan, Wang Wendong, Gong Xiangyang, Que Xirong, Cheng Shiduan, "On the placement of controllers in software defined networks", October 2012

[4] Hu Yannan, Wang Wendong, Gong Xiangyang, Que Xirong, Cheng Shiduan, "On Reliability-optimized Controller Placement for Sofware-Defined Networks", February 2014

[5] Diego Kreutz et al., "Software-Defined Networking: A Comprehensive Survey", June 2014

[6] Brandon Heller, Rob Sherwood, Nick McKeown, "The Controller Placement Problem", HotSDN '12, August 13, 2012, Helsinki, Finland.

[7] Manar Jammal, Taranpreet Singh, Abdallah Shami, Rasool Asal, Yiming Li, "Software defined networking: State of the art and research challenges", Computer Networks 72 (2014), pages 74-98

Real-World ARP Attacks and Packet Sniffing, Detection and Prevention on Windows and Android Devices

Blagoj Nenovski and Pece Mitrevski
Faculty of Information and Communication Technologies
"St. Clement of Ohrid" University
Bitola, Macedonia
cyberbaze@gmail.com, pece.mitrevski@fikt.edu.mk

***Abstract*—This paper explains the purpose and the need of the Address Resolution Protocol (ARP) and different types of attacks that can occur due to its stateless nature. We exhibit a real world example of a Man-in-the-Middle (MitM) attack by sniffing http logins of a Windows PC and an Android device, and then suggest methods of both detecting and preventing the attack.**

Keywords—ARP spoofing; ARP poisoning; DoS; MitM; packet capture; network sniffing; WireShark; network forensics

I. INTRODUCTION

Every device connected to a LAN network has two addresses: a 32-bit IP address which can be static or dynamic and a 48-bit MAC address that is static. When a device with its own IP address sends packets to another device in the network the sender needs to know the receiver's MAC address so the communication continues from IP (Layer 3) to MAC (Layer 2). This created the need of an Address Resolution Protocol (ARP).

ARP was defined by RFC 826 in 1982 [1] and has the role to broadcast a message across the network and to determine the MAC address (Layer2) of a host with a set IP (Layer 3) both static or dynamic by DHCP. This message is replied with an ARP packet that contains the MAC address. The ARP protocol has four message types:

- ARP Request – When a device asks for MAC address by sending an IP address;
- ARP Reply – When a device replies with its IP and providing its MAC address;
- Reverse ARP request – When a device asks who has a certain MAC address;
- RARP Reply – When a device replies with its MAC and providing its IP address.

In order to minimize network traffic devices remember the MAC addresses in an Arp cache. This way, devices that already had previous communication eliminate the need to repeat the ARP requests and replies. Fig. 1 illustrates arp packets that devices send to acknowledge the MAC addresses of other devices connected to the network. ARP is a stateless protocol and does not include any authentication which makes it vulnerable to various attacks.

Noteworthy is to mention that Wi-Fi Protected Access II (WPA2), the Wi-Fi Alliance branded version of the newest and most rigorous security to implement into WLAN's today, 802.11i, has still been found to have at least one security vulnerability, nicknamed "Hole196", which uses the shared key among all users to launch attacks on other users of the same Basic Service Set identification (BSSID).

II. RELATED WORK

Tripunitara and Dutta [2] discuss the Address Resolution Protocol and the problem of ARP cache poisoning and present a solution and implementation aspects of it in a Streams based networking subsystem. They also present the algorithm that is executed in the module and application to prevent ARP cache poisoning where possible, and detect and raise alarms otherwise. They discuss some limitations and present some preliminary performance numbers for their implementation.

Zdrnja [3] explains ARP attack fundamentals and analyzes recent attacks that used ARP poisoning against Web hosting companies to let attackers insert malicious code into virtually

	Time	Source	Destination	Protocol	Length	Info
239	18:01:19.897624000	IntelCor_59:8c:e7	SamsungE_1b:a4:65	ARP	42	Who has 192.168.100.3? Tell 192.168.100.1
240	18:01:19.897754000	IntelCor_59:8c:e7	Cisco-Li_aa:78:c6	ARP	42	Who has 192.168.100.1? Tell 192.168.100.3
241	18:01:19.913130000	IntelCor_59:8c:e7	AsustekC_0d:78:f2	ARP	42	Who has 192.168.100.2? Tell 192.168.100.1
242	18:01:19.913264000	IntelCor_59:8c:e7	Cisco-Li_aa:78:c6	ARP	42	Who has 192.168.100.1? Tell 192.168.100.2
243	18:01:19.927417000	IntelCor_59:8c:e7	SamsungE_1b:a4:65	ARP	42	Who has 192.168.100.3? Tell 192.168.100.1
244	18:01:19.927560000	IntelCor_59:8c:e7	Cisco-Li_aa:78:c6	ARP	42	Who has 192.168.100.1? Tell 192.168.100.3
245	18:01:19.941570000	IntelCor_59:8c:e7	AsustekC_0d:78:f2	ARP	42	Who has 192.168.100.2? Tell 192.168.100.1
246	18:01:19.941744000	IntelCor_59:8c:e7	Cisco-Li_aa:78:c6	ARP	42	Who has 192.168.100.1? Tell 192.168.100.2
250	18:01:25.749990000	IntelCor_59:8c:e7	Broadcast	ARP	42	Who has 192.168.100.1? Tell 192.168.100.4
251	18:01:25.751637000	Cisco-Li_aa:78:c6	IntelCor_59:8c:e7	ARP	60	192.168.100.1 is at 00:1d:7e:aa:78:c6

Fig. 1. WireShark capture of ARP packets

thousands of Web sites.

Kołodziejczyk and Ogiela [4] describe few, popular security mechanisms used by network protocols. In an attempt to create some basic rules and requirements which should be met by secure network protocols.

Similarly, Abad and Bonilla [5] analyze several schemes to mitigate, detect and prevent ARP spoofing attacks that have been proposed, identify their strengths and weaknesses, and propose guidelines for the design of an alternative and (arguably) better solution to the problem of ARP cache poisoning.

Lootah et al. [6] introduce the Ticket-based Address Resolution Protocol (TARP). TARP implements security by distributing centrally issued secure MAC/IP address mapping attestations through existing ARP messages. They conclude by exploring a range of operational issues associated with deploying and administering ARP security.

Bruschi et al. [7] present a secure version of ARP that provides protection against ARP poisoning. Each host has apublic/private key pair certified by a local trusted party on the LAN, which acts as a Certification Authority. As a proof of concept, the proposed solution was implemented on a Linux box.

Preventing ARP Spoofing Attacks through Gratuitous Decision Packet System (GDPS) is a mechanism suggested by Salim et al. [8] which seeks to achieve two main goals: (1) Detection of suspicious ARP packets, by implementing a real-time analyzing for received ARP packets. (2) The distinction between a legitimate and malicious host through sending a modified request packet of the gratuitous ARP packets.

Qadeer et al. [9] focuses on the basics of a packet sniffer and its working, development of the tool on Linux platform and its use for Intrusion Detection. They also discuss ways to detect the presence of such software on the network and to handle them in an efficient way. Focus has also been laid to analyze the bottleneck scenario arising in the network, using this self-developed packet sniffer. Observation has been made on the working behavior of already existing sniffer software such as wireshark (formerly known as ethereal), tcpdump, and snort, which serve as the base for the development of the sniffer software.

Pansa and Chomsiri [10] present a design of "architecture and protocols" for the LAN security preventing the process of MAC Address spoofing, ARP Spoof and MITM. In the presented system, the operation in Protocol level of DHCP is modified, and the use of ARP Protocol is canceled in order to suit and correspond with the new architecture.

Radhakishan and Selvakumar [11] address the drawbacks of TrueIP and propose a new architecture to overcome them. In addition, all sorts of man-in-the-middle attacks (MIMA) are eliminated in their TrueIP spoofing prevention technique using Identity based cryptography in which a signature scheme is used to achieve better security.

Intrusion Detection Systems (IDS) are able to prevent MitM attacks. Belenguer and Calafate [12] present a low-cost embedded IDS which, when plugged into a switch or hub, is able to detect and/or prevent MitM attacks automatically and efficiently. Since their system is limited to a micro-controller and a network interface, it can be produced at a very low cost, which is attractive for large scale production and deployment.

Callegati et al. [13] discuss Man-in-the-Middle Attacks to the HTTPS Protocol and show that it is possible to attack even Web-based connections secured via HTTPS by exploiting some properties of common LANs as well as typical behaviors of inexperienced users.

Finally, in the paper of Anaya et al. [14] a forensics network model is proposed, which allows for obtaining the existing evidence in an involved TCP/IP network. This Model uses Fuzzy Logic and Artificial Neural Networks to detect Network flows that realize suspicious activities, minimizing cost and time to process the information in order to discriminate between normal network flows and flows that have been subjected to attacks and intrusions.

Whereas Roy et al. [15] focus on using Ettercap as the tool for the attack and suggest using arpwatch to detect and prevent the attacks (both Ettercap and arpwatch are Linux based tools), this paper focuses on Windows tools that can be used for the attack and tools that can detect and prevent the attack both on Windows PCs and Android devices.

III. Types of Attack

A. *Denial of Service Attack*

Denial of service attack or DoS attack (Fig. 2) can be leveraged by sending an ARP reply to either:

- Host – by changing the gateways MAC to a non-existing MAC (ex. FF:FF:FF:FF:FF:00).

This way every time the host sends packets to the gateway, they are sent to FF:FF:FF:FF:FF:00 which results in the host not being able to communicate with the gateway.

- Gateway – by changing the victims MAC to a non-existing MAC (ex. 00:00:00:00:00:FF).

This way every time the gateway sends packets to the victim, they are sent to 00:00:00:00:00:FF which results the victim host not receiving any of the packets.

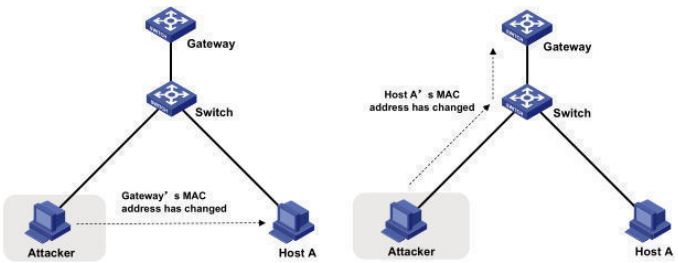


Fig. 2. DoS attack of a host (left) and gateway (right)

There are other techniques that can cause DoS of a host like ARP spoofing a vast amount of IP addresses to a single MAC address. This type of attack can be more effective on larger LAN networks and will cause the victim to be overloaded with traffic. One of the most commonly used tools to perform DoS

attacks is arpspoof, distributed in the dsniff package and Cain & Abel by oxid.

B. *Man in the Middle Attack*

Man in the middle attacks (MitM) are achieved by ARP poisoning both the victim host and the gateway (Fig. 3). The attacker, using ARP poison, sends its MAC address to the victim overwriting the gateway's MAC with its own, and an additional ARP poison to the gateway with the victim's IP its own MAC address.

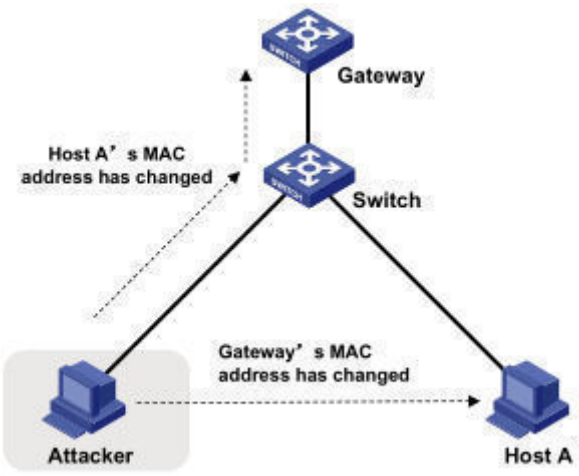


Fig. 3. MitM attack with ARP poison

TABLE I. ARP CACHE BEFORE THE MITM ATTACK

	Attacker	Gateway	Victim Host
Gateway	Attacker's MAC		Victim's MAC
Victim Host	Attacker's MAC	Gateway's MAC	

TABLE II. ARP CACHE AFTER THE MITM ATTACK

	Attacker	Gateway	Victim Host
Gateway	Attacker's MAC		Attacker's MAC
Victim Host	Attacker's MAC	Attacker's MAC	

The ARP cache after the MitM attack defines further communication between the devices in the network, meaning that any communication between the gateway and the victim host is intercepted and forwarded through the attacker's device.

Man in the middle attacks can lead to session hijacking which can grant the attacker access to all the services that victim has been logged in including bank accounts, email, social networks, forums etc. MitM attack additionally allows the attacker while intercepting and capturing the data for further examination, to modify the data in real time before it is forwarded to the victim/gateway. At first, due to the security of SSL/TLS protocol https services were immune to this type of attacks, but soon it was proven that SSL/TLS is vulnerable, too [16][17][18].

SSL/TLS protection is out of the scope of this paper and detailed information on the attack and prevention of SSLStrip can be found in [19].

IV. REAL WORLD ARP ATTACK

A. *Overview of the Network Devices*

With the use of Fing, a network scanner tool running on an Android device, we get an overview of all the devices connected on the network. At the moment the following devices are connected:

TABLE III. DEVICES CONNECTED TO THE GATEWAY

IP	MAC	Hostname	Description
192.168.100.1	00:1D:7E:AA:78:C6	Linksys	Gateway
192.168.100.2	00:1F:C6:0D:78:F2	PC	Windows PC
192.168.100.3	F0:25:B7:1B:A4:65	Samsung	Android Device
192.168.100.4	68:5D:43:59:8C:E7	Thinkpad	Notebook

At this state all of the connected devices share the same ARP cache. Before the attack we define the roles on both the attacker and the victims:

Gateway:	Linksys	IP: 192.168.100.1
Victim 1:	Windows PC	IP: 192.168.100.2
Victim 2:	Samsung	IP: 192.168.100.3
Attacker:	Thinkpad	IP: 192.168.100.4

B. *ARP Poisoning of a Windows PC and an Android Device*

The ARP attack can be conducted using either arpspoof or Ettercap on a Linux machine. In our scenario the attacker was a Windows Notebook and the ARP poisoning was achieved using Cain & Able.

C. *Wireshark Capture of the ARP Attack*

With a WireShark capture one can filter the ARP packets and inspect the ARP communication between the devices in the network (Fig. 4).

At Packet 11893 one can see the Attacker ARP poisoning the Windows PC telling that 192.168.100.1 (Gateways IP) is at 68:5D:43:59:8C:E7 which in fact is the Attackers MAC. To complete the MitM attack the attacker tells the Gateway that the Windows PC's IP 192.169.100.2 is at 68:5D:43:59:8C:E7 which can be seen at packet 11894. On packet 12043 the attacker changes the Gateway's MAC with its own in the ARP cache of the Samsung Smartphone and completes the attack with changing the Smartphone's MAC with its own on the Gateway (packet 12044).

D. *Attack Confirmation by Victim 1*

If we check Victim 1's ARP cache before the attack, using the arp –a command in cmd.exe gets us the following result (Fig. 5).

```
11893 18:05:05.773486000   IntelCor_59:8c:e7          AsustekC_0d:78:f2  ARP          42 192.168.100.1 is at 68:5d:43:59:8c:e7
11894 18:05:05.773664000   IntelCor_59:8c:e7          Cisco-Li_aa:78:c6  ARP          42 192.168.100.2 is at 68:5d:43:59:8c:e7
12043 18:05:35.760009000   IntelCor_59:8c:e7          SamsungE_1b:a4:65  ARP          42 192.168.100.1 is at 68:5d:43:59:8c:e7
12044 18:05:35.760153000   IntelCor_59:8c:e7          Cisco-Li_aa:78:c6  ARP          42 192.168.100.3 is at 68:5d:43:59:8c:e7
```

Fig. 4. WireShark capture of the ARP poisoning of Victim 1 and Victim 2

After the attack has been done, one can confirm the attack by reusing the arp –a command, which would result with the recurrent MAC address as both the Attacker and the Gateway (Fig. 6).

```
C:\Documents and Settings\b>arp -a

Interface: 192.168.100.2 --- 0x2
  Internet Address      Physical Address      Type
  192.168.100.1         00-1d-7e-aa-78-c6     dynamic
  192.168.100.3         f0-25-b7-1b-a4-65     dynamic
  192.168.100.4         68-5d-43-59-8c-e7     dynamic
```

Fig. 5. Windows PC (Victim 1) ARP cache before the attack

```
C:\Documents and Settings\b>arp -a

Interface: 192.168.100.2 --- 0x2
  Internet Address      Physical Address      Type
  192.168.100.1         68-5d-43-59-8c-e7     dynamic
  192.168.100.3         f0-25-b7-1b-a4-65     dynamic
  192.168.100.4         68-5d-43-59-8c-e7     dynamic
```

Fig. 6. Windows PC (Victim 1) ARP cache after the attack

After a successful ARP attack the ARP cache on the Victim 1's side has been altered and changed the Gateway's MAC 00:1D:7E:AA:78:C6 with the Attacker's MAC 68:5D:43:59:8C:E7. Victim 1 has 192.168.100.1 as its Gateway and therefore any further communication would continue through the Attacker's computer that has 68:5D:43:59:8C:E7 as its MAC address. At this point the Attacker can perform MitM attack by sniffing, capturing or filtering the data packets the Victim 1 both sends and receives. After the attack, the Attacker sends ARP packets that restore the MAC address for IP 192.168.100.1 to the Routers MAC.

E. Attack Confirmation by Victim 2

To check the ARP cache on an Android Smartphone we use Net Status which can be downloaded from the Play Store. Using the Get Arp Cache command in Net Status prior the attack gets the same ARP cache as all other devices (Fig. 7).

IP address	HW type	Flags	HW address	Mask	Device
192.168.100.2	0x1	0x2	00:1f:c6:0d:78:f2	*	wlan0
192.168.100.1	0x1	0x2	00:1d:7e:aa:78:c6	*	wlan0
192.168.100.4	0x1	0x2	68:5d:43:59:8c:e7	*	wlan0

Fig. 7. Android device (Victim 2) ARP cache before the attack

After the attack the ARP cache has been altered and getting the ARP cache results with the Gateways's MAC address being changed to the Attacker's MAC (Fig. 8). This way one can confirm the ARP attack, as well as who the attacker is.

IP address	HW type	Flags	HW address	Mask	Device
192.168.100.2	0x1	0x2	00:1f:c6:0d:78:f2	*	wlan0
192.168.100.1	0x1	0x2	68:5d:43:59:8c:e7	*	wlan0
192.168.100.4	0x1	0x2	68:5d:43:59:8c:e7	*	wlan0

Fig. 8. Android device (Victim 2) ARP cache after the attack

V. Sniffing Packets with WireShark

Once the ARP attack is complete the data sent and received by the users can be intercepted. This action can be done using WireShark, a free and open source packet analyzer. Wireshark is cross platform and can run on Windows, Linux, Unix and OSX. It uses pcap to capture the packets.

WireShark captures all the data transmitted to and received from the host. The MitM ARP attack makes all the communication between the Victim 1 (PC) and Victim 2 (Smartphone) and the Gateway visible in WireShark.

Capturing data can be achieved with two methods [20]:

- "Catch it as you can" – a method where all the packets are being captured and written to the storage subsequently, in batch mode. The downside of this method is the large amount of storage needed to store the data.
- "Stop, look and listen" – a method where each packet is analyzed and written to the storage only if it meets the defined criteria. The downside of this method is that is CPU intensive.

A. Capturing a Login Form

For demonstration purposes we are going to capture a login form. The login form used is the first google result on "login form" [21]. The Victim 1 will login using the login "arp-pc" and password "test" and the Victim 2 will login using the login "arp-phone" and password "test".

Between logins, WireShark was set to capture the data for 6 minutes and 9 seconds and that resulted in 14758 packets. Going through all the packets one by one can be time consuming. That leads us to filtering the packets using specific filers. User authentication is typically done using the POST

Filter: http.request.method == "POST" Expression... Clear Apply Save

No.	Time	Source	Destination	Protocol	Length	Info
1565	18:02:04.360333000	192.168.100.2	54.231.2.233	HTTP	590	POST /sn
1566	18:02:04.362581000	192.168.100.2	54.231.2.233	HTTP	590	[TCP Ret
2100	18:02:26.493112000	192.168.100.3	54.231.8.185	HTTP	761	POST /sn
2101	18:02:26.496166000	192.168.100.3	54.231.8.185	HTTP	761	[TCP Ret
2338	18:02:52.872205000	192.168.100.2	5.45.62.65	HTTP	919	POST /F/
2339	18:02:52.873455000	192.168.100.2	5.45.62.65	HTTP	919	[TCP Ret
12021	18:05:31.137572000	192.168.100.2	5.45.62.65	HTTP	919	POST /F/
12022	18:05:31.138729000	192.168.100.2	5.45.62.65	HTTP	919	[TCP Ret
12457	18:06:31.263589000	192.168.100.3	192.168.100.4	HTTP/XML	783	POST /up
12527	18:06:31.575692000	192.168.100.3	192.168.100.4	HTTP/XML	971	POST /up
12589	18:06:32.072276000	192.168.100.3	192.168.100.4	HTTP/XML	971	POST /up

Fig. 9. WireShark filter of the http packets using the POST method

method which dramatically reduces the packets displayed in WireShark. The filter used to narrow to only the packets that used the POST method is: http.request.method == "POST" (Fig. 9). To inspect the packet 1565 we will use the command "Follow TCP Stream". We need to focus on the POST paragraph which provides us with the following information:

```
POST /snippets/8-login-form/index.html HTTP/1.1
Host: s3.cssflow.com
User-Agent: Mozilla/5.0 (Windows NT 5.1; rv:34.0) Gecko/20100101 Firefox/34.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
Referer: http://s3.cssflow.com/snippets/8-login-form/index.html
Cookie: _ga=GA1.2.1069828863.1418392244; _gat=1
Connection: keep-alive
Content-Type: application/x-www-form-urlencoded
Content-Length: 39
```

This info shows us that the POST method is used to send data to: /snippets/8-login-form/index.html.

The Referrer or the side that Victims used to login is: http://s3.cssflow.com/snippets/8-login-form/index.html. At this point we can confirm that we have the right packet. The next step is to expand "HTML Form URL Encoded" in the Packet details (Fig. 10).

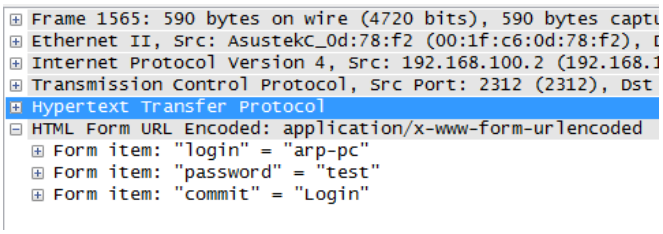


Fig. 10. Packet containing the login and password of Victim 1

Expanding the HTML Form URL encoded provides with the information of the login, password and commit form items. One can see that that login of Victim 1 is "arp-pc" and the password is "test". The Login details of Victim 2 can be inspected on packet 2100 (Fig. 11).

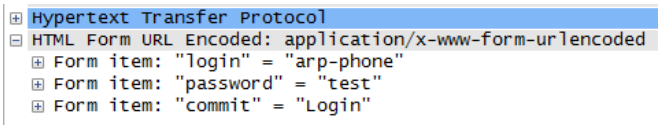


Fig. 11. Packet containing the login and password of Vitim 2

VI. Protecting from ARP Attacks

A. Detecting and Protecting Windows PCs

DecaffeinatID [16] is a Windows applications dedicated to monitoring the ARP changes in the local network. This application notifies the user for new devices that are connected to the network with a popup and stores the new events in a log. The log contains new devices that connected to the network and changes made to the ARP cache: the first and the second record identify that the Android Smartphone (Victim 2) and ThinkPad (Attacker) have connected to the network. The third record identifies that the MAC address for 192.168.100.1 (Gateway) has been changed and that the user was displayed with a warning, i.e.:

```
New IP in cache: 192.168.100.3 with MAC of f0-25-b7-1b-a4-65
New IP in cache: 192.168.100.4 with MAC of 68-5d-43-59-8c-e7
IP 192.168.100.1 changed MAC from 00-1d-7e-aa-78-c6 to 68-5d-43-59-8c-e7
WARNING!!! This IP is a Gateway!!! Something weird is up!!!
IP 192.168.100.1 changed MAC from 68-5d-43-59-8c-e7 to 00-1d-7e-aa-78-c6
WARNING!!! This IP is a Gateway!!! Something weird is up!!!
```

As one can see on the last record, DecaffeinatID does not keep a record of the original MAC of the Gateway and presents the user with a warning although the attack has stopped and the Gateway's MAC has been restored. In case the user misses the notifications the log stored as idslog.txt can be used for further network forensics.

One of the simplest methods to prevent ARP attacks is to set a static entry to the ARP cache with the Gateway's IP and MAC address. Using CMD as administrator the user can input: arp -s 192.168.100.1 00-1d-7e-aa-78-c6. This way the ARP cache will have a static entry for the Gateway which can't be altered by attackers. Static entries expire once the system reboots. To access this problem one can create a simple .bat script with the arp –s command and put the script in the "Startup" folder so the command executes every time the PC boots.

For the less experienced users, ARPFreeze [22] is an automated script that uses the ARP command to detect the devices connected on the same network and asks the user which of the entries he would like to make static.

B. Detecting and Protecting Android Devices

One of the methods in detecting and protecting Android devices is to scan the network using Fing. On Fig. 12 one can see that router's field is changed with a new MAC address and contains (+1) which indicates that another device connected on the network has the same MAC address.

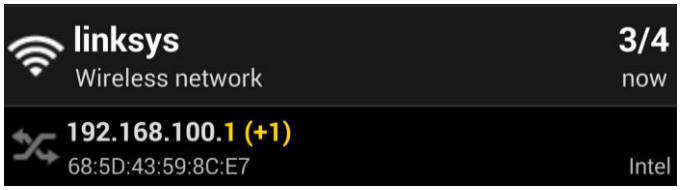


Fig. 12. Fing scan identifying additional MAC entry

Selecting the Gateway field gets that additional information so the user can identify who is that attacked, in this case the computer with the 192.168.100.4 IP and with the First seen and Last change he can easily conclude when that device has first connected on the network and more importantly when the attack occurred. Another indicator is that the previous IP and MAC address of the Gateway will be a grayed out field at the end of the list. We have to note here that the Fing network scanner has to be regularly used in order to get this specific information.

Wifi Protector [23] and WiFi ARP Guard [24] are both Android applications that automatically detect ARP spoofing

and ARP poisoning and therefore can protect the user from DoS and MitM attacks. There are many advantages of using these dedicated applications versus the Net Status and Fing methods, mainly for their automation process that checks the ARP cache of the Android device multiple times each second. Wifi Protector and WiFi ARP Guard have the option to automatically start when the wireless LAN is activated. When an attack is detected these applications can notify the user with a sound, vibration or notification. These notifications contain the MAC address of the attacker that help identify the attacker. Wifi Protector and WiFi ARP Guard store logs of the attacks and can be really useful for further network forensics. The protection that comes with these applications consists of disabling the wireless LAN in a case of attack. If the device is rooted the user can add a static ARP entry for their gateway. The disadvantage of running these types of applications is the effect on the battery life of the device.

C. *Network Based Solutions*

When it comes to small home or business networks or generally networks that are not for public use, disabling the DHCP and setting up static IP addresses for each device, bound with their unique MAC, can prove to be the best solution. Lone and Ataullah [25] have conducted a survey on various solutions of ARP attacks. The survey consists of 8 solutions including systems that use cryptography to secure the ARP protocol such as S-ARP and TARP, middleware devices installed on each host on the network to block unsolicited ARP replies and a hardware solution that consists of a server with an invite-accept and request reply protocol. The paper provides a comparative analysis of the solutions.

VII. Conclusion and Future Work

This paper shows that despite having modern applications communicating on modern hardware and software, users on local networks are vulnerable on DoS MitM attacks due to the vulnerability of the ARP protocol. We created a real world scenario of an ARP attack that serves as a platform for a packet analyzer to monitor all the data from and to the victims. Using WireShark filters we filtered the communication down to http POST methods which displayed us with the victim's login and password on a test page. The most important part, however, is protecting from ARP attacks, that detailed the detection and protection options for Windows and Android users.

Future work can complement this paper with methods and applications to detect and prevent ARP attacks. It will discuss the techniques of detecting network sniffers using ARP packets, crafting custom ping with hping3, using nmap script sniffer-detect, as well as comparative analysis and success rate of Windows applications Cain & Abel and Promqry; and Linux applications: arpwatch, Nast, ptool and sniffdet, in identifying a network card working in promiscuous mode [26].

References

[1] http://tools.ietf.org/html/rfc826

[2] M. Tripunitara and P. Dutta, "A middleware approach to asynchronous and backward compatible detection and prevention of ARP cache poisoning", 15th Annual Computer Security Applications Conf., 1999, pp. 303-309.

[3] B. Zdrnja, "Malicious JavaScript Insertion through ARP Poisoning Attacks", IEEE Security & Privacy, May/June 2009, pp. 72-74.

[4] M. Kolodziejczyk and M.R. Ogiela, "Security Mechanisms in Network Protocols", 2010 Intl. Conf. on Intelligent Systems, Modeling and Simulation, 2010, pp. 427-430.

[5] C.L. Abad and R.I. Bonilla, "An Analysis on the Schemes for Detecting and Preventing ARP Cache Poisoning Attacks", 27th Intl. Conf. on Distributed Computing Systems Workshops (ICDCSW'07), 2007, p. 60.

[6] W. Lootah, W. Enck and P. McDaniel, "TARP: Ticket-based address resolution protocol", 21st Annual Computer Security Applications Conf. (ACSAC 2005), 2005, 9 pp. – 116.

[7] D. Bruschi, A. Ornaghi and E. Rosti, "S-ARP: a Secure Address Resolution Protocol", 19th Annual Computer Security Applications Conf. (ACSAC 2003), 2003, pp. 66-74.

[8] H. Salim, Z. Li, H. Tu and Z. Guo, "Preventing ARP Spoofing Attacks through Gratuitous Decision Packet", 11th Intl. Symp. on Distributed Computing and Applications to Business, Engineering & Science, pp. 295-300.

[9] M.A. Qadeer, N. Zahid, A. Iqbal and M.R. Siddiqui, "Network Traffic Analysis and Intrusion Detection using Packet Sniffer", Second Intl. Conf. on Communication Software and Networks, pp. 313-317.

[10] D. Pansa and T. Chomsiri, "Architecture and Protocols for Secure LAN by Using a Software-level Certificate and Cancellation of ARP Protocol", Third Intl. Conf. on Convergence and Hybrid Information Technology, 2008, pp. 21-26.

[11] V. Radhakishan and S. Selvakumar, "Prevention of Man-in-the-Middle Attacks using ID Based Signatures", Second Intl. Conf. on Networking and Distributed Computing, pp. 165-169.

[12] J. Belenguer and C.T. Calafate, "A low-cost embedded IDS to monitor and prevent Man-in-the-Middle attacks on wired LAN environments", Intl. Conf. on Emerging Security Information Systems and Technologies, 2007, pp. 122-127.

[13] F. Callegati, W. Cerroni and M. Ramilli, "Man-in-the-Middle Attack to the HTTPS Protocol", IEEE Security & Privacy, May/June 2009, pp. 78-81.

[14] E.A. Anaya, M.N. Miyatake and H.M.P. Meana, "Network Forensics with Neurofuzzy Techniques", 52nd IEEE Intl. Midwest Symp. on Circuits and Systems, MWSCAS '09, 2009, pp. 848-852.

[15] http://web2.uwindsor.ca/courses/cs/aggarwal/cs60564/Assignments2Project1/RoyMoazzamiSingh_Assignment2.doc

[16] M. Marlinspike "New Tricks for Defeating SSL in Practice", https://www.blackhat.com/presentations/bh-dc-09/Marlinspike/BlackHat-DC-09-Marlinspike-Defeating-SSL.pdf

[17] D. Peifer, "SSL Spoofing", https://www.owasp.org/images/7/7a/SSL_Spoofing.pdf

[18] J. Franklin, B. Mijanovich and G. Pothier, "SSLStrip Man-In-The-Middle", http://fit.hcmup.edu.vn/~hienlth/COMP1049/Labs/04%20-%20SSL%20Spoofing/sslstrip2.pdf

[19] Y. Zhao, Y. Lei, T. Yang and Y. Cui, "A new strategy to defense against SSLStrip for Android", 15th IEEE Intl. Conf. on Communication Technology (ICCT), 2013, pp. 70-74.

[20] R. Sira, "Network Forensics Analysis Tools: An Overview of an Emerging Technology", http://www.giac.org/paper/gsec/2478/network-forensics-analysis-tools-overview-emerging-technology/104303

[21] http://www.cssflow.com/snippets/login-form/demo

[22] http://www.irongeek.com/i.php?page=security/arpfreeze-static-arp-poisoning

[23] https://play.google.com/store/apps/details?id=com.gurkedev.wifiprotector

[24] https://play.google.com/store/apps/details?id=com.mdl.arpguard

[25] I.A. lone and Md. Ataullah, "A survey on various solutions of ARP attacks", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 2, 2013, pp. 299-303.

[26] D. Sanai, "Detection of Promiscuous Nodes Using ARP Packets", http://www.securityfriday.com/promiscuous_detection_01.pdf

Wireless Charging – Concept, Theory and Applications

Nenad Trajkovski
School of Computer Science and Information Technologies
University American College Skopje
Skopje, Macedonia
nenad.trajkovski23@gmail.com

Biljana Stojcevska
School of Computer Science and Information technologies
University American College Skopje
Skopje, Macedonia
stojcevska@uacs.edu.mk

Ivana Stojanovska
School of Computer Science and Information technologies
University American College Skopje
Skopje, Macedonia
ivana.stojanovska@uacs.edu.mk

***Abstract* – In this research paper it will be explained how wireless charging works, how it can be developed and influence people`s everyday lives and maybe change the future. Wireless charging provides a convenient and safe means to reliably charge and power millions of consumer electronic and industrial devices. By eliminating the use of a physical connector or cable, wireless charging provides a number of advantages including efficiency, cost and safety. With an ongoing trend in mobile phone charger standardization, the next step to make the charging process even more user-friendly is to eliminate the physical cable connection between the charger and the mobile phone. Wireless charging works well with mobile phones, digital cameras, media players, gaming controllers and Bluetooth headsets. Other potential applications are power tools, medical devices, e-bikes and electric vehicles (EVs). To achieve wireless charging of a mobile phone, a wireless power transfer system must be designed. One such system is presented in this paper. The theory of the wireless power transfer is explained and two major wireless standards are analyzed. Wireless charging has a huge potential and it may one day replace plugs and wires similar to how Wi-Fi and Bluetooth have modernized personal communication. It can completely change the way we use or think about today`s electronic devices that use cable and give a whole new view of the future.**

Keywords— wireless charging, wireless charging theory, wireless charging concept, wireless charger device

I. INTRODUCTION

Wireless charging technology enables wireless power transfer from a power source (a charger) to a load (a mobile device) across an air gap. Wireless charging is rapidly evolving from theories towards standards, and is being adopted in commercial products, especially mobile phones and portable devices. Wireless charging improves user-friendliness as the connecting cables are removed. It enhances flexibility, especially for the devices for which the replacing of their batteries or connecting a cable for charging is costly, hazardous, or infeasible (e.g., body-implanted sensors). With a wireless charger, the battery inside any battery-powered appliance can be charged by simply placing the appliance close to a wireless power transmitter or a designated charging station. As a result, the appliance casing can be made completely sealed, even waterproof. Therefore, wireless charging can provide on-demand power, avoiding an overcharging problem and minimizing energy costs [1].

II. WIRELESS POWER TRANSFER SYSTEM AND TECHNIQUES

Nikola Tesla, the founder of alternating current electricity, was the first to conduct experiments in wireless charging. He achieved a major breakthrough in 1899 by transmitting 108 volts of high-frequency electric power over a distance of 25 miles to light 200 bulbs and run an electric motor. In 1901, Tesla constructed the Wardenclyffe Tower to transfer electrical energy globally without cords through the ionosphere. However, due to technology limitation (e.g., low system efficiency), the idea had not been widely further developed and commercialized [2][3]. Later, during the 1920s and 1930s, magnetrons were invented to convert electricity into microwaves, which enables wireless power transfer over long distance. However, there was no method to convert microwaves back to electricity [4].

A. *Wireless Inductive Power Transfer*

Inductive Power Transfer (IPT) involves the transmission of energy from a power source to an electrical load without connectors, across an air gap. The basis of a wireless power system involves essentially two coils – a transmitter and a receiver coil. The transmitter coil is energized by an alternating current to generate a magnetic field, which in turn induces a current in the receiver coil [5].

The basics of wireless power involves the transmission of energy from a transmitter to a receiver via an oscillating magnetic field. To achieve this, Direct Current (DC) supplied by a power source, is converted into high frequency Alternating Current (AC) by specially designed electronics built into the transmitter. The alternating current energizes a

copper wire coil in the transmitter, which generates a magnetic field. Once a second (receiver) coil is placed within proximity of the magnetic field, the field can induce an alternating current in the receiving coil. Electronics in the receiving device then converts the alternating current back into direct current, which becomes usable power [6].

The diagram below simplifies this process into three key steps.

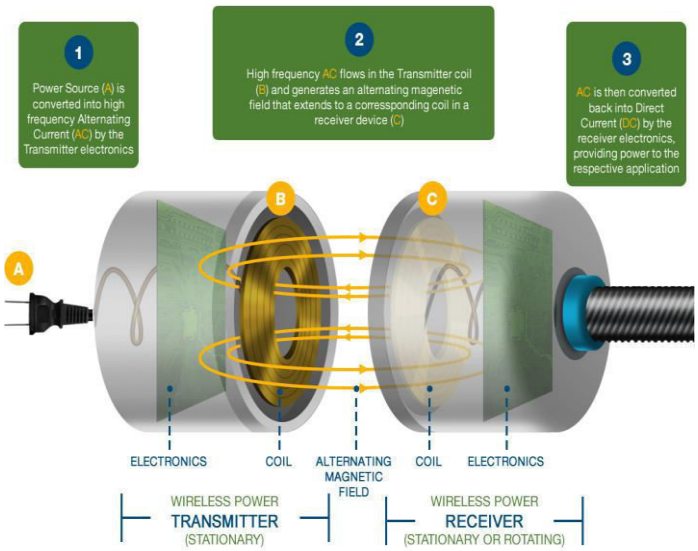


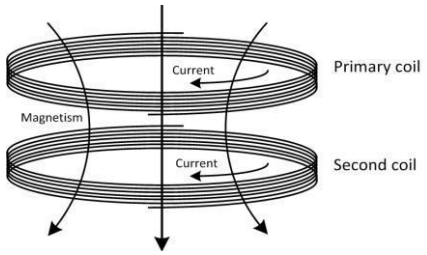
Fig.1 Wireless Power Transfer [7]

1. The main voltage is converted into an AC signal, which is then sent to the transmitter coil via the electronic transmitter circuit.
2. The AC current flowing through the transmitter coil induces a magnetic field which can extend to the receiver coil (which lies in relative proximity).
3. The magnetic field then generates a current which flows through the coil of the receiving device. The process whereby energy is transmitted between the transmitter and receiver coil is also referred to as magnetic or resonant coupling and is achieved by both coils resonating at the same frequency. Current flowing within the receiver coil is converted into direct current (DC) by the receiver circuit, which can then be used to power the device.

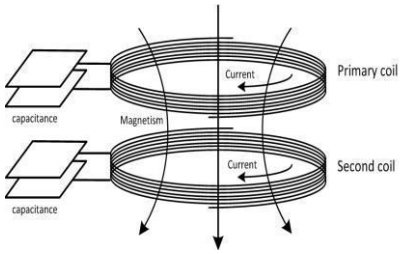
The distance at which the energy can be transferred is increased if the transmitter and receiver coils are resonating at the same frequency. This resonant frequency refers to the frequency at which an object naturally vibrates or rings – much like the way a tuning fork rings at a particular frequency and can achieve its maximum amplitude [8].

B. *Wireless Charging Techniques*

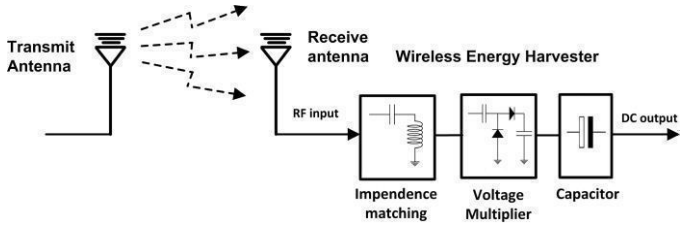
In wireless power transfer, a transmitter device connected to a power source, such as the mains power line, transmits power by electromagnetic fields across an intervening space to one or more receiver devices, where it is converted back to electric power and utilized. In general, a wireless power system consists of a transmitter device connected to a source of power such as mains power lines, which converts the power to a time-varying electromagnetic field, and one or more receiver devices which receive the power and convert it back to DC or AC electric power which is consumed by an electrical load [9]. In the transmitter, the input power is converted to an oscillating electromagnetic field by a coil of wire which generates a magnetic field, or by a metal plate which generates an electric field, or an antenna which radiates radio waves. Three major techniques for wireless charging are magnetic inductive coupling, magnetic resonance coupling, and microwave radiation [2].



a) Inductive Coupling



b) Resoning Coupling



c) Far-field Wireless Charging

Fig.2 Models of Wireless Charging System [9]

1) *Magnetic Inductive Coupling:* Magnetic inductive coupling, as shown in Fig. 2a, is based on magnetic field induction that delivers electrical energy between two coils. Magnetic inductive coupling happens when a primary coil of an energy transmitter generates a predominant varying magnetic field across the secondary coil of the energy receiver within the field, generally less than the wavelength. The near-field power then induces voltage/current across the secondary coil of the energy receiver within the field. This voltage can be used by a wireless device. The energy efficiency depends on the tightness of coupling between two coils and their quality factor. The tightness of coupling is determined by the alignment and distance, the ratio of diameters, and

the shape of two coils. The quality factor mainly depends on the materials, given the shape and size of the coils as well as the operating frequency. The advantages of magnetic inductive coupling include ease of implementation, convenient operation, high efficiency in close distance (typically less than a coil diameter) and safety. Therefore, it is applicable and popular for mobile devices.

2) *Magnetic Resonant Coupling:* Magnetic resonance coupling, as shown in Fig. 2b, is based on evanescent-wave coupling which generates and transfers electrical energy between two resonant coils through varying or oscillating magnetic fields. As the resonant coils, operating at the same resonant frequency, are strongly coupled, high energy transfer efficiency can be achieved with small leakage to non-resonant externalities. This property also provides the advantage of immunity to neighboring environment and line-of-sight transfer requirement. Compared to magnetic inductive coupling, another advantage of magnetic resonance charging is the longer effective charging distance. Additionally, magnetic resonant coupling can be applied between one transmitting resonator and many receiving resonators, which enables concurrent charging of multiple devices.

3) *Radio Charging:* Radio charging is used for charging items with small batteries and low power requirements, such as watches, hearing aids, medical implants, cell phones, MP3 players and wireless keyboard and mice. Radio waves are already in use to transmit and receive cellular telephone, television, radio and Wi-Fi signals. Wireless radio charging works similarly, with a transmitter, plugged into a socket, generating radio waves. When the receiver attached to the device is set to the same frequency as the transmitter, it will charge the device's battery.

III. WIRELESS STANDARD QI

Qi (pronounced "chee") is a wireless charging standard developed by the Wireless Power Consortium (WPC). The Qi standard specifies interoperable wireless power transfer and data communication between a wireless charger and a charging device. Qi allows the charging device to be in control of the charging procedure. The Qi-compliant charger is capable of adjusting the transmit power density as requested by the charging device through signaling. Qi uses the magnetic inductive coupling technique, typically within the range of 40 millimetres [9].

Within any Qi wireless charging system, there are two main elements:

- *Base Stations*: The Qi base station is the device that provides inductive power for the wireless transmission. As such it contains a power transmitter of which a major element is the transmitting coil. The overall Qi charger typically has a flat surface. This is normally referred to as the Interface Surface. The mobile device or devices may be placed on top of this surface.
- *Mobile Devices:* The Qi mobile devices are those which consume the wirelessly transmitted power. This is typically used to charge the battery contained within the mobile device.

Generally, a Qi wireless charger has a flat surface, referred to as a charging pad, for which a mobile device can be laid on top [10]. To keep the efficiency as high as possible, it is necessary to ensure that the coupling between the transmitter and the receive coils is as high as possible. To achieve tight coupling, a mobile device must be strictly placed in proper alignment with the charger. Qi specifies three different approaches for making alignment [11]:

- *Guided positioning*, i.e., a one-to-one fixed-positioning charging, provides a guideline for a charging device to be placed, for attaining an accurate alignment. The Qi specification achieves this by using a magnetic attractor. This approach is simple; however, it may require the implementation of a piece of material attracted by a magnet in the charging device.
- *Free-positioning with a movable primary coil* is also a one-to-one charging that can locate the charging device. This approach requires a mechanically movable primary coil that tunes its position to make coupling with the charging device.
- *Free-positioning with a coil array* allows multiple devices to be charged simultaneously, regardless of their positions. This approach can be applied based on the three-layer coil array structure.

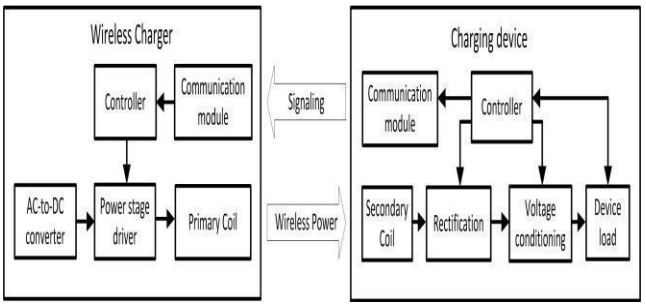


Fig.3 Qi-compliant wireless power transfer model [9]

The Qi-compliant wireless charging model supports in-band communication. The data transmission is on the same frequency band as the one used for the wireless charging. The Qi communication and control protocol is defined to enable a Qi wireless charger to adjust its power output for meeting the demand of the charging device and to disable power transfer when charging is finished.

IV. ALLIANCE FOR WIRELESS POWER STANDARD

Alliance for wireless power (A4WP) is a wireless charging and transfer technology that has been developed to enable the efficient transfer of power to devices without the use of wires. The A4WP standard is used to deliver power that would address simultaneous charging of multiple devices. The A4WP standard generates a larger electromagnetic field with magnetic resonance coupling which does not require precise

alignment and allows separation between a charger and charging devices. The maximum charging distance is up to several meters. Moreover, multiple devices can be charged concurrently with different power requirements. Different objects can be placed on an operating A4WP charger without causing any malfunctions. Therefore, the A4WP charger can be embedded in any object, improving the flexibility of charger deployment. The model for A4WP wireless charging is consisted of two components, power transmitter unit (PTU) and power receiving unit (PRU). The wireless power is transferred from PTU to PRU, which is controlled by a charging management protocol. Feedback signaling is performed from PRU to PTU to help control the charging [13].

- *Power transmitter unit* or A4WP charger is the key controller and power source for the wireless charging system and contains all the electronics to enable power to be taken from the mains of other power supply and convert it into a format where it can be picked up by the receiver unit to enable it to be charged. The PTU has three main functional units, i.e. resonator and matching circuit components, power conversion components, and signaling and control components. The PTU can be in one of following function states: Configuration, at which the PTU does a self-check; PTU Power Save, at which the PTU periodically detects changes of impedance of the primary resonator; PTU Low Power, at which the PTU establishes a data connection with PRU(s); PTU Power Transfer, which is for regulating power transfer; Local Fault State, which happens when the PTU experiences any local fault conditions such as over-temperature; and PTU Latching Fault, which happens when rogue objects are detected, when a system error or other failures are reported.
- The A4WP power receiving unit, the PRU, is the element of the A4WP charging system that is fitted into the item that required power to be transferred into it for charging. The PRU accepts power from a charger / transmitter and links with it during the charging process to enable control messages to be exchanged between the two for efficient charging to be effected. As shown in Fig.4, power enters the receiver via the resonator. It is rectified to provide a DC voltage. The rectifier needs to be able to provide efficient rectification at the 6MHz frequencies used. Once rectified, the power is converted to the required voltage using an efficient DC-DC switch mode converter. Both the rectifier and the DC-DC converter can be controlled over the 2.4GHz link to the charger to enable both charger and A4WP receiving unit to communicate and provide the most efficient power transfer. The PRU has the following functional states: Null State, where the PRU is under voltage; PRU Boot, when the PRU establishes a communication link with the PTU, PRU On, when the communication is performed; PRU System Error State, when there is over-voltage, over-current, or over-temperature alert; PRU System Error, when there is an error that forces a power shut-down.

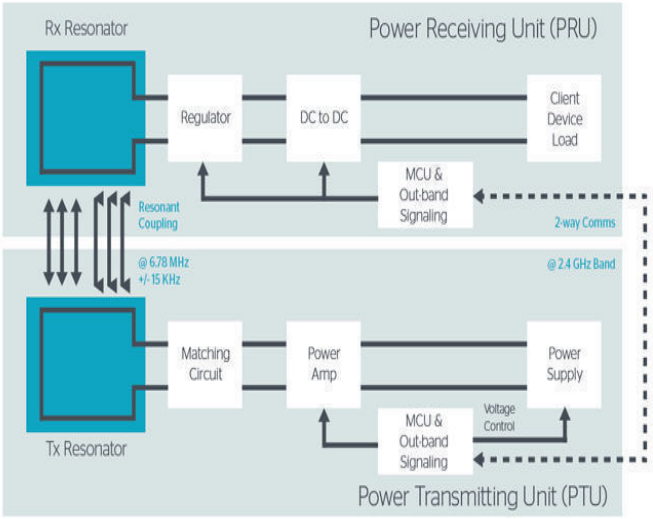


Fig.4 A4WP WPT PTU-PRU system reference model [14]

V. Advantages and Disadvantages of Wireless Charging

Wireless charging can be used for a wide variety of devices including cell phones, laptop computers and MP3 players, as well as larger objects, such as robots and electric cars. Wireless charging offers decisive advantages in the industry. It allows safe charging in a hazardous environment where an electrical spark through charge contacts could cause an explosion, or where heavy grease, dust and corrosion would make electrical contacts impractical. Wireless charging also helps when multiple insertions would wear out the battery contacts too quickly. As with any system, there are both advantages and disadvantages to wireless battery charging systems.

A. *Advantages*

- *Convenience* – it simply requires the appliance needing charging to be placed onto a charging area.
- *Reduced wear of plugs and sockets* – as there is no physical connection, there are no issues with connector wear, etc. Physically the system is more robust than one using connectors.
- *One transmitter to more receivers* – A charging transmitter would be able to charge more devices at the same time.
- *Resilience from dirt* – some applications operate in highly contaminated environments. As there are no connectors, the system is considerably more resilient to contamination.

- *Application in medical environments* – using wireless charging, no connectors are required that may harbor harmful agents such as bacteria, etc.. This makes wireless charging far more applicable for medical instruments than requiring them to be battery powered.

B. *Disadvantages*

- *Added complexity* – the system requires a more complicated system to transfer the power across a wireless interface.
- *Added cost* – as the system is more complicated than a traditional wired system, a wireless battery charger will be more expensive
- *Reduced efficiency* – there are losses on the wireless battery charging system: resistive losses on the coil, stray coupling, etc. However, typical efficiency levels of between 85%-90% are normally achieved.

VI. CONCLUSION

In this paper, we have presented an overview of the wireless power transfer system. The theory of the wireless power transfer is explained and two leading wireless charging standards are analyzed. The advantages and disadvantages can help improve and simplify the evolution and the future work of this technology. Wireless charging has a huge potential and it may one day replace plugs and wires similarly to how Wi-Fi and Bluetooth have modernized personal communication.

Wireless charging has now become a mainstream technology. Initially it was a novelty, but with its applications and advantages becoming recognized, it has now become a mainstream application. It is anticipated that wireless charging will become very widespread, if not the most common method. Wireless charging technology will become prevalent especially for consumer electronics, mobile, and portable devices.

With standard interfaces and techniques, only a single wireless battery charger will be required to charge a variety of items. No longer will a whole myriad of chargers be required. Also reliability and convenience will be improved as it is far easier to place the item to be charged on the charging mat, rather than having to use a small connector. We will enter a whole new age, an age without cables.

REFERENCES

[1] Integrated Device Technology, Inc. Introduction to Wireless Battery Charging. 2015. http://www.idt.com/products/power-management/wireless-charging-ics-wireless-power-ics/introduction-to-wireless-battery-charging (accessed on 27.02.2015)

I. Buchmann. Charging Batteries Without Wires. 2015. http://batteryuniversity.com/learn/article/charging_without_wires (accessed on 25.02.2015)

[2] N. Tesla (March 5, 1904). "The Transmission of Electric Energy Without Wires". Electrical World and Engineer (McGraw Publishing Co.) 43: 23760–23761. Retrieved November 19, 2014., reprinted in Scientific American Supplement, Munn and Co., Vol. 57, No. 1483, June 4, 1904.

[3] J. J. Schlesak, A. Alden, and T. Ohno, "A microwave powered high altitude platform," in Proc. of IEEE MTT-S International Microwave Symposium Digest, New York, NY, USA, May 1988.

[4] A. Gopinath (August 2013). "All About Transferring Power Wirelessly". Electronics For You E-zine (EFY Enterprises Pvt. Ltd.): 52–56. Retrieved January 16, 2015.

[5] N. Shinohara. Wireless Power Transfer via Radiowaves. John Wiley & Sons (2014)

[6] D. R. Schilling. WiTricity Charges Forward with Wireless Electricity. 2013. http://www.industrytap.com/witricity-charges-forward-with-wireless-electricity/574 (accessed on 26.02.2015)

[7] T. Sun,X. Xie, W. Zhihua (2013). Wireless Power Transfer for Medical Microsystems. Springer Science & Business Media. pp. 5–6

[8] Power By Proxy. Wireless Resonant Power. 2015 http://powerbyproxi.com/ (accessed on 01.03.2015)

[9] L. Xiao, N. Dusit, W. Ping, I. K. Dong, H. Zhu. "Wireless Charger Networking for Mobile Devices: Fundamentals, Standards, and Applications" 9 Dec 2014

[10] The Wireless Power Consortium. What is wireless charging? 2015. http://www.wirelesspowerconsortium.com/ (accessed on 28.02.2015).

[11] I. Poole. Qi Wireless Charging Standard. http://www.radio-electronics.com/info/power-management/wireless-inductive-battery-charging/qi-wireless-charging-standard.php (accessed on 01.03.2015)

[12] S. Y. R. Hui and W. C. Ho, "A new generation of universal contactless battery charging platform for portable consumer electronic equipment," IEEE Trans. Power Electron., vol. 20, no. 3, pp. 620-627, May 2005.

[13] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[14] R. Tseng, B. von Novak, S. Shevde, and K. A. Grajski, "Introduction to the alliance for wireless power loosely-coupled wireless power transfer system specification version 1.0," in Proceedings of IEEE Wireless Power Transfer (WPT), Perugia, Italy, May 2013

A Survey on data partitioning techniques for spatial and attribute data

Goran Nikolov
Agency for Real Estate Cadastre,
Skopje, Macedonia
g.nikolov@katastar.gov.mk

Goran Velinov, Margita Kon-Popovska
Faculty of Computer Science and Engineering
St. 'Cyril and Methodius' University
Skopje, Macedonia
goran.velinov@finki.ukim.mk
margita.kon-popovska@finki.ukim.mk

Abstract- **Current enterprise database systems contain large volume of data with large number of rows per table. Data partitioning is inevitable part of today's enterprise database systems because it offers better performance, availability and manageability. In this paper we review the recent partitioning techniques for spatial and attribute data. We also focus on the implementation of these techniques in commercial database systems like Oracle. We review spatial query optimization, spatial parallel processing, optimal space partitioning, online automatic partitioning and recent data partitioning techniques. We show recent achievements and explain why data partitioning as a concept is an attractive area for researchers.**

Keywords **– Data partitioning; spatial joins; parallel spatial join; optimal data space partitioning; online partitioning**

I. INTRODUCTION

In this paper we review data partitioning techniques and concepts that incorporate data partitioning. Data partitioning is a standard feature in database systems today for better performance, manageability and availability. It enables large tables and indexes to be subdivided into smaller pieces improving database performance, manageability and availability. Each partition can have its own characteristics, like storage and indexing independently from the other partitions. It is a powerful technique to improve the overall performance of the system. Horizontal partitioning divides a table into subsets of tuples based on criteria. For example for every new month a separate partition can be created when the criteria are based on monthly partitioning of the data. Vertical partitioning divides the table in partitions where each partition contains a subset of columns. A combination of horizontal and vertical partitioning can be applied.

Standard SQL Queries take the advantage of data partitioning because they avoid scanning of entire partitions which are not part of the result. Queries are speed up to an order of magnitude for critical systems like Online Transactional processing (OLTP) and Data Warehouses. The query optimizer attempts to determine the most efficient way to execute a given query by considering possible query plans. A challenging area for researchers is to generate more efficient execution plans for queries over partitioned tables of current query optimizers.

Spatial database systems store data in simple geometric types like points, lines, polygons. This spatial data is often large and expensive to transfer from disk. Also spatial operations are often CPU intensive because of the computational expense of the required geometry algorithms. Two dimensional spaces can be divided using different partitioning strategies and we can have different types of partitioning the data.

Parallel databases have also been investigated [17], which are beneficial because of their performance increase gained with parallel processing. Spatial databases emerge because of their popularity in the market use [16] .Combining these two areas defines a new and challenging research area of *parallel spatial databases*. The concept of spatial data partitioning enables parallel execution because of the ability to execute spatial operations on different subsets of spatial data in parallel. MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers (nodes). It is called *a cluster* if the nodes are on the same local network and use similar hardware. *A grid* is when the nodes are shared across geographically and administratively distributed systems using more heterogeneous hardware. MapReduce already uses data partitioning concept behind for parallel data access

Dynamic partitioning techniques focus on dividing the unpartitioned data [6,7]. Different execution plans can be used for different subsets of input data. Offline partitioning has limitations because it depends on the Database administrator, which reconsiders the database design at certain points in time and creates data partitioning offline. Online automatic partitioning is a concept for data partitioning that does not need human intervention. Online automatic partitioning is an emerging research area [14].

Recent achievements for spatial query optimization, spatial parallel processing, optimal space partitioning and online automatic partitioning are considered in this paper. The goal of this survey is to understand and evaluate different partitioning techniques recently proposed. We structure the paper as follows. In section II we give a short explanation of the current implementations of data partitioning in one commercial relational database system (rdbms) – Oracle. In sections III and IV we present an overview of the theory presented in recent research papers for data partitioning. In Section V we conclude our study.

II. CURRENT IMPLEMENTATION OF DATA PARTITIONING IN ONE COMMERCIAL RDBMS-ORACLE

Oracle partitioning was introduced in 1998 in Oracle version 8.0. Since then Oracle continually enhanced the functionality of Oracle Partitioning with every release by adding new functionalities, new partitioning techniques or extending manageability [1,2]. In Table I the evolution of partitioning techniques in Oracle releases by years is presented.

TABLE I.

year	functionality	performance	manageability
1997	Range Partitioning, Global range indexes	Static partition elimination	Basic Maintenance
1998	Hash and composite range-hash partitioning	partition- wise joins 'Dynamic' partition elimination	Merge partitions Operation
2001	List partitioning		Global index maintenance
2003	Composite range-list partitioning	fast partition split	
2004	Global hash indexes		Local index maintenance
2005	1M partitions per table	Multi-dimensional prunning	Fast drop table
2007	More Composite choices, Ref partitioning, Virtual column partitioning		Interval Prtitioning, Partition Advisor
2013	Zone maps		Automatic Data Optimization/ Automatic partitioning policies

The new release of Oracle 12c differs from older versions because of new functionalities like zone maps and Information Lifecycle Management presented in Figure 1.

Static partition eliminates the partitions that are not part of result at compile time while for dynamic partitioning recursively evaluates the relevant partitions at runtime. Information Lifecycle Management with Automatic data optimization can easily map partitions with older data to less expensive data storage tiers and all that can be done automatically without manual intervention required. Data compression can be added on a partition level. Appling the Oracle Database In-memory option for partitions with recent data gives us x-time faster queries [3].

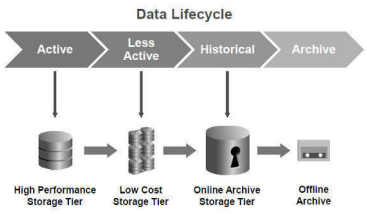


Figure 1. Information Lifecycle Management. [3]

Oracle optimizer can break the joining of two partitioned tables R and S that are equally partitioned into n one-to-one partition–wise joins. This technique is used for completing the overall join in less time which offers significant performance benefits for both parallel and serial execution. For joins where only one table is partitioned Oracle supports dynamic partitioning of the other unpartitioned table based on the primary table partitioning again creating one-to-one joins between partitions.

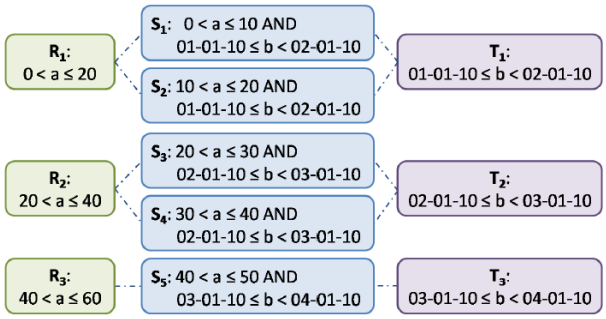
III. DATA PARTITIONING TECHNIQUES FOR ATTRIBUTE DATA

3.1 QUERY OPTIMIZATION TECHNIQUES FOR PARTITIONED TABLES

In papers [4,5] Herodotou et al. have explored the query optimization possibilities over partitioned tables. More challenges and opportunities exist today for partition aware query optimizers. Query optimizers are not progressing enough as the usage of table partitioning increase. Herodotou et al. conclude that optimizers need to be cleverer in eliminating partitions and generating efficient plans for executing queries over partitioned tables. Herodotou et al. present new partition aware optimization technique in four phases (applicability testing, matching, clustering and path selection) generating efficient plan for SQL query that contains joins over partitioned tables.

An example for partitioned tables RxSxT, query Q in Figure 2 and query execution plans P1 and P2 in Figure 3 are explained below.

In the query Q in figure 2 we have static predicates like R.a >= 5 and R.a <= 35. Most current optimizers will do partition elimination with the static predicates. This predicates will eliminate partitions from table R that have values of R.a which values do not intersect the interval [5..35] like partition R_3.




```
Q: Select *
   From   R, S, T
   Where  R.a=S.a and S.b=T.b and R.a >= 5 and R.a <= 35
```

Figure 2. Joined Partitioned tables with a simple query Q. [5]

Based on the transitive relation of join conditions partitions like T_3 and S_5 will be eliminated because there is nothing to be joined with. Most current optimizers will stop here with the optimizations of query plan executions.

Further optimization suggest the full join RxSxT can be broken up into smaller and more efficient partition wise joins that can be executed with different partition-aware join paths.

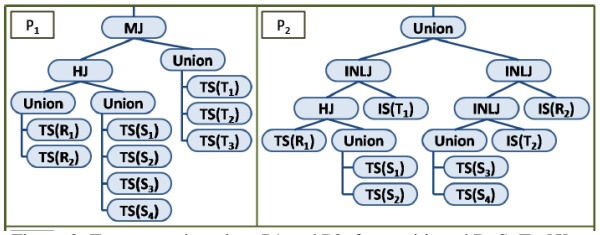


Figure 3. Two execution plans P1 and P2 for partitioned RxSxT. [5]

The optimizer may consider the best choice of join operators for R1 x (S1 U S2) x T1 different from the best join path for R2 x (S3 U S4) x T2 in Figure 3 – Plan 2. These different paths for smaller joins are because of storage , physical design differences across partitions (e.g., index created on one partition but not on another) or cardinality of partitions.
Only smart enough optimizers can eliminate partitions R3 and S5 and generate plans like P1. A plan like P2 (in figure 3) will significantly outperform plan like P1 because of breaking the larger join into smaller joins and the each small join can have its own smart path for execution of the partial join. Further Herodotou et al. explore the abilities of executing the small joins on partitioned tables in parallel. At the end an experimental evaluation with this technique of optimization prototyped in PostgreSQL optimizer using TPC–H queries is presented proving the whole concept.

These techniques focus on using existing partitioning while the techniques in next section 3.2 focus on dynamic partitioning of unpartitioned data. Our suggestion for upgrade to papers in this section 3.1 is to test this technique with spatial data. It would be interesting to see how this technique of breaking the larger join into smaller joins will behave with spatial data sets.

3.2 DYNAMIC PARTITIONING TECHNIQUES FOR UNPARTITIONED DATA THAT ENABLE DIFFERENT EXECTION PLAN ON DIFFERENT DATA

In paper [6] Polyzotis et al. present selectivity-based partitioning, a new approach to query optimization that adopts a divide and union approach. Selectivity Based partitioning takes into account the join correlations among table fragments in order to have multi joins (separate one join to many small joins which can be executed more effectively). The basic idea is to partition a table according to selectivity of the join operations, and then rewrite the query as union of queries over partitions. Polyzotis et al. present heuristic algorithm for computing effective partitioning for an input query.

In Selectivity Based partitioning taken into consideration is only joining of tables where one table is partitioned which is not the case of explained papers in previous section, papers [4,5]. The metrics for efficiency in paper [6] is the minimal total number of intermediate result, while for the papers in [4,5] the metrics for efficiency depends from the ability of the query optimizer to find good execution plan. However in paper [6] the main focus is on computing the partitioning of one relation (table) and determining the join orders that minimize the total number of intermediate result tuples for the union of queries.

Different subsets of data can have different statistics. Therefore it would be wise to process different subset of the data for a query using different plans. In paper [7] Bizarro et al. presents Content-Based routing (CBR) which is query processing technique that eliminate single plan execution. CBR automatically identifies classes – partitions of the input data and executes the query with different plans for each class–partition of data. CBR has adaptive approach to query optimization. The idea in CBR is that for each operator (filter query) we identify classifier attribute (column) and then based on the values of the column we divide the whole table to partitions on which the query is executed using different paths. The values of the classifier attribute should correlate with the selectivity of filter queries. The correlation between two distributions (values of column and selectivity of filter query) in this paper is specified using information theory (entropy).

3.3 AUTOMATIC ONLINE PARTITIONING

In paper [14] Jindal et al. present model for fully automatic online partitioning which does not require human intervention. Automatic optimal partitioning is not always reachable but however automatic (online) partitioning will in most cases be much better than the one suggested by DBA. This concept for automatic partitioning is based on the idea of affinity based partitioning. For vertical partitioning affinity between two columns is the count of their co-occurrences in a query workload. This produces affinity matrix between pairs of columns like in Figure 4.

	PartKey	Quantity	SuppKey
PartKey	8	6	5
Quantity	6	9	4
SuppKey	5	4	8

	PartKey	Quantity	SuppKey
PartKey	9	6	6
Quantity	6	9	4
SuppKey	6	4	9

Figure 4. Affinity Matrix. [14]

The core idea of affinity based partitioning is to compute affinities between every pair of attributes (columns) and then to cluster them such that high affinity pairs are close. This logic holds because high affinity columns means that both two columns will be part of many queries from the query workload and that is why they should be in same partition.
With each new query executed the affinity matrix is updated so the optimizer can decide whether new partitioning schema P' should replace the old partitioning schema P based on cost benefit model. This online repartitioning self tuning model called Autostore monitors the workload (execution of queries) and takes partitioning decisions automatically. Experimental evaluation shows that over changing query workloads Autostore outperforms other offline techniques. Our consideration regarding automatic online partitioning is that this will become more integrated in commercial databases and automatic online partitioning will be important direction of future research. Although some databases have simple online

partitioning (example every new month of data automatically goes into new partition) there is a lot to be implemented for online automatic partitioning and online repartitioning.

3.4 TRANSACTIONAL BASED PARTITIONING

Distributed transactions are expensive for shared nothing distributed systems. Data partitioning can be transactional sensitive. This means that partitioning should be considered and planned according to transactions. In paper [15] Curino et al. present Schism - new graph based data driven partitioning system which is transaction aware. The data base is represented as graph where tuples are nodes and transactions are edges between nodes like in Figure 5.

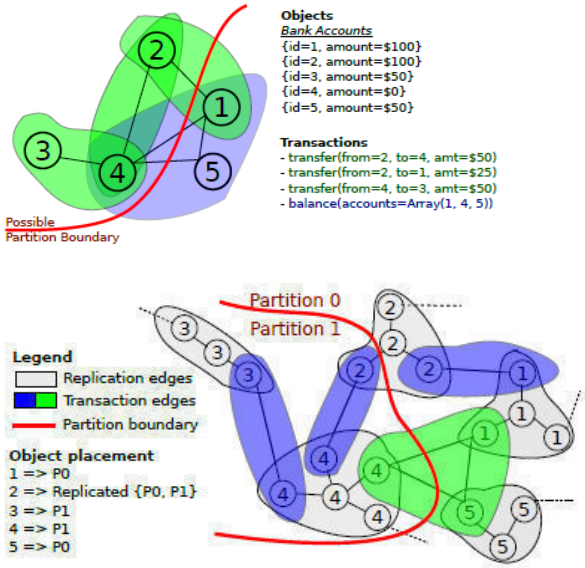


Figure 5. Example for graph representation in Schism, with replication. [15]

Graph partitioning algorithms are used to find partitions which minimize the number of distributed transactions. The basic idea is to group the data on partitions such that one transaction will be completed on only one partition which also requires replication of some tuples to many partitions.

3.5 PARTITIONING AS A KEY TO SCALABILITY OF CLOUD SYSTEMS

Cloud computing has emerged as a promising paradigm for highly scalable distributed systems. Concepts from cloud computing like infrastructure as a service, platform as a service or software as a service are very popular nowadays.
Social network applications require high scalability and availability. To provide the needed scalability and availability these systems compromise consistency. They use data storage services providing key-value data model and lack transactional guaranties. These data systems called NoSQL systems compromise consistency for better scalability and availability.
Between the traditional RDBMS and the NoSQL systems are third kind of systems for the cloud [18]. These systems rely heavily on data partitioning to provide its scalability. The way data is partitioned defines the system's ability to scale.

IV. DATA PARTITIONING TECHNIQUES FOR SPATIAL DATA

4.1 SPATTIAL JOIN ALGORITHMS IN CENTRALIZED DBMSs

Each spatial join consist of filter step and a refinement step. In filter step an approximation of spatial object like minimum bounding rectangle is used to eliminate tuples that are not part of the join result. In refinement step the candidates produced from previous step are examined for exact match using CPU intensive computational geometry algorithms.

Algorithms proposed for the filter step start from algorithm for transforming an approximation of spatial object to 1 dimensional space (Z-values algorithm). Next the nested-loop algorithm in which the tuples from one table are scanned for each tuple of the other table. Newer R-Tree join algorithms require spatial indexes on both sides of the join and represent synchronized depth first searches of both indexes.

Situation when there are no indexes could appear when the inputs of the join are intermediate results in complex query or in parallel environment where the inputs are dynamically redistributed. In Paper [8] Patel et al. present "Partition Based Spatial- Merge Join" (PBSM) - algorithm for spatial join which is effective when none of the inputs have indexes.

The basic idea in the filter step of PBSM lies in the partitioning of the 2D space into tiles and partitions like in Figure 6 and after that each partition is joined with the corresponding partition of the other table using geometry techniques.

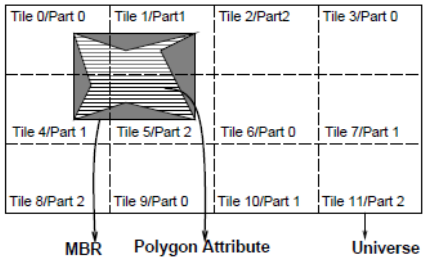


Figure 6. Spatial partitioning function using tiles. [8]

Experimental evaluation shows that PBSM is more efficient then the R Tree and indexed nested loop algorithm especially when neither of the inputs have indexes.
In paper [9] Nobari et al. present Touch In-memory spatial join algorithm. Main memory has grown so that data set can fit in. This algorithm uses hierarchical data oriented partitioning to organize bots spatial data set in one tree. Nobar et al. present approach for filtering step where the object of one data set represent leaves from the tree and object from other data set represent inner nodes from the same tree like in Figure 7.

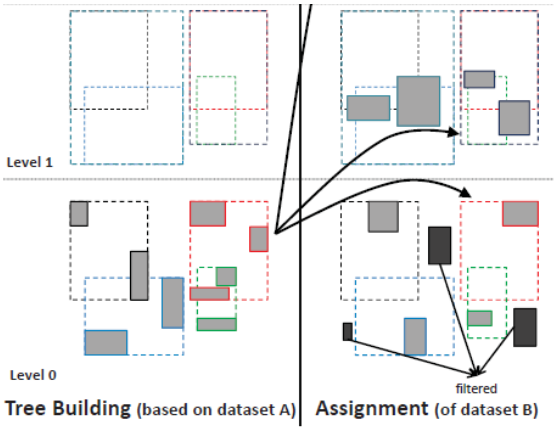


Figure 7. Three phases of touch: Tree Building, assignment and joining. [9]

Experimental evaluation confirms superiority of Touch compared to PBSM algorithm and R Tree joins. This is mostly because this algorithm uses duplicate avoidance rather then duplicate elimination like in traditional PBSM algorithm.

4.2 SPATAL JOIN ALGORITHMS IN PARALLEL DBMS

In parallel system the source of parallelism comes from partitioning. Partitioning parallelism is achieved by spreading the data across multiple nodes using some declustering function and then running operators at each node. Effective parallelism is achieved by good declastering technique and operator running at some node should access the data only on that node. In paper [10] Patel et al. present two parallel spatial join algorithms Clone Join and Shadow join. One strategy replicates entire tuples on the overlapping partitions, whereas the other strategy replicates an approximation of the spatial attribute. Experimental evaluation shows that the performance of the two algorithms depends of the characteristics of the data and the selectivity of the join.

In paper [11] Luo et al. present A Non blocking parallel Spatial Join algorithm (NBPSJ). This algorithm is import because it does not block the process waiting one phase to complete for start of the next. In paper [8] (Partition Based Spatial Merge Join) we had to wait for the declastering function to complete the spreading of all tuples to nodes and after that algorithm continues to next phase. The Non blocking algorithm enables while the tuples are redistibuted on nodes at the same time processing is done for each incoming tuple to test the join on this tuple with the corresponding tuples. This algorithm produces results continuously and the non blocking processing is done using two in-memory R Trees on both side on join relation. It also uses duplicate avoidance instead of duplicate removal using the revised-reference point method which significantly increases performance. Experimental evaluation shows that the Non Blocking Parallel Spatial Join (NBPSJ) is faster than parallel Partition Based Spatial Merge Join (PPBSM).

Map reduce is a parallel programming model proposed by Google for large scale processing on clusters of share nothing machines. An attractive quality of the MR programming model is simplicity. MR-style systems easily store and process what is known as “semistructured” data like key-value pairs. It is difficult to parallelize spatial join on MapReduce because of spatial joins process heterogeneous data while MapReduce process homogenous data. In paper [12] Zhang et al. present Spatial Join with MapReduce. This is the first parallel spatial join algorithm for MapReduce on clusters. SJMR maps every tuple into one or more partitions at the Map stage. Further partitions are divided into tiles using tile to partition mapping because of uniform distribution of data to partitions. Each partition correspond to one reduce task. The reduce step is composed of a filter step and a refinement step. The filter step is executed with a plane sweeping algorithm and duplicate avoidance method while the refinement step is executed using CPU expensive operations. This algorithm is used when neither of the inputs has index because input data set are dynamically distributed with MapReduce. Further Zhang et al. present experimental evaluation and comparison of this algorithm to traditional parallel PBSM [8] algorithm proving its efficiency.

4.3 OPTIMAL PARTITIONING FOR EFFICIENT I/O FOR SPATIAL DATA

Rectangles are most common shapes that are used for spatial indexes. Problem for partitions with circles is that space cannot be divided without overlaps and holes. Minimizing the I/O operations in spatial query processing is important for fast response time. In spatial queries the pages that have the result need to be retrieved from secondary storage. The number of page retrievals depends form the partitioning technique of the data. In paper [13] Fertatsmanoglu et al. compare different partitioning techniques for spatial data in order to minimize the expected number of partitions retrieved by query and decrease the number of disk accesses. Fertatsmanoglu et al. mathematically prove that hexagonal partitioning has optimal I/O cost for circular queries compared to other non overlapping techniques. The proof is based on the fact that the expected number of partitions that a query p intersects is minimized when the total boundary length of all partitions is minimized.

Partitioning using non convex regions with circles and wedges is also analyzed in this paper.

For the case of rectangular queries the standard rectangular grid partitioning gives optimal I/O cost.

Techniques for hashing, clustering and declustering that exists for rectangular partitioning can be adapted to hexagonal making hexagonal partitioning an effective alternative.

For traditional data base systems only rectangular partitioning is considered. Fertatsmanoglu et al. open the research space for other techniques like hexagonal partitioning which are not implemented in commercial databases and will probably give better performance for specific applications.

V. CONCLUSION

To improve performance, maintenance and availability in modern database systems, recent works have proposed data partitioning as a solution. In this review paper we show that this research area is still widely open for new concepts and methods for improvement. We have compared different techniques for data partitioning of spatial and attribute data.
In Table II we present the summary of previous techniques.

TABLE II.

Technique/Category	Practical Implemented/ theoretical	centralized DB	parallel DB	Attribute/ Spatial	Online / Offline	already partitioned tables)	Dynamic tech (for unpartitioned data)
Join Optimizati-on tech. for partitioned tables	practical imp. only for equi partition-ed tables	yes	yes	attr.	offline	yes	no
Selectivity based partitioning and content based routing	theory	yes	yes	attr.	offline	no	yes
PBSM	practical imp.	yes	yes	spat.	offline	no	yes
Touch	theory	yes	no	spat.	offline	no	yes
Clone Join / Shadow Join	theory	yes	yes	spat.	offline	no	yes
No blocking parallel Spatial	theory	no	yes	spat.	offline	no	yes
Optimal partitioning for spatial data	theory (partial practical imp.)	yes	yes	spat.	offline	yes	no
Automatic partititiong	partial imp. in Oracle ILCM	yes	yes	both	online	no	yes

Overall we conclude:
1. Query optimization techniques have not won the race with the growing usage of data partitioning. Therefore new techniques need to be implemented in commercial data base systems.
2. Partitioning is the source for parallelism in parallel database systems. We see that the concept of data partitioning is incorporated in many techniques as a fundamental point. Cloud computing uses data partitioning as a key to scalability. Most of the techniques discussed are experimentally confirmed.
3. Database systems will become more and more automated by using data partitioning. Therefore, in future releases of commercial database systems there will be less human intervention. We expect future releases of database systems to have data partitioning incorporated in the way which users will benefit from it, without knowing that the database design is partitioned. This is not the case in current settings of database systems.
The techniques we discussed in this paper are useful review for anyone moving to the next step and modifying the existing methods for more efficient database partitioning.

REFERENCES

[1] Oracle Partitioning with Oracle Database 12c, Partitioning with Oracle Database 11g R2

[2] Oracle VLDB and Partitioning Guide 11g Release 2

[3] Oracle Database 12c New Features

[4] H.Herodotou, N.Borisov, S.Babu "Query Optimziation Techniques for partitioned tables" in SIGMOD '11, 2011, Athens

[5] H.Herodotou, N.Borisov, S.Babu "Joined Optimization techniques for partitioned tables" , 2010

[6] N.Polyzotis, "Selectivity- Based Partitioning: A devide and union paradigm for Effective Query Optimization" , CIKM05 , New York

[7] P. Bizarro, S.Babu, D.DeWitt, J.Widom, "Content-Based Routing: Different plans for different data", VLDB 2005, Trondheim, Norway

[8] J.Patel, D. Dewitt, "Partition Based Spatial – Merge Join" , SIGMOD 1996, Montreal Canada

[9] S.Nobari, F.Tauheed, T.Heinis, "Touch: In memory Spatial Join by Hierarchical data oriented partitioning" SIGMOD 2013 , New York

[10] J.Patel, D.Dewitt, "Clone Join and Shadow Join: Two Parallel Spatial Algorithms" ACM GIS 2000

[11] G.Luo, J. Naughton, C. Ellmann, "A Non Blocking Parralel Spatial Join" Data Engineering 2002 , San Jose , CA

[12] S. Zhang, J.Han, Z.Liu, K.Wang, "SJMR Parallelizing Spatial Join with MapReduce" , 2009

[13] H.Fertatsmanoglu, D.Agrawal, A. Abbadi, "Optimal partitioning for Efficient I/O in Spatial Databases" , 7 International conference , UK , 2001

[14] A.Jindal, J.Dittrich, "Relax and Let the Database do the partitioning Online", VLDB 2011 Seattle USA

[15] C. Curino, E. Jones, Y. Zhang, and S. Madden. "Schism: a workload-driven approach to database replication and partitioning" VLDB 2010., Singapore

[16] X. Zho, D.J.Abel, D.Truffet, "Data Partitoning For Parallel Spatial Join Processing" , 1998 ,Journal Geoinformatica

[17] M.Stonebraker, D.Abadi, D.Dewitt, S.Madden,E.Paulson, A.Pavlo,A.Rasin, "MapReduce and ParallelDBMSs friends or Foes? " 2010

[18] F.Maia, J.Armendiras, M.I.Ruiz-Fuertes, R.Oliviera, "Scalable Transactions in the Cloud: Partitioning Revisited", OTM 2010

Better IT Administration with the help of Network Data Acquisition

Arbnor Grubi, Biljana Stojcevska, Ivana Stojanovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
arbnor.grubi@gmail.com, {stojcevska, ivana.stojanovska}@uacs.edu.mk

***Abstract*—The issues that every IT staff faces in an enterprise grow on a nearly daily basis. Keeping track of every applied fix be it hardware or software related takes time, and especially if the issue of the certain fix is more or less "sporadic". This paper suggests the following approach: gathering the data from the network workstations by using the Windows Management Instrumentation (WMI) and storing them in a centralized database. In order for this kind of data acquisition to work, means have to be implemented, for the Client, which will be installed in the end user computer, to have the ability to search, find and verify the service, regardless of whether its code has moved to another machine on the network. The usage of the UDP protocol is therefore suggested, in order to map the entire network and pinpoint the service code in its new destination, confirm its validity and authenticity and continue the data acquisition through a newly setup safe communication protocol. This way, no matter where the service or client resides, they will both be able to find and verify one another and enable the IT staff to keep track of the computers within the enterprise. The server thus becomes the point where all of the data gathered from the computers is sent and stored.**

Keywords—Windows Management Instrumentation; IT administration; Network Data Acquisition

I. INTRODUCTION

The current objective of this paper is to provide relevant insight on and around methods used for network data acquisition in developing institutions where the number of worker computers grows accordingly.

As an enterprise grows in the amount of work that it has to accomplish during the course of a day, so does the need for more complex and smart applications. In the most simple of enterprises there are at least 2-8 applications that are "a must have", each of those applications get their data from equipment that are also necessary for the completion of certain tasks, equipment that also falls under the managerial scope of the IT staff. It goes without say that as everything within the scope IT grows in complexity so does the workload that the IT staff has to be able to handle.

As the worker gets more and more sophisticated and "smart" equipment that enables him/her to finish their tasks more efficiently, the IT Staff slowly but surely drowns under a sea of procedural checks, fixes, reports that almost always have to be done manually.

The IT staff has to achieve the following:

- acquire useful information about each individual computer
- keep track of every change made to any computer
- become better equipped to support future upgrades
- prepare statistical reports which will help to quickly identify and take more efficient action in any unforeseen events.

This papers' approach leads to a certain direction. Its purpose is to develop an application for the issue of organizing an enterprises network and its worker nodes, archive the issues and fixes for those issues into a centralized database and provide an easier way for monitoring and managing the entire network in a more customizable way.

Therefore it aims to help the IT staff better understand the complexities of their working environment and the various approaches most often taken to ensure its sustainability and attempts to provide a starting ground for a solution to the issue.

The content is organized in four sections. Section two presents a brief introduction and explanation on and around the initial underlying issues. Furthermore, section three consists of a brief overview WMI – Windows Management Instrumentation and some of its features, continuing with a suggested solution in section four. Afterwards, it continues with presentation of the results from the executed tests and discussion. The paper concludes with section six, where the general conclusions are made.

II. PROBLEMS

The issues that every IT staff faces in an enterprise grow nearly on a daily basis. Keeping track of every applied fix be it hardware or software related takes time, and especially if the issue of the certain fix is more or less "sporadic".

Given that each computer is unique in terms of the user that is using it, it will face different issues be they hardware or software based on its usage. It will also maybe be in a different subnet, floor and office or even if it has any issues, an untrained person won't report them simply because they can't spot them.

So the issues that are addressed here are:

- The inability of the IT staff to be at 2 places at the same time
- Poorly keeping tabs on the performance and issues of each computer
- Lack of detailed run-time reports and statistics
- Trusting the user.

III. WINDOWS MANAGEMENT INSTRUMENTATION

Windows management instrumentation (WMI) is a suite of tools and extensions within the Windows driver model that enables scripting languages to manage PCs and servers connected remotely or locally. It provides the ability to manage distributed enterprise systems for servers and workstations powered by the Windows OS, and to automate system management tasks. [1]

Features of WMI are [2]:

- remote capabilities
- support for Queries (WQL) out of the box
- event capabilities
- code template generator
- predictability
- protect existing customer investments
- provide a logical and unified administration model

IV. SOLUTION

This paper is intended to enhance the understanding of IT staff, providing them with the analytical and assessment frameworks that will help them implement context-specific activities. There is indeed no “one size fits all” process for dealing with every issue; appropriate solutions must always be grounded in local, context-specific understanding.

The following approach is suggested:

- Utilizing the Windows Management Instrumentation (WMI);
- Network Discovery via UDP;
- A centralized relational database where the data will be stored.

WMI has the capabilities to list and change software and hardware information for any computer running Windows. However, if this is done remotely, every computer will need to be granted remote access for this, which any cautious IT administrator will deny due to the chances of an arbitrary attack on said access, such are cases spoken in papers like [3]. To alleviate this, the WMI enables the data to be retrieved from the computer that is running, this further increases the efficiency of the process while decreasing the load on the network due to the fact that no remote connection is required, and streamed via a more safe and encrypted protocol to a centralized database. This reduces the risk of a third party possibly gaining access to the computer itself.

The following problems need to be addressed:

- Different enterprise computer networks are different in architectures;
- The centralized database may have to run on a different computer due to unforeseen maintenance necessities;
- The connection between the client and server must be continuous in case of any unforeseen changes to any of the computers.

In order for this kind of data acquisition to work, means have to be implemented, for the Client, which will be installed in the end user computer, to have the ability to search, find and verify the service, regardless of whether its code has moved to another machine on the network. The usage of the UDP [4] protocol is therefore suggested, in order to map the entire network and pinpoint the service code in its new destination, confirm its validity and authenticity and continue the data acquisition through a newly set up safe communication protocol. [Fig. 1]

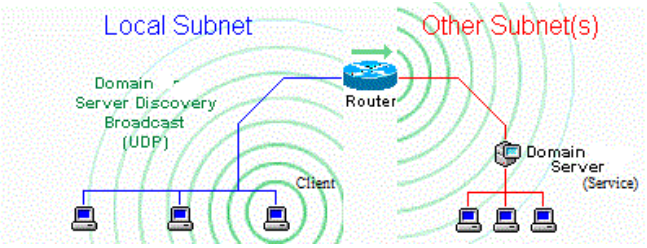


Fig. 1. UDP Broadcast Discovery([4])

A handshake protocol needs to be implemented using the chosen communications protocol, encrypted messages that suggest which is the Client and which the Service must be appointed or generated randomly for identification purposes and the usage of MAC Addresses is wisely encouraged for cases where there are more than one user accounts within a single machine.

Furthermore the Client needs to be able to tell when it’s being run for the first time and under which user account it’s been run. This is suggested in order to avoid duplicate entries of data and to further ease the process and accuracy of the acquisition. The gathered initial data as well as the changes made in accordance and comparison to the initially gathered data need to be stored locally before any other process comes to play.

A rulebook of a sort should be set for the Client to follow and methods and procedures should be coded into the client for it to abide those rules. It’s important that the Client should know when to report what as we want specific types of information as content specific reports, warnings, errors, and so on to be prioritized in order to achieve as near as a run-time presentation of any certain situation. Of course the entirety of this will prove to be a custom property for every individual IT Administrator, although certain parts of it remain the same for all.

Utilizing the WMI and its classes [5] and following the guide rules, the Client can now efficiently gather the

required and categorically set data by the IT Administrator to then transmit/upload to the Service and continue doing so till it's no longer part of the enterprise. [Fig. 2]

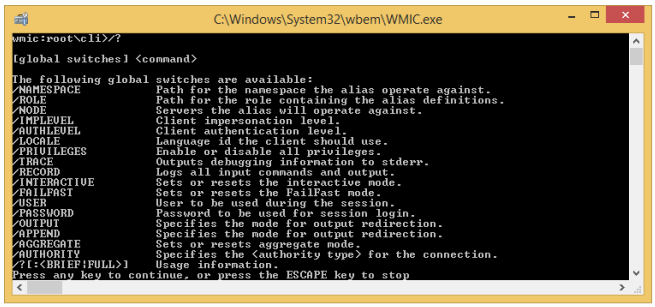


Fig. 2: WMI running on a local computer (Windows 8.1)

The service to which the Client connects should have the ability of multithreading in order to handle multiple connections simultaneously. It should be the beacon that each Client sees and identifies and be the only one within the Enterprise network in order to avoid fatal crashes. Each client broadcasts locally gathered data into the Service, so the Service has to be able to download the data and categorize it within the centralized database.

The centralized database can then be used to simplify the work of the IT staff in terms of

- keeping detailed reports of the changes;
- issues that have been reported and their causes;
- fixes that have been applied to each computer individually;
- better preparation of the enterprise for future upgrades;

This way, no matter where the service or client resides, they will both be able to find and verify one another and enable the IT staff to keep track of the computers within the enterprise. The server thus becomes the point where all of the data gathered from the computers is sent and stored.

Given that we can then monitor the network more closely, we will also manage it remotely by issuing commands to either an individual computer or a group of computers. Then, certain WMI methods and classes need to be implemented within the Client, enabling it to receive IT Administrator specific set commands and then issue them to the local computer WMI, making remote management of network, user accounts, shared folders, permissions and more, easily possible and further removing the need of power shell scripting and remote connections.

In an enterprise where the number of computers reaches a level when the IT Staff has to grow in number and further branch into different enterprise sectors in order to be able to maintain the system's sustainability, it's recommended that the data that is gathered through the acquisition be targeted specifically prior to the further development of the platform.

As noted earlier, every enterprise and its network represent their own nature and environment, thus creating the need for processing different types of data, for example if an enterprise works with large datasets and information it would normally lead towards securing safer and higher quantities of data storage drives and the IT staff would be in title to handle the transfers and backups of those datasets and information. Thus meaning that was this particular IT Staff would prioritize in would probably be a way to constantly keep watch on the remaining free space of the drives and also the drives' data-writing validity status and reports.

In an enterprise that is in a line of work where its computers more often than not break down or require repairs, it's recommended that the centralized database resembles that of a computer motherboard, accessory and repair shop. By doing so, each computer gets its history logged in details from the cause to the fix, enabling the IT staff to browse the past and thus increasing the possibility of much more efficiently finding the clues that might suggest a near-time breakdown or a much quicker fix.

By doing so we apply a near systematic approach to the way that IT Administration is done. And the way that this "system" and its characteristics are monitored, managed and sustained allows much more room for future "smarter" automations. For example, if we were to further simplify the representation of the acquired data, computer nodes, command lists and so on within a carefully designed Graphical User Interface, the IT staff would be able to view information in real time or with the press of a button making it the same for issuing commands. It would also be possible for the creation of an elaborate methodical "smart" system which could automate this systems management based on the gathered data enabling the IT staff to focus on more sophisticated tasks.

V. IMPLEMENTATION

The client, no matter where it's going to be installed needs to be able to find the Service code to then be able to initiate any type of connection in order to retrieve the data gathered, in this case we shall pull a simple CPU/Processor information from the computer or in our case laptop where the client will be running.

We first start by preparing our service to be able to identify itself on the network. We've set up a socket with the UDP Protocol to a certain port, furthermore we've made it listen to all the endpoint IP addresses aside from the one from where it is listening from. Next we prepare a reasonable response and wait for a client to hail this service.[7] [Fig. 3]

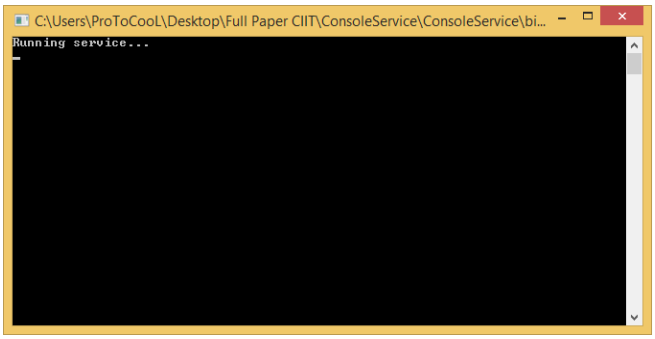


Fig. 3. Service code running on Windows 8.1

On the client side we prepare a socket to use the UDP protocol, furthermore we broadcast a packet to everyone in the network and then listen to all the responses that get returned. [8][Fig. 4] At the same time, even though this should be done in a more elegant manor, for demonstration purposes we've included the WMI management classes and called methods for retrieving detailed processor information and then print them on the screen.

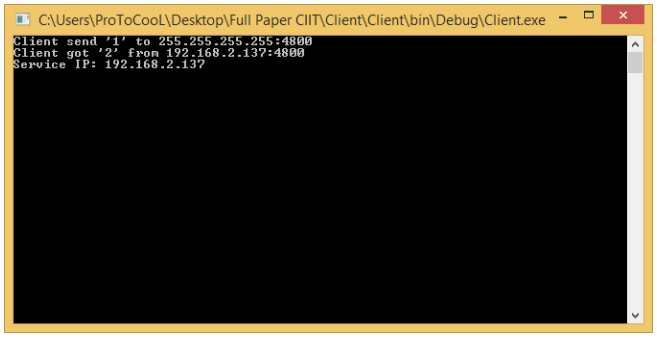


Fig. 4. Client code broadcasting-receiving on Windows 8

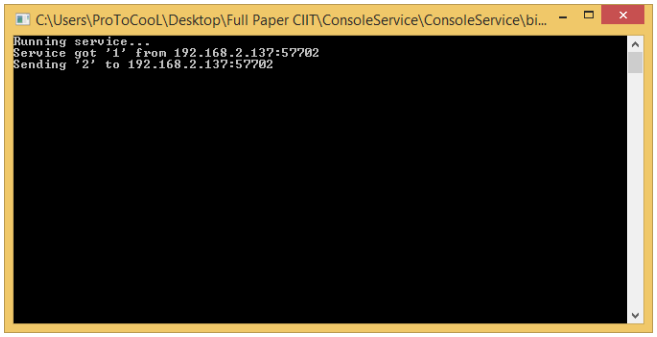


Fig. 5. Service code sending IP Address to Client on Windows 8.1

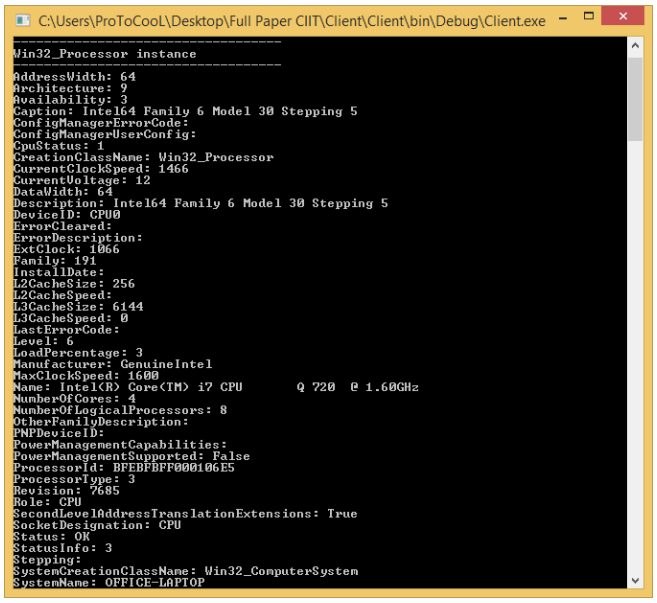


Fig. 6. WMI printing CPU Information

Once the client and the service have identified one another the client can continue to gather all the required information from the computer it's running on and send them to the newly acquired IP Address for further processing. [Fig. 6]

As seen through Figures 3-6, this small basic setup is all that is required for a test run and the information that can be acquired can then be either placed into a database, kept in files, printed.

VI. CONCLUSION

The purpose of this paper was to demonstrate how methods for data acquisition can be used in tandem with methods for IT Administration to ease the workload and further increase the accuracy and efficiency of the IT Staff. However, the reality of this is that each individual or group under the IT Staff will devise a different plan based on their dataset and situations and is left to their own creative way in utilizing the information. Another task for the future will be to further automate the tasks of an IT Administrator by utilizing more complex WMI and system specific classes.

References

[1] Windows Management Instrumentation – Microsoft Developer Network - https://msdn.microsoft.com/en-us/library/aa394582(v=vs.85).aspx .

[2] Windows Management Instrumentation – Wikipedia http://en.wikipedia.org/wiki/Windows_Management_Instrumentation [Accessed February 16, 2015]

[3] J. Dizon, L. Galang and M. Cruz "Understanding WMI Malware", A Trend Micro Research Paper – July 2010 – Available at http://la.trendmicro.com/media/misc/understanding-wmi-malware-research-paper-en.pdf

[4] Communication Networks/TCP and UDP Protocols/UDP – WikiBooks - http://en.wikibooks.org/wiki/Communication_Networks/TCP_and_UDP_Protocols/UDP

[5] WMI Classes – Microsoft Development Network https://msdn.microsoft.com/en-us/library/aa394554(v=vs.85).aspx .

[6] Remote Broadcast 1 –- https://www.greyware.com/software/domaintime/technical/network/subnets.asp.

[7] A. Grubi, Service Code - http://pastebin.com/uBfDVaHh .[Accessed March 18, 2015] .

[8] A. Grubi, Client Code - http://pastebin.com/2iJKq6pL . [Accessed March 18, 2015]

Towards Integration of Wireless Sensor Networks into Smart Home

Biljana Risteska Stojkoska
Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, Macedonia
biljana.stojkoska@finki.ukim.mk

Abstract—**Wireless Sensor Networks (WSN) brings significant advantages over traditional communication technologies for smart grid applications due to their low-cost, collaborative and long-standing nature. Still, real implementations are very rare, work separately and do not posses capabilities to be integrated into one common framework.**

In this paper we aim to overview the potential application of WSNs for smart grid and smart home systems. We propose a novel comprehensive framework that will achieve efficient integration of sensor networks into smart home. Our framework represents a three-tier system architecture that can be easily extended, modified and customized regarding users' preferences. Furthermore, it can be used as a solid base for future smart grid applications that use wireless sensor networks.

Keywords— smart home; smart grid; wireless sensor network; Z-Wave; ZigBee; smart metering; energy management

I. INTRODUCTION

The expected growth in world population would increase the demand for energy. It is predicted that world energy consumption will grow by 56% between 2010 and 2040 [1]. Current power grids, built decades ago, would not be able to fulfill the new demands. Therefore, smart grid solutions should be used as a tool to achieve optimization.

Smart grid is a new concept that integrates information and communication technologies (ICT) with grid power systems in order to achieve efficient and intelligent energy generation and consumption. It is characterized by a two-way flow of both electricity and information. Smart grid approaches search for novel solutions that would effectively explore existing power grid. Hence, the blackouts, voltage sags and overloads should be reduced or eliminated. Reliability, quality, security and safety of the power grid would increase.

Combination of information technologies and advanced communication and sensing systems, creates a variety of new potential application. The new advanced concepts, such as pervasive or ubiquitous computing [2], impose huge potential to be integrated into smart grid application [3]. Smart devices, capable of computing and communicating are present everywhere around us, from simple sensor nodes to sophisticated smart phones. Such a heterogeneous wireless sensor networks (WSN) composing of home appliances and smart gadgets, if combined with smart metering, can transform residential houses, homes and offices into energy-aware environments. Using intelligent power scheduling algorithms, residents would be able to make optimal a priory choices about how to spent electricity in order to decrease energy consumption [4]. On the other hand, utilities would benefit as load demand would be balanced in critical situations.

In this paper, the potential state-of-the-art WSNs applications for smart grid and smart homes have been presented. A novel comprehensive framework for integrating WSN into smart home has been introduced. To the best of our knowledge, this is the first work that is suggesting hierarchical architecture for smart home systems using WSN. Furthermore, we briefly discussed designing issues. Knowing that WSN nodes have constraint resources (energy, memory and processing), intelligent techniques are needed to face these challenges.

The rest of this paper is organized as follows. Section 2 and Section 3 survey WSN applications for smart grid and smart home, respectively. In the forth section we present our novel WSN framework. Finally, this paper is concluded in Section 5.

II. WIRELESS SENSOR NETWORKS IN SMART GRID

Smart sensing devices that communicate with each other are very popular as Wireless Sensor Networks (WSN), a paradigm that was tremendously explored by the research community in the last decade [5]. Nowadays, this concept is being widely integrated within smart grid. Table 1 separates the potential WSN application regarding different sides (parts) of the power grid [3][6].

Many of the pioneering researches impose a lot of challenges for WSN designers, especially when dealing with the first three sides (generation, transmission and distribution) of the power grid. The reason is mostly due to harsh conditions in which the WSN is deployed. Experimental results using IEEE 802.15.4-compliant sensor networks show that wireless links (including both line-of-sight (LOS) and non-LOS (NLOS) scenarios) in smart grid have high packet error rates and variable link capacity due to electromagnetic interference, equipment noise, obstruction, etc. [7].

Fortunately, most of these challenges are not present in the fourth, consumer side of the smart grid, i.e. smart home. For example, sensors are usually connected to the home appliances

TABLE I. WSN APPLICATIONS FOR SMART GRID

Energy providers Energy generation	Transmission	Distribution	Consumers
Real-time generation monitoring	Transmission lines controlling	Underground cable system monitoring	Wireless automatic meter reading (smart metering)
Power plants controlling	Power monitoring	Transformers stations controlling	Home (Residential) energy management
Alternative energy sources controlling			Solar panels management
Residential (distributed) production monitoring			Predicting future solar panels and wind turbine production (using sensor data like temperature or humidity)

and the problem with battery recharging shouldn't be considered, as the devices have steady power supply. Additionally, strong electromagnetic fields are not associated with home grid infrastructure. Still, smart home WSNs pose challenges like reliability, security and resource constraints (due to memory and processing limitations of WSN nodes) [8] [9].

III. WIRELESS SENSOR NETWORKS IN SMART HOME

Smart devices, capable of computing and communicating are present everywhere around us, from simple sensor nodes to sophisticated smart phones. Modern homes equipped with smart meters, smart appliances, smart power outlets and sensing devices enables the development of energy-aware smart homes (Fig. 1).

Although smart home has been a dream for both utilities and consumer for a long time, such implementations are still very rare.

Novel architecture regarding state-of-the-art software technologies with focus on domestic environments and habitat monitoring are proposed in [10] and [11] respectively. In [10], the authors advocate design guidelines toward collecting and integrating household data, thus enabling data interoperability. In [11], a web interface is developed in order to increase the interaction between the deployed WSN and its end users. Authors of [12] proposed a solution for a Web-based energy-aware smart home framework that enables smart appliances to the Web. They have developed a graphical user interface to easier the interaction. The evaluation was done with WSN

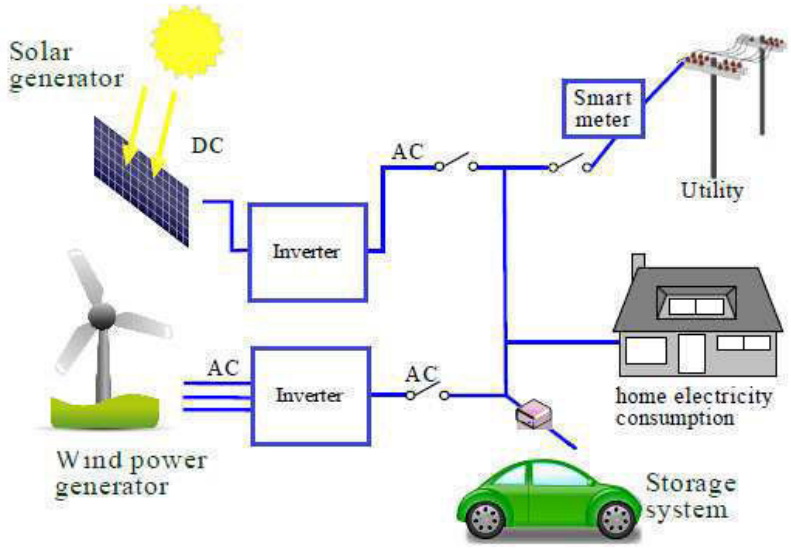


Fig. 1. Smart home

organized as star topology and multihop topology (up to three hops) for larger apartments (smart homes of around 100 m2).

VillaSmart [13] is associated with ECOGRID EU [14] project. The authors have installed a modular and extensible WSN in a test and reference household called VILLASMART. The authors are modeling the energetic behavior of the building. These thermal models are improved using indoor and outdoor WSN readings (air and water temperature, solar radiation sensor, weather conditions and power consumption information), thus achieving more precise predictions of indoor temperature. Using standard resistance-capacitance (RC) model, the maximum prediction error was 1.790C. IEEE 802.15.4 standard in the 2.4 GHz was used for indoor communication, while grey-box estimation method was used for model parameter determination. In [15], the authors present a framework for temperature regulation inside commercial and administrative buildings, with focus on design and implementation regarding network topologies and node localization.

Currently, the main issue toward development of generic smart home solutions is the cost associated with integrating smart home devices. Leading companies in the world that are producing smart devices are working toward to achieve full interoperability that will ensure easy integration with the exiting Internet [16][17]. It is expected that smart objects will invade the market in the next few years and will become vastly present in consumers households, which will impose the need for new and enriched services. For these reasons, the need for smart home solutions will be inevitable.

IV. WSN-BASED FRAMEWORK FOR SMART HOME

WSN integration within smart homes is partly achieved as sensor devices at homes are widely present and attached to the home appliances. For example, smart boiler and smart TV are equipped with communication module and have own IP addresses. Therefore, they can be easily controlled remotely. Smartphones are equipped with different sensors, like accelerometer, microphone, proximity sensor, GPS and ambient light sensor. Smart phones can serve as processing units in smart home, as they can easily communicate with smart meters and other appliances which are equipped with communication interfaces. Hence, they can receive data and perform sophisticated algorithms for optimal load balancing, or create task schedulers in order to achieve reduction in energy consumption.

A. *System architecture*

In this section we are going to explain in details our WSN-based framework for smart home. The system has simple architecture (Fig. 2).

1) First level (Smart home): All household devices, equipped with interfaces for wireless communication, compose home WSN. Each home has its own WSN, and the sensed data are collected in a central station (sink) that is represented with the **home sink**. Home sink can be a smart meter or any other device that can perform data storage and local processing (PC, tablet or smartphone). Each node of the network (home device) can accomplish advanced computational and communicational operations. Mesh is the most suitable choice for network topology due to presence of obstacles in the home, like walls, furniture, etc.

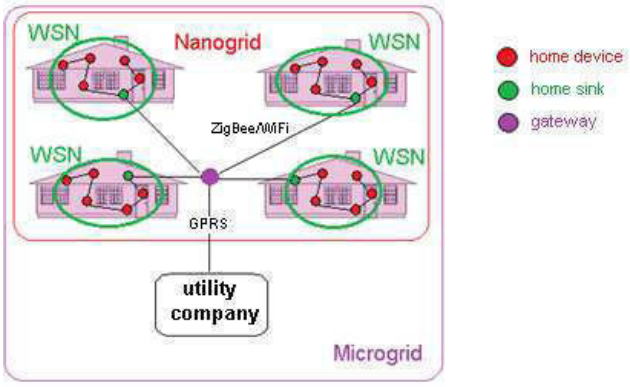


Fig. 2. Three-tier smart home architecture

The communication protocol is based on ZigBee standard. The extension of this standard known as ZigBee PRO Green Power [18] is widely accepted as the global standard for indoor WSNs. Other technology is Z-Wave, which can be used if the network consists of less than 70 nodes [19]. Table 2 compares these two protocols.

2) Second level (Nanogrid): On the next level, all home sinks from the building can communicate with each other and exchange information. Network topology on this level can be both cluster-based or mesh, depending of the residential complex. In case of smart building, mesh topology is more suitable, as smart meters can not always sent its data directly to the gateway due to obstacles in the building. If there are detached houses in the smart residential complex, than star topology or cluster topology is more appropriate solution, which can be accomplished using WiFi.

3) Third level (Microgrid): The gateways of all residential buildings communicate with the utility (through GPRS, 3G or optical fiber). This can be realized using cloud computing, as a state-of-the-art paradigm that is already commonly accepted for such problems. The typical information that can be exchanged between the gateway and the utility are: price of the electricity, current and future consumption of the microgrid, current and future production of the distributed production sources associated with the microgrid, etc.

B. *Smart home management systems*

In-home (residential) energy management system presents an interface between a utility company and smart devices present at home. It aims to provide benefits for both parties (utilities and consumers), but with focus on consumers.

Demand Side Management (DSM) is a set of technologies that enable monitoring and controlling the consumption/production at customer level in order to perform power balancing in future energy systems [20]. In our

TABLE II. COMPARISON OF ZIGBEE AND Z-WAVE PROTOCOLS

	ZigBee	Z-Wave
Frequency	2.4 GHz (worldwide), 915 MHz (Americas and Australia) and 868 MHz (Europe)	< 1GHz, (868.42 MHz Europe; 908.42 MHz United States; 916 MHz Israel; 919.82 MHz Hong Kong; 921.42 MHz Australian/New Zealand)
Network topology	mesh	mesh
Vendors	Texas Instruments, Atmel, Silcon Labs, Freescale, etc.	Sigma Designs
Maximum number of nodes	65000 (theoretically), 500 (in practice)	256 (theoretically), 70 (in practice)
Range	10 to 100 meters line-of-sight	30m open-air, < 30m indoor
Modulation	Binary phase-shift keying (BPSK) for 868 and 915 MHz bands, or offset quadrature phase-shift keying (OQPSK) for 2.4GHz	Gaussian, frequency-shift keying (GFSK) manchester channel encoding
Routing protocol	Ad-hoc On demand Distance Vector (AODV)	Source-based routing

framework, there is two-tier energy manager that consists of home manager (HM) at the first level associated with the home sink and residential manager (RM) at the second level associated with the gateway. HM and RM of the nanogrid are connected with interface that is capable of:

- Data reduction inside the WSN
- Data extraction
- Decision making
- Visualization

C. Designing issues

Large and complex collection of sensed data are increasingly being gathered by ubiquitous information-sensing mobile devices and Wireless Sensor Networks (WSN) deployed everywhere around us. Transmitting, storing and analyzing of the large datasets is not a trivial task, as traditional methods usually fail to respond with desired scalability. Most of these traditional operations aren't energy harvesting and doesn't meet the "green" paradigm by means of networking and computation. Hence, different mechanisms are proposed to solve these issues.

Reducing the number of transmissions is very important in order to avoid latency issues and saturation of the wireless channels, especially if using Z-Wave communication protocol [19]. For these reasons, different data reduction techniques should be employed. To meet the network requirements, instead of sending all sensed data, only critical data could be sent to the gateway [21]. If data are not needed in real time, data compression (like delta compression) can be used. Furthermore, local data readings can be store only at the gateway side, while metadata can be transmitted further to the main cloud centers.

The most important issue of the emergent requirements facing the smart grid development is related to cyber security of the systems. Smart grid can be a target for cyber terrorists, which implies a critical concern for system designers. WSN require security but demand lower cost for providing it. Therefore, different approaches that provide security through cryptography should be investigated [22].

V. CONCLUSION

The vision is that the residential buildings would shift themselves toward modern houses that would be an evolution of the passive house. They would have its own solar panels and small wind turbine to produce its own energy, thus they would be able to buy and sell energy to the power grid.

This paper overviews WSNs applications for smart grid and smart home systems. We propose a comprehensive WSN framework for smart home applications. Our framework has simple architecture which is designed for hierarchically organized WSN. We believe that our framework can serve as a solid base for future WSN-based smart grid applications, as it can be easily extended to other part of the smart grid, i.e. production and distribution parts.

ACKNOWLEDGMENT

This work is supported by the project "INLOC - INdoor LOCalization for service oriented mobile applications", financed by the Faculty of Computer Science and Engineering University "Ss. Cyril and Methodius" in Skopje, Macedonia.

REFERENCES

[1] US Energy Information Administration (EIA). International Energy Outlook 2013; EIA, Washington D.C., WA, USA, 2013.

[2] Greenfield, A. (2010). Everyware: The dawning age of ubiquitous computing. New Riders.

[3] Parikh, P. P., Kanabar, M. G., & Sidhu, T. S. (2010, July). Opportunities and challenges of wireless communication technologies for smart grid applications. In *Power and Energy Society General Meeting, 2010 IEEE* (pp. 1-7). IEEE.

[4] Koutsopoulos, I., & Tassiulas, L. (2011, May). Control and optimization meet the smart power grid: scheduling of power demands for optimal energy management. In *Proceedings of the 2nd International Conference on Energy-Efficient Computing and Networking* (pp. 41-50). ACM.

[5] Oppermann, F. J., Boano, C. A., & Römer, K. (2014). A decade of wireless sensing applications: Survey and taxonomy. In *The Art of Wireless Sensor Networks* (pp. 11-50). Springer Berlin Heidelberg.

[6] Gungor, V. C., Lu, B., & Hancke, G. P. (2010). Opportunities and challenges of wireless sensor networks in smart grid. *Industrial Electronics, IEEE Transactions on*, *57*(10), 3557-3564.

[7] Gungor, V. C., & Korkmaz, M. K. (2012). Wireless link-quality estimation in smart grid environments. *International Journal of Distributed Sensor Networks*, *2012*.

[8] Iyer, S. (2011). Cyber security for smart grid, cryptography, and privacy. *International Journal of Digital Multimedia Broadcasting*, *2011*.

[9] Li, F., Qiao, W., Sun, H., Wan, H., Wang, J., Xia, Y., ... & Zhang, P. (2010). Smart transmission grid: Vision and framework. *Smart Grid, IEEE Transactions on*, *1*(2), 168-177.

[10] Monacchi, A., Egarter, D., & Elmenreich, W. (2013, May). Integrating households into the smart grid. In *Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES), 2013 Workshop on* (pp. 1-6). IEEE.

[11] B Stojkoska, B., & Davcev, D. (2009, August). Web interface for habitat monitoring using wireless sensor network. In *Wireless and Mobile Communications, 2009. ICWMC'09. Fifth International Conference on* (pp. 157-162). IEEE.

[12] Kamilaris, A., Trifa, V., & Pitsillides, A. (2011, May). HomeWeb: An application framework for Web-based smart homes. In *Telecommunications (ICT), 2011 18th International Conference on* (pp. 134-139). IEEE.

[13] Caracas, A., Mueller, F. L., Sundstrom, O., Binding, C., & Jansen, B. (2013, October). VillaSmart: wireless sensors for system identification in domestic buildings. In *Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2013 4th IEEE/PES* (pp. 1-5). IEEE.

[14] EcoGrid EU, http://www.ecogrid-eu.net

[15] Risteska Stojkoska, B., Popovska Avramova, A., & Chatzimisios, P. (2014). Application of wireless sensor networks for indoor temperature regulation. *International Journal of Distributed Sensor Networks*, *2014*.

[16] Aman, S., Simmhan, Y., & Prasanna, V. K. (2013). Energy management systems: state of the art and emerging trends. *Communications Magazine, IEEE*, *51*(1), 114-119.

[17] Karnouskos, S. (2011, October). Crowdsourcing information via mobile devices as a migration enabler towards the smartgrid. In *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on* (pp. 67-72). IEEE.

[18] http://www.zigbee.org/Specifications/ZigBee/GreenPower.aspx

[19] http://www.z-wave.com/

[20] Atzeni, I., Ordóñez, L. G., Scutari, G., Palomar, D. P., & Fonollosa, J. R. (2013). Demand-side management via distributed energy generation and storage optimization. *Smart Grid, IEEE Transactions on*, *4*(2), 866-876.

[21] Stojkoska, B., Solev, D., & Davcev, D. (2011, August). Data prediction in WSN using variable step size LMS algorithm. In *SENSORCOMM 2011, The Fifth International Conference on Sensor Technologies and Applications* (pp. 191-196).

[22] Dimitrievski, A., Stojkoska, B., Trivodaliev, K., & Davcev, D. (2006). Securing communication in WSN trough use of cryptography. *NATO-ARW, Suceava*.

Primitive polynomials as a tool in generation of $(t, s)-$ sequences

Vesna Dimitrievska Ristovska
Faculty of Computer Sciences and Engineering
Univ. Cyril and Methodius, Skopje, Macedonia
vesna.dimitrievska.ristovska@finki.ukim.mk

Vasil Grozdanov
Faculty of Natural Sciences and Mathematics
Department of Mathematics
Univ. Neophit Rilski, Blagoevgrad, Bulgaria
vassgrozdanov@yahoo.com

***Abstract*—In this paper we present some numerical simulations which are based on our new algorithm, proposed and presented in previous paper [1]. This algorithm generates uniformly distributed sequences.**
Primitive polynomials are used as a tool in the algorithm to get linear homogeneous recurrence relations with constant coefficients in the finite field in base b (b is an arbitrary prime). In this paper we present some results when in $s-$ dimensional case we set different basis b_j for each coordinate $j = 1, \ldots, s$ where b_j is arbitrary prime. Obtained experimental results and their visualizations have verified our theoretical expectations.

Keywords—primitive polynomial, irreducible polynomial, linear homogeneous recurrence relations, (t,m,s)-sequence.

I. Introduction

Following Niederreiter [6] we will give the concept of a class of sequences with well distribution of their points in $[0,1)^s$. So, let $b \geq 2$ be a fixed integer and will denote the base in which are constructed the considered sequences. When it is necessary b to be a prime number we will specially indicate. In the following we give the definition of a $(t, m, s)-$net and a $(t, s)-$sequence.

Definition 1 *Let $0 \leq t \leq m$ be integers. A point set P consisting of b^m points in $[0,1)^s$ forms a $(t, m, s)-$net in base b, if every subinterval $J = \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}}, \frac{a_j+1}{b^{d_j}} \right)$ of $[0,1)^s$, with integers $d_j \geq 0$ and integers $0 \leq a_j < b^{d_j}$ for $1 \leq j \leq s$ and of volume b^{t-m}, contains exactly b^t points of P.*

Definition 2 *Let $t \geq 0$ be a given integer. The sequence $(\mathbf{x}_n)_{n \geq 0}$, $\mathbf{x}_n \in [0,1)^s$, is a $(t, s)-$sequence in base b if for all $l \geq 0$ and $m \geq t$ the point set $\{\mathbf{x}_{lb^m}, \ldots, \mathbf{x}_{(l+1)b^m-1}\}$ is a $(t, m, s)-$net.*

It is obvious that a $(t, m, s)-$net is extremely well distributed if the parameter t is small.

This paper is based on two principles:

1) generalization of the constructive approach of Sobol' [7], which generate classes of (t,s)-sequences over the field $\mathbf{F_2}$ by using monocyclic difference operators over $\mathbf{F}_2$, and
2) previous paper [1] where the monocyclic difference operators are over $\mathbf{F}_b$, where b is a prime number.

The main work in this paper are simulations with different basis b_j on each coordinate $j = 1, \ldots, s$ by using linear homogeneous recurrence relations with constant coefficients in the finite field in base b_j, where b_j is an arbitrary prime.

II. Linear homogeneous recurrence relations and primitive polynomials over $\mathbf{F}_b$

In the following we will give some notations and statements about the linear homogeneous recurrence relations and primitive polynomials over the field $\mathbf{F}_b$.

A relation of the form

$$Lu_i = u_{i+m} + a_{m-1}u_{i+m-1} + \ldots + a_1 u_{i+1} + a_0 u_i, \quad (1)$$

where for $i \in \mathbf{Z}$, $u_i \in \mathbf{F}_b$, for $0 \leq i \leq m-1$ $a_i \in \mathbf{F}_b$, $a_0 \neq 0$ is called linear homogeneous recurrence relation over $\mathbf{F}_b$ of order *m* with constant coefficients (LHRR).

A solution of the equation

$$Lu_i = 0 \quad (2)$$

we will call the sequence $\ldots u_{-2}, u_{-1}, u_0, u_1, u_2, \ldots$, of elements from $\mathbf{F}_b$ which satisfied this equation for all integer i. The solution will be denoted with $\{u_i\}$. The solutions of the equation $Lu_i = 0$ have cyclic character. It is easy to show that the solution $\{u_i\}$ is periodical with a period ω_1 such that $\omega_1 \leq b^m - 1$.

A cycle of the solutions of the equation (2) we will call a set of solutions which are different by the shifting of the numeration, i.e. if $\{u'_i\}$ is a solution of the equation (2), then an arbitrary another solution $\{u_i\}$ of the cycle can be presented of the form $u_i = u'_{i+\alpha}$ for some integer α.

The relation Lu_i is called monocyclic if the equation $Lu_i = 0$ has only one solution with period $\omega = b^m - 1$.

To every LHRR (1) corresponds a polynomial of degree m over $\mathbf{F}_b$ of the form

$$Lu_i \leftrightarrow P(x) = x^m + a_{m-1}x^{m-1} + \ldots + a_1 x + a_0, \quad (3)$$

where for $i = 0, 1, \ldots m-1$ $a_i \in \mathbf{F}_b$.

Following Lidl and Niederreiter [5] we note the next definitions and results: *An irreducible polynomial* is a non-constant polynomial that may not be factored into the product of two non-constant polynomials. It is well known that of a monocyclic relation of the form (1) corresponds an irreducible polynomial. In other side this condition is not sufficient. Zierler [9] proved that necessary and sufficient condition that the relation (1) to be monocyclic is the polynomial (3) to

be primitive. A polynomial is primitive if it is irreducible, a divisor of the binomial $x^{\omega}-1$ (ω is the period of the corresponding monocyclic relation), and it is not a divisor of a binomial $x^{q}-1$ of degree $q<\omega$.

Following Sobol' [7] we will recall the theoretical bases of the construction of sequences of $b-$adic rational type, the so-called $BR-$ sequences. So, the details are as follows:

Definition 3 *Let $V_1, V_2, \ldots, V_j, \ldots$, be an arbitrary sequence of $b-$adic rational numbers, where for $j \geq 1$ we have that $0 < V_s < 1$. The numbers of this sequence we will call direction numbers. A $BR-$ sequence $\{r(i)\}_{i\geq 0}$ which corresponds to the direction numbers $\{V_j\}_{j\geq 1}$ is defined as: if an arbitrary integer number i has the $b-$adic presentation*

$$i = e_m e_{m-1} \ldots e_2 e_1,$$

then we replace

$$r(i) = e_1 V_1 * e_2 V_2 * \ldots * e_m V_m,$$

*where * is the operation digit-by-digit summation modulo b and $e_j V_j = \underbrace{V_j * \ldots * V_j}_{e_j - times}$.*

We can represent the direction numbers V_j in the form of $b-$ adic fractions:

$$V_j = 0, v_{j1} v_{j2} \ldots v_{ji} \ldots,$$

where all $v_{ji} \in \mathbf{F}_b$. In this sense the setting of the sequence $\{V_s\}$ is equivalent to setting an infinite matrix, which elements are from the field $\mathbf{F}_b$.

$$(v_{ji}) = \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1i} & \ldots \\ v_{21} & v_{22} & \ldots & v_{2i} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ v_{j1} & v_{j2} & \ldots & v_{ji} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}. \quad (4)$$

This matrix is called direction matrix. Using results in our paper [1], the following two theorem hold:

Theorem 1 *Let in the direction matrix (4) we have that for $j = 1, 2, \ldots$ $v_{jj} \neq 0$, i. e. $v_{jj} \in \{1, 2, \ldots, b-1\}$, and for $i > j$ $v_{ji} = 0$. Then the corresponding $BR-$ sequence $\{r(i)\}_{i\geq 0}$ is a $(0,1)-$ sequence in base b.*

Let Lu_i be an arbitrary monocyclic LHRR over the field $\mathbf{F}_b$ of order m. We will generate the direction numbers $V_1, V_2, \ldots, V_i, \ldots$ as a solution of the equality

$$V_{i+m} * a_{m-1} V_{i+m-1} * \ldots * a_1 V_{i+1} * a_0 V_i = b^{-m} V_i, \quad (5)$$

i.e. $LV_i = b^{-m} V_i$, as in the relation Lu_i should the symbol + to be replaced by the operation *. The initial conditions $V_1, V_2, \ldots, V_m$ of the equation (5) can be chosen in different manners, but for our purposes it is necessary to satisfy the next conditions: if the $b-$adic presentation of the number V_j is

$$V_j = 0, v_{j1} v_{j2} \ldots v_{ji} \ldots,$$

we assume that for all $j \geq 1$ $v_{jj} \neq 0$ and for $i > j$ $v_{ji} = 0$. In this way the matrix

$$B = \begin{bmatrix} v_{11} & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ v_{m1} & v_{m2} & \ldots & v_{mm} \end{bmatrix}$$

is triangular and nonsingular. On the main diagonal there are nonzero numbers and over it stand zeros. We will say that the $BR-$ sequence $\{r(i)\}_{i\geq 0}$ with these direction numbers $\{V_i\}_{i\geq 1}$ corresponds to the relation Lu_i.

In the next theorem [1] it is shown the possibility to use monocyclic LHRR as a tool to construct $(t,s)-$sequences.

Theorem 2 *Let b be a prime number and $L_1, L_2, \ldots, L_s$ are different monocyclic LHRR of orders $m_1, m_2, \ldots, m_s$ over $\mathbf{F}_b$. For $k = 1, 2, \ldots, s$ let $(P^{(k)}(i))_{i\geq 0}$ be the $BR-$ sequence which corresponds to the relation L_k. Then the sequence $(P(i))_{i\geq 0}$ of points of the form*

$$P(i) = (P^{(1)}(i), P^{(2)}(i), \ldots, P^{(s)}(i))$$

is a $(t,s)-$ sequence in base b with a parameter

$$t = \sum_{k=1}^{s} (m_k - 1). \quad (6)$$

To construct $(0,s)-$sequence for a dimension $s \geq 3$ we must work over other field $\mathbf{F}_b$, where b is a prime number.

The formula (6) for quality parameter t shows us that if we desire to obtain a parameter $t = 0$, we must choose all monocyclic relations L_k of order $m_k = 1$.

It is clear that for every dimension s, we can find enough big prime base b such that there are at least s primitive polynomials over the field $\mathbf{F}_b$.

III. Algorithm and Numerical simulations

A. Algorithm

According to Theorem 2 and exposed algorithm in [1], in this paper we propose a new algorithm for construction of sequences, when in $s-$ dimensional case we set different basis b_k for each coordinate $k = 1, \ldots, s$ where b_k is arbitrary prime.

1) **Input the dimension s, the base b_k, the number of the points of the net and the coefficients of the monocyclic relations L_k, $k = 1, \ldots, s$, where b_k is arbitrary prime for each coordinate k.**
2) **For $k = 1, \ldots, s$, for initial direction numbers V_j we use arbitrary random numbers which satisfies the condition of Theorem 1.**
3) **For $k = 1, \ldots, s$ a construction of the direction matrix which corresponds to the LHRR L_k.**
4) **For $k = 1, \ldots, s$ a generation of the one-dimensional sequence $(P^{(k)}(i))_{i\geq 0}$ which corresponds to the relation L_k.**
5) **Construction of the sequence $P(i) = ((P^{(1)}(i), P^{(2)}(i), \ldots, P^{(s)}(i)))_{i\geq 0}$, which is $s-$ dimensional sequence.**

For the purposes of this work to visualize the $(t,s)-$ sequences constructed by the proposed algorithm, a computer programs is written. It constructs one-dimensional $(t,s)-$ sequences, when in $s-$ dimensional case we set different basis b_k for each coordinate $k = 1, \ldots, s$ where b_k is arbitrary prime.

B. Software simulations and visualizations

In the following we present some results of this program for different dimensions s and different bases b_k.

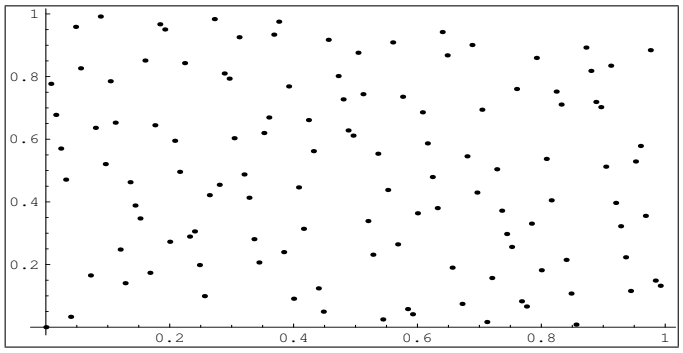


Fig. 1. s=2 ,n=121, b1=5,a1={2}, b2=11, a2={3,1,9}

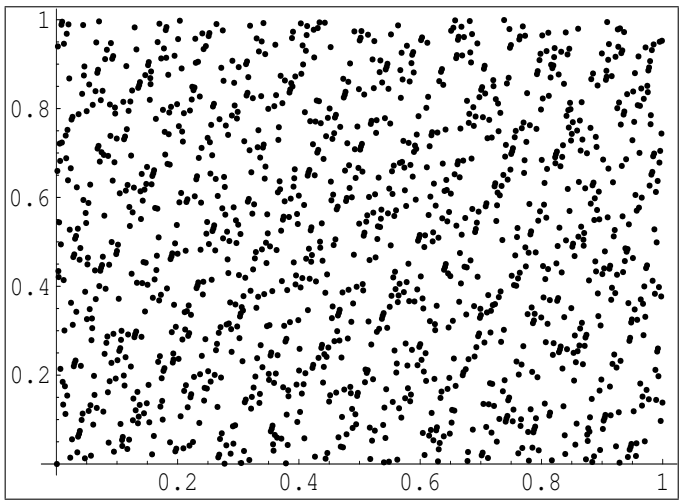


Fig. 2. s=2 ,n=1331, b1=7,a1={2,3,4}, b2=11, a2={3,1,8}

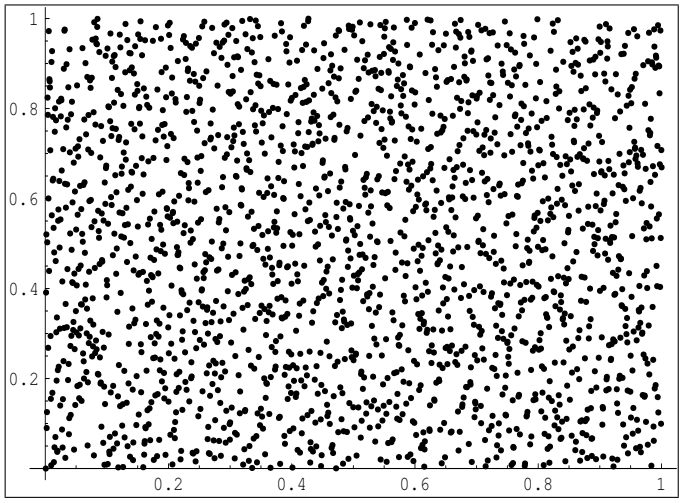


Fig. 3. s=2 ,n=2048, b1=2,a1={1}, b2=7, a2={5,5}

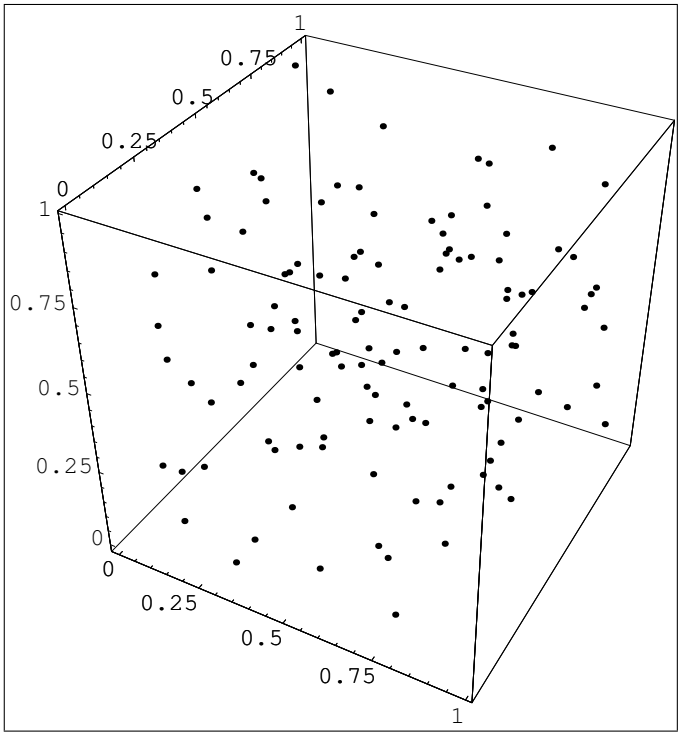


Fig. 4. s=3, n=121, b1=11,a1={3,2,2}, b2=23, a2={13},b3=7, a3={3,0,4,5}

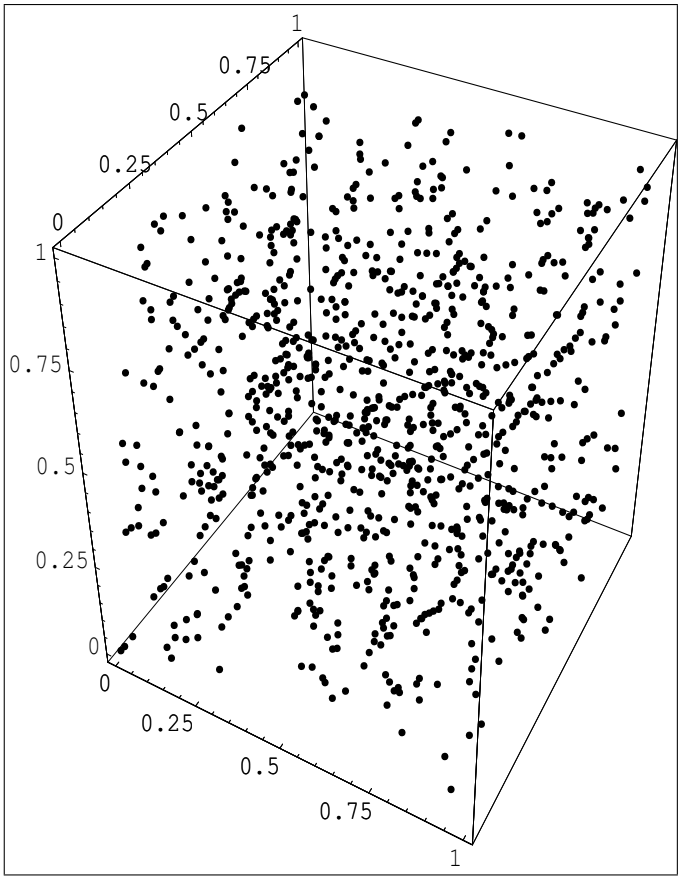


Fig. 5. s=3, n=1000, b1=2,a1={1}, b2=19,a2={5,17}, b3=11,a3={6,10,2,3,1}

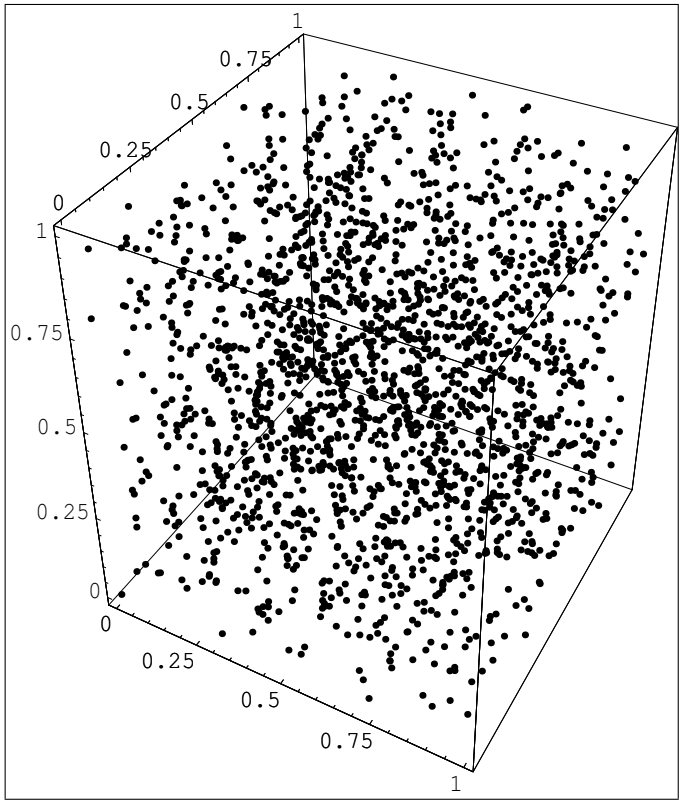


Fig. 6. s=3, n=2000, b1=7,a1={2}, b2=11,a2={6,10,2,3,1}, b3=19,a3={5,17}

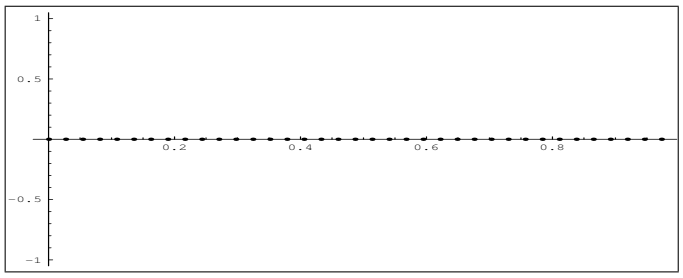


Fig. 7. s=1, n=37, b1=37, a1={35}

IV. Conclusion

Obtained experimental results and their visualizations have verified our expectations. The next our step will be the proof that the analog of Theorem 2 holds, when in $s-$ dimensional case we set different basis b_k for each coordinate $k = 1, \ldots, s$ where b_k is arbitrary prime, i.e. we set this hypothesis:

Let $b_1, b_2, \ldots b_s$ are arbitrary prime numbers and $L_1, L_2, \ldots, L_s$ are different monocyclic LHRR of orders $m_1, m_2, \ldots, m_s$ over corresponding bases $\mathbf{F}_{b_1}, \mathbf{F}_{b_2}, \ldots \mathbf{F}_{b_s}$. For $k = 1, 2, \ldots, s$ let $(P^{(k)}(i))_{i\geq 0}$ be the $BR-$ sequence which corresponds to the relation L_k. Then the sequence $(P(i))_{i\geq 0}$ of points of the form

$$P(i) = (P^{(1)}(i), P^{(2)}(i), \ldots, P^{(s)}(i))$$

is a $(t, s)-$ sequence.

Acknowledgment

This research was partially supported by FINKI, UKIM, Skopje.

References

[1] V. Dimitrievska Ristovska, V. Grozdanov, A. Atanasov, *An effective algorithm for constructing of (t,s)-sequences over* **F***b*, Proceedings of the V Congress of UMM, Ohrid, 2014 (in print)

[2] H. Faure, *Discrépance de suites associées* à *un systeme de numeration (en dimension s)*, Acta Arithmetica, XLI, (1982), 337-351.

[3] H. Faure, *Variation on (0,s)-Sequences*, Journal of Complexity, 17, 2001, 741-753.

[4] L. Kuipers, H. Niederreiter, *Uniform distribution of sequences*, John Wiley & Sons, New York., 1974.

[5] R. Lidl, H. Niederreiter, *Introduction to finite fields and their applications*, Cambridge University Press, New York, USA, 1986.

[6] H. Niederreiter, *Random Number Generator and Quasi-Monte Carlo Methods, CBMS - NSF Series in Applied Mathematics*, 63, SIAM, Philadelphia, 1992.

[7] I. M. Sobol', *Mnogomernye kvadraturnye formuly i funktsii Haara*, Izdat. Nauka, Moscow, 1969.

[8] J. G. Van der Corput, *Verteilungsfunktionen*, Proc. Kon. Akad. Wetensch. Amsterdam, 1935, 38, No 8, 813-821.

[9] N. Zierler, *Linear recurring sequences*, J. Soc. Industr. Appl. Math., 7 (1), (1959), 31-48.

Comparative Study of Two Approaches for Solving the Torsional Schrödinger Equation: Fourier Grid Hamiltonian Method and Hamiltonian Diagonalization Method

Bojana Koteska, Anastas Mishev
Faculty of Computer
Science and Engineering,
Ss. Cyril and Methodius University,
Rugjer Boshkovikj 16, P.O. Box 393,
1000 Skopje, Republic of Macedonia
Email: {bojana.koteska, anastas.mishev}@finki.ukim.mk

Ljupčo Pejov
Institute of Chemistry,
Faculty of Natural Sciences
and Mathematics,
Ss. Cyril and Methodius University,
Arhimedova 5, P.O. Box 162,
1000 Skopje, Republic of Macedonia
Email: ljupcop@iunona.pmf.ukim.edu.mk

***Abstract*—In this study, we have compared the computational performance of two methods implemented to solve the Schrödinger equation for intramolecular torsional motions. The first approach is the Fourier grid Hamiltonian (FGH) operator method, which is based on fragmentation of the total torsional Hamiltonian into kinetic energy part, which is diagonal in momentum representation, and the potential energy part, diagonal in coordinate representation. The second approach is the standard diagonalization technique, based on variational principle of quantum mechanics. Torsional energy eigenvalues are further used to compute the torsional correlation times in the framework of BPP (Bloembergen-Purcell-Pound) approach. The results show that diagonalization technique performs much faster than the FGH algorithm. Besides that, the convergence of eigenvalues with the number of basis functions appears to be achieved faster with Hamiltonian diagonalization**

Keywords*—torsional Shcödinger equation, Fourier Grid Hamiltonian method, diagonalization technique, computational performance

I. Introduction

Solution of the torsional Schrödinger equation is a problem that arises in many areas of contemporary science and technology, and is therefore a rather relevant computational task. If one wants to treat rigorously (*i.e.* quantum-mechanically) the intramolecular hindered rotations of particular atomic groups, it is necessary to go beyond the classical approximation in treating such motions. For example, in case of molecular systems which are potential candidates for molecular switches or transistors, one of the possible mechanisms that underlay the switching behavior is the conformational transition between two (or in principle, even more) possible conformations in the considered molecule. The potential energy function $V(\varphi)$ for this motion is often conveniently represented as Fourier series expansion, which allows one to easily account for the inherent symmetry of intramolecular motion:

$$V(\varphi) = V_0 + \sum_{n=1}^{\infty}[V_n \cos(n\varphi) + V'_n \sin(n\varphi)] \qquad (1)$$

The corresponding torsional Hamiltonian is:

$$\hat{H}_{\text{torsion}} = -\frac{\hbar^2}{2I_\varphi}\frac{\partial^2}{\partial\varphi^2} + V(\varphi) \qquad (2)$$

In the last equation, φ is the torsional variable (angle), I is the moment of inertia corresponding to the motion considered, while ($\hbar$ being Planck's constant). In principle, the torsional transitions may occur as a result of classical "over the barrier" flip, by intrawell transitions (between different states within the same well of the potential), and due to tunneling between adjacent wells. Accounting for all these types of torsional transitions is essential if one wants to describe the actual realization and functioning of a molecular device under realistic conditions. To achieve this aim, however, one has to solve the torsional Schrödinger equation as a first step. In the present study, we consider a solution of such equation for a realistic molecular system.

The goal of our research is to find the computational performance of the Fourier Grid Hamiltonian method and standard diagonalization technique of the Hamiltonian matrix to solve the Schrödinger equation for intramolecular torsional motions. The Fourier Grid Hamiltonian method is diagonal in momentum representation, and the potential energy part, diagonal in coordinate representation. The transformation between the two representations is achieved via Fourier transformation algorithm using a suitably chosen grid of points.

In a case of the standard diagonalization technique, the Hamiltonian operator is written in matrix representation within a suitably chosen basis set, consisting of the free rotor wave functions. The Hamiltonian matrix is further diagonalized, giving energy eigenvalues and eigenfunctions.

When applying the Fourier Grid Hamiltonian operator method, the dependence of computational performance and the convergence of obtained torsional energy levels on the number of grid points used are analyzed, while when applying the standard diagonalization technique, the dependence of

computational effectiveness and the convergence of obtained torsional energy levels on the number of basis functions is studied. Special attention was paid to the high-energy bound levels (close to the torsional barrier) and their convergence in both cases.

The paper is organized as follows. The related work is given in Section 2. The methods for solving the torsional Schrödinger equation are described in Section 3. Section 4 presents the experiments and results of the experiments for measuring the computational effectiveness of the applied methods. The conclusion and future work are exposed in Section 5.

II. Related Work

Solution of the torsional Schrödinger equation has been a research subject for several authors. Turovtsev et al. [1] find the characteristics of the internal rotation based on the solution of the torsion Schrödinger equation. They solved the one-dimensional torsional Schrödinger equation with a general periodic potential and found the relations for the calculation of eigenvalues and eigenfunctions of the Hamiltonian of internal rotational motion in molecules in the basis of plane waves. In [2], the authors present a numerical method for solving the approximate Schrödinger equation (SE) for a single internal motion. They pay special attention to computer programs for calculations and their applications to torsional studies in areas of spectroscopy, thermodynamics, and reaction rates.

A comparison of the methods for solving the vibrational Schrödinger equation is given in [3]. The authors apply three numerical methods to compute the anharmonic OH stretching vibrational frequencies of the free and aqueous hydroxide ion on the basis of one-dimensional vibrational potential energies: simple Hamiltonian matrix diagonalization technique, based on representation of the vibrational potential in Simons-Parr-Finlan (SPF) coordinates, Numerov algorithm and Fourier grid Hamiltonian method (FGH). According to their research, the diagonalization technique performs remarkably well in a very wide range of frequencies and frequency shifts (up to 300 cm1) and FGH method, showed a very good performance, exhibits more significant (and nonuniform) discrepancies with the Numerov algorithm, even for rather modest frequency shifts.

III. Methods for Solving the Torsional Schrödinger Equation

A. Fourier Grid Hamiltonian Method

The Fourier grid Hamiltonian (FGH) method is a special case of a discrete variable representation method (DVR) described in [4] [5]. The FGH method generates the wavefunctions of the Hamiltonian operator as amplitudes of the wave function on the grid points. It is simple because the wavefucntions are not given as a basis functions or as a linear combination [6]. Discretization is achieved when the continuous range of the coordinates values x is replaced by a grid of discrete values x_i. The uniform grid x_i is defined as:

$$x_i = i\Delta x \tag{3}$$

where Δx is the uniform spacing between the grid points.

The state function (property of a system which depends only on the current state of the system) $|\psi\rangle$ can be given as a vector on a discretized grid of points in coordinate space represented as:

$$|\psi\rangle = \psi^x = \sum_i |x_i\rangle \cdot \Delta x \cdot \psi(x_i) = \sum |x_i\rangle \cdot \Delta x \cdot \psi_x^i \tag{4}$$

where $|x_i\rangle$ are basis functions or in momentum space represented as:

$$|\psi\rangle = \psi^k = \sum_i |k_i\rangle \cdot \Delta k \cdot \psi(k_i) = \sum |k_i\rangle \cdot \Delta k \cdot \psi_k^i \tag{5}$$

where Δk is the reciprocal grid size in momentum space and it is defined as:

$$\Delta k = \frac{2\Pi}{N\Delta x} \tag{6}$$

and $N\Delta x$ is the total length of the coordinate space covered by the grid. The grid points are evenly distributed about $k = 0$ which is the central point in the momentum space [3] [6].

The transformation from one to other representation is performed by using the Fourier transform technique (FTT) that decomposes a function of time (signal) into the frequencies that it is made of [7]. Transformation between the two representations can be written as:

$$\psi^k = U * \psi^x \tag{7}$$

where U is an unitary matrix.

We also have to define a column vector matrix ϕ_n where each element is 0 except the unity element in the n-th row:

$$\phi_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix} - \text{n-th row} \tag{8}$$

The $n_t h$ column of the Hamiltionan matrix can be represented as follows by implementing a forward and reverse FFT:

$$H_{in} = [(U^{-1}TU + V)\phi_n]_i \tag{9}$$

where T and V are the diagonal kinetic energy and potential energy $[V(x_i)]$ matrices. The complete Hamiltionian matrix H can be generated by repeating this process for all possible N vectors ϕ_n.

B. Standard Diagonalization Technique Method

Schrödinger equation can be solved by using the standard matrix diagonalization technique. In this case, the wavefunction is represented as a finite set of basis functions. The eigenfunctions and eigenvalues are found from the diagonalization of the Hamiltonian matrix [8].

The Schrödinger equation can be written in the following form [9]:

$$\hat{H}\psi(q) = E\psi(q) \tag{10}$$

where $\hat{H}$ is the Hamiltonian operator, $\psi(q)$ are eigenfunctions and E are eigenvalues of $\hat{H}$. This equation is solved by finding the eigenvalues and eigenfunctions of the Hamiltonian operator.

If we split the function $\psi(q)$ in basis functions, we can write the $\psi(q)$ as:

$$\psi(q) = \sum_n a_n \varphi_n(q) \tag{11}$$

and if we replace the $\psi(q)$ in the eq. 10:

$$\hat{H}\sum_n a_n \varphi_n(q) = E\sum_n a_n \varphi_n(q) \tag{12}$$

Hamiltonian operator is a linear operator and it can be embedded in the sum:

$$\sum_n a_n(\hat{H}\varphi_n(q)) = E\sum_n a_n \varphi_n(q) \tag{13}$$

If (13) is multiplied scalarly by $\varphi_m^*(q)$ and integrated by q, then we get the following:

$$\sum_n a_n \int \varphi_m^*(q)\hat{H}\varphi_n(q)\mathrm{d}q = \sum_n a_n E \int \varphi_m^*(q)\varphi_n(q)\mathrm{d}q \tag{14}$$

$$\int \varphi_m^*(q)\hat{H}\varphi_n(q)\mathrm{d}q = H_{mn} \tag{15}$$

$$\int \varphi_m^*(q)\varphi_n(q)\mathrm{d}q = \delta_{mn} \tag{16}$$

If we substitute in (14), we have:

$$\sum_n a_n H_{mn} - \sum_n a_n E\delta_{mn} = 0 \tag{17}$$

Equation (17) can be written in the following form:

$$\sum_n a_n(H_{mn} - E\delta_{mn}) = 0 \tag{18}$$

where H_{mn} is the matrix element of the Hamiltonian operator $\hat{H}$.

An homogeneous algebraic system of equations is defined by (18). It has a nontrivial solution if its determinant is equal to zero:

$$|H_{mn} - E\delta_{mn}| = 0 \tag{19}$$

Equation (19) is a polynomial whose roots are the possible values of E. Thus, the eigenvalues of $\hat{H}$ are roots of (19). The eigenfunction, which corresponds to a given eigenvalue is determined such that the given eigenvalues is inserted in (17). Thus, the system of coefficients a_n is determined ant that helps to find the corresponding eigenfunction.

The elements H_{mn} determined by (15) are called matrix elements of the Hamiltonian operator $\hat{H}$ in the basis functions $\varphi_n(q)$ and they constitute the matrix $[H]$:

$$[H] = \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1n} & \dots \\ H_{21} & H_{22} & \dots & H_{2n} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ H_{n1} & H_{n2} & \dots & H_{nn} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{20}$$

From (12), it follows that the functions $\psi(q)$ have a corresponding column vector:

$$\psi(q) \rightarrow \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_n \\ \dots \end{bmatrix} \tag{21}$$

If the matrix $[H]$ is diagonal, then its diagonal elements determine the eigenvalues of the operator $\hat{H}$.

IV. Experiments and Results

A. Testing environment

The testing of these two different approaches for solving the torsional Schrödinger equation was performed by using the Fourier Grid Hamiltonian 1D Program (FGH1D) [10] and Basis-set Expansion solver for 1-Dimensional Schrödinger equation (BEx1D) [11].

The FGH1D program calculates the energy levels (eigenvalues) and wavefunctions (eigenvectors) for a given one-dimensional potential. It solves the Shrödinger equation variationally by using the Fourier Grid Hamiltonian method which is described in the subsection III-A. This method uses a basis set of delta functions and it requires an even number of grid points (basis functions). The accuracy increases with increasing the number of grid points. The most computationally intensive part is the matrix diagonaization. We used the option $cos(nx)$ potential which is useful for describing torsional rotations. The range is set automatically to be 0 to $2p(p-1)/p$ where p is the number of grid points, x is an angle in radians and the moment of inertia (in our case 22.84588830778) is expressed in amu * Å^2 (Å=Ångströms, 1Å = 10^{-10} m, amu = atomic mass unit.

The BEx1D program is a package of calculation tools for finding eigenstates and partition functions of intramolecular nuclear motions, for which harmonic oscillator approximation is inadequate. We used the bx1HRsol file which is Schrödinger-equation eigen-problem solver for a intramolecular one-dimensional hindered Rotation on a Fourier series potential. The basis functions used by bx1HRsol are the free-rotor eigen functions. We only had to change the number of basis functions maxAbsQN which is the maximum absolute J. For example, if maxAbsQN= 50, the program will use

101 basis functions in total from $J = -50$ to $+50$. Another parameter that we have specified is the rotational constant (rotConst) which in our case was set to 0.738488849 cm^{-1}.

Both programs were run under the Windows operating system on a single node computer with Intel Core 2 Duo processor with speed of 2.26 GHz.

B. Results

The computational performances of the used programs are obtained by measuring the time required for the program execution for a different number of basis functions (grid points). The maximum number of basis functions when using the BEx1D program is 3001 (maxAbsQN= 1500) and the maximum number of grid points when using the FGH1D program is 710. The execution times (in seconds) of the programs are shown in the fig. 1. The x-axis denotes the number of grid points or basis functions. The y-axis denotes the execution time in seconds.

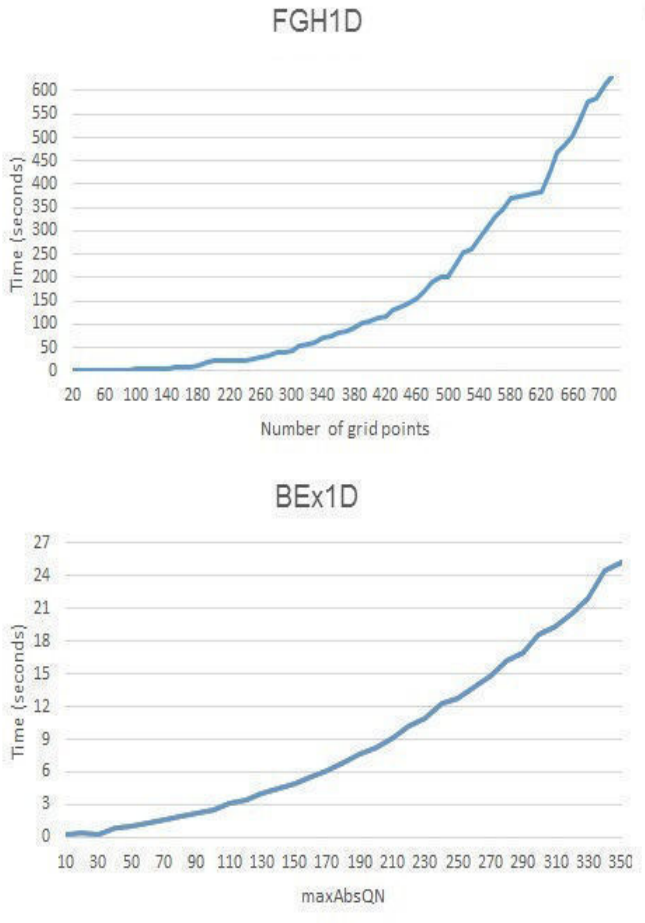


Fig. 1. Comparison of the execution times of BEx1D (up) and FGH1D program (down)

The execution times of the BEx1D program, starting for 101 (maxAbsQN= 50) to 3001(maxAbsQN= 1500) basis functions are shown in fig. 2.

The convergence with an error of 1% relative to the converged one has been achieved with the FGH1D method

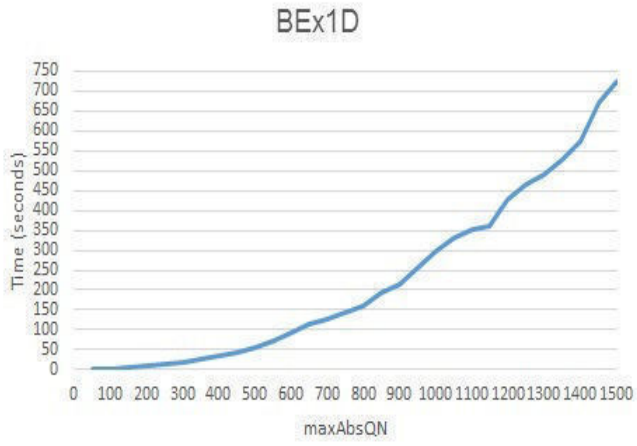


Fig. 2. Execution times of BEx1D program

for 102 grid points. In the case of diagonalization technique, on the other hand, such convergence is achieved with 47 basis functions.

V. Conclusion and Future Work

In conclusion, we can summarize the results of the present work as follows. Making an analogy between the number of grid points used to discretize the operators in the FGH1D methodology and the number of basis functions in diagonalization technique, one can see from Fig. 1 and Fig. 2 that the diagonalization technique performs much faster than the FGH1D algorithm. Besides that, the convergence of eigenvalues with the number of basis functions appears to be achieved faster with Hamiltonian diagonalization in the basis of free-rotor wavefunctions than by discretization technique. Further improvements of both techniques are certainly possible from a purely computational viewpoint, either by implementing more efficient large matrix diagonalization parallel algorithms and more efficient Fourier transformation algorithms that allow transition from coordinate to momentum basis in FGH technique.

References

[1] V. Turovtsev, A. Belotserkovskii, and Y. Orlov, "Solution of a one-dimensional torsion schrdinger equation with a general periodic potential," Optics and Spectroscopy, vol. 117, no. 5, 2014, pp. 710–712. [Online]. Available: http://dx.doi.org/10.1134/S0030400X14090264

[2] A. Chung-Phillips, "Torsional energy levels and wave functions," Journal of Computational Chemistry, vol. 13, no. 7, 1992, pp. 874–882. [Online]. Available: http://dx.doi.org/10.1002/jcc.540130712

[3] J. Petreska and L. Pejov, "Comparison of methods for solving the vibrational schrdinger equation in the course of sequential monte-carlo-quantum mechanical treatment of hydroxide ion hydration," Macedonian Journal of Chemistry and Chemical Engineering, vol. 29, no. 2, 2010. [Online]. Available: http://www.mjcce.org.mk/index.php/MJCCE/article/view/167

[4] D. O. Harris, G. G. Engerholm, and W. D. Gwinn, "Calculation of matrix elements for one dimensional quantum mechanical problems and the application to anharmonic oscillators," The Journal of Chemical Physics, vol. 43, no. 5, 1965, pp. 1515–1517. [Online]. Available: http://scitation.aip.org/content/aip/journal/jcp/43/5/10.1063/1.1696963

[5] D. Baye and P. H. Heenen, “Generalised meshes for quantum mechanical problems,” Journal of Physics A: Mathematical and General, vol. 19, no. 11, 1986, p. 2041. [Online]. Available: http://stacks.iop.org/0305-4470/19/i=11/a=013

[6] C. C. Marston and G. G. Balint Kurti, “The fourier grid hamiltonian method for bound state eigenvalues and eigenfunctions,” The Journal of Chemical Physics, vol. 91, no. 6, 1989.

[7] J. S. Walker, Fast fourier transforms. CRC press, 1996, vol. 24.

[8] M. Feit, J. F. Jr., and A. Steiger, “Solution of the schrdinger equation by a spectral method,” Journal of Computational Physics, vol. 47, no. 3, 1982, pp. 412 – 433. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0021999182900912

[9] S. D. University of California, “Quantum physics,” http://quantummechanics.ucsd.edu/ph130a/130_notes.pdf, April 2003, [Online; accessed 06-March-2015].

[10] T. N. I. of Standards and T. (NIST), “Fourier grid hamiltonian 1d program,” [Online; accessed 17-March-2015].

[11] A. Miyoshi, “Bex1d: One-dimensional schrdinger equation solver,” [Online; accessed 17-March-2015].

Symbolic calculations of the static properties of Bose-Einstein condensates

Mihaela Carina Raportaru and Alexandru Nicolin
Department of Computational Physics and Information Technologies
National Institute for Physics and Nuclear Engineering (IFIN-HH)
Magurele, Romania
mraportaru@nipne.ro, alexandru.nicolin@nipne.ro

***Abstract* — In this article we review a series of analytical results concerning the ground- and excited-state properties of Bose-Einstein condensates for which the symbolic computations were crucial. We address two distinct physical settings, namely a condensate loaded into an optical lattice and a condensate confined in a standard magnetic trap, and show that the analytical reconstruction of the ground- and excited state properties was impossible without the help of a computer algebra system.**

Keywords — variational equations, Gross-Pitaevskii, Bose-Einstein condensates

I. Introduction

Investigations into the nonlinear dynamics of ultra-cold quantum gases have become particularly attractive after the first experimental achievement of an atomic Bose-Einstein condensate in 1995 [1, 2]. We note on the experimental side that the system is robust and quite flexible, as one can, for instance, modify in time and/or space the frequencies of the underlying magnetic trap which confines the condensate and the strength and spatial profile of the interactions between atoms, while on the theoretical side we note that the so-called Gross-Pitaevskii equation (GPE) offers an excellent quantitative description of the properties of the condensate in both the static (i.e., equilibrium) and dynamic regime (see the review in Ref. [3]). The GPE is nonlinear three-dimensional partial differential equation which has been successfully used to describe a wide series of linear and nonlinear phenomena using both analytical and numerical treatments [1]. *In silico* investigations have been catalyzed by the numerous numerical recipes that are available for the GPE (in one-, two-, and three-dimensional condensates with various symmetries, see in particular Refs. [4, 5, 6]), one of which [6] is OpenMP-parallelized and offers running times substantially smaller than the average codes even on a multi-core desktop computer. Analytical investigations have been appealing in their own way, as one could use a computer algebra system (such as MATHEMATICA, Maple, MATLAB, etc.) to resolve even the most difficult problems. In this article we will focus on the variational treatment of the dynamics of the condensate which allows us to reduce the equations which describe the condensate to the level of a few ordinary differential equations. To this end, one usually starts from a trial wave function that depends on a few (possibly time-dependent) parameters which approximates the exact wave function of the condensate. Following the integration of the spatial degrees of freedom in the Lagrangian functional which describes the condensate one is left with an algebraic expression that can be then minimized through the standard Euler-Lagrange equations. These equations are either algebraic or ordinary differential and have a computing cost orders of magnitude lower than that of the original partial differential equation [1]. The difficult part in every variational treatment is to reach the optimal balance between the accuracy (and therefore the complexity) of the trial wave function and the analytical tractability: a very complex (though accurate) trial wave function may not allow us to compute analytically the symbolic integrals in the Lagrangian functional, while very simple trial wave functions may not describe the physics accurately enough. Only very few of these variational treatment rely on paper-and-pencil calculations and it is quite common to use a computer algebra system to find the optimal trial wave function. MATHEMATICA and Maple are the most used for problems which involve complex trial wave functions (with or without special functions), while Matlab is favored for variational treatments which yield large set of algebraic equations which have to be solved numerically.

In this article we show two problems connected to the properties of a Bose-Einstein condensate, namely the band structure of a condensate loaded into an optical lattice and the ground state of a one-dimensional condensate confined in a magnetic trap, for which the use of MATHEMATICA [7] was crucial in reaching the desired solution.

II. Band structure of a Bose-Einstein condensate

One popular experiment in ultra-cold gases concerns the loading of a condensate into a so-called optical lattice, that is, an effectively (usually) one-dimensional egg-carton trapping potential produced by two counter-propagating lasers [8, 9]. In the case of classical gases each well of the egg-carton will hold the same number of atoms, while for Bose-Einstein condensates it is possible to have a different load in each well.

The simplest example is one-dimensional: the standard solution has all wells carrying the same number of atoms, while the novel (nonlinear) solution which is particular to Bose-Einstein condensates has only each second well carrying the same number of atoms.

On the technical side to obtain the band structure of the condensate we have to start from the Gross-Pitaevskii energy functional, written either for one- or two-periods of the optical lattice depending on the periodicity of the solution that we are interested in. In the case of state of equal period to that of the lattice the energy takes the form

$$E(\psi)=\frac{1}{d}\int_{-d/2}^{d/2}\left[\frac{\hbar^2}{2m}\left|\frac{d\psi(x)}{dx}\right|^2+V_0\cos\left(\frac{2\pi x}{d}\right)\left|\psi(x)\right|^2\right]dx$$
$$+\frac{1}{d}\int_{-d/2}^{d/2}\left[\frac{U_0}{2}\left|\psi(x)\right|^4\right]dx \qquad (1)$$

where d is the period of the optical lattice, U_0 is the strength of the two-body interactions, V_0 is the strength of the optical lattice, m is the mass of an atom and $\hbar$ is the reduced Planck constant. The linear density of particles is defined as

$$n=\frac{1}{d}\int_{-d/2}^{d/2}\left|\psi(x)\right|^2dx \qquad (2)$$

and is a constant of the system.

As the wave function of the condensate is periodic the most natural option is to consider a Fourier expansion whose (time-independent) coefficient are obtained from the Euler-Lagrange equation. In the Fourier expansion k is the so-called wave vector which can be looked at as a parameter which distinguishes between the types of solution of equal spatial period. The band structure of the condensate amounts to determining all physical configurations of the condensate (i.e., mathematically, all solutions of the Euler-Lagrange equations) for different values of k. For a simple truncated Fourier expansion

$$\psi=\sqrt{n}\exp(ikx)\left[a_0+a_1\exp\left(\frac{i2\pi x}{d}\right)+a_{-1}\exp\left(-\frac{i2\pi x}{d}\right)\right] \qquad (3)$$

we can compute by hand the resulting energy functional which consists of three terms, namely

$$E_{kin}=\frac{\hbar^2}{2m}\left[k^2+2k\left(\frac{2\pi}{d}\right)\sin^2(\theta)\left(\cos^2(\phi)-\sin^2(\phi)\right)\right]$$
$$+\frac{\hbar^2}{2m}\left[\left(\frac{2\pi}{d}\right)^2\sin^2(\theta)\right], \qquad (4)$$

$$E_{pot}=V_0\sin(\theta)\cos(\theta)\left(\cos(\phi)+\sin(\phi)\right) \qquad (5)$$

and

$$E_{\text{int}}=nU_0\left[\frac{1}{2}+\sin^2(\theta)\cos^2(\theta)\left(\cos(\phi)+\sin(\phi)\right)^2\right]$$
$$+nU_0\left[\frac{1}{4}\sin^4(\theta)\sin^2(2\phi)\right], \qquad (6)$$

for the kinetic, potential and interaction contribution to the energy of the condensate.

While this case is simple and the Euler-Lagrange equations

$$\frac{\partial E}{\partial\phi}=0,\frac{\partial E}{\partial\theta}=0 \qquad (7)$$

follow easily for anything more complex the computations are done in MATHEMATICA. In Figs. 1 and 2 we depict the band structure of the condensate for physical configurations of the condensate which have the same spatial period as that of the lattice (Fig. 1) and twice the spatial period (Fig. 2).

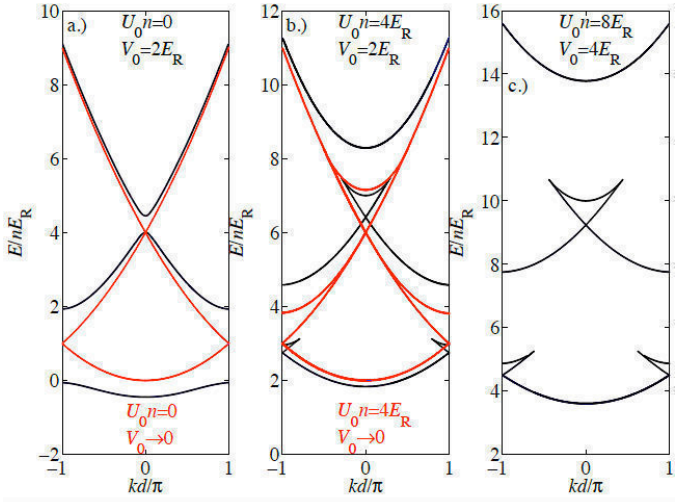


Fig 1. Typical band structures of the condensate depicted here in terms of the so-called recoil energy E_R. The three plots depict the band structure of the condensate for three distinct experimental configurations considering only states which have the same period as that of the underlying optical lattice.

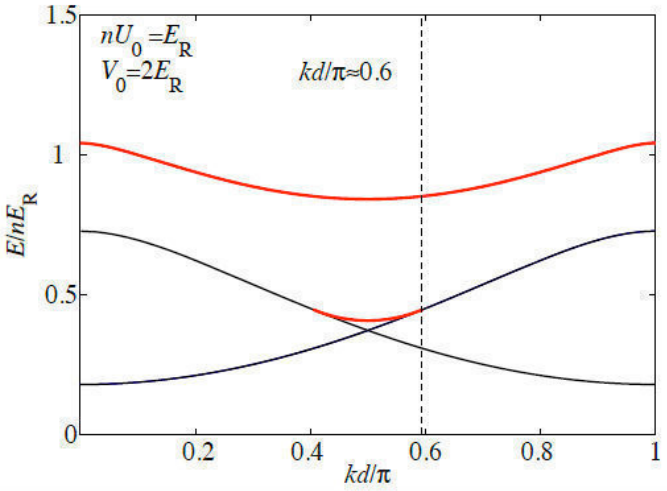


Fig. 2 Typical band structures of the condensate depicted here in terms of the so-called recoil energy E_R. The black curves show the states which have the same period as that of the underlying optical lattice, while the red curves show period which have a spatial period equal to twice that of the lattice.

III. Q-GAUSSIAN-BASED VARIATIONAL TREATMENTS

A regular Bose-Einstein condensate confided in a magnetic trap exhibits two distinct density regime: the low-density regime in which the wave function is well approximated by a

Gaussian function and a high-density regime in which the wave function is well approximated by a square root function. Given the differences between these two regimes one of the quests of variational treatments has been to offer a unified approach that describes not only these two distinct regimes but also the relatively large intermediate regime. Two trial wave functions have been particularly attractive: the S_n function proposed in Ref. [10] and the q-Gaussians introduced in Ref. [11]. While both of these two wave functions appear analytically untractable as far as the integrations in the Lagrangian are concerned a detailed investigation shows that this is not the case. Indeed, the S_n function defined as

$$S_n(x) = \exp\left(-\sum_{k=1}^{n}\frac{x^{2k}}{k}\right), \tag{8}$$

which has the limit

$$\begin{aligned} S_\infty &= \exp\left(-\sum_{k=1}^{\infty}\frac{x^{2k}}{k}\right), \\ &= \exp\left(\ln\left(1-x^2\right)\right) = 1-x^2, \end{aligned} \tag{9}$$

immediately produces special functions that cannot be further integrated symbolically. To see this let us recall that

$$\int_{-\infty}^{\infty} x^2 exp(-ax^2)dx = \frac{\sqrt{\pi}}{2a^{\frac{3}{2}}} \tag{10}$$

$$\int_{-\infty}^{\infty} exp\left(-ax^2 - \frac{ax^4}{2}\right)dx = \frac{exp\left(\frac{a}{4}\right)K_{\frac{1}{4}}\left(\frac{a}{4}\right)}{\sqrt{2}} \tag{11}$$

$$\int_{-\infty}^{\infty} x^2 exp\left(-ax^2 - \frac{ax^4}{2}\right)dx = \frac{\sqrt{\pi}\,U\left(\frac{3}{4},\frac{1}{2},\frac{a}{2}\right)}{2(2a)^{\frac{3}{4}}} \tag{12}$$

where the special and hypergeometric functions preclude all other computations.

The q-Gaussian function, however, is, in fact, quite flexible. To see this more clearly we consider the trial wave function for a trapped one-dimensional condensate

$$\psi(x,t) = f\left[q(t)\right]\sqrt{N}\left(1-\frac{\left(1-q(t)\right)x^2}{2\omega(t)^2}\right)^{\frac{1}{1-q(t)}}\exp\left(ix^2\beta(t)\right) \tag{13}$$

normalized to N, the number of atoms in the condensate.

From the normalization condition we have that

$$f\left[q(t)\right] = \frac{\left(1-q(t)\right)^{1/4}}{2^{1/4}\omega(t)^{1/2}B^{1/2}\left(\frac{1}{2},\frac{q(t)-3}{q(t)-1}\right)} \tag{14}$$

where the integration domain is taken to be between the two zeros of the wave function, that is

$$D = \left[-\frac{\sqrt{2}\omega(t)}{\sqrt{1-q(t)}},\frac{\sqrt{2}\omega(t)}{\sqrt{1-q(t)}}\right]. \tag{15}$$

For a Lagrange density of the form

$$\mathcal{L}(x,t) = \frac{i}{2}(\psi\psi_t^* - \psi^*\psi_t) + \frac{1}{2}|\psi_x|^2 + V(x)|\psi|^2 + \frac{U(t)}{2}|\psi|^4 \tag{16}$$

we have after integrating the spatial degrees of freedom, that is after integrating over x, and energy functional of the form

$$L(t) = \int_D \mathcal{L}(x,t)\,dx \tag{17}$$

The resulting expression is of the form

$$\frac{L(t)}{N} = L_1(t) + L_2(t) + L_3(t) + L_4(t) \tag{18}$$

where each individual term is given below:

$$L_1(t) = \frac{2\omega(t)^2\beta(t)}{7-3q(t)} \tag{19}$$

$$L_2(t) = \frac{NU(t)}{\omega(t)\sqrt{2}}\Delta[q(t)] \tag{20}$$

$$L_3(t) = \frac{\omega(t)^2\Omega^2}{7-3q(t)} \tag{21}$$

$$L_4(t) = \frac{5-q(t)}{8\omega(t)^2(1+q(t))} + \frac{4\omega(t)^2\beta(t)^2}{7-3q(t)} \tag{22}$$

The only term which is problematic is $L_2(t)$ which includes the function

$$\Delta[q(t)] = \frac{\sqrt{\pi}\Gamma\left(1-\frac{4}{q(t)-1}\right)\sqrt{1-q(t)}}{2B^2\left(\frac{1}{2},\frac{q(t)-3}{q(t)-1}\right)\Gamma\left(\frac{3}{2}-\frac{4}{q(t)-1}\right)} \tag{23}$$

which looks very complex but is, in fact, an almost linear function with respect to $q(t)$. To see this more clearly we consider the following two approximations which comes from the linear expansions around $q(t) = 1$ and $q(t) = -1$, respectively, namely

$$\Delta[q(t)] = \frac{41-9q(t)}{64\sqrt{\pi}} \tag{24}$$

and

$$\Delta[q(t)] = \frac{51-9q(t)}{100\sqrt{\pi}}, \tag{25}$$

and depict in Fig. 3 these three functions.

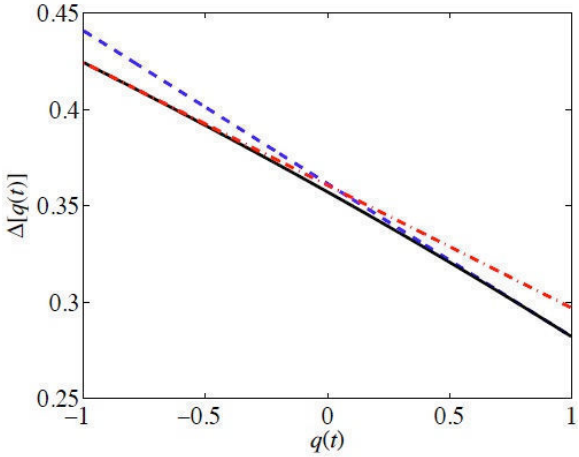


Fig. 3 The full function in equation (9) (depicted in black) along with the two approximations shown in equation (10) (in red) and (11) (in blue).

The Euler-Lagrange equation now follow easily for both analytical approximations in equations (24) and (25) [11].

The main message conveyed by these computations is that through the help of MATHEMATICA (and other computer algebra systems) one can quickly determine the best trial wave function which is well-suited for a particular variational treatment. More importantly, it allows us to decide quickly on the usefulness of a particular trial wave function, sometimes with surprising results, as it is the case of the q-Gaussian function (see also the applications of q-Gaussian functions for the effective equations used to describe the dynamics of strongly elongated or highly oblate condensates in Refs. [12, 13, 14]).

IV. CONCLUSIONS

We have reviewed a series of analytical results which demonstrate the usefulness of computer algebra systems for the variational treatment of Bose-Einstein condensate. We have focused on two distinct physical settings, i.e., a condensate loaded into an optical lattice and a condensate confined by a typical magnetic trap, and have showed that the study of ground- and excited states is substantially simplified by the possibility to compute automatically all integrations. We emphasize the idea that theoretical physics calculations can be easily taken beyond paper-and-pencil calculations through the use of a computer algebra systems.

ACKNOWLEDGMENT

For this work A.I.N. was supported by a grant of the Romanian Ministry of Education, CNCS-UEFISCDI, project PN-II-RU-PD-2012-3-0154.

REFERENCES

[1] C.J. Pethick and H. Smith, Bose-Einstein condensation in dilute gases, Cambridge: Cambridge University Press, 2008.

[2] P.G.Kevrekidis, D. J. Frantzeskakis, and R. Carretero-González, Eds., Emergent nonlinear phenomena in Bose-Einstein condensates, Berlin: Springer-Verlag, 2008.

[3] V.S. Bagnato *et al.*, "Bose-Einstein condensation: Twenty years after," Rom. Rep. Phys., vol 67, 5-50, 2015.

[4] P. Muruganandam and S.K. Adhikari, "Bose-Einstein condensation dynamics in three dimensions by the pseudospectral and finite-differece methods," J. Phys. B: At. Mol. Opt. Phys., vol 36, 2501, 2003.

[5] P. Muruganandam and S.K. Adhikari, "Fortran programs for the time-dependent Gross-Pitaevskii equation in a fully anisotropic trap," Comp. Phys. Comm., vol. 180, pp. 1888-1912, 2009.

[6] D. Vudragović, I. Vidanović, A. Balaž, P. Muruganandam and S.K. Adhikari, "C programs for solving the time-dependent Gross-Pitaevskii equation in a fully anisotropic trap," Comp. Phys. Comm., vol 183, 2021-2025, 2012.

[7] S. Wolfram, The Mathematica Book, Wolfram Media, 2003.

[8] A.I. Nicolin, Spatial period-doubling in a Bose-Einstein condensate in an optical lattice, MSc Thesis, University of Copenhagen, 2004.

[9] M. Machholm, A. Nicolin, C.J. Pethick, and H. Smith, "Spatial period doubling in Bose-Einstein condensate in an optical lattice," Phys. Rev. A, vol. 67, 053613, 2004.

[10] M. Keceli, F.O. Ilday, and M.O. Oktel, "Ansatz from nonlinear optics applied to trapped Bose-Einstein condensates," Phys. Rev. A, vol. 75, 035601, 2007.

[11] A.I. Nicolin and R. Carretero-Gonzalez, "Nonlinear dynamics of Bose-condensed gases by means of a q-Gaussian variational approach," Physica A, vol. 387, 6032, 2008.

[12] A.I. Nicolin and M.C. Raportaru, "Faraday waves in high-density cigar-shaped Bose-Einstein condensates," Physica A, vol. 389, 4663, 2010.

[13] M.C. Raportaru and A.I. Nicolin, "Nonlinear dynamics of Bose-Einstein condensates by means of symbolic computations," ICT Innovations 2012 Web Proceedings ISSN 1857-7288, 9-15.

[14] A.I. Nicolin and M.C. Raportaru, "Effective low-dimensional polynomial equations for Bose-Einstein condensates," Rom. Rep. Phys., vol. 67, 143-147, 2015

Implementation of well known metrics and their practical use in PHP

Elena Jovanovska and Nevena Ackovska
Faculty of Computer Science and Engineering
"Ss. Cyril and Methodius" University
Skopje, Macedonia

***Abstract*—Software metrics have significant role in the process of evaluating systems design. They quantify the systems' characteristic providing better ground for identifying their problems. Various authors suggest different sets of metrics in order to evaluate the systems' design, but not all of those metrics have practical implementation in most popular programming languages. This paper provides description of an implementation of common metrics that do not traditionally have implementation for systems implemented in PHP.**

Keywords—software metrics; system evaluation; refactoring

I. INTRODUCTION

In order to achieve better evaluation of the software implemented with PHP, or any other programming language for that matter, quantifying software complexity, coupling, inheritance and duplication is necessary. Quantifying software characteristic by measuring them is called metric [1]. Another significant component of system evaluation is determination what needs to be measured in order for successful software design assessment. Even though various authors suggest different set of metrics for software evaluation and assessment, the main focus in this paper is on the practical implementation and use of metrics and evaluation strategies elaborated in [1].

Tools that are currently available, for PHP programming language, either provide limited set of metrics for measuring software entity structure, complexity, coupling and inheritance, for example tool "PHP_Depend" [2], or are developed as part of academic research and are not commercially available, as "Object Oriented Metrics Calculator" [3]. The lack of this type of tools might be because this approach of design evaluation is not strongly used in process software analysis done prior to the refactoring.

Although "Object Oriented Metrics Calculator" have full implementation of all traditionally used metrics and it is language independent, this tool is not commercially available. Additionally, it requires software analysis to be performed remotely which is a big issue for proprietary software. On the other hand, "PHP_Depend" is commercially available open-source application that is most commonly used for measuring system metrics for PHP systems. This tool provides most accurate and largest list of metrics measurements that can be used for proprietary and open-source applications. As a results this tool provides metrics exported as Overview pyramid, Abstraction Instability (A-I) diagram and full metrics measurements exported in xml format. Some of the metrics supported by this tool are: Cyclomatic Complexity Number (CCN), Extended Cyclomatic Complexity Number (CCN2), Afferent Coupling (CA), Efferent Coupling (CE), Number of Fanouts (FANOUT), Lines Of Code (LOC) and many other.

Dealing with duplications is the most common problem of the legacy systems. "PHP Copy/Paste Detector" [4] is a tool that is available for PHP for identification of code duplications. However, this tool contains only basic logic for identification of meaningful duplicates. Its logic is limited to filtering of the duplicates according their size, by lines of code or number of tokens.

Nevertheless, using only these tools does not provide sufficient knowledge of the code quality. Therefore additional changes and analysis need to be made. The next section elaborates on the problem in focus and the tool that was developed in order to meet these limitations.

II. THE PROBLEM

The application in focus is a commercial Customer Relationship Management (CRM) application developed using PHP as a programming language and supporting several types of databases (MySQL, Microsoft SQL Server, IBM DB2 and Oracle). This commercial open source application is widely used as a framework for developing more complex, customer and domain related, customized systems. Having good design that will allow the system to be easily extended and maintained is crucial for this type of application.

In order to evaluate and improve its code structure, the following metric were measured or calculated in its original version:

- Ratio between CCN (or CYCLO) and LOC is 0.234, which indicates that the application's complexity is uniformly dispersed across the application

- Ratio between Number of Methods (NOM) and LOC is 22.549. This means that the application tends to have large class methods

- Ratio between Number of Packages (NOP) and Number of Classes (NOC) is 122.222, which signals that the application has no granulated structure, classes are organized in few packages

Those measurements signal that there are potential design flaws that make the process of customization and maintenance more difficult. Measurements of the previously mentioned metrics together with other metrics, part of the "Overview pyramid" diagram, are shown in the Fig. 1.

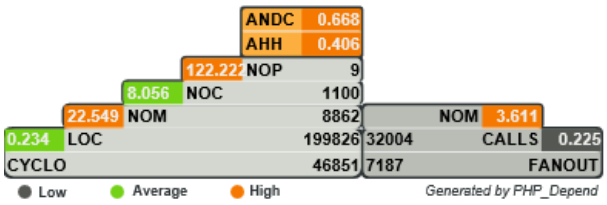


Fig. 1. CRM's Overview pyramid

Fig. 2 presents the Abstraction Instability (A-I) diagram generated with "PHP_Depend", where the level of Abstraction and Instability of the application is shown. The values displayed in this diagram are calculated according the Affective Coupling (CA) and Effective Coupling (CE) metrics [2] [5]. From this diagram same conclusion can be made as from the ratio between NOP and NOC, i.e. the application has no granulated structure and the classes are organized in few packages. This diagram also shows that the level of abstraction in the application is very low.

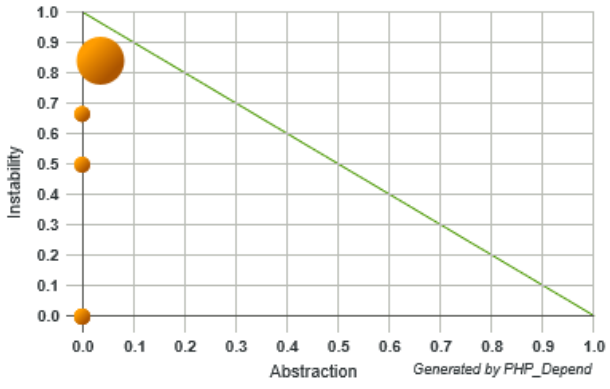


Fig. 2. CRM's A-I diagram

Analyzing those results, it can be concluded that the application might have serious design flaws related with its inheritance structure. Same conclusion can made from the values of the metrics Average Number of Derived Classes (ANDC) and Average Hierarchy Height (AHH) displayed in Overview pyramid, which indicates that the application tend to have flat and wide class hierarchy.

Extending the knowledge of the system, elaborated previously, with measurement results from other metrics supported by "PHP_Depend", like Weighted Method Count (WMC), Non Private Properties (VARSNP) and Number Of Added Methods (NOAM), does not provide enough information which part of the code suffers from design flaw.

As already explained, dealing with large, but limited set of metrics and no possibility to identify code duplications that are potential design flaws, implies the need of additional metrics and tool for extended duplication analysis to be developed. Those metrics are implemented as extension of previously mentioned tool "PHP_Depend" [2] called "PHP_Depend_Extension". This, newly defined tool, provide more information of the application's structure, complexity and coupling needed for evaluation and refactoring process.

Even through this tool is developed to provide implementation of some metrics needed for measuring CRM's characteristics for evaluation of its design, this tool can be used and for other applications developed using PHP programing language.

III. METRIC'S DEFINITION AND USAGE

Each software, in order to be flexible and easily maintained, needs to be characterized with the following heuristics:

- Each entity in the system needs to have its own purpose, its own identity in the system – also known as Identity harmony

- Each entity needs to collaborate with the other system's entity in order to achieve its purpose – also known as Collaboration harmony

- Each entity needs to harmonically collaborate with its ancestors and descendants – also known as Classification harmony [1] [6]

The software that does not implement the heuristics stated above, can be assessed as software with potential design flaws and disharmonious design.

More than one matric is used to assess software design more precisely and to provide more accurate analysis and identification of the potential design flaws. The combination of the metrics, their thresholds and logical condition between them is called detection strategy. Detection strategy is one of the technics used to evaluate software design [1].

Detection strategies in the rest of this paper are going to be defined using following syntax:

(input term 1) (logical AND/OR operator) (input term 2) (1)

where input term can be: primitive; statement construct from metric, logical operator (>, ≥, =, ≤ or <) and metric's threshold value, or derived; construct from several primitive input terms related with AND/OR logical operators.

Thresholds used in the detection strategies, as part of the input terms, can be statistic-based or meaningful. In the tables below are shown threshold used for the metrics from both types: the Table 1 shows the Statistic – based thresholds, and the Table 2 shows the meaningful thresholds.

TABLE I. STATISTIC-BASED THRESHOLDS

Metric	PHP			
	Low	***Average***	***High***	***Very High***
CCN	1	5	7	10
CYCLO/ lines of code	0.16	0.20	0.24	0.36
LOC/ Method	7	10	13	19.5
NOM/ Class	4	7	10	15
NOC/ Package	6	17	26	-
LOC/ Class	28	70	130	195
CALLS/ Operation	2.01	2.62	3.2	-
FANOUT/ CALLS	0.56	0.62	0.68	-
ANDC	0.25	0.41	0.57	-
AHH	0.09	0.21	0.32	-
WMC	5	14	31	47
AMW	1.1	2.0	3.1	4.7

TABLE II. MEANINGFUL THRESHOLDS

Numeric Value	**Sematic Label**
0	None
1	One/Shallow
2-5	Two, Three/Few/Several
7-8	Short memory capacity

For identification of potential design flaws caused by violation of the system harmonies, previously mentioned, detection strategies are used. The detection strategies used in this research are described in the following sub – sections.

A. *Identity disharmonies*

Classes that tend to centralize system's intelligence [1] are characterized by: accessing many attributes from other classes, the class itself is very complex and its cohesion is low. Those classes are also called God Classes and can be identified by use of detection strategy stated below:

$$(ATFD > few) \wedge (WMC \geq very\ high) \wedge (TCC < 1/3) \quad (2)$$

On the other hand, methods that tends to centralize class's intelligence [1] are large and very complex methods that are using a lot of variables. They can be identified by use of following detection strategy:

$$(LOC > high_{(class)}2) \wedge (CYCLO \geq high) \wedge (MAXNEST \geq several) \wedge (NOAV > many) \quad (3)$$

For identification of methods that are accessing a lot of attributes from other classes, following detection strategy is used:

$$(ATFD > few) \wedge (LAA < 1/3) \wedge (FDP \leq few) \quad (4)$$

Refactoring of those methods helps minimizing system cohesion.

Classes that are data providers, classes that are simple data holders with few methods that use those data, are identified by following detection strategy:

$$(WOC < 1/3) \wedge ((VARSNP + NOAM > few \wedge WMC < high) \vee (VARSNP + NOAM > many \wedge WMC < very\ high)) \quad (5)$$

The existence of large number of data classes indicates that the system does not have object-oriented design.

B. *Collaboration harmony*

In order to identify methods that are not complex and are invoking many methods that belongs to few classes, following detection strategy is used:

$$((CINT > short\ memory\ capacity \wedge CDISP < ½) \vee (CINT > few \wedge CDISP < ¼)) \wedge (MAXNEST \leq shallow) \quad (6)$$

Those methods indicate tight system coupling.

In the other hand, methods that invoke many methods from many other classes are identified by:

$$(CINT > \text{short memory capacity}) \wedge (CDISP \geq ½) \wedge (MAXNEST > \text{shallow}) \quad (7)$$

C. Classification harmony

For identification of classes that are refusing their parent bequest, the following detection strategy is used:

$$(NProtM > few \wedge BUR < 1/3 \wedge BOvR < 1/3) \wedge (AMW > average \wedge WMC > average \wedge NOM > average) \quad (8)$$

Classes identified by this detection strategy are potentially misplaced, they do not belong to their current class hierarchies.

In order to identify classes that tend to implement services that are unrelated to those provided by its base class, the detection strategy stated below is used:

$$(NOAM \geq average(NOM) \wedge NOAM/NOM \geq 2/3 \cdot 100) \wedge ((AMW > average \vee WMC \geq very\ high) \wedge NOM \geq high) \wedge (AMW_{(parent\ class)} > average \wedge NOM_{(parent\ class)} > high/2 \wedge WMC_{(parent\ class)} \geq (very\ high)/2) \quad (9)$$

Measurements and calculations for some of the metrics mentioned in the detection strategies above are not included in the "PHP_Depend". This paper provides description of their implementation as extension of the tool "PHP_Depend" called "PHP_Depend_Extension".

IV. CUSTOM TOOL DEVELOPPED – "PHP_DEPEND_EXTENSION"

As previously mentioned, for this research, a custom made extension of the tool "PHP_Depend" was created. It includes implementation of some metrics used for identification of design flaw of the system that currently are not part of the "PHP_Depend" implementation. The following metrics are included in this extension: Maximum Nesting (MAXNEST), Access To Foreign Data (ATFD), Access To Local Data (ATLD), Locality of Attribute Access (LAA), CALLS (on method level), Coupling Intensity (CINT), Coupling dispersion (CDISP), Number of Protected Methods (NProtM), Number of Protected attributes (VARSProt), Base Class Overriding Ration (BOvR) and Weight Of a Class (WOC). The description of those metrics is given below:

A. Complexity metrics

One of the metric that are used for better evaluation of application complexity is MAXNEST. This metric calculates maximum nesting level of a control structures within some artifact, functions and class's methods. It is used in process of identification of methods that tend to centralize class intelligence or methods that invoked many other methods that belongs to few or many classes [1].

Other complexity metric developed as part of this tool is WOC. It calculates number of functional public methods (NPM) divided by total number of a public members [4] [7]. This metric is used in identification of classes that are data providers.

$$NPM/(VARSNP + NPM) = WOC \quad (10)$$

B. Coupling and inheritence metrics

The following metrics provide better overview of the coupling in the system: ATFD, ATLD, LAA, CALLS (on method level), CINT and CDISP.

ATFD metric counts number of attributes from unrelated classes that are accessed directly or by invoking accessor methods from measured method or class [7]. It is used in detection strategies for identification of classes that tend to centralize system intelligence [1] or methods that access a lot of attributes from other classes.

ATLD metric is used to count number of attributes from defining class and its parent classes that are accessed directly by the method. This metric is used for calculating LAA metric.

LAA metric presents the ratio between the number of total attribute invokes within the method and the locally used attributes [1]. This metric is used in detection strategy for identification of methods that access a lot of attributes from other classes. This metric is calculated with following formula:

$$ATFD/(ATLD+ATFD)= LAA \quad (11)$$

Metric CALLS (on method level) counts how many methods are invoked form measured method. This metric is used for prioritization of methods that invoked many other methods that belongs to few or many classes.

CINT specifies the number of methods from unrelated classes that are invoked from the measured method [1]. Is used for identification of methods that invoked many other methods that belongs to few or many classes.

CDISP specifies the number of classes in which methods counted in the CINT metrics belong [1]. Is used for identification of methods that invoked many other methods that belongs to few or many classes.

BOvR calculates number of methods of the measured class that override methods from the base class, divided by total number of methods in the class [1]. This metric is used in detection strategies for identification of classes that are refusing their parent bequest.

$$NOOM/NOM = BOvR \quad (12)$$

C. Class level metrics

Metrics that provide extended overview of the entity structure of the application are: NProtM and VARSProt.

NProtM counts number of protected methods in the class and it is used in identification of classes that are refusing their parent bequest.

VARSProt counts number of protected attributes in the class.

V. Results of the Custom Tool Implementation

By use of newly implemented metrics, with the tool "PHP_Depend_Extension", in the combination with existing metrics from "PHP_Depend" the following results are calculated:

- 9 classes tend to centralize application intelligence – God Classes (Table 3). For identification of those classes, since TCC is not supported by "PHP_Depend", metrics CA is used instead. Measured values for the metrics that are used in identification of the this types of classes are shown in the table below:

TABLE III. God Classes

Class Name	Metrics		
	WMC	***ATFD***	***CA***
SugarBean	1171	13	37
ListView	383	15	14
OutboundEmail	86	7	10
SearchForm	164	6	11
SugarEmailAddress	253	7	14
Email	512	40	18
Lead	88	6	16
ModuleBuilderController	183	8	28
User	257	27	64

- 194 methods are strongly depended from other classes attributes
- 27 classes are categorized as data providers
- 23 methods tend to centralize application intelligence
- 513 methods are invoking many methods from few unrelated classes
- One method is invoking many methods from many different unrelated classes
- One class has implementation that refuses parent bequest. Measurements for the metrics used in the detection strategy for this disharmony are shown in the Table 4

TABLE IV. Classes that refuses parent bequest

Metric	Measured Value
NPROTM	6
NOOM	16
NOM	22
WMC	83
BOvR	0.72
AMW	3.09

Those entities together with duplications identified with the tool - "PHP_Depend_Extension", provide solid starting point in the process of refactoring that will result with better application design.

VI. Additional Tool Improvements – Future Work

In order to perform more accurate assessment of system design following metrics should be included as part of the future "PHP_Depend" extensions:

- Number of Parameters (NOPAR) of the method - this metric can be used for better assessment of method complexity [3] [9]
- Foreign Data Providers (FDP) – this metric counts number of classes in which the attributes are access. It should be used for better identification of methods that are accessing a lot of attributes from other classes [1].
- Tight Class Cohesion (TCC) – counts relative number of method pairs of a class that access in common at least one attribute of the measured class [1] [10]. This metric should be used for better identification of classes that centralize system intelligence [1].
- Base Class Usage Ration (BUR) – is a number of used inheritance-specific members by the measured class, divided by the total number of inheritance-specific members from the base class [1]. This metric should be used for better identification of classes that are refusing their parent bequest

Additionally, as improvement of this tool new way of outputting calculated metrics can be implemented. The new

tool output should extract data in format defined by some metric ontology, for example ontologies mentioned in [8] or SEON Metrics ontology [9].

VII. CONCLUSION

Limited set of tools that analyze the code and poor result sets should not be a discouragement while working with large systems. Effective process of system evaluation and refactoring can lead to a large benefit for people that participate in the process of development and support of the legacy systems. Therefore development and extensions or additional tools that will provide more information about system's structure, complexity and inheritance can be developed, as in this particular case. For this particular research the tool named - "PHP_Depend_Extension" was created as an extension to "PHP_Depend".

The measuring and calculation of a single metric is not enough for system's evaluation. Combination of metrics need to be used in order to provide better evaluation and identification of design flaws. Those metric's combinations, logical expression, are called detection strategies [1]. Detection strategies as technique can be used for identification of potential system entities that need to be refactored in order to improve system design, but not all of the identified entities really suffers from design flaws.

Even though there are additional metrics that can be implement to improve system evaluation, metrics that are implemented as extension of "PHP_Depend" and elaborated in this paper, provide sufficient information about software design and identification of entities that potentially suffer from design flaws.

In the extended analysis that need to be done over design flaws candidates, in the refactoring process later on, prioritization of the entities that need to be in focus is very helpful. Additional, related, metrics can be used to prioritize and organize process of analysis and refactoring.

REFERENCES

[1] M. Lanza, R Marinescu, "Object-Oriented Metrics in Practice: Using Software Metrics to Characterize, Evaluate, and Improve the Design of Object-Oriented Systems", Springer-Verlag, 2006

[2] M. Pichler, "PDepend," [Online]. Available: http://pdepend.org/.

[3] "Object Oriented Metrics Calculator", [Online]. Available: http://habanero.ifi.uzh.ch/famixMetrics/

[4] S. Bergmann, "Copy/Paste Detector (CPD) for PHP code," [Online]. Available: https://github.com/sebastianbergmann/phpcpd/tree/master

[5] R. C. Matrin,"OO Design Quality Metrics - An Analysis of Dependencies"; 1994

[6] A. J. Riel, "Object-Oriented Design Heuristics", Addison – Wesley, 1996

[7] R. Marinescu, "Measurement and Quality in Object-Oriented Design", IEEE International Conference on Software Maintenance, pp. 701 – 704, September 2005

[8] M. de los Angeles Martin, L. Olsina, "Towards an Ontology for Software Metrics and Indicators as the Foundation for a Cataloging Web System", First Latin American Web Congress Proceedings, 2003.

[9] Department of Informatics of the University of Zurich, "SEON", [Oneline]. Available: http://www.se-on.org/

[10] J.M. Bieman, B.K. Kang,"Cohesion and Reuse in an Object- Oriented Systems", Proceeding ACM Symposium on Software Reusability, April 1995

Applying Soft Computing Methods to Discriminant Analysis

Milan Djordjevic
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria

***Abstract*—The purpose of this research was to present advantages when one of the soft computing techniques is used, where the value (intensity) of relation is an element from a real interval [0,1] and not only from $\{0,1\}$ as in a classical case, and apply it on the statistical tool for discriminant analysis. Specific application in order to test hypothesis has been developed. This paper will present several examples. One of them will be used to demonstrate the power of new approach. Also, several benchmark data sets from UCI machine learning repository will be used. Special emphasis will be given on one complex example used in banking for loan approval. When we talk about banks, the proper selection of clients to whom a loan will be granted is the essence of successful business, and discriminant analysis is an excellent tool that can be used for this purpose. This example demonstrates precisely this multivariate analysis, with special reference to the power of this method when fuzzy logic joins in. Finally, the power of new approach to discriminant analysis is clearly presented, since the results achieved in this way were better in every case.**

Keywords—Soft computing, Discriminant analysis, Fuzzy logic, Classification, Cluster analysis, Loan approval, MATLAB, UCI data base

I. Introduction

One of the most frequently used tools in statistics is the discriminant function analysis or simply discriminant analysis (DA). This method is based on a model creation which allows us for the values of interval variables (the set of attributes of a data set) to determine belonging to the certain group. Discriminant analysis has been widely used in various fields of natural science [1] as well as economy and humanistic. Linear discriminant analysis (LDA) was first introduced by R.A. Fisher [2] who tried to find a good projection line in order to obtain well-separated classes. In discriminant analysis, also known as supervised classification, known classifications of some observations ("training set") are used to classify others [3], [4]. The number of classes is assumed to be known. The method separates two classes of objects by maximizing the ratio of between-class variance to within-class variance in any particular data set as follows.

$$J(A) = \max_{A} \frac{A^T S_B A}{A^T S_W A} \tag{1}$$

where J(A) is called as objective function and S_B and S_W denote the between- and within- class scatter matrix, respectively. According to (1), one must find a linear transform matrix A, which maximizes the objective function besides maximizing the projected class means and minimizing the class variances [5].

In order to obtain a good model, which is presented by the discrimination function, it is necessary that the input data contain in addition to the attribute values also belonging to a group. If we want to achieve good results by using discriminant analysis, it is necessary to fulfill the following preconditions which are related to input data:

1) Groups (clusters, classes) should have at least approximately normal distribution;
2) Participants of the sample are assumed to be randomly sampled;
3) Matrices of variation and covariation should for each group be the same;
4) The initial data must be correct qualified (divided into groups).

The main strength, but also the weakness of classical discriminant analysis is based on the fact that belonging to a group is determined by a linear combination of interval variables (attributes of the problem). How to overcome disadvantages of this approach and at the same time to keep the tool in usage linear, suitable because of mathematical simplicity and small number of calculations, we will see after we take into account the basic assumptions of [0,1] valued logic as natural generalizations of logical reasoning [6]. One recent paper presents one way of modeling candlestick patterns using interpolative Boolean logic [7]. The fuzzy sets [8] represent mathematical theory suitable for modeling imprecision and vagueness. The fundamental fact that lies behind fuzzy logic is that any field and any theory may be fuzzified by using the concept of fuzzy set. Through fuzzzification we archive greater generality, higher expressivity, an enhanced ability to model real-world problems, and a methodology for exploiting the tolerance for imprecision [9].

Technically, main goal of discriminant analysis is formation of linear combinations of independent variables which will be discriminated between predefined groups so that the error of incorrect classification of observations will be minimized, or in other words, to maximize the variance relative ratio between and within groups. By linear combination of independent variables for each subject or object the number is determined, called discriminative score, which is then transformed into a posterior probability that the subject or the object originates from the one of the groups. Nevertheless, classical $\{0,1\}$ valued Boolean logic is suitable for solving "black and white" or bi-valued problems. However, many real world problems from the logical point of view are more-valued even infinite-valued and/or [0,1] - valued problems [10].

The remainder of the paper is structured as follows. In Section 2 methodology, theoretical background and difference

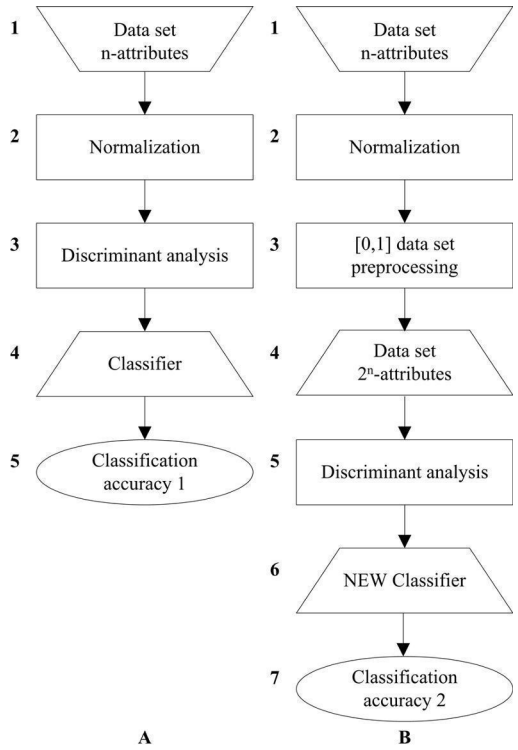


Figure 1: Difference between old (A) and new approach (B) to discriminant analysis

between old and new approach to discriminant analysis is presented. Description of used data sets is given in Section 3. Main experimental results are presented in Section 4. Section 5 is reserved for conclusion and future work.

II. Methodology

In this paper one of the soft computing [11] methods which uses [0,1] valued logic as a natural generalization of logical reasoning is presented. This method detects and uses interrelationships of the attributes of the given problem in order to obtain preprocessed data on which a well-known apparatus of discriminant analysis will be applied. We will show that output data of the soft computing method give a better insight into observed phenomena, because the analysis uses attributes like the classical approach does, as well as the logical combination of attributes. The results obtained in this way are much better than those obtained over a non-preprocessed data. The application is developed using "The MathWorks MATLAB®"[1]. Application is able to compare results obtained by classical and new approach to discriminant analysis in order to get better interpretation of results. Furthermore, the application includes the cluster analysis which determines the new objects belonging to the group after the model has been created.

A. Fuzzy aspects and description of the experiments

In [6], a new [0,1] valued logic is presented. It can be seen as a basis of decision making and reasoning complementary to $\{0,1\}$ or "black and white" reasoning. Furthermore, it is proven that all properties of classical $\{0,1\}$ logic are preserved in [0,1] valued logic. The developed application is able to solve linear problems using $\{0,1\}$ valued logic and [0,1] valued logic. The difference between old and new approach to discriminant analysis and the steps taken during one single experiment can be seen in Figure 1.

Data set is entered in the form of matrix in which the rows are representing the values of observations, and the columns - attributes. Vector which shows belonging to the class of all observations needs to be entered separately as well. This one column matrix will allow model creation. The next step, required for further analysis, is normalization of the input data. By default, the normalization is linear and all the attributes are normalized between the closed interval [0,1], but also we can perform different types of normalization for each data set attribute.

Next step is crucial and it is depicted as step 3 in Figure 1B. The application allows to expand the input data (which are represented as a matrix where rows are observations and columns - attributes), on which the discriminant analysis will be applied. A new matrix with new set of attributes is obtained by this expansion (step 4 in Figure 1B). This new matrix will contain 2^n set of attributes where n is the number of attributes of initial data set. These new attributes are representing interrelationships between the initial attributes and are obtained as it is presented in Table I. Firstly, the atomic structural vectors are calculated (2^n), and then based on atomic structural vectors and logical structural functions the expressions for data set extension are obtained. Finally, every observation in data set is extended to 2^n attributes. The application uses $AND(\&)$ operator as a logical structural function which is implemented as a product $(*)$ and as a minimum (min). The resulted expressions obtained in this way are applied to attribute values of every observation in initial data set and thereby the new input data set which contains initial values of attributes and the values of their dependencies is obtained.

Further step that we can do is to apply some linear statistical method (in this case that is discriminant analysis) on the new input data set with 2^n attributes. As a result a discrimination function is obtained and the graphical representation of

TABLE I: Attributes of data set and components of the structural vector in partial lexicographic order used for initial data set expansion

no.	set of attributes	structural vectors
1.	ø	$\mu(\emptyset) = \mu_0$
2.	$\{a_1\}$	$\mu(\{a_1\}) = \mu_1$
3.	$\{a_2\}$	$\mu(\{a_2\}) = \mu_2$
4.	$\{a_3\}$	$\mu(\{a_3\}) = \mu_3$
5.	$\{a_1, a_2\}$	$\mu(\{a_1, a_2\}) = \mu_{1,2}$
6.	$\{a_1, a_3\}$	$\mu(\{a_1, a_3\}) = \mu_{1,3}$
7.	$\{a_2, a_3\}$	$\mu(\{a_2, a_3\}) = \mu_{2,3}$
8.	$\{a_1, a_2, a_3\}$	$\mu(\{a_1, a_2, a_3\}) = \mu_{1,2,3}$

[1] http://www.mathworks.com

the results. The obtained discriminant function allows us to classify new observations correctly. In addition, we are able to see in figure graphically depicted initial matrix and how every single observation is distributed. In order to estimate classification accuracy, the number of errors is calculated, which shows how many observations are classified in wrong class.

As illustrated in Figure 1A, the application is able to apply discriminant analysis over initial data set, so we can analyse output of the classical approach to the problem as well.

III. Data sets

A. Basic example

The first illustrative example is designed to be demonstrative. Furthermore, it will show the power of new approach. All characteristics of this basic data set are given in Table II.

TABLE II: Basic example data set characteristics

data set name	no. of instances	no. of attributes	no. of classes
BasicExample	50	3	2

B. Benchmark data sets

Several appropriate data sets from UCI[2] machine learning repository were used, which is the most widely used data base in the classification literature. Data sets names and its common characteristics are shown in Table III.

TABLE III: Benchmark data sets characteristics from the UCI data base repository

data set name	no. of instances	no. of attributes	no. of classes
Haberman	306	3	2
Transfusion	748	5	2
BanknoteAuthentication	1372	4	2

C. Banking example - loan approval data set

This data set contains 1500 instances which are presenting real bank clients who had taken a loan from the bank. Data set is composed of two classes. In first class are the clients who have regularly returned their loans and the second who have not. Data set basic characteristics are shown in Table IV.

TABLE IV: Banking example data set characteristics

data set name	no. of instances	no. of attributes	no. of classes
BankingExample	1500	3	2

IV. Experimental results

A. Basic example

The new approach which is based on the [0,1] - valued logic can solve the problems which so far have been unsolvable by liner discriminant analysis. To support this claim the following example is presented. In Figure 2a, we can see that each object is described with two criteria and their class membership have a given shape. The classes are marked with red and blue points. In Figure 2b, the results of initial data using classical approach to discriminant analysis are depicted. The green points on a Figure 2b are representing the objects which are incorrectly classified. Also, Figure 2c shows that significant difference between the two groups was not found since the clear separation line between two groups was not created. Unlike the traditional approach, which has classified 24 objects incorrectly out of 50 given objects, on Figure 3b we can see that new approach based on [0,1] - valued logic was able to separate classes completely. Also, Figure 3c shows that significant difference between two groups was revealed and number of incorrectly classified objects is equal to zero.

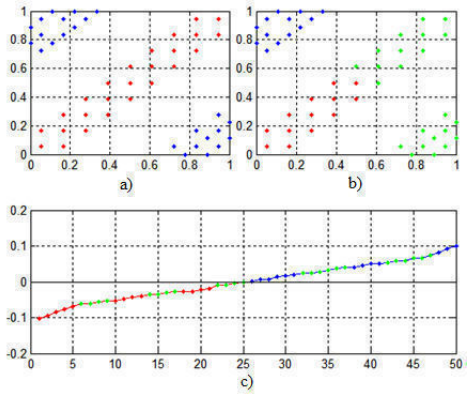


Figure 2: Classical approach

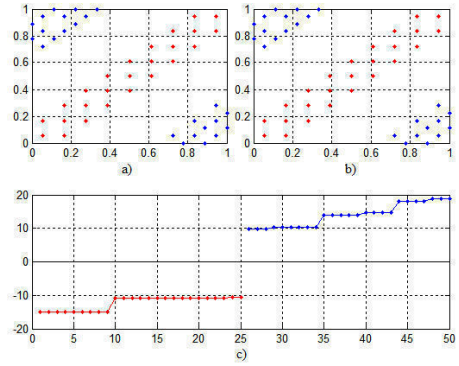


Figure 3: New approach

B. Banking example - loan approval

Let us assume that a bank has collected data for 1500 customers who have so far received credit, as follows:

1) Income;
2) Readiness to take risks;
3) The attitude towards family.

In Figure 4a with blue color are marked the clients who have returned their loans on time and with red color the clients who have not. So, several groups of those who are not appropriate for loan approval clearly can be recognized. After

[2]http://archive.ics.uci.edu/ml/datasets.html

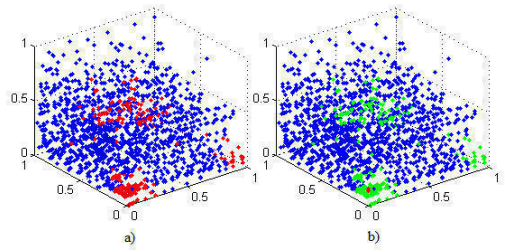


Figure 4: Loan approval - Classical approach

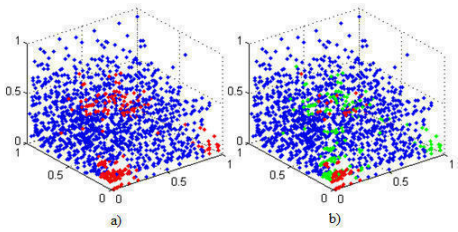


Figure 5: Loan approval - New approach

processing the data using classical approach, results depicted in Figure 4b were obtained. In this figure it is clearly visible that classical discriminant analysis was not able to divide the observed data set into appropriate groups. Again, with green color the objects that are misclassified are marked. It is easy to see that all the objects that belong to the group of those who are not appropriate for loan approval are assigned to the wrong group. Discriminant function obtained in this way is totally useless.

After processing the data using new approach, the obtained results were much better. In Figure 5b it is easy to see that other groups are recognized. These groups are important since they represent the clients who are not appropriate for loan approval. These objects are marked with red color in the Figure 5b. With green color are marked objects which are misclassified. It is evident that the number of such objects is significantly reduced in comparison to the previous method. Hence, discriminant function obtained in this way could already serve employees to determine whether the client is eligible to get a loan, in other words they will be able to get fast calculation of client eligibility for loan approval.

C. Benchmark data sets and overall results

In experimental work 5 data sets were used. According to the Table V, we can see that when new approach to discriminant analysis is used, better results are achieved.

Even when difference in obtained classification results is not too big, as it is in banking example, the focus is on identifying appropriate classes in data set. Proudly, it can be stated that new approach to discriminant analysis shows better results and also was able to separate classes more clearly.

V. Conclusion and Future Work

The purpose of this research was to present advantages when one of the soft computing techniques is used, where the value (intensity) of relation is an element from a real interval

TABLE V: Classification accuracy of all data sets

data set name	classical DA	fuzzy DA $(*)$	fuzzy DA (min)
BasicExample	52%	100%	100%
Haberman	74,84%	77,12%	75,16%
Transfusion	77,14%	79,28%	77,54%
BanknoteAuthentication	98,83%	100%	100%
BankingExample	86,33%	88,4%	86,8%

[0,1] and not only from $\{0,1\}$ as in a classical case, and apply it on the statistical tool for discriminant analysis. The idea that expansion of classical $\{0,1\}$ - valued logic approach with [0,1] - valued logic and at the same time to keep the same statistical apparatus will provide much better results, represents the essence of this research. It has been shown that certain problems were not solvable by classical approach to discriminant analysis, while the solutions obtained by new approach were not only better, but also contained smaller percentage of errors. The developed application provides a comparative overview of the results obtained by the classical and the new approach. Both methods have processed the same input data. The output results obtained by applying the techniques of soft computing provide a better insight into the observed phenomena, because the analysis uses attributes like the classical approach does, as well as the logical combination of attributes. Method by itself is independent of use case, so it can serve in many other use cases.

References

[1] C. Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society - Series B*, 10(2):159–203, 1948.

[2] R. A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8(4):376–386, 1938.

[3] Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.

[4] G.J. McLachlan. *Discriminant Analaysis and Statistical Pattern Recognition*. Ney York: Wiley, 1992.

[5] R. A. Johnson and D. W. Wichern, editors. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[6] Dragan G. Radojevic. (0,1)-valued logic: A natural generalization of boolean logic. *Yugoslav Journal of Operations Research*, 10(2):185–216, 2000.

[7] Ivan Nesic, Pavle Milosevic, Aleksandar Rakicevic, Bratislav Petrovic, and Dragan G. Radojevic. Modeling candlestick patterns with interpolative boolean algebra for investment decision making. In *Soft Computing Applications - Proceedings of the 5th International Workshop Soft Computing Applications, SOFA 2012, Szeged, Hungary, 22-14 August 2012*, pages 105–115, 2012.

[8] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338 – 353, 1965.

[9] George J. Klir and Bo Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995.

[10] L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[11] Lotfi A. Zadeh. Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, 37(3):77–84, March 1994.

Uper Bound on *q*-ary Codes

Dejan Spasov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
dejan.spasov@finki.ukim.mk

***Abstract*—Given the code length *n* and minimum distance *d* we show the following bound over a nonbinary alphabet of *q* elements**

$$A_q(n,d) \le q^{n-d+3-\log_q((q-1)(n-d+3)+1)}$$

Since this bound improves the Singleton bound, we discuss the implications on the code parameters of the MDS codes and the MDS conjecture.

Keywords—error-correcting codes, code bounds, MDS codes, Singleton Bound, MDS Conjecture.

I. INTRODUCTION

Let F_q be a finite field of q elements and let F_q^n denote n-dimensional vector space over F_q. Then a *code* C is a subset of F_q^n of M elements. Elements of the code $c_i \in C$ are called *codewords* and M is known as the size of the code C.

Let $d(x,y)$ denote the *Hamming distance*, i.e. the number of coordinates in which two vectors x and y differ, and let $wt(x)$ denote the (*Hamming*) *weight*, i.e. the number of nonzero coordinates of x. Then we say that the code C has *(minimum) distance d* if

$$d = \min\left\{d(c_i,c_j)\right\}, \forall c_i, c_j \in C, i \ne j. \quad (1)$$

A code C is linear if its codewords form k-dimensional linear subspace in F_q^n. We will write $[n,k,d]_q$ to denote that the code C is linear over the field F_q. For linear codes there exist k basis vectors that are kept as rows in a matrix G called the *generator matrix*. For each linear code there is a generator matrix of type $G = [I \quad A]$ for which we say that is in *standard form*. It is well-known that for linear codes there exist a so-called *parity check matrix* H, such that $\forall c_i \in C$ $Hc_i^T = 0$. Let $G = [I \quad A]$ is the generator matrix, then $H = [-A^T \quad I]$ is the parity check matrix of the same code.

The following theorem is fundamental result in coding theory:

Theorem 1: A code C with parameters $[n,k,?]_q$ and parity check matrix H has minimal distance d if every linear combination of $d-1$ columns of H is linearly independent, and there exist a linearly dependent combination of d columns of H.

The *covering radius* of a code is the largest possible distance between the code C and a vector from F_q^n, i.e. [1]

$$\rho = \max_{x \in F_q^n} \min_{\forall c \in C} d(x,c). \quad (2)$$

II. UPPER BOUNDS ON THE CODE SIZE

Given q, n, and d, let $A_q(n,d)$ denote the largest number of codewords at distance d over F_q^n. Important problem in coding theory is finding a code of largest size, i.e. $A_q(n,d)$. This problem is well known to be intractable, since the only known method that works for arbitrary q, n, and d is super-exponential search over all possible orderings of F_q^n.

Thus we want to provide approximate results on $A_q(n,d)$ known as *upper bounds* and *lower bounds*. Lower bounds can be estimates of code parameters of a code family described with certain constructive technique. Upper bounds exclude existence of a code with parameters better than the bound. If an $(n,M,d)_q$ code meets an upper bound then $A_q(n,d) = M$.

The most famous example of an upper bound is the Hamming bound:

Theorem 2: Let $t = \left\lfloor \frac{d-1}{2} \right\rfloor$. Then

$$A_q(n,d) \le \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i} (q-1)^i} \quad (3)$$

Codes that meet (3) with equality do exist and are known as *perfect codes*. Examples of non-binary perfect codes are the $[11,6,5]$ ternary Golay code and the q-ary Hamming code with parameters

$$\left[\frac{q^r-1}{q-1}, \frac{q^r-1}{q-1}-r, 3\right]. \quad (4)$$

Interestingly, no other perfect codes exist [2].

A. The Singleton Upper Bound

Singleton Bound is another well-known bound for codes over non-binary alphabet. We will start with the following simple results

Theorem 2: Let $n > d > 1$. Then

i. $A_q(n,d) \le A_q(n-1,d-1)$ (5)

ii. $A_q(n,d) \le qA_q(n-1,d)$ (6)

Both results are proved using *puncturing* – a technique for obtaining new codes from old – and appropriate pigeonhole argument.

The following result can be found in [1]

Theorem 3: $A_q(n,n) = q$ (7)

Solving the recurrence relation (6) with (7) as initial conditions, we obtain the Singleton bound.

Theorem 4: (Singleton Bound)

$$A_q(n,d) \le q^{n-d+1} \quad (8)$$

or in case of linear codes

$$k \le n-d+1 \quad (9)$$

Suppose we are interested in asymptotical behavior of codes. Thus a single code is of no interest to us, but an infinite family of codes C_i, $i \to \infty$, of increasing length $n_i \to \infty$. The main concern here is whether $k_i = \log_q A_q(n_i, d_i)$ will follow n_i by a constant ratio or it will lag behind. Thus infinite code families are more convenient to be described in terms of the *code rate* $R = \lim_{i\to\infty} \inf\left(\frac{k_i}{n_i}\right)$ and the *relative distance* $\delta = \lim_{i\to\infty} \inf\left(\frac{d_i}{n_i}\right)$. From (8) we obtain the asymptotical Singleton bound

$$R \le 1-\delta \quad (10)$$

B. New Upper Bound

Combining the recurrence relation (5) with (4) as initial condition, we obtain the following bound:

Theorem 5:

$$A_q(n,d) \le q^{n-d+3-\log_q((q-1)(n-d+3)+1)} \quad (11)$$

Proof: Starting from (5) we obtain

$$A_q(n,d) \le A_q(n-1,d-1) \le A_q(n',3)$$

where $n' = n-d+3$. From (4) we have $n' = \frac{q^r-1}{q-1}$, thus $r = \log_q((q-1)n'+1)$. Since the dimension of the Hamming code is $k' = n'-r$, we conclude $A_q(n,d) \le q^{k'}$. □

III. MDS CODES

MDS codes are a class of codes that meet Singleton bound (9) with equality. Reed-Solomon codes, one of the most studied family of codes, are in fact MDS codes. The popularity of MDS codes stems from the fact that they exhibit many useful properties [3]. However, one mystic property that still accompanies the MDS codes is their length.

Though (8) does not restrict the length of the MDS codes, even from our bound (11) it is obvious that MDS codes can exist if

$$2-\log_q((q-1)(n-d+3)+1) \ge 0$$

Solving this inequality we will obtain the following well-known restriction on code size

$$k \le q-1. \quad (12)$$

One of the remaining conjectures from the "old" coding theory is the MDS conjecture:

Conjecture 1: If C is an $[n,k,n-k+1]$ MDS code, then $n \le q+1$, with exceptions when q is even and $k=3$ or $k=q-1$.

A natural question arises whether recurrence relations (5) and (6) with appropriate initial conditions can settle this conjecture. We believe that for solving the MDS conjecture additional recurrence relation is needed:

Conjecture 2:

$$A_q(n,d) \le q^{-1}A_q(n,d-1) \quad (13)$$

Assuming (13) holds true, using (4) as initial condition, we can obtain the following bound:

$$A_q(n,d) \le q^{n-d+3-\log_q((q-1)n+1)} \quad (14)$$

Following the previous reasoning for (12) we conclude that

$$n \le q+1. \quad (15)$$

In order to prove (13) we can start from $A_q(n,d-1)$ where we know that we have a subset of $\frac{A_q(n,d-1)}{q}$ codewords that have the same symbol α, $\alpha \in F_q$, in the i-th coordinate. Instead of α, in the i-th coordinate we can write all symbols from F_q and thus we hypothesize that we have a $\left(n, q^{-1}A_q(n,d-1), d\right)$ code. Now we have to prove that this

construction is an upper bound on $A_q(n,d)$. So far, no simple argument appears to be useful.

The difficulty of proving (13) can be better viewed in the case of linear codes. Given *n*, *q*, and *d,* let $\lfloor n, \log_q A_q(n,d), d \rfloor_q$ be the best linear code with parity check matrix *H.* Can we add a column-vector to *H* and obtain an $[n,k,d-1]_q$ code without increasing the codimension? If we can answer this affirmatively, we have proved (13). However, adding a column-vector to *H* without increasing the codimension is only possible if the covering radius ρ of the $\lfloor n, \log_q A_q(n,d), d \rfloor_q$ is exactly $d-1$. Covering radius is notoriously difficult topic in coding theory. So far, for the best codes it is only known that $\rho \le d-1$ [5,6].

IV. Conclusion

There are few upper bounds for codes over arbitrary alphabet F_q. In fact the Singleton bound is still the best known bound in the asymptotically large environment (10).

We have shown asymptotical improvement of the Singleton bound in terms of code sizes. However, in terms of code rate *R* and relative distance δ our bound converges to the asymptotical Singleton bound (10), thus showing no improvement.

We believe that the future of this approach is in providing better bounds on the code length of the MDS codes.

References

[1] W. C. Huffman, V. Pless, Fundamentals of Error Correcting Codes. Cambridge University Press, 2003.

[2] A. Tietäväinen, A. Perko, There are no unknown perfect binary codes Ann. Univ. Turku Ser., AI148 (1971), pp. 3–10

[3] A. Vardy, "What's New and Exciting in Algebraic and Combinatorial Coding Theory?," in ISIT, 2006.

[4] T. Jiang and A. Vardy, "Asymptotic Improvement of the Gilbert–Varshamov Bound on the Size of Binary Codes," IEEE Transactions on Information Theory, vol. 50, no. 8, pp. 1655-1664, Aug. 2004.

[5] A. Barg, Complexity Issues in Coding Theory.

[6] A. M. Mc Loughlin, "The Complexity of Computing the Covering Radius of a Code," IEEE Transactions on Information Theory, vol. 30, no. 6, pp. 800-804, Nov. 1984

Application of Machine Learning Techniques Towards Speech Separation

Tamara Dimitrova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
1000 Skopje, Republic of Macedonia
Email: tamaradimitrova@gmail.com

***Abstract*—There are different approaches towards the problem of separation of words from the audio signal. Some include previously obtained data or external knowledge to process the speech, and those are aided segmentation methods, and others are blind, where there is no pre-existing knowledge regarding linguistic properties. Speech separation is useful in many real-world applications though it is a challenging problem. In this paper we propose a blind method for separating the words. We filter the signal using a butterworth filter to eliminate the noise, extract features such as short-time energy, intensity, pitch, zero crossing rate from the audio and employ different machine learning techniques in order to provide for successful separation of the speech. Separation can be formulated as a classification problem, so we train a model that determines in which point in the sequence there is a segment that contains word. From theoretical studies, it has been observed that energy and magnitude for voiced segments are high, whereas ZCR rate is low for voiced signals. Therefore, these methods are proved to be effective in separation of words. Experimental results have been presented in this paper to verify the theoretical studies.**

***Keywords*—Speech Separation, Speech Recognition, Short-Time Energy, Zero Crossing Rate, Classification**

I. Introduction

One of the many unanswered questions is "Which is the mystery of speech?". Research has shown that infants are capable of distinguishing between phones of practically any language, but as they mature, their discrimination abilities endure strongly in languages they constantly hear. The infant brain gathers statistical cues of the environment and continuously, as they grow up the statistical representation starts to enable higher level structuring meaningful direction of the language[1].
In order to extract and distinguish bits of information from speech we need to distinguish and organize different parts of the signal and divide the speech signal into meaningful and coherent parts so the further processing for creating something that can be defined as information and knowledge could be obtained. The aim is to divide the continuous signal into discrete, non overlapping entities. In other words, to divide the continuous speech signal into smaller parts, where each segment has phonetic or acoustical properties that distinguishes it from neighboring segments. There are different methodological approaches to segmentation. They are divided into two groups, blind and aided, based on how much the segmentation algorithm uses previously obtained data or external knowledge to process the extracted speech. Some systems can learn and adapt statistically to signals they are being fed with, or can be taught before with varying techniques and emphasis. Blind segmentation methods are those when there is no pre-exsisting knowledge regarding linguistic properties such as orthography or phonetic annotation of the signal. Usually, real-time processing and speaker or language independence can usually be fulfilled only by systems that use no prior or external knowledge. Tracking the behavior of some chosen parameters can lead to cues for possible segment boundary discovery. Aided segmentation algorithms use some external knowledge of speech stream to segment it into corresponding segments of a desired type [2]. Implementation is done by using orthographic or phonetic transcription as a parallel input with the speech or train algorithm with such data. On the other hand, as any other pattern recognition problem, we can use top down or bottom up approach to proceed with processing. In the top down pathway towards the problem of speech separation, the aim is to group features, where each group would represent well-defined sections of the original speech signals into semantically (words, sentences) coherent groups. On a very abstract and conceptual level, human data processing is similar to speech data clustering in speech recognition systems, where finding the best of several good clusters for each vector is actually the process of matching new input with something already known. The top down information can consist of two different types: to use the prior knowledge, such as grammar or lexicon of the language and the semantic framework that our working memory deals with, so while we are hearing someone talking our brain continuously analyzes the semantic content of the speech and compares the signal to existing models of grammar and performs word

matching exploiting or lexical memory; or to consider lower level unconscious neural feedback systems, which are seen as ”built-in systems” and are result of normal development of the individual, most of them being partially or fully operational in infancy. The bottom up processing of speech signals is split into two parallel processes: blind segmentation of speech into phone sized units, and feature extraction process to provide descriptive data for each segment. The reason for this division is because the segmentation algorithm needs to find robust and salient cues that help in estimation of phone boundaries, while feature extraction has to build congruent and normalized parametric representations of the segments that describe each phone. The general idea is to form clusters of the parts that should be combined. If the distance is smaller than merging threshold, it will be part of the previous cluster, otherwise it will form a new one, as discussed in [3] [4]. In this paper a method is proposed which contains all the phrases of a pattern recognition system. At first, signal filtering is performed using different types of audio filters and it is determined which is the best one for the problem. After which features are extracted from the audio that will help in the next phase in which a classification task will be presented, in which we will determine if in some point of time a cut should be made and begin new word, or continue building the previous one.

II. Proposed Method

The method that is proposed in this paper consists of several phases that are part of the pattern recognition process. At first, preprocessing is done, by filtering the signal; following step is the feature extraction and selection, in which information is extracted from the audio so at the end classification can be performed. The classification is performed in a way so that the signal is passed through and depending on the algorithm in a certain moment decides if a cut should be made or not.

A. Filtering

When working with audio, important part of processing is the filtering step. One of the major problems in working with signals is the noise that distorts(corrupts) the signal and can make problems in processing. That is why, filters are applied. Filtering is a process that removes some unwanted components or features of the signal, or reduces the background noise. There are different types of filters, which are initially separated into two groups: digital and analog. Digital filters are used to eliminate the noise and extract the signal of interest from the other signals. Analog filters are used for the same cause, just in different conditions. There are different types of filters, such as butterworth, chebyshev, elliptic filter approximation, and many others. Each filter is characterized by few parameters:

- Cut-off frequency is the frequency that separates the pass band and the stop band frequency. In other words this parameter is the frequency at which the response changes from pass to stop band.
- Stop band

For this task, low pass butterworth filter was used, with cut-off frequency of 0.2 and of fourth order.

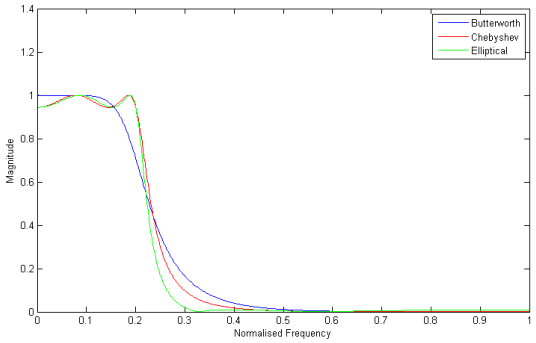


Fig. 1: Visual representation of pass and stop band on the different filter types.

B. Feature Extraction

A conversion of the speech waveform to some parametric representation is needed, which can be further used for analysis and processing. We need feature extraction that would reduce the variability in the speech which comes from many sources such as: different speakers, speaking rates, content, acoustic conditions so we could eliminate various source of information such as voiced/unvoiced sound. In this way, we eliminate the effect of periodicity of speech/pitch and amplitude. Short-time energy is a parameter that is extracted from the audio and distinguishes between voiced and unvoiced speech segments, since the unvoiced segments have significantly lower short-time energy. It is a way to represent and reflect the amplitude variations, which are much lower of the unvoiced segments than the voiced. Short-time energy can be defined as

$$E_n = \sum_{m=-inf}^{m=+inf} [x(m)w(n-m)]^2. \qquad (1)$$

The choice of the window determines the nature of the short-time energy representation. In this model, Hamming window is used. The window gives much

greater attenuation outside the bandpass than the comparable rectangular window. The attenuation is independent of the window duration. Increasing the length, N, decreases the bandwidth. If N is too small, E_n will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, E_n will change very slowly and thus will not adequately reflect the changing properties of the speech signal. For this task, hamming window was used, with window length equal to 300 samples, so it is not too detailed and everything can be seen.

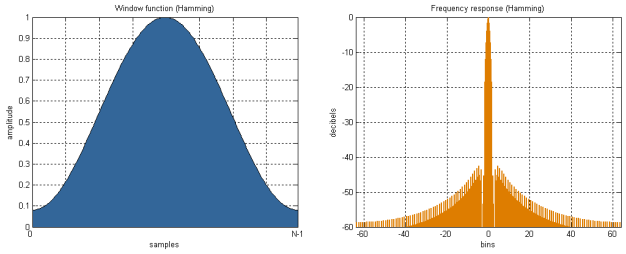


Fig. 2: Hamming window and its Fourier transform.

In Fig. 3 are shown sample of a signal and the corresponding short time energy.

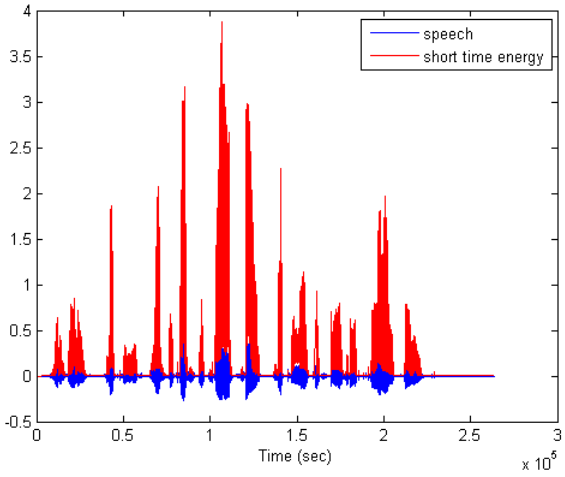


Fig. 3: Signal and the corresponding short-time energy.

For better performances and results, normalization of the short-time energy was performed. Rescaling the range was made, using Feature scaling, to translate the features to range [0, 1], by calculating

$$x' = \frac{x - min(x)}{max(x) - min(x)}, \tag{2}$$

where x is the original value and x' is the normalized value.

Also, Hilbert transform and envelope was applied to the short time energy. The Hilbert transform can be considered to be a filter which simply shifts phases of all frequency components of it's input by $-\pi/2$ radians. More specifically, this envelope involves creating the analytic signal of the input using the Hilbert transform. An analytic signal is a complex signal, where the real part is the original signal and the imaginary part is the Hilbert transform of the original signal. Mathematically, the envelope $e(t)$ of a signal $x(t)$ is defined as the magnitude of the analytic signal as shown by the following equation:

$$e(t) = \sqrt{(x(t)^2 + \hat{x}(t)^2)} \tag{3}$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$.

The envelope of the signal can be found by taking the absolute value of the analytic signal. The envelope is a low frequency signal compared to the original signal. To reduce its sampling frequency, to eliminate ringing and to smooth the envelope, this signal is downsampled and the result is passed through a lowpass filter.
In Figure 4. we can see how the energy looks like after Hilbert envelope and normalization are applied on the already calculated short time energy. A change can be easily seen from the normalized and non normalized short time energy.

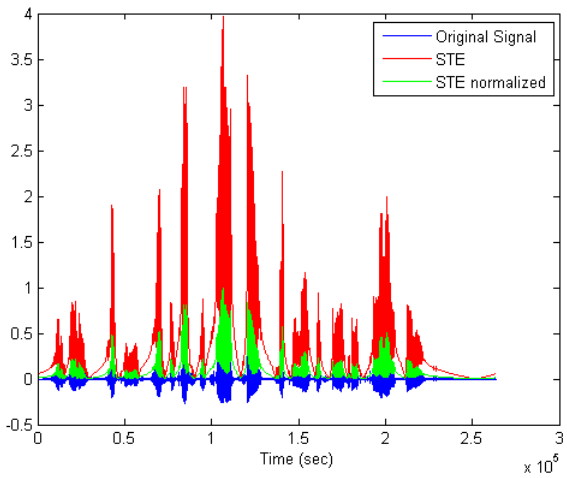


Fig. 4: Short-time energy with applied normalization and Hilbert envelope.

C. Separation

Having the short-time energy extracted from the signal, an algorithm was constructed to separate the signal into segments. Separation task is represented as a classification task: go through the signal frame by frame and determine if at certain point a cut should be

made or not; if the current frame should be included in the already started segment, or to make a cut and start new segment. This decision is determined by having a function for adaptive setting the low and high threshold, that are 10th percentile for the low and the 80th percentile for the high threshold, from the energy from the sample that is reviewed. In future, it would be more efficient if the thresholds are precomputed, so the processing would go faster.

In Figures 5 and 6, results are given from the separation from the algorithm without and with preprocessing of the extracted feature from the audio. It can be concluded that after applied Hilbert envelope, we get better results, since the segments are separated in a better way.

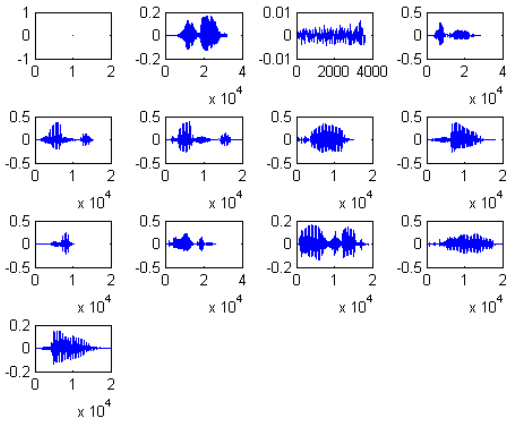


Fig. 5: Separated segments, with none preprocessing on the features applied.

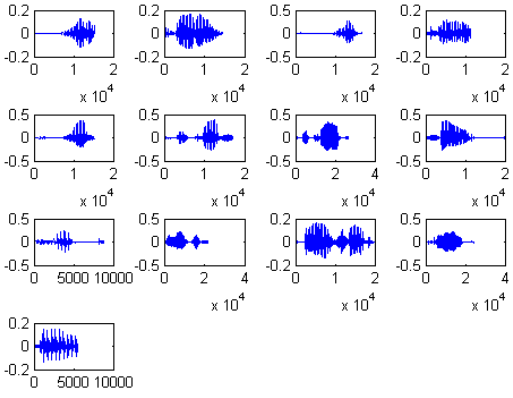


Fig. 6: Separated segments after Hilbert envelope and normalization are applied.

III. Experimental Results and Discussion

The Audio/Visual Emotion Challenge and Workshop (AVEC 2011) was the first competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and audiovisual emotion analysis, with all participants competing under strictly the same conditions. The dataset that AVEC provides contains audio data(wav), video data(avi), aligned transcripts with word timing information, labels for the train and development partitions, video and audio features. Having the audio data and aligned transcripts, we see an opportunity to test our algorithm on this dataset. The results can be seen in the table, and they vary from sample to sample, depending on the noise and audio quality the audio has. Some samples reach to 70-80 percent of correctly segmented samples. We also tested the algorithm with recorded samples in different environments and got relatively good separation of the words. Since this is relatively new method, the previous research gives results of approximately 70 percent successful performance on the border detection, but this does not include the successful word detection in a signal.

TABLE I: Experimental results from the word separation on AVEC dataset.

Threshold (ms)	200	225	250	275
Accuracy (%)	24.14	29.18	34.75	40.4

IV. Conclusions and Future Work

This paper gives experimental verification of the short-time analysis techniques for classification of a signal in order to separate it in different segments which have phonetic or acoustical properties that distinguishes them from the neighboring segments. A simple techniques were used for calculating short time energy in order to perform the classification. Note could be that a better choice of window and window length gives appropriate results, so in these problems the window length should be chosen carefully.

As future work, an improvement is planned by introducing another parameters extracted from speech, such as zero crossing rate, pitch, formant frequencies and introducing boosting techniques in order to get better results. Another step in the future work would be using the transcripts from the AVEC database and try to learn the threshold for separation in advance, so the segmentation could be performed in a better way.

V. Acknowledgments

I would like to thank Alessandro Rozza, who suggested and motivated me to work in this field, for his help, support and extensive feedback.

REFERENCES

[1] K. Gold and B. Scassellati, "Audio speech segmentation without language-specific knowledge," in *Cognitive science society*. Citeseer, 2006, pp. 1370–1375.

[2] M. A. Zissman *et al.*, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, p. 31, 1996.

[3] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[4] D. Wang, L. Lu, and H.-J. Zhang, "Speech segmentation without speech recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–468.

Semantic Web Access Control Aspects

Slobodanka Trajkova
Faculty of Computer Science and Engineering
Skopje, Macedonia
slobodanka.trajkova@gmail.com

Riste Stojanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
riste.stojanov@finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*—The data exchange among the application is popular via the various APIs provided by the vendors. The main reason for development of proprietary APIs instead of using the available semantic web standards is the lack of access control mechanism in this area. This paper surveys the state of the art semantic web security research and analyses the key aspects that should be covered by the access control model that is used for protection. The aspects are backed up by a use-case scenario in the university domain, where the user's data is stored in multiple cooperating systems, and there is a need to control the access and interaction of that data by the users. In this paper we show that the semantic web standards are appropriate for providing interface for interaction with this data if the access control system covers the defined aspects.**

Keywords*—*Semantic Web, Access Control, Semantic Security

I. Introduction

There is emerging trend of the application vendors to expose their data and functionalities via proprietary application programming interfaces (APIs)[1]. This is because their data is more valuable if it is shared and interconnected with other application, which will bring additional features to the and users, and will make them more satisfied. However, if one wants to connect with multiple applications, a separate connector must be implemented for all of them, due to the lack of standardization. This means, that the developers must learn and understand all APIs, and than integrate with them, which slows them significantly.

The semantic web[1] defines standards for resource description is machine readable formats, such as RDF[2], RDFS[3] and OWL[4], as well as standards for interaction with this data, such as Linked Data Platform (LDP)[5] and SPARQL[6]. The main reason why application vendors don't use the semantic web standards is the lack of security mechanisms and standards in its definition. Even though there is a trust, cripto and proof layers in the semantic web stack (Figure 1), the vendors' main concerns are the mechanisms for authentication and authorization. Even though the semantic web integrates, automates and reuses data from various sources, it is not applicable in the business environments the access to the semantic resources can't be controlled.

[1]https://developers.facebook.com/docs/graph-api, https://dev.twitter.com/rest/public, http://www.omdbapi.com/
[2]http://www.w3.org/TR/rdf-primer/
[3]http://www.w3.org/TR/rdf-schema/
[4]http://www.w3.org/TR/owl-guide/
[5]http://www.w3.org/TR/ldp/
[6]http://www.w3.org/TR/sparql11-query/

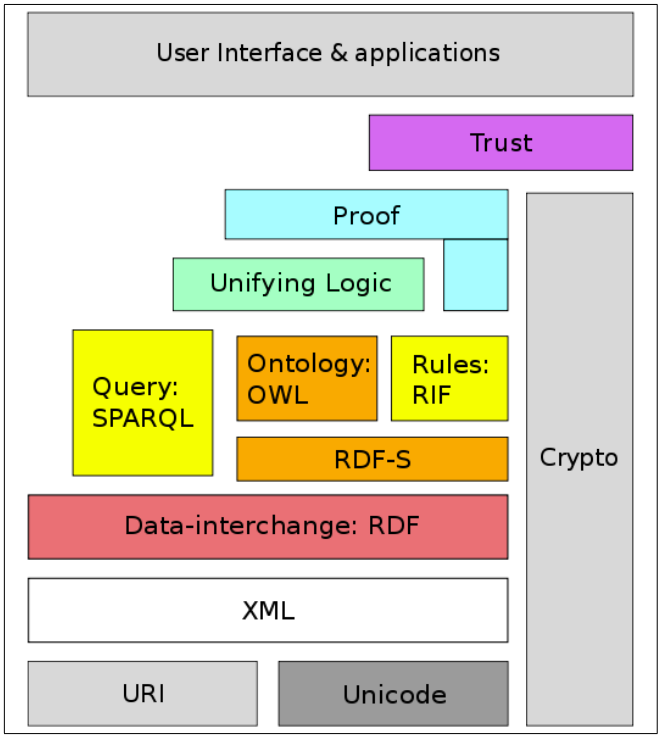


Fig. 1. *The Semantic Web layers.* [7]

The semantic web is intended to provide a structured, machine readable, representation of the human knowledge. Thus, protecting this data should correspond to the way people protect their knowledge. We, as a human, are in control of our knowledge, and one can decide which information will share with the others to the tiniest bit, and can also decide which information will accept form the others. The expectation from the semantic security system are nothing less from this, since they should protect the machine readable representation of the human knowledge.

There are multiple articles that threat the security of the semantic web[2][3][4][5][6][7], but all of them are mainly concerned about the access control models and the performances of their systems. None of them discusses the aspects that should be covered when protecting the semantic data. This will be the main goal of this paper: to define the aspects that should be covered by the access control systems for the semantic web.

[7]Source: http://commons.wikimedia.org/wiki/File:SW$_layercake_2$006.svg

II. Related work

Existing research initiatives in the area of semantic web security are mainly differentiated based on the used access control model, which defines data operation protection via the policy for given authorization model. The access control model is used for policy selection based on the user authorization privileges.

The authorization models define the connection between the authenticated user and the stored resources. There are two widely used authorization models: Role Based Access Control (RBAC)[8] and Attribute Based Access Control (ABAC)[9]. The systems using the RBAC model[4][5][6][10] uses user grouping with explicit roles, and the data privileges are assigned to the roles, while the systems[3][7][2] are using the Attribute Based Access Control authorization and they select the permissions based on the authenticated user attributes. The RBAC model provides rigid user classification and requires adding mechanisms for handling the static and dynamic separation of duty[4], and doesn't provide flexibility in the protection of the data. The ABAC authorization model combines the user attributes and the requested data in the process of permission selection, and thus provides flexibility and overcomes the need of static or dynamic separation of duty.

The policies are used as a mechanism for definition of the user permissions for operations with the data. The access control lists[**?**] are the most common format for policy definition. In the semantic security field, beside the definition of custom policy formats[3][5], the policies are often defined in RDF[6][4] or in combination of RDF and SPARQL[2][7].

The definition format of the policy directly influences the permission granularity and the definition flexibility. The permission (or policy) granularity defines the smallest amount of data that can be protected by the model. Fine grained policies allow better and more accurate data protection, but there are cases when they reduce the definition flexibility because of the high configuration effort. In the semantic security field for data protection granularity units are used the RDF documents[10][7][6], triples[3][5], named graphs[2] and resources[4]. The RDF documents and the named graphs approaches have coarse grained protection with less configuration effort, while resource and triple protection provide fine grained protection, and it's definition cost depends on the policy language.

The permission definition flexibility directly depends on the policy format. The systems with custom format, the flexibility depends on the tools available for policy modeling, but they require additional training for their using and understanding. In the cases of RDF policies, the possibility for finner granularity reduces the flexibility due to the large number of entries for protection of all the possible resources. The SPARQL query language overcomes this gap, and provide mechanism for selection of large number of resources with a query, which significantly lowers the configuration effort. But the downside of SPARQL is that the language doesn't provide mechanism for declaring permission features, such as if the query grants or denies access of the selected data. Thus, the combination of the RDF and SPARQL provides selection of many resources with less configuration i.e. one query, and possibility for adding additional permission meta-data.

The access control model also defines the way of enforcing policies for the operations over the data. Almost all current systems support only protection of the read operation, while the write, modify and delete operations are less researched. The work that covers the protection of write operations is mainly for coarse granularity RDF documents[6][10][7]. In [2] is presented system that covers the CRUD operations. However, more of these systems[10][2][4] use restrictive permission enforcement, where the operation is permitted only if the user has permission for interaction with all the data in the result. This means that if the result contains a fraction that is not accessible, the user won't be permitted for the whole batch of data, even though he/she has permissions for a part of it. The selective permission enforcement[3][5] works in a way that the operations are allowed to interaction only with the granted data for the user, while they ignore the the denied data fraction. In practice, this approach separates a subset of the data permitted for the user, and all the operations are executed against this portion.

III. Semantic Web Access Control Aspects

The process of evaluation of the semantic web security systems should focus on the aspects for protection of the semantic data. Most of the current systems are not defining and discussing the aspects for protection and their coverage, but instead they are concerned with the performances of their systems. Further more, the evaluation procedure is not standardized and every system is evaluated against separate dataset.

In this paper, we are trying to identify the key aspects for protecting semantic data, and we are defining a set of use-cases that can be used for testing the potential of the semantic security systems.

The idea behind the semantic web is to provide machine readable representation of the human knowledge. As such, it has to support all the human ways for protecting the data, such as contextual decision and selection of the granted data based on the user that is trying to interact with it. We, as a human, are able to identify the agent that is requesting interaction with our data. This is the process of authentication, and this paper doesn't cover this topic. We will assume that the systems provide a separate services for user authorization, such as Central Authentication Service (CAS)[11], or the WebID[12] authentication mechanism for the semantic web.

The second aspect is the authorization model. The authorization model constrains the flexibility for policy selection. The RBAC model[8] is more rigid way of authorization, and it needs to be extended in order to support contextual dependency and dynamic rules based on separate users. It also requires mechanisms for static and dynamic separation of duties[13]. The ABAC model[9] provides flexible definition of the permission selection since the users are not categorized statically, and the policy activation can also depend on the protected data and the user context. Also, since all semantic resources are represented through a set of relations toward the property values or other resources, the ABAC model is more natural for permission selection of the semantic data.

The semantic web data is represented with a resources connected among each other with their properties, forming

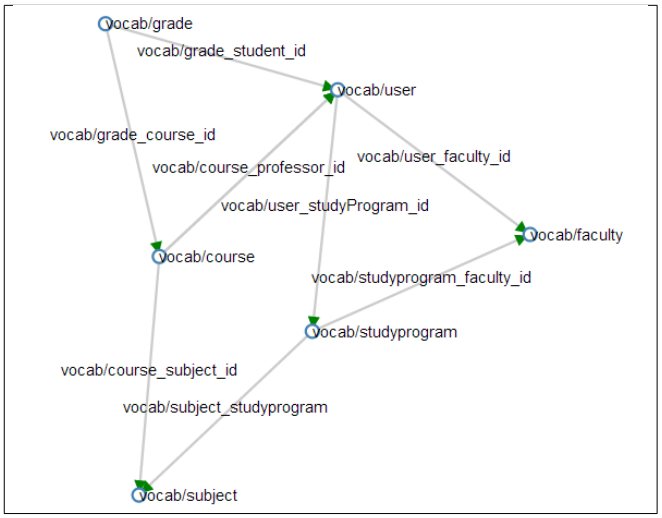


Fig. 2. *University domain ontology.*

triples composed of subject, predicate and object. The subject from a triple defines the resource that is being described, the predicates represent the properties (or attributes) of the subject and the object represent the value of the given property for the resource in the subject. Additionally, the resources can be organized in classes through the rdf:type property. Next, the triples can be grouped in a logical units called graphs, and the graphs are persistent in the datasets.

In the rest of the paper, we will use an examples modeled by an ontology from the university domain, where the agents can be students or professors. The ontology is shown in Figure 2.

Having this structure, and with the premise that the policies should provide protection to the thinnest bit of data according to the human needs, the semantic security systems should provide fine grained policies. However, the protection of the smaller units of data, can require significant amount of configuration effort in order to protect the data, if each peace of information is protected separately. However, we as a human often classify the information in groups and than decide how to share it with the others. In our ontology, this can be represented by the fact that the professor knows his subjects and courses, so this data is categorized according to some rules. If one should provide separate configuration about each grade professor gave to a student separately, this will overhead will overwhelm the configuration and the system will loose its purpose. This why, there is a need of providing flexible way of defining and selecting these groups of categorized data. SPARQL is a query language that provides all these features, and thus it is used in policies for easier definition of the protected data portion[2][5].

Given the semantic data structure and the human needs for data protection, a mapping among them should be defined in order to provide secure data access. Here we will enumerate the structural needs and we will describe their connection to the real life examples.

- If one wants to provide selection of resources having some property with a certain value, such that the professor will be able to interact with the resources that have property *vocab/studyprogram_faculty_id* with value *vocab/studyprogram/5*, meaning that he will be able to interact with all the user resources form that study program.
  - Resource protection based on their property value.
  - $?x < ex : Prop > value$
- The oposite direction should also be supported, because the one should be able to select all resources that are value to a property of another resource, such as interaction with all subjects from a study program.
  - $< res >< ex : Prop >?x$
- In the cases when there is a need to constrain the values of some properties of the resource, such as do not display the phone number of anybody, there is a need to constrain that property of the resource
  - $FILTER(?p! =< ex : someProp >)$
- The contextual protection is one of the key aspects when the human protect the data. That is why there must be a way of saying that, for example, the professor can interact only with his subjects
  - $< me >< ex : userProp >?x$
  - $?x < ex : userProp >< me >$
- At the end, the user data protection needs are never composed with one rule, but instead they are composition of multiple simple rules. Thus, the policy should support combining of multiple selection patterns.
  - $< me >< ex : userProp >?x.$
    $?x < ex : otherProp >< res >$

Each of the items in the previous list contains a SPARQL like expression that provides a selection or filtering pattern. The next question is how far can we protect the data with SPARQL i.e. is it enough for defining policy that can protect all the aspects of the data. The answer for this question is no, since the purpose of SPARQL is to provide a query language, not resource definition language. For this reason the SPARQL should be used for data selection, but information such as to whom this policy applies, or is it granting or denying access should be added additionally. This is why most of the approaches use the RDF as a main format for policy definition, where the policy contains a property with SPARQL query as a value, for defining the portion of the protected data[2][7]. Additionally, the SPARQL is not enough to provide contextual dependence, since it doesn't provide mechanism to define the current user in the policy query, which is a variable. One way of solving this problem is pre-processing the query by the policy enforcing system and replacing some predefined expression with the current user, and the other is do extend the data with predefined ontology that contains the user information and to incorporate selection of that information in the query itself.

Further more, the SPARQL language is not suitable for definition of information such as weather or not the query is granting or denying access for the selected data. Also it doesn't provide information about the data source for which the permission holds. This is why it is often used in combination with RDF, the format used to provide machine readable form of the human knowledge. The RDF provides flexibility for

extensible feature definition, and can store all the information of the policy.

The agent interacts with the data through a series of actions, and the usual approach of implementing security is in the business layer or in the web layer[14]. However, no mater how complex those actions are, they can be decomposed into a four basic operations: Create, Read, Update and Delete. Thus, each policy should define permissions for each of these operations.

IV. Discussion

In order to support the previously defined aspects, we defined synthetic dataset for the university domain, where we have students an professors as agents. The dataset is not defined in order to test the performances of the system, but to test the flexibility of the policy model definition and the possibilities for data protection.

Even though we favor the ABAC authorization model, we defined a two groups of users: Students and Professors, and defined a set of use-cases that cover all the previously discussed aspects.

Due to the space limitation, we will only describe few of these use-cases, and we will connect them with the previously defined aspects.

One of the requirements is that everyone can se the subjects and the study programs from their faculty. This case, emphasizes both, the data protection by its property value, as well as the contextual dependency with the current agent. Similar case is that everyone can see and modify their own data, which additionally shows the need for definition operation in the policy. However, here we don't consider the data ownership strategies, and we assume that the data ownership can be defined appropriately.

Next, we consider the cases where the students are able to see everything about their professors, but the phone numbers. With this, we cover the aspect for protection of property values for some data entities.

Some of the users, can be students and professors at the same time, which since in our professor generalization include the PHD teaching assistants. With this cases, we cover the static and dynamic separation of duties in the process of data protection.

Also, there are rules that define that the professor can grade only the students enrolled in his subject, which is a requirement with higher complexity and requires combination.

The key goal of these use-cases is to provide an evaluation mechanism for the granularity and flexibility of the access control models for protection of semantic data.

V. Conclusions

The decisions about the policy language, the authorization model, accepted operations and other similar features significantly influence the possibilities and constrain the flexibility of the data protection.

In this paper we defined the key aspects for protection of semantic data, so that one can optimize the protection of the data with flexible permissions. The flexibility of the access control model took the central part in our analysis, since the semantic web should provide protection of the human knowledge. We described the ABAC authorization model provides more flexible permission assignment based on the user interaction with the data, while the SPARQL language is most suitable for selection of the protected data. Additionally, the SPARQL language should be extended with additional metadata information, in order to provide better protection and policy definition flexibility.

References

[1] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[2] L. Costabello, S. Villata, O. R. Rocha, and F. Gandon, "Access control for http operations on linked data," in *The Semantic Web: Semantics and Big Data*. Springer, 2013, pp. 185–199.

[3] F. Abel, J. L. De Coi, N. Henze, A. W. Koesling, D. Krause, and D. Olmedilla, *Enabling advanced and context-dependent access control in RDF stores*. Springer, 2007.

[4] T. Finin, A. Joshi, L. Kagal, J. Niu, R. Sandhu, W. Winsborough, and B. Thuraisingham, "R owl bac: representing role based access control in owl," in *Proceedings of the 13th ACM symposium on Access control models and technologies*. ACM, 2008, pp. 73–82.

[5] G. Flouris, I. Fundulaki, M. Michou, and G. Antoniou, "Controlling access to rdf graphs," in *Future Internet-FIS 2010*. Springer, 2010, pp. 107–117.

[6] J. Hollenbach, J. Presbrey, and T. Berners-Lee, "Using rdf metadata to enable access control on the social semantic web," in *Proceedings of the workshop on collaborative construction, management and linking of structured knowledge (CK2009)*, vol. 514, 2009.

[7] O. Sacco, A. Passant, and S. Decker, "An access control framework for the web of data," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*. IEEE, 2011, pp. 456–463.

[8] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," *Computer*, vol. 29, no. 2, pp. 38–47, 1996.

[9] T. Priebe, E. B. Fernandez, J. I. Mehlau, and G. Pernul, "A pattern system for access control," in *Research Directions in Data and Applications Security XVIII*. Springer, 2004, pp. 235–249.

[10] H. Muhleisen, M. Kost, and J.-C. Freytag, "Swrl-based access policies for linked data," *Procs of SPOT*, p. 80, 2010.

[11] B. C. Neuman and T. Ts'o, "Kerberos: An authentication service for computer networks," *Communications Magazine, IEEE*, vol. 32, no. 9, pp. 33–38, 1994.

[12] H. Story, B. Harbulot, I. Jacobi, and M. Jones, "Foaf+ ssl: Restful authentication for the social web," in *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2009)*, 2009.

[13] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed nist standard for role-based access control," *ACM Transactions on Information and System Security (TISSEC)*, vol. 4, no. 3, pp. 224–274, 2001.

[14] C. Scarioni, *Pro Spring Security*. Apress, 2013.

Exploitation and distribution of setuid and setgid binaries on Linux systems

Eva Tanaskoska
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Rugjer Boshkovikj 16, Skopje, Macedonia
evamolotow@zeroscience.mk

***Abstract*—In an era of Internet freedom, lack of control and supervision, every system is exposed to various attackers and malicious users which, given the right circumstances, are able to cause colossal damage. A single security vulnerability can be the reason for a business' downfall, therefore significant attention needs to be paid to said systems' security to avoid such issues. Unix-like filesystems define certain access rights flags, named setuid and setgid, which allow users to execute files with the permissions of the file's owner or group. This can be exploited to gain unprivileged access using buffer overflow attacks. I performed tests by running a script to collect the files in Ubuntu, Debian, Slackware, Fedora and CentOS to find the files with the setuid and setgid bits set. My aim is to determine which distribution is the most secure one and whether Slackware, considering it's known for its secure design and characteristics, will prove its reputation. The results show that Debian and CentOS have the least amount of exploitable binaries, while Slackware and Fedora have the most.**

Keywords—setuid, setgid, Linux, permissions, security, buffer overflow.

I. INTRODUCTION

The vast majority of servers in the world utilize Linux operating systems. Improper system administration can open up a whole lot of security concerns which can impact a business' reputation and result with extensive damage to company and customer data. With the increase of Internet use, the need for enhanced fortification and improved security increases proportionally, leaving system administrators with the daunting task of defending the company's systems and valuable information. Unix-like systems use a simple system for managing permissions, as well as access control lists. However, my research is currently focused on Linux systems. On a multiuser system, every file and directory permissions must be set appropriately to the sensitivity of the file or directory in question. Failure to do so can result with unprivileged access to sensitive information and exploitation.

The read, write and execute permissions define the access rights for the owner, group and others with the help of the chmod command by specifying the desired permission settings. Permissions are represented by a set of 3 bits for each of the corresponding groups, each bit representing the presence or absence of a permission. The three bits represent a digit, ranging from 0 to 7 which defines the permissions for a certain entity. Read permission has a value of 4, write has a value of 2 and execute translates to 1. Adding the given permissions defines the permission set. It is also possible to change a file's owner and group using the chown and chgrp commands.

Beside the classic read, write and execute permissions on files and directories, the special permission flags (setuid, setgid and sticky bit) found in Linux filesystems can pose a security treat if they aren't properly used. When set, the setuid bit allows an user to execute a file with the permissions of the owner of the file, regardless of who executes it.

Similarly, the setgid bit causes the file to be run as a member of the group that owns it. If the setuid or setgid bit is set on a file owned by root or a privileged user, the special permission can constitute a security risk. The problem was addressed [1] a long time ago, but never sanitized. If a root-owned executable is compromised using a buffer overflow attack, the attacker will gain superuser privileges. Fortunately, buffer overflows are a rather complex attack method and require extensive knowledge of CPU/memory architecture, assembly language, reverse engineering and programming. It requires years for a security researcher or attacker to learn and practice before they are able to find and execute a buffer overflow in software. Only the most determined attackers resort to searching for buffer overflows in targeted attacks. Newbies and script kiddies are in most cases unable to locate such security flaws by themselves and instead run automated software for vulnerability scanning which isn't a reliable option and locates only publicly disclosed vulnerabilities that in most cases have already been patched. If such a vulnerability is found to be exploited, it is needed to update the compromised system to avoid further exploitation. Outdated software is one of the leading causes for security breaches nowadays.

Another security concern is the fact that all users can set special permissions for their own files. In reality, the vast majority of the users don't pay much attention to security, especially when it comes to public servers everybody uses or servers they do not own in general. In environments where security is of critical importance, the users should be educated in regards to security.

Most Linux distributions have a different structure and filesystem, implying that a portion of the files found on the systems differ, while some are mutual. Due to their differences, the security measures in the systems vary as well, indicating that every distribution is rather unique in respect to security. Even though Slackware's default options enhance its security, it might not necessarily be as secure in terms of suid/guid binaries.

II. Related work

The lack of documentation on the workings of the setuid and setgid mechanisms result with insufficient research done on the subject as well. Hao Chen, David Wagner and Drew Dean [2] researched the workings of setuid and setgid by analyzing the system calls in Linux, Solaris and FreeBSD, as well as building a formal model of user IDs as a finite state automaton. Matt Bishop [3] wrote a guide in 1987 on how to write setuid programs and the potential vulnerabilities which may arise, however, Unix systems have changed a great deal since then.
Most of the information used in this paper originates from Internet sources, whitepaper, as well as the Unix `man` pages, which provide some information on the subject.

III. Permissions and exploitation

In Unix everything is a file, including directories, which are actually files containing a list of other files. Other special types of files, such as pipes, are used to transfer data between processes. The extensive use of files implies that filesystem security should be one of the priorities for a system administrator. Unix access control is based on user IDs associated with processes spawned by an executed file. The process permissions and available resources are determined and inherited from the owner's user. A process which is owned by the root user with ID 0 has access to every system resource. In order to create a safe environment where no process will be able to access resources it isn't supposed to, it is necessary to have certain knowledge about the way Unix permissions work. The basic way of limiting a process' available resources is by setting the basic permissions of a file. To view a file's permissions, the user has to run a long listing (`ls -l`) of the directory in which the file is located.

A. Basic permissions

User and group accounts represent someone or something able to use the files in the filesystem. These accounts can be used by both people and processes, determining the capabilities of the user. Each file has two owners, [4] one of them being a user and the other one a group. The permission bits regulate the access for the file's owner, group and the others which do not fit in any of the other two categories. Those three sets consist of three possible basic permissions on a file: read, write and execute. The presence or absence of a certain bit in the set denotes whether the user(s) in the category are able to manipulate the file, as can be seen in Figure 1. An exception of this rule is the root user, who can do everything, however that doesn't mean that everything can be manipulated by the root user without a prompt or a warning. For example, in order to execute a script which does not have the execute bit set, even if they were the owner, a root user must supply the execute permission with the `chmod` command.
A read permission allows a user to read the file, the write permission permits writing and the execute permission allows execution. Directory permissions work slightly different. Read permissions of a directory means the user can browse the directory, the write permission lets them modify it by adding/removing files, whereas the execute permission allows a user to change working directory to this directory, and read and modify the files contained within (if the files' permissions

```
-rwsr-xr-x 1 root    root          27256 2010-01-29 00:02 /bin/fusermount
-rwsr-xr-x 1 root    root          78096 2009-10-23 09:58 /bin/mount
-rwsr-xr-x 1 root    root          35600 2009-05-12 03:13 /bin/ping
-rwsr-xr-x 1 root    root          31368 2009-05-12 03:13 /bin/ping6
-rwsr-xr-x 1 root    root          36864 2009-07-31 19:29 /bin/su
-rwsr-xr-x 1 root    root          56616 2009-10-23 09:58 /bin/umount
```

Fig. 1. Long listing on a Linux system.

allow that). Without an execute permission, a user cannot browse nor use a directory, regardless of the files' permissions. The `umask` command specifies the default permissions for new files with the help of a mask which defines which permissions will be substracted from newly created files. The default permissions for new files are 666 octal, or read and write permissions for user, group and others, while the default umask on most systems is set to 0022, resulting in newly created files with a 0644 mask, meaning the user has read and write permissions, and the group and others have read permissions only.

B. Special permissions

Where necessary, the special file permissions allow files to be executed with the privileges of the owner or the group of the file. There may be some cases where a file needs to be executed with elevated privileges and the invoking user's privileges are insufficient. A classic example for this is the `ping` utility which needs access to low level system interfaces, such as `socket`, `PF_INET`, `SOCK_RAW`, etc. When the `ping` command is used, it inherits root permissions in order to access those interfaces and without them, wouldn't be able to function properly. Some server applications require root privileges to set up their environment. After performing the privileged tasks, the server process drops the elevated privileges to evade the possibility of an attacker gaining root access if the process was to be compromised. The *uid-setting* system calls grant this functionality. This flexibility is a double-edged sword, since an improper use of the mechanism will result in severe security vulnerabilities, but when used properly they can enhance the security of a system [5].
There are a few types of user IDs: real user ID (ruid), which is the user ID of the user that started the process, an effective user ID (euid) - the user ID of the user with whom privileges the process is executed, filesystem user ID (fsuid) user for access control to the file system and a saved user ID (suid) used when a program running with elevated privileges needs to perform unprivileged tasks. Beside the basic permissions set with the `chmod` command, an additional bit can be appended to the beginning of the octal string, denoting whether the setuid or setgid bit is set or not. The digit 2 implies that the setgid is set and the process executes with the group's permissions, while the digit 4 means the setuid bit is set and the process utilizes the privileges of the file's owner. A permission digit 6 means both permissions are active. An example of this scenario is if the file is owned by the root user, and the group wheel, upon execution, the file will run as `root:wheel` regardless of who executes it.
When set on a directory, these permissions have a different purpose. Should a directory have its setgid bit set, all the newly created files will inherit the group ID of the directory itself, instead of the group of the user creating a file. Files created

before the bit was appended are not affected. The setuid bit has no effect on directories on Linux systems.
However, setting setuid and setgid permissions manually is rarely needed and most permissions are set during the installation process while performing the installation as a root user. This permission bit can almost always pose a minor security threat, even when the process is not compromised. An utility which is not supposed to elevate privileges, but does, might disclose sensitive information. If applications which have the ability to modify the filesystem contents have the setuid/setgid bit set, they might give an unprivileged user the ability to alter the files after successful exploitation of the said process. The worst case is having programs which are able to execute shell commands elevate their privileges. Such an example is the `Vim` text editor. If an unprivileged user were to run commands from the command line within `Vim`, the commands will execute with root privileges. To check a file's permissions, a user has to run the `ls -l` command and search for `s` or `S` in the execute permission bit for the owner and group categories. A lowercase `s` denotes the program [6] has both setuid (or setgid) and execute permissions, while an uppercase `S` means that the execute bit is absent. A severe security risk pose privileged shell scripts, because it would be trivial for an attacker to modify the content of the shell script and execute arbitrary commands. Most modern operating systems ignore the setuid/setgid bit set on shell scripts.

C. Exploiting SUID binaries

Many people underestimate the risk of buffer overflow exploits [7], with the excuse that once a vulnerability is discovered, a simple patch will secure the system. However, the real life examples aren't as perfect. Unfortunately, many vulnerabilities take years to sanitize and there is also a great risk of 0day vulnerabilities which haven't been reported publicly and circle around through the black markets, sold by cyber criminals in exchange for money, cryptocurrencies, 0days and other goods. Plenty of vulnerabilities remain undiscovered. Vulnerability assessments aren't reliable enough to determine these 0days, fooling the tester or the company being tested, that it is secure enough and no additional security measures are needed. Signature-based intrusion detection systems cannot detect 0day attacks because they only have signatures for known attacks and even a slight variation in an already known attack can fool a signature-based intrusion detection system. Anomaly-based intrusion detection systems may be able to detect 0days, given that the person examining the potential attack is skillful enough.
Buffer overflow attacks are the main cause of suid-based problems. A precisely-executed buffer overflow attack on a privileged suid binary can provide an attacker with root access. There are plenty of different targets for buffer overflow attacks. The operating systems themselves contain a number of potentially exploitable vulnerabilities, as well as the protocols, services and applications written in unsafe languages, such as C and C++. The most popular attack method are stack-based buffer overflows, which exploit the stack structure. Every time a function is called, a frame is pushed onto the stack which holds local variables, return addresses and function arguments.
Buffer overflow occurs when a variable gets provided more user-supplied data than it can hold. The excess data floods into the neighbour memory, overwriting valid data. If the user-supplied malicious data is set to properly overwrite the return address, forcing it to point to attacker-controlled memory, the malicious code gets executed and in most cases, results with privilege escalation. There are other types of buffer overflows, such as heap-based buffer overflows. Heap-based buffer overflows occur in the heap memory and are significantly harder to detect and exploit. The payload which is inserted into the buffer is called shellcode and often contains instructions for spawning a root shell. Shellcode is often writen in Assembly, but the format needs to correspond to the language in which the exploit is written because most languages have a different format for representing the instructions. The instructions are written in hexadecimal opcodes. Due to the low-level nature of shellcode, it is architecture-specific. Shellcode payloads can be generated [8] with the Metasploit Framework for a variety of languages. Obfuscation techniques should be used in order to avoid detection by network intrusion detection systems of the shellcode. After constructing the shellcode, the attacker begins fuzzing the application with random input to crash it and reveal more information, an example can be seen in Figure 2.

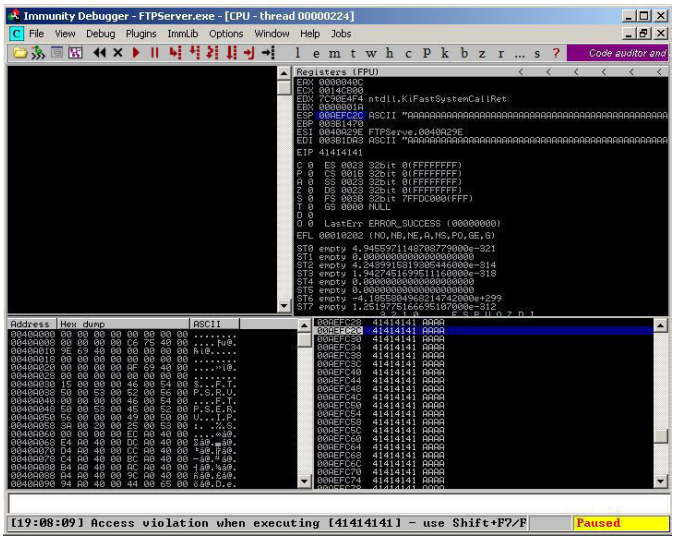


Fig. 2. Fuzzing an application.

Segmentation faults represent invalid memory access which could be potentially exploited. Once the attacker overwrites the instruction pointer, they are able to control the flow of the program.The exploitation methodology is as follows:

1) Create an accessible file with shellcode.
2) Prefix the shell code with a NOP sled and store it in an environmental variable.
3) Find the address of the previous variable containing the shellcode.
4) Overflow the vulnerable function call by repeating an address within the NOP sled of the environment variable containing the shellcode. The Extended Base

Pointer (EBP) and the Extended Instruction Pointer (EIP) will be overwritten.

5) when the EIP is overwritten with an address belonging to the NOP sled, the execution will slide down to the code.
6) The shellcode will be executed, spawning a root shell.

An unsuccessful attempt results in a core file on the target system which is created as a result of segmentation fault when an illegal attempt to access memory has been noted. Most attackers remove these core files after successful exploitation to avoid detection. These core files are only created when remote buffer overflows are executed, but not when the buffer overflow is executed on the target system directly. After exploitation, in order to avoid having to execute the buffer overflow every time they wanted access to the server, most attackers leave backdoors on the filesystem for easy access. One of the methods used for post-exploitation access is creating a new suid binary which allows the attacker to execute commands or switches the user to root.

As previously mentioned, significant attention should be paid to programs written in unsafe languages which do not do boundary checks by default. Functions such as `strcpy()` and `strcat()` are vulnerable by design by accepting arbitrary inputs whose length is not checked and can overflow into nearby memory. Their safe versions, `strncpy()` and `strncat()` require an additional parameter indicating the maximum length of the input. The unsafe variants' input must be checked additionally and sanitized accordingly.. Such a task should not be left as a responsibility to the programmers because a single omitted input could prove disasterous.

Since it is practically impossible to prevent every single attack and be fully secure, it is necessary to monitor for pre-attack scenarios to detect attacks timely. Buffer overflow attacks, like any other attack, include a reconnaissance phase which is a passive inspection of the target system, as well as a scanning phase where the vulnerabilities are actually found. If an attacker executed a script which searches for vulnerable and executable suid binaries, it would help him perform the attack with greater precision and reliability. The `strings` command proves very helpful in this task, since it locates the printable strings within a binary file, along with the *awk* command which uses regular expressions and can be used to search for lines which contain potentially vulnerable functions.

The scanning phase allows an attacker to determine which binaries are best to exploit and probe. Standard binaries originating from Unix vendors have been throroughly researched numerous times, so it's not practical to try to attack them. Instead, the attacker can focus on third-party installed binaries with the setuid bit set.

D. Remediation

To minimize the risk of buffer overflow exploits, the matter should be looked at from two aspects. From an administrator's side, detection systems can be set up so that they alert the administrator of such scans and monitor for new suid files in the filesystem as well as keeping track of the old ones, however, there's rarely anyone doing it. It is a good practice to schedule a `cron` job to run weekly which counts and lists the suid binaries in the system, and alerts the administrator if a change is detected. Certain processors support the No-eXecute (NX) bit which prevents code execution within the stack, thus preventing stack-based buffer overflows. The NX bit is supported by the Linux kernel. Another kernel feature which provides buffer overflow defense is address space layout randomization (ASLR). This feature is supported in the Linux kernel and enabled by default since 2.6.12. It randomizes the address space by assigning different memory addresses every time the program is executed.

Many attackers leave bot software on the server which gives them full remote control at any time using reverse shells. This is why outgoing and ingoing connections should also be monitored at any time and every suspicious connection should be analyzed. If any suspicious binary is found, it should be analyzed with software like `ltrace`, which traces library and system calls, and `strace` to trace system calls and signals. Any user accounts associated with malicious user activity should be locked and inaccessible. There is also the possibility of having system configurations and sensitive files being altered by the attacker, or even infected by rootkits, which is why regular scans for md5 checksums should be run or scheduled weekly. A change in the digital signature of a file indicates that it has been tampered with and must be remediated.

Software developers should also pay attention to their development process and avoid using vulnerable functions, such as `strcat()`, `strcpy()` and many others. Use of functions which do boundary checks and perform the same tasks is recommended. Avoiding insecure languages like C and C++ is also recommended, or used for small applications if necessary. The greater the complexity of the native code, the greater the attack surface will be. Some newer sophisticated compilers check for vulnerable functions and add boundary checks by themselves. Including vulnerable libraries in the code also opens up opportunities for exploitation, which is why they should be avoided whenever possible. Certain languages, such as Java and C# perform boundary checks and runtime checks for whether the code attempts to access protected memory. The use of a canary value also provides buffer overflow protection, by the placement of an integer before the return address in the stack. After execution, a check is performed whether this value has been overwritten by an attacker.

IV. EXPERIMENT

The purpose of the experiment is to test the default configuration for several popular Linux distributions that are often used as servers. Without installing additional software, five different Linux distributions (Ubuntu, Debian, CentOS, Fedora and Slackware) were tested in a virtual environment. The command used was `find / -xdev -perm -4000 -type f -print0 -exec ls -l {} \;` in order to locate the files within the filesystems with the setuid bit set. `-xdev` prevents the `find` command to descend to lower directories, while the *-perm* option specifies the file permissions. The `-type` option defines the type of files the command is searching for, in this case normal files. The `-print0` option prints a null character at the end of a line without breaking certain filenames. `-exec` executes the `ls -l` command and feeds the output back to the `find` command.

The setgid files were found with the same command, with the exception of the parameter `-perm`, which had `-2000` as an argument. The `-6000` argument was used to locate the files which had both bits set. The host machine is running

Kali Linux 1.1.0 with 3.14 kernel and the test systems were virtualized in the VMware Player software. The test systems were CentOS 7.0 with 3.10 kernel, Slackware 14.0 and Debian 7.5 with 3.2 kernel, Fedora Server 21 on 3.17 and Ubuntu Server 12.04 LTS with 3.8 kernel. X Window and desktop environments were excluded from the installations for a more precise result. Most of the discovered binaries were shared among the systems, but there were also some files which aren't used as often and their privilege escalation mechanisms should resort to a different method to avoid introducing additional vulnerabilities. After finding the offending binaries on every system, I calculated the percentage of root-owned files for both setuid and setgid binaries.

V. Results and Discussion

The gathered list of binaries with the setuid/setgid bit set shows that contrary to my own belief, as can be seen in Figure 3, Slackware prove to be the least secure distribution, with leading 38 suid binaries. Second comes Fedora with 24 binaries which pose a potential security threat, followed close by Ubuntu with 22. The most secure distributions are Debian and CentOS, both utilizing 17 suid binaries.
Most of these binaries are mutual among the different systems and are commonly used, but there are also some which aren't used too often. As a security measure, it is recommended that these binaries be removed or disabled altogether. A greater attack surface means a higher chance of having your systems exploited. The reason why Slackware is the leading distribution using the most suid binaries could be because Slackware comes with a rather long list of tools and services preinstalled for system administration. Even without the desktop environment, Slackware offers a plethora of different applications which can be used to ease the life of a system administrator.

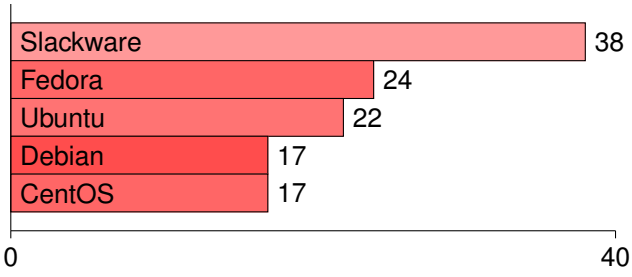


Fig. 3. Amount of suid binaries on each system

As for the setgid bit, the results shown in Figure 4 imply that Ubuntu is the leading distribution with 14 setgid binaries, second being Debian with 13, the third place is taken by Slackware with 10 binaries, fourth comes Fedora with 9 and CentOS with 8 guid binaries.
Most distributions share one or two binaries between the two groups, except CentOS, which does not have a binary which has both setuid and setgid flags set. The vast majority of these binaries do not pose a security threat, due to the fact that almost all of them are owned by users other than the root user. They're used to elevate a set of very limited privileges in order to perform a specific task and have limited access to system resources. Fedora and CentOS are the only systems that contain one root-owned binary with the setguid bit set. This binary is used to detect changes in the network interfaces.

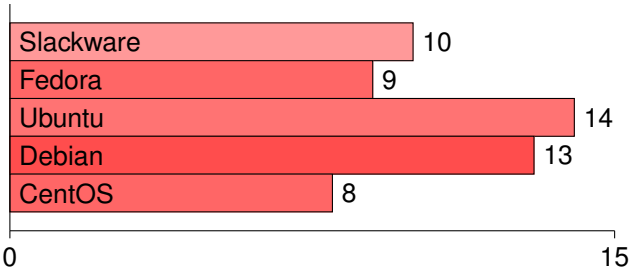


Fig. 4. Amount of guid binaries on each system

Even though Slackware has the most suid binaries, the percentage of these files which were owned by the root user showed that surprising 60% of the files (23 out of 38) pose a potential threat, which means that both Slackware and Fedora's filesystems suid/guid binaries amount to a total of 23 root-owned files. The other distributions' dangerous binaries, as shown in Figure 5, were mostly owned by the root user, with very few exceptions. The risk of a binary being exploited is lesser when it comes to non-root suid owners. These users are mostly used for specific tasks and have very few permissions beside the ones they require. If an attacker was to exploit such a binary, they would get the permissions of the user owning the binary. When the owner has access to limited resources, the attacker would have to resort to further exploitation methods, such as local root exploits for the kernel, in order to gain root privileges. This does not render the system completely immune, but slows down the attacker and might give a system administrator or intrusion detection system enough time to detect and sanitize the attack.

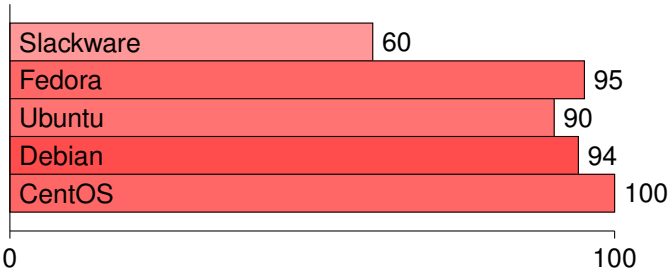


Fig. 5. Percentage of root-owned suid binaries

After analysing the results, I traced some of the system calls of the said binaries and managed to achieve privilege escalation by executing a buffer overflow on a binary I found to be vulnerable on a different system. Using the Metasploit Framework for penetration testing, I loaded a module that abuses the nmap executable with a setuid bit set, which was installed as a root user on the system, and successfully exploited the binary, gaining privileged access.

VI. Conclusions and future work

The used ID-based access control in Unix systems poses a security threat due to the poor design and documentation of system calls that modify the user IDs. The risk escalates with the increase of the number of files that utilize UID-setting system calls. setuid escalation is a common attack vector and should be seriously taken in order to avoid security breaches. Considering it is an necessary feature for the binaries to execute properly, steps must be taken to remediate the existing risks, such as patching and upgrading the kernel with the GRSecurity [9] tool. This tool has access control capabilities, prevents certain types of memory exploits, auditing options and many others.

I concluded that the lack of information on the ancient setuid and setgid mechanisms and the mystery that shrouds them is the reason for confusion which resulted in an influx of security vulnerabilities exploiting these mechanisms. They should be replaced by a more suitable mechanism which isn't prone to exploitation as much. Software development tools and utilities can also be used by a malicious attacker, especially debuggers and disassemblers, therefore they should not be available to everyone on the systems. If there really is a need for such an application, access should be limited to the root user only. Examples such as `nm`, `objdump`, `gcc`, `gdb`, `as`, `elfdump`, `strace`, `ltrace`, `nasm` and many others, can only help an attacker to analyze the attacked system. Bhushan Jain, Chia-Che Tsai, Jitin John and Donald E. Porter [10] introduced a few techniques and a framework for enforcing least-privilege policies in the kernel, mitigating the risk of buffer overflows.

In the future, I would be glad to run the same tests on other *nix operating systems, such as Solaris and FreeBSD. I'm curious as to which operating system in general pays the most attention to this issue. I'm already assuming that FreeBSD, famous for its secure design and software implementation, will perform as the most secure operating system out of all three. I shall also delve further into the methods of exploitation of these vulnerabilities to examine the inner workings of the mechanisms which are left undocumented, but exposed nonetheless on thousands of systems around the world.

References

[1] C.Torek, C.H.S. Dik. *setuid mess*, Berkeley Software Design, 1995.

[2] H.Chen, D.Wagner, D.Dean. *setuid demystified*, USENIX Security Symposium. University of California at Berkeley, SRI International. 2002.

[3] M. Bishop. *How to write a setuid program*. ;login:, 12(1):5-11, 1987.

[4] M.Bauer. *Paranoid Penguin - Linux Filesystem Security, Part I*, Linux Journal, 2004.

[5] C.Perrin. *Understand the setuid and setgid permissions to improve security*, Tech Republic. 2009.

[6] One Course Security. *Linux Special Permissions*. 2013.

[7] J. Pike. *Discovering a Local SUID Exploit*, SANS Institute. May 2004.

[8] *Metasploit Unleashed - Generating Payloads*, Offensive Security. Web. March 6, 2015.

[9] *Grsecurity documentation*. Web. 8 March, 2015.

[10] B.Jain, C. Tsai, J.John, D.E. Porter.*Practical Techniques to Obviate Setuid-to-Root Binaries*, Stony Brook University.

Federating Any on-premise established identity management system with Office 365 cloud services

Vasko Sazdovski
Faculty computer center,
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia

Boro Jakimovski
Faculty computer center,
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia

Abstract— **Office365 is a group of software plus services subscriptions that provides productivity software and related services to its subscribers as a SaaS cloud service. The service allows the use of Microsoft Office online apps, provides storage space on Microsoft's cloud storage service OneDrive. Additionally for enterprise and academic users, Office 365 offers plans including e-mail and social networking services through hosted versions of Exchange Server, Lync, SharePoint and Office Web Apps, integration with Yammer, as well as access to the Office software. The new service model allows easier and more diverse usage of the applications such as tablets and smartphones or any other web enabled rich client with the comfort of using a standard PC.**

The user management of Office365 relies on Microsoft Azure Active Directory that is integrated into the Office365 admin portal. This administrative model presents a big overhead when adopting the service in an existing environment. Office365 allows federation of identity management using Active Directory Federation Services and SAML2 protocol.

In this paper we introduce a model of federation of any on-premise established identity management system with Office 365 cloud services using SAML2 protocol. This model describes the process of assessment, laying the groundwork and the process of federating the already established identity management system with Office365 cloud service.

Keywords – Microsoft Office365, Shibboleth, Federation Identity Management, Single Sign On

I. INTORDUCTION

Office365 introduces new cloud based services for Microsoft productivity software that is implemented as a SaaS. For easier adoption or migration of on-premise deployments of these services to Office365 cloud based services, Microsoft has developed an integration/migration tools and services. Microsoft has releases tutorials and guides, how to establishes this federation, but only limited to Microsoft on-premise technologies. This presents a challenge for enterprises to adopt Office365 since it can involve double management of users, that incurs bigger expenses, can have a slowdown in introduction of new services and can be also time consuming. In this paper we introduce a model of federation of any on-premise established identity management system with Office 365 cloud services using OASIS Security Assertion Markup Language (SAML2) protocol. This model describes the process of assessment, laying the groundwork and the process of federating the already established identity management system with Office365 cloud service Figure 1.

II. IDENTITY MANAGEMENT

Identity management (IdM) describes the management of individual principals, their authentication, authorization, and privileges within or across system and enterprise boundaries with the goal of increasing security and productivity while decreasing cost, downtime and repetitive tasks. The Identity Provider (IdP) provides Single Sign-On services and extends reach into other organizations and new services through authentication of users and securely providing appropriate data to requesting services. In addition to a simple yes/no response to an authentication request, the Identity Provider can provide a rich set of user-related data to the Service Provider (SP). This data can help the service provide a more personalized user experience, save the user from having to manually enter data the service requires, and refresh the data each time the user logs onto the service [1].

Identity management systems have been developed for more than 3 decades, mainly focusing on technologies based on X.500 series of standards for electronic directory services and its implementations. Usually companies choose one of the two most widely used IdM systems based on LDAP/Kerberos protocols, Microsoft Active Directory and OpenLDAP. Also, there are many cases of custom developed IdM systems based on web CRM or ERP/HR solutions that build and maintain large user databases. We call these systems on-premise established IdM.

On-premise IdM systems usually do not provide direct access to IdP functionalities. There are several IdP technologies that provide integration or can be adapted to authenticate/attribute exchange against legacy IdM systems. Some of them are open-source solutions, like in our case Jasig CAS [3] that can be easily extended to integrate with any on-premise IdM system for authentication purposes. shows a model of on premise IdP based on existing IdM system.

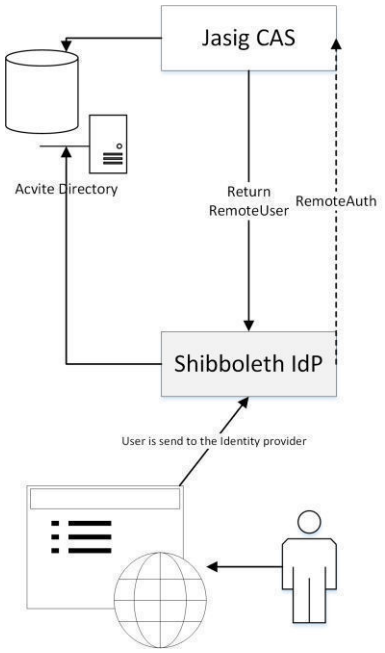


Figure 1. Model of integration of IdP and existing IdM

III. OFFICE365 FEDERATION

Office365 is a deployment of MS Office, MS SharePoint, MS Exchange and MS Lync based in the Microsoft Azure Cloud. All services in Office365 are integrated with Azure Active Directory that maintains all users and groups information.

In order to ease the process of adoption of Office365 in large Enterprises, Microsoft enables integration of on-premise Active Directory with Azure Active Directory. The integration is implemented as both federation and synchronization of both Active Directories. The process of synchronization is usually one way user and group objects exchange, since there is a need for users to be maintained in the local Active Directory and their information to be sent to Azure Active Directory. The sync does not include all attributes, and does not include authentication information. Microsoft has developed an Active Directory Sync tool (DirSync) that is a subsystem of Forefront Identity Management. This tool enables scheduled Active Directory synchronization using a set of rules.

For this purpose Office365 uses federation between Active Directories. Federation of local Active Directory is usually based on Active Directory Federation Services 2.0 (ADFS 2.0) server that is an on-premise system that enables user web authentication using Active Directory Kerberos credentials. The protocol for federating both Active Directories is based on SAML 2.

This implementation of Office365 federation is limited only to on-premise Active Directory IdM that limits the potential use cases.

IV. INTEGRATION OF ANY ON-PREMISE IDM

In order to enable Office365 integration with any on-premise IdM, we need to establish both synchronization and trust federation between local IdM and Azure Active Directory.

Synchronization using Microsoft DirSync tool presents a complex workflow that is very hard to manage and define, and thus can easily lead to problems such as: locked users, wrong attribute information, unsuccessful synchronization due to conflicts with GUID and UPN attributes. Also DirSync does not provide proper debugging information that can help in troubleshooting problems.

Microsoft provides API interface for Azure Active Decretory only for Azure subscribers. This is a limitation for users of Office365 that do not plan to purchase Azure subscription. This presents a challenge for implementation of synchronization with a non-Active Directory on-premise IdM.

Microsoft has developed a Power-shell library that enables management of Office365 users and groups without the need of Azure subscription. Using this libraries we have developed a custom user synchronization tool that can replace DirSync and fully synchronize any on-premise IdM with Office365 as shown in Figure 2. This Sync Tool is based on dictionary logic. It takes the user from the local database and from the Azure AD and makes matches. If a user exists on the both databases, checks if that user has a suitable license assigned; if it has the right license ignores that user, and if the license is not suitable or is missing, the Sync tools assigns the right license for that user. Figure 3 shows an example of Power shell code for manipulation of Office365 account information.

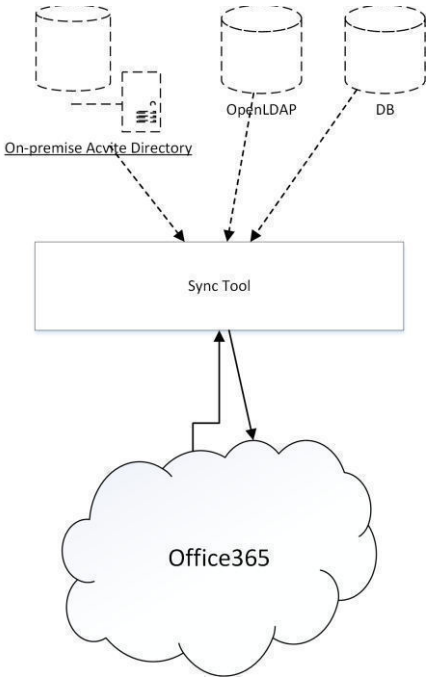


Figure 2. Synchronization tool scenario

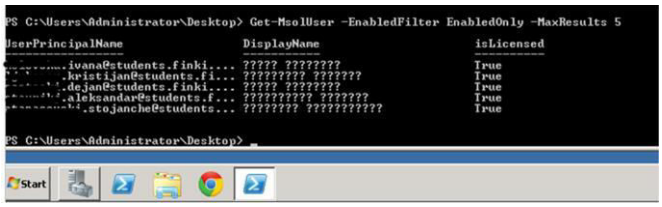


Figure 3. Power Shell script for manipulation with Office365 Users

A next step in establishing the federation is the process of federating on-premise IdM and Office365. As Microsoft only provides Active Directory Federation Services 2.0 as a tool that federates Active Directory IdM, this does not present a generic solution for other authentication and authorization services.

Since ADFS 2.0 is based on SAML 2, that is an open standard, Office365 supports federation with other SAML 2 IdP providers, having that synchronization to Azure Active Directory is already established.

Enterprises can establish a Single sign on service based on different technologies in order to integrate their different web application with single user identity and authentication database managed on a central IdM. Some of these solutions provide the possibility to federate several administrative domains by creating a trust relationships between services and identity providers. SAML2 is on such protocol with many implementations, one of which is Shibboleth.

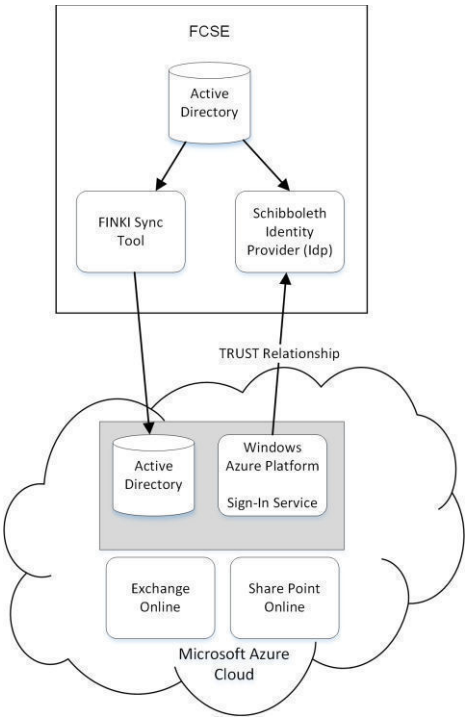


Figure 4. Architecture for integration of on-premise IdP and Office365

Shibboleth is a standards based, open source software package for web single sign-on across or within organizational boundaries. It allows sites to make informed authorization decisions for individual access of protected online resources in a privacy-preserving manner. The Shibboleth software implements widely used federated identity standards, principally the SAML2, to provide a federated single sign-on and attribute exchange framework. A user authenticates with his or her organizational credentials, and the organization (or identity provider) passes the minimal identity information necessary to the service provider to enable an authorization decision. Shibboleth also provides extended privacy functionality allowing a user and their home site to control the attributes released to each application. Shibboleth IdP can easily integrate with different backend authentication services and IdM attribute databases. Having this we have implemented a Shibboleth based solution that integrates with our Sync tool and allow full federation of on-premise IdM/IdP with Office365. On Figure 4is shown the model that we propose for integration of on-premise IdP and Office365.

In order to successfully integrate Shibboleth IdP with Office365 we need to:

- Define trust relationship in Shibboleth IdP metadata Figure 5.
- Define attribute exchange profile for the Office365 service

```
<!-- Release userPrincipalName as Windows Azure AD User ID -->
  <afp:AttributeFilterPolicy id="UserId">
    <afp:PolicyRequirementRule
xsi:type="basic:AttributeRequesterString"
value="urn:federation:MicrosoftOnline"/>
    <afp:AttributeRule attributeID="UserId">
      <afp:PermitValueRule xsi:type="basic:ANY"/>
    </afp:AttributeRule>
  </afp:AttributeFilterPolicy>

  <!-- Release Immutable ID to Windows Azure AD -->
  <afp:AttributeFilterPolicy id="ImmutableID">
    <afp:PolicyRequirementRule
xsi:type="basic:AttributeRequesterString"
value="urn:federation:MicrosoftOnline"/>
    <afp:AttributeRule attributeID="ImmutableID">
      <afp:PermitValueRule xsi:type="basic:ANY"/>
    </afp:AttributeRule>
  </afp:AttributeFilterPolicy>
```

Figure 5. Shibboleth IdP metadata

V. CONCLUSION AND FUTURE WORK

In this paper we have shown a model and real implementation of integration between an existing on-premise IdM/IdP and Office365. The solution described has been implemented in the Faculty of Computer Science and Engineering for several years and several other National level Office365 instances for High schools integrated with the national eGradebook system login system.

The solution can be further extended by building direct Azure Active Directory API interaction in order to be able for faster and richer User and Group synchronization.

VI. ACKNOWLEDGMENT

This project and partly funded by the Faculty of Computer Science and Engineering in Skopje.

VII. REFERENCES

[1] Morgan, R. L., et al. "Federated Security: The Shibboleth Approach." Educause Quarterly 27.4 (2004): 12-17.

[2] Hughes, J. ,Cantor, S. ,Hodges, J. ,Hirsch, F. ,Mishra, P. ,Philpott, R. ,Maler ,E .(2005). “Profiles for the OASIS Securiy Assertion Markup Language(SAML) V2.0”. OASIS standard. http://docs.oasis-open.org/security/saml/v2.0/saml-profiles-2.0-os.pdf

[3] Ionita,M.G. “Secure Single Sign-On using CAS and OpenID”. Journal of Mobile, Embedded and Distributed Systems, 4(3) (2012), 159-167.

Learning models of abiotic influence on the biodiversity indices in Lake Prespa

Andreja Naumoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Republic of Macedonia
andreja.naumoski@finki.ukim.mk

***Abstract* – The abiotic factors are correlated with the biodiversity indices in a complex way. However it is possible to find the connection that binds these variable. This will reveal information on how abiotic factors influence on the diversity of community. This is very important information for both ecologist and decision makers. By quantifying the diversity and evenness with mathematical indices, this process can provide better understanding on how the community structure functions. Therefore, this paper aims to understand this influence on the community structure by modelling the relationship between these two variables by using machine learning methods. Several decision trees methods are consider to models this relationship. In order to find which one is the best for building prediction models, we compare their performance. Furthermore, we compare the regression models with the classification models. Later, one obtained model from the regression task and one from the classification task is presented and discussed. Based on the obtained results, further investigation with other machine learning method is suggested. This will also further expand the methods used in machine learning for analysis of environmental data.**

Keywords - Diversity indices, evenness indices, machine learning, decision trees.

I. INTRODUCTION

Diatoms are ideal bio-indicators and they rapidly react on the environmental changes. Therefore, they are strongly related to the abiotic factors of the environment. The modelling of this relationship, between the abiotic factors and the diatoms that inhabit that environment is the main focus of habitat suitability modelling [1, 2]. This process tries to obtain models that will discover the direct interaction, in many times a non-linear relationship, between the environment and the abundance of a given organism or the entire community. They are different approach of representation of the biological information that is contained in the measurement of the abundance of the organisms. One is to take the raw measured data as it is, or another possibility is to mathematically calculate variables that integrates several ecological parameters or different species of organisms into one. One such possibility are the biodiversity indices (diversity and evenness indices) derived from the measured abundance of several selected species or entire community. These indices reflects the interactions that are effected by the abiotic factor of the environment [3]. Also, they provide more information about community composition than simply species richness.

The diversity indices are widely used in environmental studies [4, 5, 6]. They are applied in different environments ranging from forests, pastures, lakes, oceans and etc. [7, 8, 9]. Many of the diversity indices are introduced the last century, while some of them are adapted from other disciplines [10, 11, 12, 13]. Some of the biodiversity indices, like the Shannon and the Brillouin index are used in information theory [14, 15] and letter adopted in ecology [16]. One of the many examples is the Brillouin formula discussed in [17] or calculating evenness among species [18]. Several research studies emphasis the differences between diversity and evenness indices [19, 20, 21].

In environmental studies, a typical approach of measuring diversity, is to take the measured sample and then calculate the biodiversity indices from the raw data. Then based on the obtain results manually found out which organisms are influence by what abiotic factors. In some cases gradient analysis is used, however, this type of analysis can be subjective, because it is based on existing knowledge about the studied species and the availability of data about these species that are organised along the gradient of a factor [22]. Therefore, the hypothesis that species from the community can be arranged in some sequence along a gradient of some environmental factor does not necessary lead to a conclusion that this particular factor found by the gradient analysis is the most important one.

In order to overcome this problem and further improve the set of learning tools used to build biodiversity models, in this work we apply different regression and classification machine learning methods. These methods are widely used in the field of eco-informatics [23, 24, 25, 26], because they obtain models that can be easily interpreted with satisfactory accuracy. Until now, using variant of these machine learning methods, several studies were done in order to investigate the influence of the abiotic factors on the organisms' community [27, 28, 29, 30, 31]. However, they do not attempt to describe the relationship between the biodiversity indices and the abiotic factors. First attempt to correlate the biodiversity indices and the abiotic factors for Lake Prespa was made by [32, 33], by using regression trees. These models provided beneficial inside information about the environmental influence on the biodiversity, however the obtained accuracy was medium and for many biodiversity indices no models were obtain.

In this direction, this paper aims to further improve the model induction by analysing both approaches (classification and regression) and selecting the best one. In order to apply classification task, each biodiversity index will be discretize. Categorisation of certain physic-chemical parameters is done similar like the water quality and trophic state index classes found in the literature.

The rest of the paper is organized as follows: Section II provides mathematical definitions of the diversity and evenness indices. Section III presents the dataset description and the experimental setup. In section IV we present the experimental results and based on these results one biodiversity model is selected and discussed. Finally, Section V concludes the paper and research direction is outlined.

II. DATA DESCRIPTION AND EXPERIMENTAL SETUP

Experimental analysis are performed on a dataset that is collected in 16 month period during the EU project [33]. The following physico-chemical properties of the water samples were measured: temperature, dissolved oxygen (DisO), Secchi depth, conductivity (Conduc), alkalinity (pH), nitrogen compounds (NO_2, NO_3, NH_4, organic and total nitrogen), sulphur oxide ions (SO_4), phosphorus P, sodium (Na), potassium (K), magnesium (Mg), copper (Cu), manganese (Mn) and zinc (Zn) [34]. Additionally, the relative abundance of 116 different diatom species were measured.

The physico-chemical properties of the samples provides the abiotic information, while the biological samples provided information on the relative abundance of the studied diatoms. The 116 different diatom species represented with abundance are converted and characterized with the 4 diversity and 4 evenness indices. Biological experts do not express most of the biological knowledge, when it comes to water quality or trophic state as single numbers, they are represented with water quality classes. Similar, each calculated diversity and evenness indices will be transformed or categorized in three categories (low, medium, high). The evenness indices are in range between 0 and 1, so the low category represent the first 33%, while the upper 33% of the range belong to the high category. The medium class is represented by measured sample that are between 33% and 66%. The same categorization is made on the diversity indices, with respect on the higher and lower limit of each diversity index. The output class of each of the models consists from three classes: low, medium and high, according [6, 7, 8, 9, 10].

Consequently, the evaluation dataset consists from 29 input parameters. From them, 21 attributes (the physico-chemical parameters) represents input in the model, while the 4 diversity and evenness indices are the output parameters of the model. Based on this information, two types of experiments are made, which are setup as follows:

1) **Train**: Train model is obtained based on the raw measured data for each of the four diversity and evenness indices.

2) **Test**: Standard 10-fold cross validation is used to evaluate the classification accuracy of each classification and regression algorithm.

We use several different algorithms that all belong to group that produce interpretable model in a hieratical tree-like structure, known as classification and regression trees. The algorithms that are used for classification experimental evaluation are: BF Tree [35], FT [36], J4.8 [37], LAD Trees [38], LM Tree [39], NB Tree [40] and REP Tree [41]. While the regression analysis is conducted using M5P [42] method. All algorithms are used with their default settings.

III. EXPERIMENTAL RESULTS

A. Performance Analysis

The results from the classification experimental evaluation are given in Table 1 and Table 2, while the results from the regression evaluation are given in Table3. Separately we present the classification analysis of the diversity indices (Table 1) and for the evenness indices the results are presented in Table 2.

As we can note from the results in Table 1, we have variable classification accuracy of the models for the diversity indices. The best train and test classification accuracy is achieve by Shannon index models.

Table 1. Performance evaluation of the classification algorithms for diversity correctly classified instances (in %). The highest value per index for each algorithm is bolded.

Algorithms	Train				Test			
	Hill's	Brillouin	Simpson	Shannon	Hill's	Brillouin	Simpson	Shannon
BF Tree	62.71	74.36	74.53	**81.63**	52.74	76.42	48.23	**81.54**
FT	78.62	**86.40**	66.48	84.42	56.32	**77.41**	55.12	77.23
J48	83.75	**89.25**	67.56	88.12	58.41	**75.63**	49.12	71.12
LAD Tree	80.21	87.23	72.34	**92.65**	61.23	71.23	46.90	**75.12**
LM Tree	63.96	**82.52**	67.42	71.78	65.43	76.12	51.70	**82.33**
NB Tree	62.48	83.21	67.98	**82.12**	58.23	77.51	44.21	**81.20**
REP Tree	75.72	77.34	77.8	**84.74**	64.41	63.51	54.90	**80.92**

If we analyses the results for the evenness indices the train classification accuracy is relatively same (small decrease for Shannon index), but there is a small increase of the test accuracy, compared with the Table 1. According Table 2, it is very obvious that the Shannon index have the highest accuracy from the most of the algorithms, followed by the Brillouin index. In regression analysis the highest accuracy is achieved using LMT method using for the Shannon index of 92.65 % in train, and 82.33% in test using the LM Tree method. If we compare the results from the regression analysis (Table 3) the Hill's Evenness index achieved highest train and test correlation, while the Brillouin and Shannon indices models achieved highest train and test correlation, respectively. The mixed accuracy show that is hard to pin point a single biodiversity index that is suitable for high accuracy model.

Table 2. Performance evaluation of the classification algorithms for evenness correctly classified instances (in %). The highest value per index for each algorithm is bolded.

Algorithms	Train				Test			
	Hill's	Brillouin	Simpson	Shannon	Hill's	Brillouin	Simpson	Shannon
BF Tree	61.37	79.36	**87.83**	82.54	58.20	78.31	46.56	**82.54**
FT	79.36	**87.30**	65.08	85.18	55.02	**79.36**	52.38	75.66
J48	84.65	**90.47**	61.37	87.83	57.15	75.13	48.67	**77.78**
LAD Tree	82.01	**89.95**	70.37	86.65	62.96	72.49	45.50	**83.54**
LM Tree	68.78	82.00	66.67	**83.60**	62.43	**80.42**	51.32	73.01
NB Tree	61.38	85.20	62.43	**87.83**	57.67	74.60	43.91	**80.60**
REP Tree	77.78	79.36	76.20	**82.30**	60.32	75.13	56.09	**80.95**

Table 3. Performance evaluation of the regression analysis for both diversity and evenness indices with correlation coefficient (in %). The highest value per index for each algorithm is bolded.

Algorithms	Train				Test			
	Hill's	Brillouin	Simpson	Shannon	Hill's	Brillouin	Simpson	Shannon
M5P (Evenness)	**87.82**	85.19	75.15	85.86	**63.54**	60.62	31.17	59.80
M5P (Diversity)	78.48	84.07	43.42	**84.62**	25.50	**55.78**	23.74	52.36

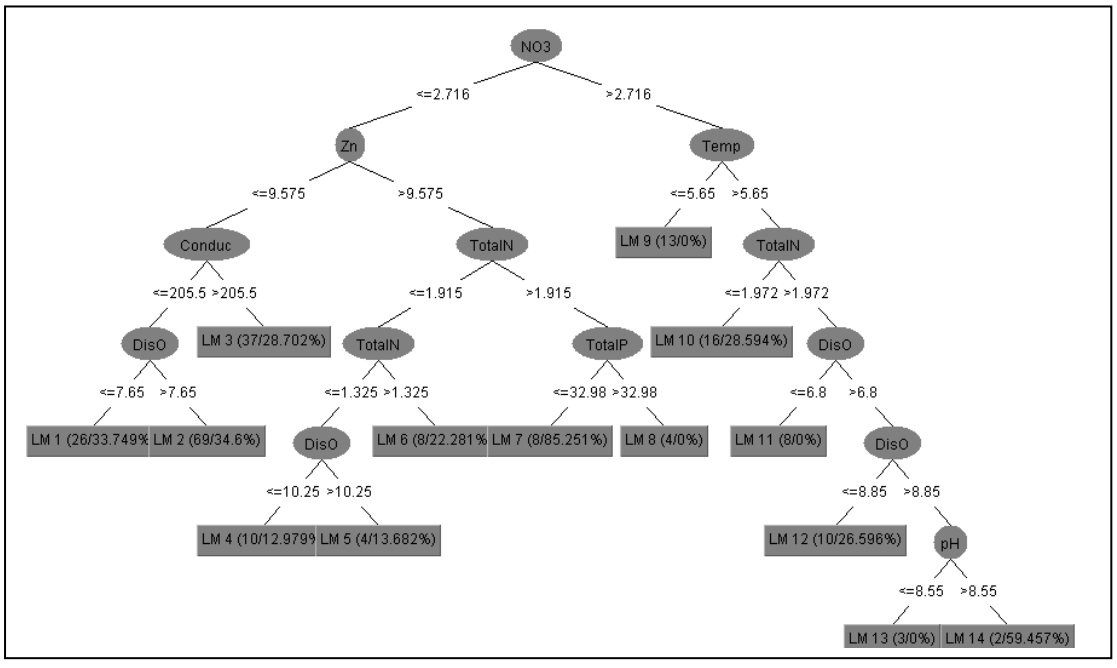


Fig. 1. Regression model for Hill's Evenness index obtained using M5P algorithm

The Brillouin index achieve equal and in some cases better accuracy compare to the Shannon index, because the number of species and the proportional abundances are constant. Additionaly, the Hill's index achieved highest accuracy in regression, because it represents the number of uniformly distributed species. However, the test accuracy evaluation in regression presented weaker results related to the train accuracy, which is not the case using the classification methods.

B. *Indices models*

Two models are presented in this section: one model obtained using the regression method, and one model with classification method. The first model shown with Fig. 1, depicts the diversity information correlated with several abiotic factors. NO_3 in combination with Zn and Temperature factors are the most influence factors on the Hill's biodiversity index according to classification model. These two abiotic factors are followed by the Nitrates and Potassium parameters.

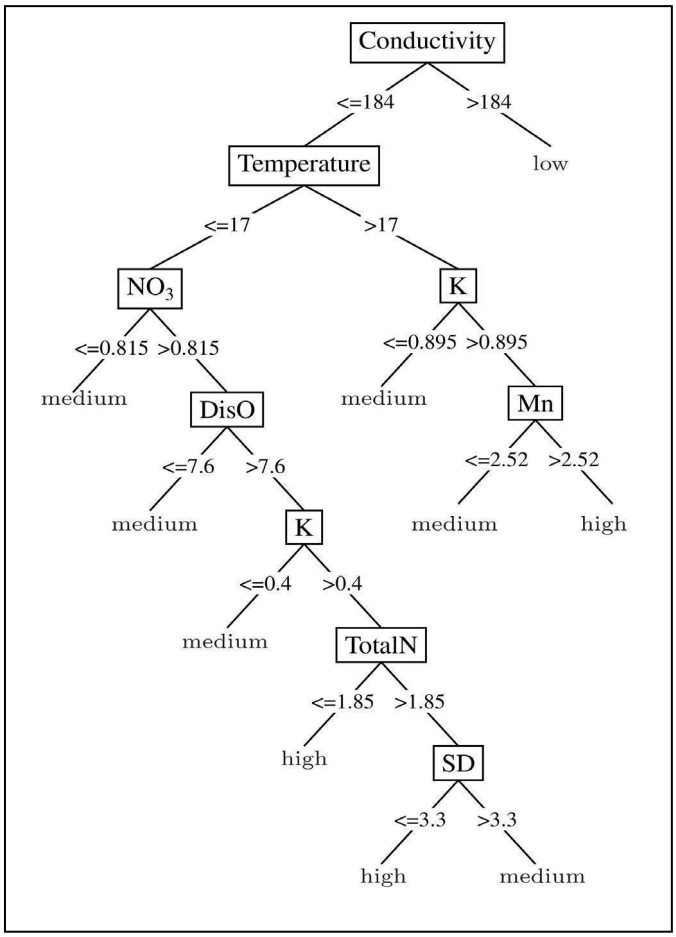


Fig. 2. Classification model for Brillouin evenness index obtained using J48 method

The second model given in Fig. 2, presents the influence of the abiotic factors on the Brillouin evenness index. According this model, the conductivity and the temperature parameters have the greatest influence on this index. As it is point out in many research studies, the abiotic parameters found by both models, play metabolic role in the diatom life, as the biological expert confirmed in many previous research studies [29, 31]. The Potassium, Manganese and the Nitrates are a crucial part of the enzymes that play an important role in the life of diatoms [32].

IV. CONCLUSION

In this paper we experimentally evaluate the implication of the machine learning methods, both classification and regression decision tree methods, on building biodiversity index models, individually for four diversity and evenness indices. The two presented models have reveal that Nitrates, Conductivity, Temperature and metal parameters are effecting the biodiversity distribution of the diatoms. Both classification and regression model results are correlated with the findings in [29, 31, 32]. Therefore, the paper shows that obtaining such models can provide useful information about the relationship between the organisms and the abiotic factor in the environment.

Since these are the first classification models for the biodiversity models of Lake Prespa, we have only compared with the similar regression models obtain for the first time for Lake Prespa in [30]. We also compared the experimental evaluation results for both methods. The experimental results obtained using the classification algorithms can be a subject for future research since the number of classes (discretization level) can vary. In this paper we used 3 classes, however further research should focus on experimental evaluation regarding lower or higher number of discretization values.

Further research should be directed on classifying more diversity and evenness indices and try more different classification algorithms. Another direction for future study is to obtain models based on larger set of samples coming from a diverse environments (not just a single lake).

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] S. Džeroski, "Applications of symbolic machine learning to ecological modelling," *Ecological Modelling* 146 (1-3), 263 – 273, 2011.

[2] S. Džeroski, "Machine learning applications in habitat suitability modeling," In: Haupt, S. E., Pasini, A., Marzban, C. (Eds.), *Artifcial Intelligence Methods in the Environmental Sciences*, Springer, pp. 397-412, 2009.

[3] P.A. Colinvaux, "Introduction to Ecology," *Wiley*. ISBN 0-471-16498-4, 1973.

[4] S. Villéger, N.W.H. Mason, D. Mouillot, "New multidimensional functional diversity indices for a multifaceted framework in functional ecology," *Ecology* 89, pp. 2290–2301, 2008.

[5] S.R. Biswas, A.U. Mallik, "Species diversity and functional diversity relationship varies with disturbance intensity, " *Ecosphere* 2(4), article 52, 2011.
[6] T. Leinster, C.A. Cobbold, "Measuring diversity: the importance of species similarity:, *Ecology* 93, pp. 477–489, 2012.
[7] S. Ferrier, G. Manion, J. Elith, K. Richardson, "Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment," *Diversity and Distributions* 13, pp. 252–264, 2007.
[8] L. Nagaoka, "Using diversity indices to measure changes in prey choice at the Shag River Mouth site, southern New Zealand," *Int. J. Osteoarchaeol* 11, pp. 101–111, 2001.
[9] T. Yamada, S. Aiba, Y. Kubota, K. Okubo, I. Miyata, E. Suzuki, H. Maenaka, M. Nagano, "Dynamics of species diversity in a Japanese warm-temperate secondary forest," *Ecosphere* 2, 2011.
[10] E.H. Simpson, "Measurement of diversity," *Nature* 163, pp. 688, 1949.
[11] E.C. Pielou, "Shannon's formula as a measure of specific diversity," Its use and misuse, *American Naturalist* 100, pp. 463-465, 1966.
[12] E.C. Pielou, "Species-diversity and pattern diversity in the study of ecological succession," *Journal of Theoretical Biology* 10, pp. 370-383, 1966.
[13] L. Brillouin," Science and information theory," *Academic Press*. New York. pp. 320, 1956.
[14] M.O. Hill, "Diversity and evenness: a unifying notation and its consequences," *Ecology* 54, pp. 427-432, 1973.
[15] C.E. Shannon, W. Weaver, "The mathematical theory of communication," *Univ. Illinois Press. Urbana* 117, 1949.
[16] J.A. Hendrickson, P.R. Ehrlich, "An expanded concept of "species diversity."," *Not. Nat. Phila*. 439, pp. 1-6, 1971.
[17] K.P. Peet, "Relative Diversity Indices," *Ecology* 56 (2), pp.496-498, 1975.
[18] A.E. Magurran, "Ecological Diversity and its measurement," *Princeton University press, Princeton*. NJ, 1988.
[19] E.C. Pielou, "Ecological diversity," *Wiley Interscience, New York*, 1975.
[20] A. E. Magurran, "Measuring Biological Diversity," *Blackwell*, *Oxford* UK, 2004.
[21] M.O. Hill "Diversity and evenness: a unifying notation and its consequences," *Ecology* 54, pp. 427-431, 1973.
[22] Levatić. J., Kocev. D., Debeljak. M., Džeroski. S., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecological Modelling 2014 (in press).
[23] C. Kampichler, R. Wieland, S. Calmé, H. Weissenberger, S. Arriaga-Weiss, "Classification in conservation biology: A comparison of five machine-learning methods," *Ecological Informatics* 5 (6), pp. 441 – 450, 2010.
[24] C.A. Drew, Y.F. Wiersma, F. Huettmann, "Predictive Species and Habitat Modelling in Landscape Ecology," *Springer*, 2011.
[25] S. Oppel, A. Meirinho, I. Ramirez, B. Gardner, A.F. O'Connell, P.I. Miller, K. Louzao, "Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds," *Biological Conservation* 156, pp. 94 – 104, 2012.
[26] G. De'ath, K.E. Fabricus, "Classification and regression trees: a powerful yet simple technique for ecological data analysis," *Ecology* 81, pp. 3178-3192, 2000.
[27] A.H. Fielding, "Cluster and Classification Techniques for the Biosciences," Cambridge: Cambridge University Press, 2007.
G. De'ath, "Multivariate regression trees: a new technique for modeling species environment relationships," *Ecology* 83, pp. 1105-1117, 2002.
[28] A. M. Prasad, L.R. Iverson, A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological predictions," *Ecosystems* 9, pp.181-199, 2006.
[29] A. Naumoski, D. Kocev, N. Atanasova, K. Mitreski, S. Krstić, S. Džeroski, "Predicting chemical parameters of water quality form diatoms abundance in lake Prespa and its tributaries," In: *4th International ICSC Symposium on Information Technologies in Envi-ronmental Engineering - ITEE* 2009, (ed.) I. N. Athanasiadis, P. A. Mitkas, A.E. Rizzoli & J. M. Gómez, Springer Berlin Heidelberg, pp. 264-277, 2009.
[30] A. Naumoski, "Multi-target modelling of diatoms diversity indeces in Lake Prespa," *Applied Ecology and Environmental Research* 10(4), pp. 521-529, 2012.
[31] A. Naumoski, K. Mitreski, "Influence of climate change on diatoms diversity indices in Lake Prespa," *Journal of E-Analytical - Energy and Climate Change - Southeast Europe in Focus*, 2(1), 2009.
[32] Z. Levkov, S. Krstić, "Diatoms of Lakes Prespa and Ohrid (Macedonia)," I*conographia Diatomologica*, 16 (Iconographia Diatomologica), pp. 603, 2006.
[33] S. Krstić, "Description of sampling sites," 2006.
[34] H. Shi, "Best-first decision tree learning," *Hamilton*, NZ, 2007.
[35] J. Friedman, T. Hastie, R. Tibshirani, "Additive logistic regression:A statistical view of boosting," *Annals of statistics*. 28(2), pp. 337-407, 2000.
[36] J. Gama, "Functional trees," *Machine Learning*. 55 (3), pp. 219–250, 2004.
[37] J.R. Quinlan,"C4.5: Programs for Machine Learning," *Morgan Kaufmann*, *Sanateo*, CA, 1993.
[38] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, M. Hall, "Multiclass alternating decision trees," *European Conference in Machine Learning*, pp. 161–172, 2001.
[39] N. Landwehr, M. Hall F. Eibe, "Logistic Model Trees," *Machine Learning*. 95(1-2), pp. 161-205, 2005.
[40] R. Kohavi,"Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," In: *Second International Conference on Knowledge Discovery and Data Minin*g, pp. 202-207, 1996.
[41] J.R. Quinlan, "Simplifying decision trees," *International Journal of Man-Mahine Studies*, 27 (3), pp. 221-248, 1987.
[42] J.R. Quinlan, "Learning with Continuous Classes," In: *5th Australian Joint Conference on Artificial Intelligence*, Singapore, 343-348, 1992.

Text Classification of Macedonian News

Dimitar Venov*, Igor Trajkovski‡
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
Email: *venov.dimitar@students.finki.ukim.mk, ‡igor.trajkovski@finki.ukim.mk

***Abstract*—With the growth of text documents available on the internet, their manual processing has become an impractical task. One of them is the task of *text classification*. Text classification (also known as *text categorization*), which is the task of assigning text documents to a predefined set of classes, would be an infeasible task without automatic text classification. We were motivated to conduct empirical testing in order to find out which text classification algorithm and which pre-processing techniques will give the best results on dataset composed of short Macedonian documents. In this paper we present the process of dataset creation. Then we will present the pre-processing techniques that we apply to the dataset. Later on, we will discuss the text classification algorithms which we use for automatic text classification of the Macedonian news. These include: Naïve Bayes, Logistic Regression, Nearest Centroids (when used for text classification it is known as Rocchio), Support Vector Machines (SVM) and Logistic Regression and SVM with stochastic gradient descent (SGD). The results from the experiments are presented at the end where we compare them and elaborate them. From the conducted experiments we concluded that SVM has the highest precision and Logistic Regression is just around the corner.**

***Index Terms*—Text Classification, Text Categorization, Topic Spotting, Machine Learning, Logistic Regression, Naïve Bayes, Nearest Centroids, Rocchio, Support Vector Machines, Stochastic Gradient Descent**

I. Introduction

With the growth of online text documents available on the Internet, digital libraries, news sources and company-wide intranets, their manual processing has become an impractical task. It is forecasted that these unstructured data will become the dominant data found online. Text mining, by analyzing these large quantities of unstructured data throughout detecting usage patterns, tries to extract high-quality information. Text classification (also known as text categorization or topic spotting), which is the task of assigning text documents to a predefined set of classes and is a subfield of text mining, would be an infeasible task without automatic text classification.

In this paper we present the process of automatic text classification. We will start by giving a brief overview of the related work of this field in Section II. Then, in Section III we will present the process of dataset creation (417,412 Macedonian news text classified in 13 different categories crawled from 5 different Macedonian news portals). In Section IV we will present the pre-processing techniques that we apply to the dataset and the process of feature vector extraction. Later on, in Section V we will discuss the text classification algorithms which we use for automatic text classification of the Macedonian news. These include: Naïve Bayes, Logistic Regression, Nearest Centroids (when used for text classification it is known as Rocchio), Support Vector Machines (SVM) and Logistic Regression and SVM with stochastic gradient descent (SGD). And, at the end, in Section VI we will present the conducted experiments. We will present the results from those experiments, we will compare them and elaborate them.

II. Related Work

The main approaches to text classification (categorization) that fall within the machine learning paradigm are discussed in [1]. Firstly, they elaborate about document representation and dimensionality reduction, where are given different ways of representing a document and are given different approaches to dimensionality reduction. Then, they discuss the problem of classifier construction from a ”training” set of pre-classified documents. At the end, there are presented techniques for classifier evaluations such as precision, recall, accuracy, error.

According to [2] the text classification process is separated in six steps which are Document Collection, Pre-Processing, Indexing, Feature Selection, Classification and Performance Evaluations.

The multivariate Bernoulli model and the multinomial Bernoulli model are described in [3]. The authors describe the details and the differences of this two models by empirically comparing their classification performance on five text corpora. By comparing their results, the authors find out that the multivariate Bernoulli performs well with small vocabulary sizes, but the multinomial Bernoulli usually performs even better on larger vocabulary sizes.

In [4], the authors compare discriminative and generative learning as typified by Logistic Regression and Naïve Bayes. To evaluate both classifier they have tested them on 15 datasets. They have shown that even though the generative model has higher asymptotic error than the discriminative model, the generative model may reach its (higher) asymptotic error much faster than the discriminative model. Thus, one may choose the generative Naïve Bayes if the training dataset is low in numbers, but as the training examples are increased, one would choose the discriminative Logistic Regression.

Nearest Centroid (when used for text classification with tf-idf vectors it is known as Rocchio) is extensively studied and analyzed in [5]. The performed experiments by the authors show that Nearest Centroid substantially outperforms other algorithms such as Naïve Bayesian, K-Nearest-Neighbors, and C4.5, on a wide range of datasets.

An extensive tutorial on the basic ideas behing Support Vector Machines (SVMs) is given in [6]. The tutorial starts with an overview of VC (Vapnik-Chervonenkis) dimension. Then, they describe linear SVMs for separable and non-separable data. Later on, in detail it is discussed the kernel mapping technique which is used to construct non-linear SVMs. At the end of this tutorial, the limitations are given and explained by the authors.

In [7] it is given a detailed description of the Stochastic Gradient Descent (SGD) algorithm. The authors present SGD algorithms for a number of classic machine learning schemes such as Perceptron, K-Means, and SVM. Furthermore, they discusses the asymptotical efficiency of estimates obtained after a single pass over the training set. And, at the end, empirical evidences are presented.

The process of collecting a large text collection of Macedonian news documents (103,637 classified in 6 categories), by crawling them from news archive and boiler plate removal, is presented in [8]. Furthermore, the authors conducted experiments on this dataset using state-of-the-art text classification algorithms such as Naïve Bayesian, K-Nearest-Neighbors, Decision Trees, Neural Networks and Ensembles of classifiers. The experiments show that the Ensembles of Naïve Bayesian classifiers achieves best results (77.8%) with a slight improvement over the single Naïve Bayesian classifier (75.6%).

III. Dataset Creation

Our process of document collecting and dataset creation consists of three steps. The first one is the web pages crawling. Then comes the process of extracting the wanted news text from the crawled HTML documents (boilerplate removal). And, at the end, is the process of combining the news text into categories.

A. Crawling

The dataset was comprised of news text that were obtained from 5 different Macedonian news portals. Firstly, we have gathered the HTML documents that contained the news text by crawling the news portals one by one, category by category, news article by news article. For the task of crawling the Macedonian news portals, we have developed a multithread crawler application written in the programming language Python. At any time, there were 5 threads opened that downloaded the HTML documents, where each thread sends a HTML request to a given site that contains the news text, receives it and writes it to a file. The files were organized in folders, where each folder corresponds to a given category.

B. Boilerplate removal

When the crawling of the web pages was done, we had to extract the news texts and the news titles from the HTML documents. For this task, extracting the desired content and the boilerplate removal (removing the unnecessary content and the HTML tags), we have used a Python library called BeautifulSoup [9]. We have extracted the news text articles from the crawled HTML documents by identifying the HTML tags that contains the news text and the news title and taking out the content using BeautifulSoup.

C. Dataset Categories

After the news texts and the news titles are extracted from the HTML documents, we had to organize them into adequate categories. Since the news that were crawled from the Macedonian news portals, were previously categorized in some classes by the news portals themselves, we only had to merge all of the news texts from the same category. The whole dataset consists of 417,412 news texts classified into 13 categories: Macedonia, World, Economy, Scene, Health, Culture, Technology, Life, Football, Basketball, Handball, Tennis and MotoSports. Table I displays the whole dataset with all of its categories and the number of news texts in them.

TABLE I
CATEGORIES

Category	# of News
Macedonia	161,671
World	125,542
Economy	41,685
Scene	22,839
Health	9,462
Culture	12,247
Technology	5,420
Life	14,931
Football	12,187
Basketball	5,283
Handball	3,596
Tennis	1,988
MotoSports	561

IV. Pre-Processing The Texts And Feature Vector Extraction

Since texts can not be directly interpreted by a classifier, we need a pre-processing step that maps document (text) d_i into compact representation of its content [1]. We have pre-processed the texts by first cleaning the punctuations marks from every text, then tokenizing the texts and finally removing the stop words. For the task of tokenizing the texts and cleaning the punctuations marks we have used a regular expression that splits every text into a list that contains only the words from the text. The list of the stop words that we removed from the texts is composed of the roughly 100th most frequent words from the dataset. With these process, every text was represented by a vector containing its words.

We wanted to try out and compare the classification results from different feature vector representations of the texts. These are the different representations that we tried:

1) For every term (here term is word in that text) in every text, we have calculated its weight using the TFIDF function. Later on, we will refer to this representation with WOL (Without Lemmatization).

2) First, every word was lemmatized. Then, for every term (here term is lemmatized word) we have calculated its weight using the TFIDF function. Later on, we will refer to this representation as WL (With Lemmatization).
3) From every text, only the most frequent words in that text were picked for the feature vector. Then, for these terms (here term is one of the most frequent words) their weight was calculated using the TFIDF function. We have picked the 7th, 14th and 21st most frequent words. Later on, we will refer to this representations as WOL_First_7, WOL_First_14 and WOL_First_21.
4) First, every word was lemmatized. Then, only the most frequent terms (here term is lemmatized word) in that text were picked. For these terms, their weight was calculated using the TFIDF function. We have picked the 7th, 14th and 21st most frequent words. Later on, we will refer to this representations as WL_First_7, WL_First_14 and WL_First_21.

The TFIDF function is defined as:

$$tfidf(t_k, d_j) = \#(t_k, d_j) * \log \frac{|T_c|}{\#_{T_r}(t_k)}, \quad (1)$$

where $\#(t_k, d_j)$ denotes the number of times term t_k occurs in document d_j, and $\#_{T_r}(t_k)$ denotes the document frequency of term t_k, that is the number of documents in T_r in which t_k occurs [1].

So now, every text was represented with 8 different feature vectors and we had 8 different representations of the dataset.

V. Text Classification Algorithms

We will now give a brief theoretical background of the text classification algorithms that we use for automatic text classification of the Macedonian news in our experiments. These include Naïve Bayes, Logistic Regression, Nearest Centroid (Rocchio), Suppport Vector Machines and Logistic Regression and SVMs with Stochastic Gradient Descent.

A. Naïve Bayes

The Naïve Bayes classifier belongs to the Bayesian family of classifiers which make strong assumptions about how the data is generated and assume a probabilistic model that embodies these assumptions. Naïve Bayes classifier also assumes that all attributes of the examples are independent of each other. This is the so-called "Naïve Bayes assumption" [3].

The probability of a document d with a feature vector x, being in class c_k is computed as:

$$P(c_k|x) = \frac{P(c_k) * P(x|c_k)}{P(x)}, \quad (2)$$

where $P(x)$ is identical for all classes and therefore can be ignored. We can estimate $P(c_k)$ and $P(x|c_k)$ from a training set using a *maximum likelihood estimate* (MLE) [10]. So for $P(c_k)$ we have:

$$P(c_k) = \frac{N_{c_k}}{N}, \quad (3)$$

where N_c is the number of documents in the class c_k, and N is the total number of documents in the training set. For the estimate of $P(x|c_k)$ we have:

$$P(x|c_k) = \prod_{i=0}^{n_d} \frac{T_{c_k x_i} + 1}{T_{c_k} + |V|}, \quad (4)$$

where $T_{c_k x_i}$ is the number of occurence of the term x_i in class c_k, T_{c_k} is the total number of terms in class c_k. We use Laplace smoothing to overcome zero probabilities. New examples are classified in the category with maximum $P(x|c_k)$.

B. Logistic Regression

Discriminative classifiers model the posterior $P(y|x)$ directly, or learn a direct map from inputs x to class labels, in contrary to generative classifiers which learn a model of the joint probability $P(x, y)$ [4]. The Logistic Regression classifier belongs to the discriminative family of classifiers.

Logistic Regression relies heavily on the logistic function $g(t)$ for the discriminant function

$$g(t) = \frac{1}{1 + e^{-t}}. \quad (5)$$

The discriminant function of the Logistic Regression is defined as:

$$h(z) = \frac{1}{1 + e^{-z}}. \quad (6)$$

where z is a linear combination of the input feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and the parameter weight vector $\mathbf{w} = (w_0, w_1, w_2, \ldots, w_n)$:

$$z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n. \quad (7)$$

The probabilities $P(Y = 1|z)$ and $P(Y = 0|z)$ are:

$$P(Y = 1|z) = h(z), \quad (8)$$

$$P(Y = 0|z) = 1 - h(z). \quad (9)$$

Let us consider a loss function $l(h(z), y)$, that measures the cost of predicting $h(z)$, when the actual answer is y. The empirical risk

$$E(h) = \frac{1}{m} \sum_{i=1}^{m} l(h(z_i), y_i), \quad (10)$$

measures the training set performance. In order to calculate optimal values for **w**, we need to minimize the empirical risk. The empirical risk can be minimized using gradient descent [7].

C. Nearest Centroid (Rocchio)

The Nearest Centroid classifier is based on the Rocchio relevance feedback algorithm [11], [12]. In this classifier, the documents of each class are represented by a centroid (prototype) vector, and new document is assigned to the class that corresponds to its most similar vector, as measured by the cosine function [5].

Every document $d \in D$ is represented as vector $\overrightarrow{d} = (x_1, x_2, \ldots, x_n)$. The centroid vector, $\overrightarrow{c}$, for each class is calculated as:

$$\overrightarrow{c} = \frac{1}{|D_c|} \sum_{\overrightarrow{d} \in D_c} \overrightarrow{d}, \tag{11}$$

where D_c is the set of documents in class c. The similarity between a document and a centroid vector is calculated using the cosine function [5]

$$\cos(\overrightarrow{d}, \overrightarrow{c}) = \frac{\overrightarrow{d} * \overrightarrow{c}}{||\overrightarrow{d}|| * ||\overrightarrow{c}||}. \tag{12}$$

So every new document d' is classified to the class that corresponds to its most similar centroid:

$$\arg\max_{\overrightarrow{c} \in C} \cos(\overrightarrow{d}, \overrightarrow{c}). \tag{13}$$

D. Support Vector Machines (SVMs)

Support Vector Machines were proposed by Vapnik in [13], and in [14] it is shown that linear SVMs are among the most accurate classifiers for text categorization.

A linear SVM is a hyperplane that separates a set of positives examples, from a set of negative examples with maximum margin. This hyperplane insures that

$$y_k * (w^T * x_k + w_0) \geq 1 \qquad k = 1, \ldots, n, \tag{14}$$

where (x_k, y_k) for $k = 1, \ldots, n$ is the training data and $y_k \in -1, 1$. In order to maximize the margin, we need to minimize

$$\frac{1}{2}||\overrightarrow{w}||^2, \tag{15}$$

subject to constraint

$$y_k * (w^T * x_k + w_0) \geq 1 \qquad k = 1, \ldots, n. \tag{16}$$

For the purpose of text classification, in our experiments we have used an implementation of SVMs called *LIBSVM* [15] with linear kernel.

E. Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is an optimization technique that uses approximate gradients estimate from subsets of the training data and updates the parameters in an online fashion [16].

Let us consider a basic supervised learning step where each sample is a pair (x_k, y_k) composed of input vector x_k and output vector y_k. Let us consider a loss function

$$l(f_w(x_k), y_k) \tag{17}$$

that measures the cost of predicting $f_w(x_k)$, which is discriminant function parameterized by the weight vector **w**, when the actual answer is y. The empirical risk

$$E(f) = \frac{1}{m} \sum_{i=1}^{n} l(f_w(x_i), y_i), \tag{18}$$

measures the training set performance. In order to find a optimal value for the weight vector **w**, we need to minimize the empirical risk $E(f_w)$, thus we use SGD which in each iteration estimates the gradient of $E(f_w)$ on the basis of a single example (x_k, y_k) and updates the weights according to

$$w_{r+1} = w_r - \alpha * \frac{\delta}{\delta w} l(f_w(x_k), y_k), \tag{19}$$

where w_{r+1} are the new weights.

VI. Experiments & Results

For every text classification algorithm that we briefly explained in the previous section, these include Naïve Bayes, Logistic Regression, Nearest Centroid (Rocchio), SVMs and Logistic Regression and SVMs with SGD, we have conducted experiments using 6-fold cross validation for training a model and measuring its results. The training and the evaluating of each model was done using the previously built dataset and its 8 different representations.

To evaluate, test and compare the results from the classifiers we decided to use three measures: precision, recall and F1 score. Even though, in [1] there are many other evaluation criteria suggested, we decided to use these because they are among the most popular evaluation criteria for text classification. Precision is the proportion of items placed in the category that are really in the category, recall is the proportion of items in the category that are actually placed in the category and F1 score is a combination of precision and recall that gives a single global measure of performance [8], [14].

From the conducted experiments we concluded that the SVMs classifier achieved the best results. Its best results were achieved both on the WL and WOL representations of the dataset, with precision of 0.90, recall 0.90 and F1 score of 0.90. When training on the WL_First_7, WL_First_14, WL_First_21 and on the WOL_First_7, WOL_First_14, WOL_First_21 representations of the dataset, the performance of the SVMs dropped, with precision of 0.85, recall of 0.85 and F1 score of 0.85 on the WL_First_21 and WOL_First_21, precision of 0.81, recall of 0.82 and F1 score of 0.81 on the WL_First_14 and WOL_First_14 and on the representations WL_First_7 and WOL_First_7 it achieved precision of 0.70, recall 0.70 and F1 score of 0.68.

The Logistic Regression classifier achieved the second best results. Its best results were achieved on the WOL representation of the dataset with precision of 0.89, recall of 0.90 and F1 score of 0.89. On the WL representation of the dataset it achieved precision of 0.88, recall of 0.89 and F1 score of 0.88. Its perfromance dropped significantly when training on the WL_First_7, WL_First_14, WL_First_21 and on the WOL_First_7, WOL_First_14, WOL_First_21 representations of the dataset. It achieved precision of 0.83, recall of 0.84 and F1 score of 0.83 when training both on the WL_First_21 and WOL_First_21. On the WL_First_14 and WOL_First_14 it achieved precision of 0.80, recall of 0.81 and F1 score of 0.80. And on the WL_First_7 and WOL_First_7 it achieved precision of 0.69, recall of 0.69 and F1 score of 0.69.

SVMs and Logistic Regression with SGD both achieved precision of 0.88, recall of 0.89 and F1 score of 0.88

TABLE II
RESULTS FROM THE EXPERIMENTS

	NB			LR			NC			SVMs			LR with SGD			SVMs with SGD		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
WL	0.87	0.86	0.86	0.88	0.89	0.88	0.79	0.78	0.78	0.90	0.90	0.90	0.88	0.89	0.88	0.88	0.89	0.88
WOL	0.87	0.87	0.87	0.89	0.90	0.89	0.80	0.79	0.79	0.90	0.90	0.90	0.88	0.89	0.88	0.88	0.89	0.88
WL_First_7	0.70	0.69	0.70	0.69	0.69	0.69	0.55	0.41	0.46	0.70	0.70	0.68	0.63	0.51	0.55	0.65	0.55	0.58
WL_First_14	0.81	0.80	0.80	0.80	0.81	0.80	0.72	0.63	0.67	0.81	0.82	0.81	0.73	0.70	0.71	0.73	0.71	0.72
WL_First_21	0.84	0.84	0.84	0.83	0.84	0.83	0.77	0.70	0.73	0.85	0.85	0.85	0.77	0.78	0.77	0.79	0.79	0.79
WOL_First_7	0.69	0.68	0.68	0.69	0.69	0.69	0.69	0.40	0.48	0.70	0.70	0.68	0.68	0.67	0.65	0.69	0.68	0.66
WOL_First_14	0.81	0.82	0.81	0.80	0.81	0.80	0.74	0.61	0.65	0.81	0.82	0.81	0.79	0.79	0.78	0.79	0.80	0.78
WOL_First_21	0.84	0.84	0.84	0.83	0.84	0.83	0.78	0.69	0.72	0.85	0.85	0.85	0.82	0.83	0.82	0.83	0.83	0.83

on the WOL and WL representation of the dataset. There is big drop of performance for both SVMs with SGD and Logistic Regression with SGD when training on the WL_First_7, WL_First_14, WL_First_21 and on the WOL_First_7, WOL_First_14, WOL_First_21 representations of the dataset. These performance are presented in Table II.

The Naïve Bayes classifier achieved its best performance on the WOL representation of the dataset. It achieved precision, recall and F1 score of 0.87. On the WL representation of the dataset it achieved precision of 0.87, recall of 0.86 and F1 score of 0.86. The scores from the WL_First_7, WL_First_14, WL_First_21 and from WOL_First_7, WOL_First_14, WOL_First_21 representations of the dataset are presented in Table II.

The worst performance were achieved by the Nearest Centroid (Rocchio) classifier. It scored a precision of 0.80, recall of 0.79 and F1 score of 0.79 on the WOL representation of the dataset. On the WL representation of the dataset it score precision of 0.79, recall of 0.78 and F1 score of 0.78. The scores that Rocchio achieved on the WL_First_7, WL_First_14, WL_First_21 and on the WOL_First_7, WOL_First_14, WOL_First_21 representations of the dataset are reported in Table II.

The training time of the SVMs was approximately 62,000 second (17 hours) and its testing time was 11,000 seconds (3 hours) on the WL and WOL representation of the dataset. The training and testing times of the Logistic Regression on the WL and WOL representations of the dataset were a lot faster with training of 1,200 seconds and testing 10 seconds. The training time of both the Naïve Bayes classifier and the Nearest Centroid classifier were approximately 70 seconds and their testing were approximately 7 seconds. Time needed for training for the SVMs and Logistic Regression with SGD was 180 seconds and time needed for testing was 100 seconds.

Table II summarize the results from the conducted experiments.

VII. CONCLUSION & FUTURE WORK

In this paper we presented the task of building a text classifier for the Macedonian language. Firstly, we presented the process of dataset creation. Then, we described the preprocessing techniques that we apply to the dataset and the process of feature vector extraction. And, at the end, we presented the results from the conducted experiments. From the experiments we concluded that the SVMs classifier has the highest precision of 0.90 and Logistic Regression, with its precision of 0.89, is just around the corner. However, the training time of the SVMs, which is 17 hours, is far greater than the training time of the Logistic Regression, which is 20 minutes.

In future work, we plan to conduct experiments with Deep Neural Networks and compare its results with the results from the conducted experiments in this paper. We also plan to try out some more advanced feature vector extraction techniques and dimensionality reduction algorithms.

REFERENCES

[1] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[2] Bhumika, S. Sukhjit, and N. Anand, "A review paper on algorithms used for text classification," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 2, pp. 90–99, 2013.

[3] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[4] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[5] E.-H. S. Han and G. Karypis, "Centroid-based document classification: Analysis & experimental results," 2000.

[6] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[7] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[8] M. Saveski, I. Trajkovski, and J. Pehcevski, "Classification of macedonian news articles," *World*, vol. 18, pp. 2–122.

[9] L. Richardson, "Beautiful soup," *Crummy: The Site*, 2013.

[10] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," vol. 1, 2008.

[11] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." DTIC Document, Tech. Rep., 1996.

[12] J. J. Rocchio, "Relevance feedback in information retrieval," 1971.

[13] V. N. Vapnik and V. Vapnik, "Statistical learning theory," vol. 1, 1998.

[14] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*. ACM, 1998, pp. 148–155.

[15] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[16] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 477–485.

Chalenges, Benefits and Expectations from Implementing an E-health System in Macedonia

Ristana Mitrovska, Veno Pachovski, Adrijan Bozhinovski
School of Computer Science and Information Technologies
University American College Skopje
mitrovskif@hotmail.com, pachovski@uacs.edu.mk, bozinovski@uacs.edu.mk

***Abstract*—The goal of this study is to analyze the conditions of implementing e-Health in the health system in Macedonia. In every country, health is national-scale issue and governments are looking for solutions to improve the health state of its citizens. Information technology in the Macedonian health system has been rapidly developing in the passing years and in 2013 and 2014 the first e-services for citizens and healthcare institutions (HCI) were implemented. But this doesn't mean that the e-Health system is completed. To achieve this, a national integrated health information system (HIS) is required. Centralized electronic health records (EHR) must be implemented in this system, as well as electronic medical records (EMR) and personal health records (PHR) of citizens, all the while taking into account their privacy and safety of information.**

Keywords—e-Health, electronic health card, electronic health record, healthcare system

I. INTRODUCTION

Technological progress and the Internet created new opportunities and challenges to the traditional health care information technology industry. E-health characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology [1].

There are many definitions for term and concept of e-Health, but all they cover the usage of ICT in health products, services and processes [2], which demand organizational change in healthcare systems, as well as new skills, in order to improve the health of citizens, efficiency and productivity in healthcare delivery, and the economic and social value of health. E-Health covers the interaction between patients and health-service providers, institution-to-institution transmission of data, or peer-to-peer communication between patients and/or health professionals. Health information about patients has to be transferred in digital form, transmitted, stored and retrieved electronically, for clinical, educational and administrative purposes, both at the local site and at a distance. An E-Health system may be divided to major application areas: a) public health policy and prevention, b) information services to citizens, c) integrated patient management and patient health records, d) telecare and independent living services [3].

HCIs meet with a large amount and complex data of health-related information, and a major component is their information processing. Their ICT system increasingly depends on information and communication technologies and these technologies support progress in medical research, better management, and diffusion of medical knowledge, and a shift towards evidence-based medicine. E-Health tools support the aggregation, analysis and storage of clinical data in all of their forms, while information tools provide access to the latest findings, and communication tools enable collaboration among many different organizations and health professionals [4].

People have greater awareness about their health and the health of their family and they expect quality and fast services from the health system. As can be seen from Fig. 1, [5] the number of people who use the Internet for find some information about health is growing from year to year.

In the past several years Macedonia has started building a new healthcare system, but without having adopted any national policy or strategy for it. Therefore, in this study, an attempt will be made to allocate the most pressing steps and activities that need to be taken for the implementation of this system in Macedonia, using the experiences of other countries and information from the current status of the healthcare system.

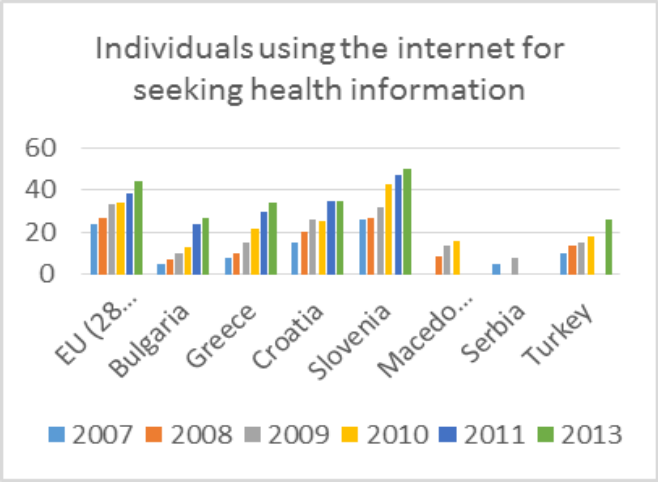


Fig. 1. Individuals using the Internet for seeking health information

II. THE CURRENT STATUS OF ICT IN MACEDONIA

In Macedonia, ICT in the healthcare system can be analyzed at three levels [6]:

- Primary HCIs (General Practitioners-GPs and pharmacies), all in private ownership. They use software from more than ten different providers on different platforms. They all have Internet connections and there are rare cases of extensive networks;
- Secondary HCIs (specialist outpatient clinics), in private and public ownership. They use software from different providers and different platforms. Public HCIs have some network architecture;
- Tertiary HCIs (hospitals) are predominantly in public ownership (there are less than ten in private ownership). The situation is similar with the secondary HCIs because most of them are giving secondary and tertiary healthcare services.

But are the information systems and the available infrastructure in HCI (public and private) ready to fulfill the technical conditions for implementing e-Health? HCI IT systems are being developed by various companies or in-house IT-departments. The integration of different health information systems has to provide administrative support, health insurance management, code tables, data access and privileges, privacy and security management etc.

Large and various data from health services given to patients are collected every day in HCIs. Data can be differentiated as medical data, administrative data and financial data. Only a small part of medical data is in electronic form, while most of them are still in paper form. The results from a number of analyses (laboratory, radiology, etc.) are usually transmitted from the source to the specialist/GP in written form. Many times the same analysis is performed repeatedly because some of the analyses are lost or unavailable.

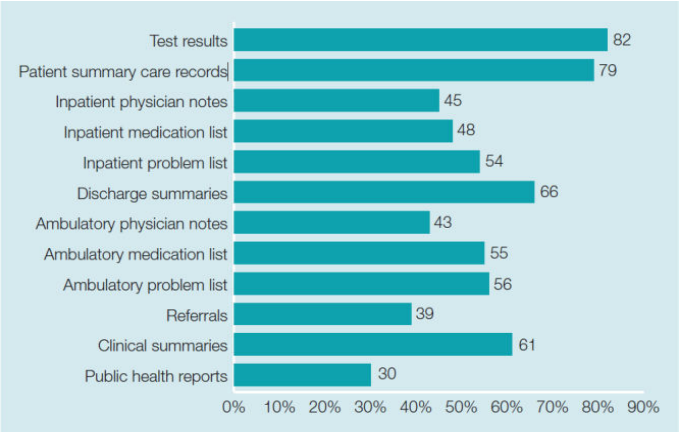


Fig. 2. Percentage of operational HIE efforts exchanging various types of data [7]

Discharges and results from secondary and tertiary healthcare services in practice are still unavailable in electronic form for GPs. Fig. 2 shows some results by analyzing the exchange of various types of data in healthcare systems.

Today in a healthcare system, during the visit to the doctor, patients must carry healthcare insurance booklet. One part of data from health examinations, prescriptions, diagnosis, orthopedics devices are recorded manually in the booklet by GPs, pharmacies and health insurance fund. The other data from specialists and tertiaries examinations and services are recorded manually on different paper patterns. At this way most of medical data is only available for healthcare workers in specifics HCI. Software which is used in HCI records medical data for financial and accounting purposes.

These results from examinations need to be digitally stored in one central place, organized in electronic health records and electronic medical records, and thus become available to other health workers who would have the privilege to use them. This will save time to establish the right diagnosis, because the doctors will have access to historical data for their patient. Also this can reduce a number of doctor visits and hospitalizations.

Introducing the electronic health card (EHC) in the healthcare system in the Republic of Macedonia has opened the gates for implementing e-services for its citizens. Electronic communications between the health information systems, such as the Health insurance fund of Macedonia (HIFM), Ministry of health (MH) and HCI, require the introduction of the EHC on the central level, as a sort of personal ID for the citizen. By December 2014, more than 85% of the citizens in Macedonia have an EHC or more than 1.6 million EHCs have been distributed to citizens from almost 1.8 million insured people in the country [8].

Carriers of the integration of the health systems at the national level should be the Health insurance Fund (HIF) and the Ministry of Health (MH) of the Republic of Macedonia. In this direction, they implemented two major projects: E-prescriptions and E-referrals by the MH and E-doctor, E-contracts and E-invoice by the HIF.

In the primary healthcare level, all 2.335 GPs are working with E-prescriptions from GPs to pharmacies and E-referrals from GPs to secondary healthcare institutions. With service E-prescriptions GPs are prescribing drugs to patients but with parallel use of paper prescription. All 716 pharmacies are using this service for issuing drugs. On a monthly basis in 2014, GPs had prescribed more than 1.5 million prescriptions [9]. Service E-referrals is used in 746 secondary HCIs. With this service GPs are referring patients to upper healthcare level with parallel use of paper referral too. But results from examinations are not evaluable for GP at this point of service. There are 69 tertiary HCIs with 347 units which have to be included in the E-health

system in the near future. Medical data from hospitals have to be available to GPs and specialists.

In 2013, the HIF had introduced E-services for a chosen doctor. Today, more than 98% of GPs who have an EHC use this service. With this service patients choose a personal doctor in health offices and this information is automatically recorded in HIFMs electronic system. In January 2014, the HIF signed electronic contracts with 3.208 public and private health care institutions, where a digital signature from EHC, was used by an official from the health care institution to sign the contract. Since January 2014, the HIF introduced service E-invoice to all 3.208 HCIs. Every invoice from a HCI is electronically signed by an official from the HCI with the EHC and every month financial and medical data from health services, given to insured persons, are transferred in electronic form to the HIF system. This service concerns financial and insurance data.

Introducing of these e-services doesn't mean that in Macedonia e-health is fully implemented. Data which are collected on central level by these services are not organized as patient record. These services are the first step of implementation of e-health. It's necessary collected patient's data on central location to be available for all healthcare providers. Next steps of implementation are described in the following section.

III. System Design and Implementation

Implementation of an e-Health system must be maintained by health and non-health stakeholders. Success in implementing e-Health is heavily dependent on having the continued support and guidance of stakeholders [10].

A system implementation of a health system considers several components such as: policy maker and financing, implementing electronic health records (EHR), electronic medical records (EMR), and personal health record (PHR), interoperability of ICT solutions and standardizations of health data exchange, ensuring privacy and safety.

The design principles for achieving integrated healthcare information systems relate to the principles of flexibility, expansion from the installed base through modular strategies, and identification of key healthcare actors to provide them with immediate benefits [11].

A. *Policy, Financing and Security*

In time when there is no accepted strategy for the development of an e-Health system, it is essential to build a national platform aimed to bring necessary regulatory and financial environment. Managing and monitoring of the implementation of the system should be coordinated on a central level. This central body should have a regulatory status to enable standardization of communication and data collection. Because of the nature of the data, privacy and safety of health data has to be covered technically and by law. Implementing such a system demands substantial financial support from the government and other financial sources.

Medical data is mainly confidential and patient rights for secrecy of medical data are protected by law [12]. Using this system, patients must be assured that security of health data should be provided by the system.

On Fig. 3, the major key stakeholders of the implementation of the e-health are shown: MH responsible for political support, HIF responsible for implementation, a Regulatory Body responsible for interoperability and standardization and HCIs.

Successful implementation of an e-health system depends by its managing. Because this has to be a national project, it is necessary to be monitored and managed by one center. This project needs full political and financial governmental support. HCI's must invest in modernization of their software, hardware and network infrastructure.

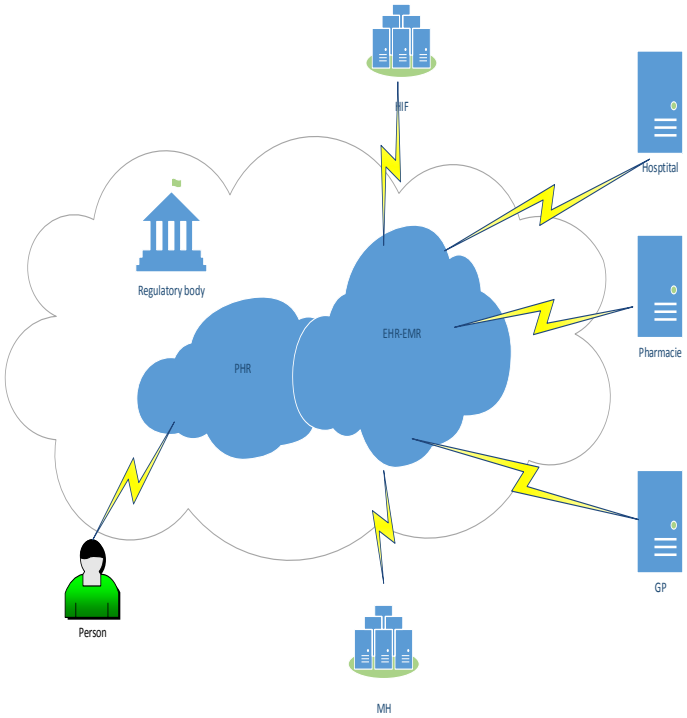


Fig. 3. Stakeholders of e-Health in RM

B. *Electronic Records*

Digitalization of patients' health data in electronic way has to enable easy exchange between healthcare providers. The purpose of collecting and storing patient health data is to make it available for decision-making. It is important to realize that patients sometimes have more than one care point, a primary general care provider and specialist care providers. It is essential

that patient data be available at all points of care and all sites for analysis. Distributed HCI networks should have the ability to collect and exchange data mutually and with the HIF, MH and other relevant institutions.

An electronic medical record (EMR) is a real-time patient health record with access to evidence-based decision support tools that can be used to aid clinicians in decision-making. The EMR can also support the collection of data for uses other than clinical care, such as billing, quality management, outcome reporting, and public health disease surveillance and reporting. EMRs are said to be interoperable if they are able to exchange data using standardized data transmission formats. The electronic health record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. The EHR automates and streamlines the clinician's workflow. It has the ability to generate a complete record of a clinical patient encounter – as well as supporting other care-related activities directly or indirectly via an interface – including evidence-based decision support, quality management, and outcomes reporting. A third type of record and one which is controlled and managed by the citizen is the personal health record (PHR). The PHR is owned, managed, and shared by the individual or his or her legal proxy(-ies) and must be secure to protect the privacy and confidentiality of the health information it contains. It is not a legal record unless so defined and is subject to various legal limitations [13]. Interoperable electronic health record systems have to be capable of storing structured results and transmitting those results to other care providers [14].

HCIs have different medical devices and equipment which produces results in different formats (text, video, pictures etc.) and all this data has to be uniformed and collected in one central system.

C. Standardization and Interoperability

Integration of health information system depends of integration of health information system at all three levels: primary, secondary and tertiary and their interconnection. This process requires defining and adopting ICT standards of interoperability, standards for electronic cards and health related standards.

Standardization and interoperability ensures that the two systems understand the data in the same way. The most common messaging standard used in the EU is the Health Level 7 International (HL7). Founded in 1987, the Health Level Seven International (HL7) is a not-for-profit, ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services. A Clinical Document Architecture (CDA) is a document markup standard that specifies the structure and semantics of "clinical documents" for the purpose of exchange between healthcare providers and patients. It defines a clinical document as having the following six features: 1) Persistence, 2) Stewardship, 3) Potential for authentication, 4) Context, 5) Wholeness and 6) Human readability [15]. ISO TC 215 is standardization in the field of health informatics, to facilitate the coherent and consistent interchange and use of health-related data, information, and knowledge, to support and enable all aspects of the health system [16].

An integration of heterogeneous information system can be reached with their standardization and interoperability. Interoperability has to be defined, supported, maintenance and implemented on national level. This will lead to changing current healthcare process and services and developing new. It should be considered that the HCIs have a different level of information systems and thereby transparency of the process is limited. Investing in ICT in healthcare system is strategic investment and has to be realized in stages.

However, experience [17] shows that implementation and integration of a healthcare system is challenged by:

- lack of awareness of, and confidence in e-health solutions among patients, citizens and healthcare professionals;
- lack of interoperability between e-Health solutions;
- limited large-scale evidence of the cost-effectiveness of e-Health tools and services;
- lack of legal clarity for health and wellbeing mobile applications and the lack of transparency regarding the utilization of data collected by such applications;
- inadequate or fragmented legal frameworks, including the lack of reimbursement schemes for e-Health services;
- high start-up costs involved in setting up e-Health systems;
- regional differences in accessing ICT services, limited access in deprived areas.

IV. CONCLUSION

The benefits of an e-Health system should be felt by citizens and healthcare professionals. Easy access to relevant health data for the patients will incur savings for not having to make analyses and examinations that were once made. Indirectly, this will reduce costs in healthcare institutions. Aggregate data collected in one place give an opportunity to do analysis for various diseases, diagnoses and the general health status of the population. Using ICT can increase health service efficiency and will provide equality in healthcare system.

This system will need strong financial and legal support. Such a robust system can be successfully implemented only with strong national support with power to enforce all necessary activities. In a situation when MH and HIFM have started

separately few e-services in the health system, the process of implementation should be continued and headed by a national regulatory body. This body has to develop programs for education, training and raising awareness of key participants.

V. REFERENCES

[1] G. Eysenbach, "What is e-Health?" Journal of Medical Internet Research, 3(2), e20, p. 3, 2001.

[2] European commission, "E-health action plan 2012-2020-innovative healthcare for the 21st century", 2012. [Online]. Available: http://ec.europa.eu/digital-agenda/en/news/ehealth-action-plan-2012-2020-innovative-healthcare-21st-century. [Accessed 21.02.2015].

[3] E. Christodoulou, A. Dunbar, P. Gaspar, R. Jaksa, K. Krapez, "The Development of eHealth in an Enlarged EU: Synthesis Report," Office for Official Publications of the European Communities, Luxembourg, 2008.

[4] European commission, "e-Health - making healthcare better for European citizens: An action plan for a European e-Health area," 2004. [Online]. Available: https://ec.europa.eu/digital-agenda/en/eu-policy-ehealth. [Accessed 21.02.2015].

[5] European commission, Eurostat 2015 [Online]. Available: http://ec.europa.eu/eurostat/. [Accessed 21.02.2015].

[6] Law of healthcare protection, Official Gazette of the Republic of Macedonia Nb. 43 from 29.03.2012.

[7] J. Adler-Milstein, A. K.Jha, "Health Information Exchange under HITECH: Progress and Challenges," in Health Information Technology in US: Better Information Systems for Better Care, Robert Wood Johnson Foundation, pp. 44-56, 2013.

[8] Health Insurance Fund of Macedonia, Realisation of the Project "Electronic health card" (in Macedonian) 2014 [Online]. Available: http://www.fzo.org.mk/WBStorage/Files/Soopstenie%20za%20EZK%2026-06-2014.pdf. [Accessed 21.02.2015].

[9] Health Insurance Fund of Macedonia, "Semi-annual report on prescription drug use from the list for drugs for primary health care PHC (January-June 2014 year)", (in Macedonian) 2014 Available: http://www.fzo.org.mk/WBStorage/Files/Polugodisen%20izveshtaj%202014.pdf. [Accessed 21.02.2015].

[10] WHO, "National e-Health strategy toolkit - establishing a national e-Health vision," Geneva, 2012.

[11] T. Bleging, "Desing principles for achieving integrated healthcare information systems," Denmark Health Informatics Journal, vol. 19, no. 1, pp. 29-45, 2013.

[12] Law for protection the rights of patients, Official Gazette of the Republic of Macedonia Nb. 82 from 08.07.2008.

[13] World Health Organization, Management of patient information Trends and challenges in Member States - Based on the findings of the second global survey on eHealth, 2012. Available: http://apps.who.int/iris/bitstream/10665/76794/1/9789241504645_eng.pdf?ua=1. [Accessed 21.02.2015].

[14] M. Samwald, J. A. Minarro-Gimenez, K. Blagec, K.-P. Adlassnig, "Towards a Global IT System for Personalized Medicine: the Medicine Safety Code Initiative," Studies in Health Technology and Informatics, vol. 198, pp. 25-31, 2014.

[15] Health Level Seven International, About HL7, 2015 [Online]. Available: http://www.hl7.org/about/index.cfm?ref=common. [Accessed 21.02.2015].

[16] International Organization for Standardization, ISO/TC 215 Health informatics, 1998 [Online]. Available: http://www.iso.org/iso/iso_technical_committee?commid=54960. [Accessed 21.02.2015].

[17] Commission of the European Communities, "e-Health - making healthcare better for European citizens: An action plan for a European e-Health Area", 2004, [Online]. Available: http://www.epsos.eu/uploads/tx_epsosfileshare/EHealth-Action-plan_01.pdf. [Accessed 06.03.2015].

Publishing Skopje Air Quality Data as Linked Data

Kostadin Mishev[1], Angjel Kjosevski[1], Nikola Kalemdzhievski[1], Nikola Koteli[1], Milos Jovanovik[1], Kosta Mitreski[1], Dimitar Trajanov[1]
[1]Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia

Abstract— **Publishing raw data as Linked Open Data gives an opportunity of data reusability and data understandability for the computer machines. Today, the air pollution problem is one of the biggest in the whole world. Republic of Macedonia, especially its capital Skopje, has big problems with the PM2.5 and PM10 particles in the air approved by several measurement stations positioned on several locations in Skopje. In this paper, we demonstrate the process of centralizing of all the data collected from different measurement stations in one database. Also, we enable interpolation of collected data providing information about the current air quality state in the area between the measurement stations using previously implemented eco models. Interpolated data is saved in the same database providing interfaces that transform saved data into four-star and five-star data, by reusing the existing ontologies from the domain and linking them to the physical places where the measurements were taken and the interpolations were calculated. As a use case scenario, we provide and heat map about the values from various pollutants in the areas in Skopje providing information about the regions that have problems with air pollution.**

Keywords— air quality, indicator, measurements, measurement interpolation, ecoinformatics, open data, linked data

I. INTRODUCTION

Linked Open Data and Semantic Web principles are the main contributors in realizing the idea of data reusability and data scalability [1][2]. It gives the opportunity of linking the information from various fields and enabling simple access to them. With this approach, data becomes understandable not only for humans, but also for computer machines [3].

On the other hand, the air pollution in Macedonia, especially in its capital Skopje, is one of the biggest problems that the citizens of Skopje have. It can be described as the pollution of the atmosphere with gases, or dust of solid materials, particulate matter as other substances whose amounts are constantly increasing [4]. This information is approved by the multiple measurement stations positioned on several locations in Skopje and its environment. Some of them offer public domains for data access.

In this paper, we demonstrate the process of transforming the collected data from several measurement stations into four- and five-star Linked Open Data. We aggregate all the data collected from the services provided by stations of The Ministry of environment and physical planning, The Institute of Eco informatics at Faculty of Computer Science and Engineering, and the CO_2 measurements provided by the project Skopje Green Route, into one centralized database. They provide different air quality indicators. Afterwards, we transform the collected data into four- and five-star data by reusing the existing ontologies from the domain, necessary for the transformation and annotation process and linking them to the physical places where the measurements were taken [5].

The network of monitoring stations is very important in urban environment because it provides information of actual quantity of air quality indicators. The main problem is impossibility of obtaining appropriate values for all points of interest. There are several reasons and as most important we can mention the price of the measurement stations. Consequently, only the most important points could be monitored. Air dispersions models are solving this problem by providing estimations and predictions of the pollutants in the air using mainly emissions and meteorological data. These models include on mathematical algorithms based on combinations of physical and chemical simulating the spread of pollutants in the air. We describe the process of interpolation of the air quality data obtained from the measurement stations. This process of interpolation is repeating on constant time intervals to obtain approximate values of all air quality parameters [6]. The number of measurement stations is upgradable and it is directly proportional with the accuracy of the approximations.

In the final section, we propose and demonstrate several use-case scenarios querying published data set and represent the results on a heat map providing information about the current pollution from various pollutants in the area of Skopje.

II. RELATED WORK

The problem with air pollution is emerging almost all urban cities in the world. It is estimated that worldwide, 2 million people and more than half of them are in developing countries, die every year from air pollution. By releasing of the air indicators from measurement stations as open data provide contribution of the public understanding and dialogue around far-reaching and potentially data-rich aspects of life in the city. Consequently, air quality measurement datasets are already part of the LOD cloud. Home Weather ontology is intended for weather phenomena and exterior conditions providing property hasAirPollution which express an index of air pollution depending on the current air quality measurement values (Fig. 1).

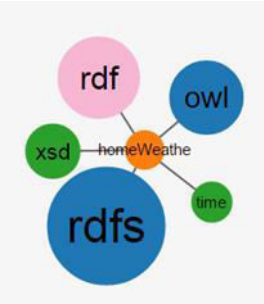


Figure 1. HomeWeather dataset

AirQuality+ is a project[1] which gathers real-time air quality measurements from different points in Sheffield, England, providing open licenses for communities and organizations to access and re-use. This includes near real-time data on pollutant levels in Sheffield collected by a network of monitoring stations, as well as data related to the issue of air quality, such as industrial activity, traffic and transport, public health, weather and land use.

The PESCaDO Ontology[2] is a modular application ontology exploited for personalized environmental decision support, that enables to formally describe:

- the user decision support request
- the environmental data relevant to process the request
- the decisions and conclusions to be produced

The PESCaDO Ontology was thoroughly developed following state of the art best practices, and it is matched with a comprehensive and detailed documentation.

Air dispersion models are based on mathematical algorithms providing probabilistic values about the current air quality at each point of the city. They are related to the city infrastructure, the current weather conditions and the real-time measurements from several measurement stations. Currently, the Institute of Eco informatics has developed air dispersion models about the capital of R. Macedonia, Skopje. They use the measurements from The Ministry of environment and physical planning providing interpolation values for each point in the central region. These data is kept as 1 star data providing interpolated monochromatic visualizations (Fig. 2) on Skopje's map, generated from ArcGIS server. In this paper, we will convert the information from this visualization into 5 star data and we will provide useful information mining the gathered dataset.

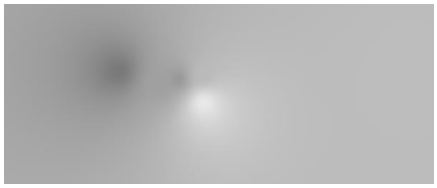


Figure 2. CO interpolated monochromatic visualization

[1] http://betterwithdata.co/portfolio/air-quality-plus/

[2] https://dkm.fbk.eu/technologies/pescado-ontology

III. Data from Skopje's Air Quality Measurement Stations

A. *Centralizing data from multiple services*

There are multiple measurement stations distributed in the region of Skopje providing different air quality indicators. Most of them, provide open access REST services which return real-time measurements. In our paper, we will use the services provided by:

- Ministry of environment and physical planning

The JSON service which provides measurement about CO, NO_2, PM10, PM2.5, SO_2, O_3 air quality indicators providing data refresh each hour during the day. There are multiple measurement stations over Macedonia, but for purposes of this paper, we will use only the stations located in Skopje: Centar, Karpos, Lisitche, Gazi Baba and Rektorat.

- Measurement CO_2 stations provided by the project Skopje Green Route

These measurement stations are placed on the most frequent crossroads in Skopje: Justice Palace, Red Cross and Faculty of Agriculture, providing measurement about CO2 air indicator refreshing the information each 5 minutes.

- Measurement station maintained by the "Laboratory of Eco informatics at Faculty of Computer Science and Engineering" providing information about the same air indicators like the measurements stations enabled by Ministry of environment and physical planning

All services are RESTful and provide open URL location which can be accessed with GET parameters. The log of all services is kept on our database which centralizes all information about all air quality indicators for all measurement stations (Fig. 3). It runs scheduled processes which poll the JSON services asking for new fresh data from sensors. It appends timestamp to the measurement information and saves in MySQL database whose EA diagram is represented on Fig. 4.

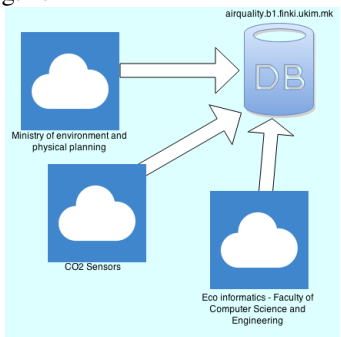


Figure 3 Centralized data polling architecture.

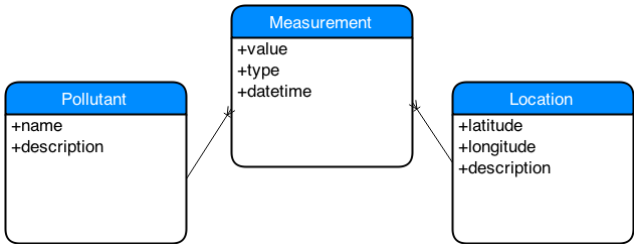


Figure 4. EA diagram of the centralized database.

B. *Interpolation of the measurements in the area of Skopje*

The process of interpolation is based on the newest "up to time" data as the average of the air parameter concentration per hour for each parameter. The data are provided by the network of measurement stations described in section A stored in one centralized database. They are used for generation of the grid raster layers by interpolating techniques. The model depends on the weather conditions and the infrastructure of the area taking as references the values from the nearest real weather and pollution measurement stations. The interpolation data is calculated on ArcGIS Simulation Server which implements the pollution model [6], gets the measurements from weather and pollution stations and provide interpolated information about the area of Skopje. The output of the model is a raster image, so we create algorithm for data transformation from raster monochromatic image to numerical format about the pollution state. The darker positions represent greater values of pollution. After the transformation process, we provide RESTful services which could be easily accessed by setting the latitude and longitude of the required position. This RESTful services is accessed by the interfaces of our application. They append appropriate timestamp and current weather conditions, and save in the centralized database.

C. *PESCaDO Ontology*

In order, to transform the measurement 3 star data, from the centralized database, into RDF, we need ontology. Among multiple ontologies that we reviewed in our research, the PESCaDO Data ontology proved as the most useful for our needs. It is developed by Data & Knowledge Management research group [7] which is part of the Information and Technology Center in Fondazione Bruno Kessler and it is provided for mapping of measurement data from sensors of PM10, PM2.5, CO and other air pollutants. It also provides mapping of the weather conditions so we concluded that this ontology is satisfying our needs.

We divided the properties of the ontology in two main types: weather conditions properties (Table 1) and air pollution indices (Table 2).

Table 1. Weather condition properties

Property	Type	Description
HumidityValue	Datatype Property	Air humidity value
TemperatureValue	Datatype Property	Ambient Temperature value
WindSpeedValue	Datatype Property	Wind speed value

Table 2. Air pollution indices properties

Property	Type	Description
PM10IndexValue	Datatype	Provides information about the value of the PM10 particles in the air
PM2.5IndexValue	Datatype	Provides information about the value of the PM2.5 particles in the air
COIndexValue	Datatype	Provides information about the value of the CO concentration in the air
NO2IndexValue	Datatype	Provides information about the value of the NO_2 concentration in the air
SO2IndexValue	Datatype	Provides information about the value of the SO_2 concentration in the air
O3IndexValue	Datatype	Provides information about the value of the O_3 concentration in the air

D. *Geo Ontology*

To provide mapping of the geographical location of the measured or interpolated instance, we used the Geo ontology which is one of the most used (Fig. 5). We have the correct positions of the static measurement stations and we link the measurement with the exact position of the station. Afterwards, we divide the area of Skopje in zones providing interpolated information about each zone separately. Each zone has own latitude and longitude enabling linking to the appropriate instance from the Geo ontology [8].

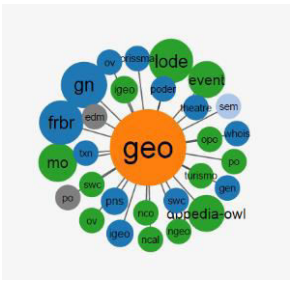


Figure 5. Geo ontology.

Table 3. Geo ontology

Property	Type	Description
Location	ObjectType	Description of the geographic entity
Lat	Datatype	Latitude of the mapped object
Lon	Datatype	Longitude of the mapped object

E. *Mapping the data from 3-star to 5-star data*

After defining the ontologies, we need to transform the data saved in database to RDF. In order to accomplish this, we

decided to use D2RQ server[3] which is compatible with MySQL databases and provides accessing relational databases as virtual, read-only RDF graphs without replicating into an RDF store. Using D2RQ we can query a non-RDF database using SPARQL, access to the content of the database as Linked Data over the Web, create custom dumps of the database in RDF formats for loading into the RDF store and provides access to a non-RDF database using the Apache Jena API.

The mapping process consisted of two steps. The first step provides wrapping of the relational database, in our case MySQL, with the interfaces provided by the D2RQ providing access to the stored data. Afterwards, we should define a mapping file using the D2RQ mapping language[4] to map the relational database schemas to RDF vocabularies and OWL ontologies. The mapping file defines a virtual RDF graph that contains information about the database. This graph contains RDF terms using d2rq:ClassMaps and d2rq:PropertyBridges. The class map specifies how URIs (or blank nodes) are generated for the instances of the class. It has a set of property bridges, which specify how the properties of an instance are created.

Our database, referencing to Figure 4, stores information about Pollutant, Location and Measurement. As the image represents, it is designed in 3rd normal form. As defined in section D, we use PESCaDO and Geo ontologies so we need to decompose the database in 2nd normal form providing the table pollutant be part of the measurement. To solve this problem, we change the mapping configuration using the D2RQ mapping language, so we need not to make any changes in the model of the relational database only by using the property d2rq:condition. The property d2rq:condition provides the SQL WHERE condition so an instance of this class will only be generated for database rows that satisfy the condition.

```
map:measurements_CO a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern "measurement/@@T_MEASUREMENT.id@@";
        d2rq:class pescadoData:COIndexValue;
        d2rq:join "T_MEASUREMENT.pollutant_id =>
T_POLLUTANT.id";
        d2rq:condition "T_MEASUREMENT.pollutant_id = 3";
        d2rq:propertyDefinitionLabel "Measurements CO";
```

Transforming the data to 5 star data is provided by the geo:Location property linking the measurement information to specified Location where it is measured or interpolated. So, in the global graph, we provide air pollution indicator for the specified location.

F. USE CASE EXAMPLE

In this section we will demonstrate that transformation of data into Linked Data, can provide useful use-case scenarios. The result of use-cases gives opportunities for visual representation of the pollution on a heating map caused by all pollutants separately.

By following query, we can obtain information about CO measurements for the area of Skopje in determined time:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pescado: https://ontohub.org/fois-ontology-competition/PESCaDO_Ontology/pescadoData.owl#
PREFIX prov: <http://www.w3.org/ns/prov#> .
SELECT DISTINCT ?lat ?lng  ?value WHERE {
?s rdf:type pescadoData:COIndexValue;
rdf:value ?value;
prov:atLocation ?location;
prov:generatedAtTime "2015-03-03T19:15:46"^^xsd:dateTime.
?location geo:lat ?lat.
?location geo:lng ?lng.
}
```

This query starts executing over the local RDF graph providing the measured and interpolated measurements in a determined time from the area of Skopje. This query returns similar data shown on the table 4:

Table 4. Partial result from the SPARQL query

Lat	Lng	Value
"42.05"^^xsd:fkiat	"21.32"^^xsd:fkiat	0.3
"41.96"^^xsd:fkiat	"21.31"^^xsd:fkiat	0.33
"41.94"^^xsd:fkiat	"21.29"^^xsd:fkiat	0.37
"42.03"^^xsd:fkiat	"21.3"^^xsd:fkiat	0.31

The result from the query could be used as input of a heat map obtaining the visual representation of the concentration of the CO in the air of the area of Skopje. The results from the previous query are represented on Figure 5.

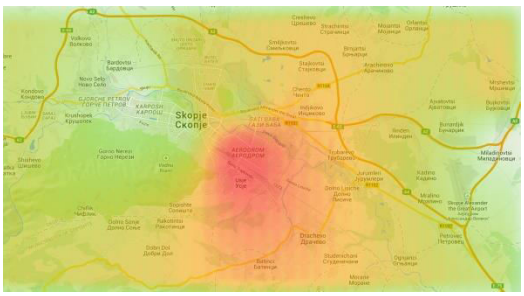


Figure 6. Visual heatmap representation of the CO measurement in the area of Skopje.

[3] http://d2rq.org/d2r-server
[4] http://d2rq.org/d2rq-language

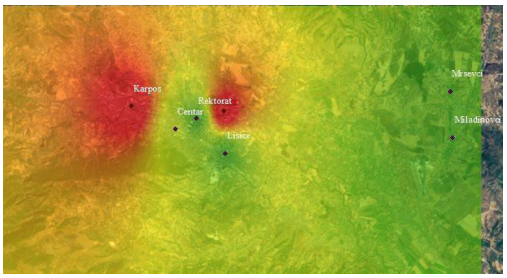


Figure 7 Visual heatmap representation of the PM10 measurements in the area of Skopje.

Analyzing the results on the heat map, we can conclude that the municipality of Aerodrom has the highest values of CO pollution.

The SPARQL endpoint for reviewing and analyzing the results of measurements is available on the following url:

http://airpollution.b1.finki.ukim.mk/

IV. Conclusion

The concept of Linked Data represents a big advantage in representation and retrieval of structured data from distributed parts of the Web. A large number of communities, companies and other interested stakeholders are taking part in the initiative and are contributing to the expansion of the LOD Cloud.

The type of the data that we contribute to open and to link is providing interesting analyzes about the current state of the air in the area of Skopje. We are allowing measurements about CO, CO_2, SO_2, O_3, NO_2, PM10 and PM2.5 air pollutants. The measurements are provided by 7 air pointer stations and 3 CO_2 measurement stations. We provide interpolated values for the areas that are not covered by the measurement stations. Interpolated values are created by sophisticated models of pollution spreading taking as parameters the infrastructure and the model of spreading of the appropriate pollutant. We save all of the data, measurements and interpolated values, in centralized database that is wrapped by D2RQ server providing mapping to RDF triples and linking to appropriate locations in Skopje. We provide a URL for accessing the data and reviewing the results using SPARQL query.

This type of data can help the citizens to find the best places for their activities or for living in Skopje. Also it can be used for retrieving the best eco routes for travelling in Skopje.

REFERENCES

[1] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, “Linked data on the web,” 17th International conference on World Wide Web, ACM, 2008, pp. 1265-1266

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - the story so far," International Journal on Semantic Web and Information Systems 5, no. 3, 2009

[3] A. Naeve “The Human Semantic Web, Shifting From Knowledge Push To Knowledge Pull”, International Journal on Semantic Web and Information Systems (IJSWIS), 2005

[4] L. Barandovski, V. Urumov “Air pollution studies in Macedonia using the moss biomonitoring technique, NAA, AAS and GIS technology”, INIS, 2006

[5] M. Oprea “Mapping Ontologies in an Air Pollution Monitoring and Control AgentBased System”, Lecture Notes in Computer Science Volume 4265, 2006, pp 342-346.

[6] N.Koteli, K. Mitreski, D. Davcev “Monitoring, Modeling and Visualization System of Traffic Air Pollution – A Case Study for the City of Skopje”, International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2014

[7] V. Epitropou, L. Johanson, K.D. Karatzas, A. Bassouckos, A. Karppinen, J. Kukkonen, M. Haakana, “Fusion of Environmental Information for the Delivery of Orchestrated Services for the Atmospheric Environment in the PESCaDO project”, International Environmental Modelling and Software Society (iEMSs), 2012

[8] A. Patil, S. Oundhakar, A. Sheth, K. Verma, “Meteor-s web service annotation framework”, Proceedings of the 13th international conference on World Wide Web, pp 553 - 562

Improving learning in mathematics through 3D digital game play

Jovan Shalamanoski
Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, Macedonia
shalamanoski.jovan@students.finki.ukim.mk

Biljana Risteska Stojkoska
Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, Macedonia
biljana.stojkoska@finki.ukim.mk

Abstract— **Teaching mathematics is a challenging goal for primary and secondary math teachers. Learning through game has been proven to increase kids desire to acquire new knowledge.**
This paper presents the design and implementation of a three dimensional (3D) educational game that aims to help pupils to improve their mathematical skills. The game is simple to use and easy to navigate, but at the same time the kids feel rich game environment, which makes the experience better. To increase availability, the game requires simple installation and can be played directly on the web browser.
Game evaluation was made in elementary school environment. Results of the evaluation showed that kids that played the game have shown interest in learning math with this educational game. Their experience was positive; they liked the learning through game playing, which made the learning process more competitive and fun.

Keywords— Game-Based Learning; 3D Education Game; Digital Game-Based Learning; Mathematics; Primary Education; Secondary Education

I. INTRODUCTION

Most of the kids have difficulties in learning particular mathematical concepts. In order to learn and master those concepts, a lot of practice and dedication is needed, which is hard to be accomplished. Children quickly lose interest for solving mathematical problems and want to do something else, something easier and more appealing. On the other side, children spend most of their free time on their tablets, smart phones, or in front of their computers, playing their favorite games.

There are different ways to stimulate learning. For student at university, project based learning has been proved to increase students' interest for the learning course material, especially for engineering subjects [1]. Project-based learning is learner centered. Students have a significant role in selecting the content areas and nature of the projects that they do. There is considerable focus on students understanding what it is they are doing, why it is important, and how they will be assessed.

Recently, there has been an ongoing interest in the use of computer games for educational purposes. The studies found that video games can impact positively on problem solving skills, i.e. they can improve engagement and motivation. This learning approach, which combines digital game-based activities and educational content, is often referred to as Digital Game-Based Learning (DGBL) [2].

Video games are powerful tool for teaching, especially with their interactive nature [3]. They create virtual worlds that cannot be done in reality, and allow us not only to tell the story, but also to experience the story and learn from it. Hence, creating and designing video games is not a trivial task. Moreover, generating game that is educational and entertaining at the same time is a challenging task. DGBL constitutes a relatively recent and still evolving learning approach that needs more empirical evidence in order to be applied more effectively.

The goal of our project was to combine mathematical problems with something more entertaining like playing video games. We aimed to find and show solution that will help kids to solve mathematical problems or review mathematical problems learned in school while spending their time on video games.

The most challenging problem was to make the smooth experience transition from educational to entertainment game, i.e. to make the academic content integral to the game rather than an add-on. Solution for that question is important because that is the way how kids can accept educational game, which have to give them fun experience that they are used to while playing entertainment commercial games. When we are trying to solve this problem, we should explore how to include educational content and activities in the games so they will remain fun and attractive. If the transition is fast, for example, simple implementation of educational methods in entertainment games, kids will recognize the game as educational math game and will discard it as unattractive. Therefore, it is very important the transition to be done correctly, to be slow and not noticeable by the players, i.e. the kids.

This paper presents the design and implementation of 3D digital game which helps pupils from elementary and secondary school to increase their knowledge in mathematics. In this paper we propose a slow transition approach that uses already existing game mechanics and adapts it for educational goals for learning mathematics. Game mechanics [4] are usually methods, actions and rules which motivate the player

to play the game. Collecting items (coins, diamantes, etc.), players' health, magic power, avoiding obstacles, etc. are only a few examples of game mechanics. This approach were used to create our educational game for learning mathematics.

The paper is organized as follows. The second section provides an overview of different solution for learning through game play. The third section provides implementation details. Section four evaluates the game. This paper is concluded in section five.

II. Related work

Part of this idea is not new and there are other educational games for learning math as well. The major problem with most of the existing educational games for learning mathematics is that the graphics, story, game content, details of the characters and other details, which are very important for keeping the attention and the interest of the player, are not done correctly. That is the reason why kids who will start playing that games lose their interest after short period of time and start to play more appealing entertainment commercial games. Kids always choose better graphics, details and games with better story.

Most of the games for learning mathematics are web based short games. Some of them are Cheese Capers [5], Bus Math Multiplication [6] and Feed Fribbit Addition [7]. Most of the other web base games for learning mathematics look very similar to the abovementioned games.

There are many mobile applications for learning mathematics. Graphics in those games are better than graphics in web based games. Still, there are a lot of improvements that can be done in order to make them more entertaining and with better functionalities. The content of these games should be improved, enhanced with better and more useful math's problems and examples that are synchronized with the lessons from school. Examples of popular mobile math's games that worth looking at are Math Training for Kids[8], Dinosaur Park Math[9] and Mathly Hollows[10].

Other interesting game for learning mathematics is Timez Attack [11]. This game is a good example how a fun educational games should look like. However, the story and the graphics for Timez Attack need to be improved in order to match the quality from good entertainment games.

There is nice, carefully created, educational project from Rovio, the company that is the creator of the globally successful Angry Birds[12], called Angry Birds Playground [13]. This is a good example how a popular entertainment games can be used in education.

The difference between our game and other games for learning mathematics is that we let the player to experience the fun like playing entertainment game. At the same time, the part of the game where the player solves math`s problems is nicely integrated with the story of the game, so it feels like it`s part of the game story. We also tried to adapt the educational content to be more useful, i.e. math problems to be more similar to those learnt at school. We believe that it will increase the motivation to solve the math problems in the game.

III. Game design and implementation

This game fits in the category of small games and can be played on browser with Unity player and standalone on Windows, Mac and Linux. When playing the game on browser, there is a background image that fits in the mood of the game play, providing better playing experience.

More details about the design of this game can be obtained from Fig. 1, which actually presents the activity diagram.

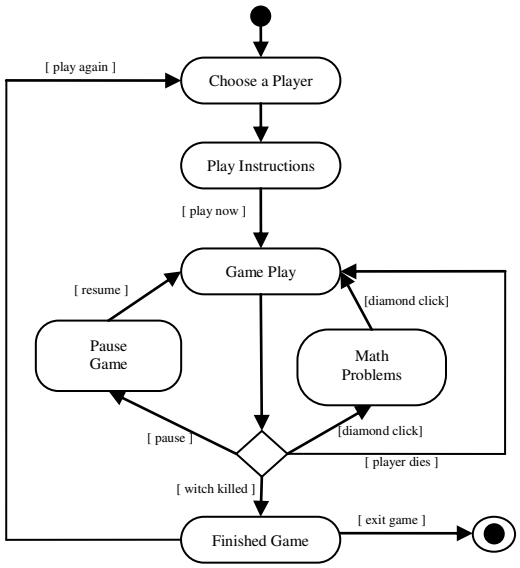


Fig. 1. Activity Diagram

For creating this game, Unity3d [14] game engine was used to bring all together and to build the final application. Photoshop [15] was used as a supporting tool for image editing, while 3ds max was used for creating 3d models and animations [16]. Most of the techniques we used are extensively described in [17].

Type of this game is platformer, i.e. it involves guiding an avatar to jump between suspended platforms (or over obstacles) to advance the game. The player controls the jumps to avoid letting the avatar fall from platforms or miss necessary jumps. This game is a 3d game, but the camera is positioned like 2d platformer to make the game easier to play and navigate. For this game we created simple story so the kids have motivation to play the game. The story is about a wizard who fights against an evil witch using magic powers. Hence, the game has a goal that need to be accomplished.

The first option that shows up when the game is started, is to choose a type of player. There is an option to choose to play as a boy or as a girl (Fig. 2). Characters in the game are very important as game mechanics. Allowing players to choose different type of game characters and customizing them makes players engaged and identified with the game character. This is a small detail which helps player to get more attached to the game.



Fig. 2. Game start screenshot

After choosing the game character, simple explanation is provided in order to introduce the players with the game rules Only few controls are used while playing the game, which makes the game easier for playing.

After the game starts, the player can enjoy playing nice designed game and feel like any good entertainment game. There are standard game options for pause and restart the game. The player can enjoy special effects sounds. Background classical music makes nice ambient and helps player to feel the quality of classical music.

In order to defend from the enemies, the player uses magic power to attack and destroy the enemies. Each time the player uses magic, the magic power goes down and the only way to get more magic power is to click on the diamond icon, which brings new state where player can focus on solving a math's problem (see Fig. 3), and for reword gets a magic power. With current settings of the game, player gets 10 magic power for each solved problem which allow him/her to use the magic for 10 times. This setting can be adjusted.

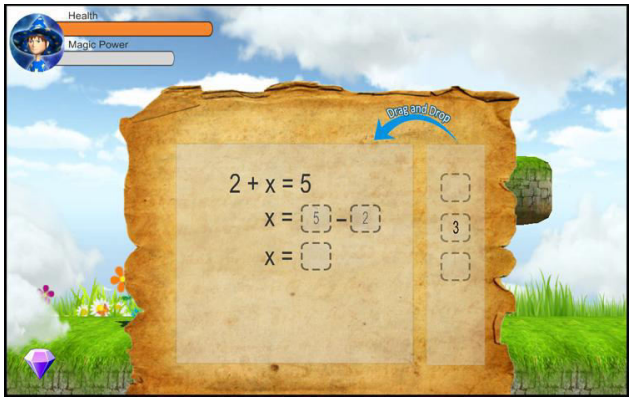


Fig. 3. In-game Math Screenshot

In this way, the player discretely is forced to solve the math's problems in order to achieve the game`s goal. It is also important to mention that the appearance of the educational content and activities in our game is very similar to the game mechanics in entertainment games and they are connected smoothly with the story of the game. Educational contents should be included in that way so the player feels like they are part of the gameplay.

Another important thing to mention is that educational game should contain fun, entertainment part of a game play, because if the game is full of activities for solving problems, it may not be appealing for the player. So the fun part of the games must exist and as addition to it, educational activities to be included. This is the strategy we used in our game.

The graphics for games are very important, because graphics are first things that player sees and according to them he/she decides whatever to play the game or not. No matter how good the game mechanics are, if graphics of the game are with low quality there are big chances that the player will quit the game. Keeping that in mind we tried to create nice graphics for our game (Fig. 4 and Fig. 5). Graphics are combination of 2d and 3d graphics. Most of the environment (clouds, grass, flowers, etc.) is 2d graphic, while the platforms and the characters are 3d graphic. For the magic spell actions we used 3d particle system.

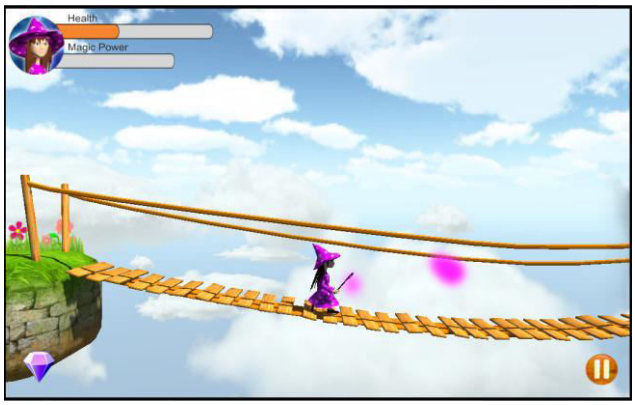


Fig. 4. In-game screenshot for female player

Two different animations were used in the game. One type is the environment animation like "dark clouds disappearing" animation and "witch explosion" that happens in the end of the game. Other type of animation is the character`s animation [18]. We tried to work on details on these animations, because they are what players see most of the time while playing the game. Character animations were split into smaller animations like character actions (run, jump, magic spell, attack, etc.). Each of these parts were carefully animated and imported into the game engine with an appropriated sound, i.e. when the animation is active, the corresponding sound is played.



Fig. 5. In-game screenshot for male player

IV. Evaluation

In order to validate the actual educational value of our game, we performed a simple evaluation.

The evaluation was done in two schools, one private (QSI International School) and one public school (Vera Ciriviri Trena) in Skopje, Republic of Macedonia. In order to evaluate this project, 45 students (22 males and 23 females) were asked to play the game for 15 minutes. Only one student, a girl, refused to play the game. All students were enrolled in fifth and sixth grade in primary school.

The first impression of the students about the game was positive. They were excited about the game and the idea. All students had the opportunity to play the game and after that, they were asked to fill a questionnaire.

Students have answered the following questions:

1) Do you like the game?
2) Is the game easy to play?
3) Can this game help in studying mathematics?
4) After playing the game, do the math's problems like that in the game look more fun and easy for you?
5) Do you have some ideas to make the game better?

The results of the evaluation (see Fig. 6) showed that kids who played the game have shown interest in learning math skills with this educational game. Their experience was positive; they liked the learning through game playing, which made the learning process more competitive and fun.

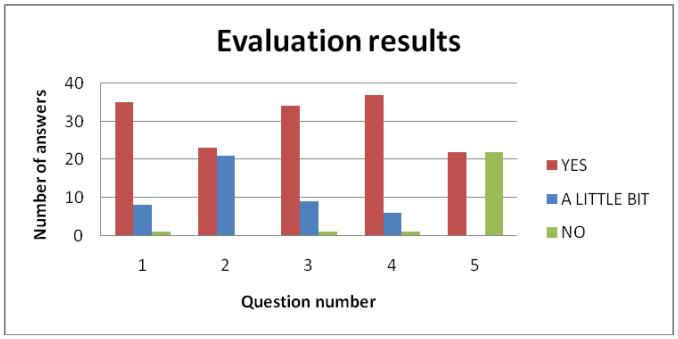


Fig. 6. Results from the questionnaire

V. Conclusion

This paper presents the design and implementation of 3D digital game which helps pupils to increase their knowledge in mathematics. We adopted existing game mechanics from commercial games to make slow transition from entertaining to educational game.

In the future, we plan to improve and extend this game with: diversity of rewards and achievements, implementation of timers, levels, multi-player gaming, etc. These extensions should be carefully integrated within the game, in order to keep the amusing component dominant in the game.

Additionally, we intend to implement educational games for high school pupils and university students. In this case, the mathematical concepts must be explained correctly and without hiding details. This process should include extensive consultations with math teachers. Moreover, the game should be frequently updated to take pace with the changes in the course syllabus.

References

[1] Davcev Danco, Biljana Stojkoska, Slobodan Kalajdziski, and Kire Trivodaliev. "Project based learning of embedded systems." In *Proc. 2nd WSEAS Int. Conf. Circuits, Syst., Signal Telecommun*, pp. 120-125. 2008.

[2] Kiili, Kristian. "Foundation for problem-based gaming." *British Journal of Educational Technology* 38.3 (2007): 394-404.

[3] U.Ritterfeld, R.Weber, "Video Games for Entertainment and Education". In P. Vorderer, J. Bryant (Eds.), Playing Video Games-Motives, Responses, and Consequences ,pp. 399-413, Mahwah, NJ: Lawrence Erlbaum, Inc. 2006

[4] Fabricatore Carlo, "Gameplay and Game Mechanics: A Key to Quality in Videogames" in Proceedings of OECD-CERI Expert Meeting on Videogames and Education, Santiago de Chile,Chile, 2007.

[5] Cheese Capers
www.learninggamesforkids.com/math_games/multiplication/math-capers.html accessed (Feb.2015)

[6] Bus Math Multiplication
http://www.learninggamesforkids.com/math_games/multiplication/bus-drivers-math-multiplication.html accessed (Feb.2015)

[7] Feed Fribbit Addition
www.coolmath-games.com/0-feed-fribbit-addition
accessed (Feb.2015)

[8] Math Training for Kids
play.google.com/store/apps/details?id=com.divmob.eastermath.game.GameActivity&hl=en accessed (Feb.2015)

[9] Dinosaur Park Math
itunes.apple.com/us/app/dinosaur-park-math/id531951016?mt=8
accessed (Feb.2015)

[10] Mathly Hollows
itunes.apple.com/us/app/mathly-hollows/id571479644?mt=8
accessed (Feb.2015)

[11] Timez Attack
www.bigbrainz.com,accessed(Feb.2015)
www.youtube.com/watch?v=WqPLAEDDr58
accessed (Feb.2015)

[12] Vilhelmiina Harju and Jari Multisilta,"Angry Birds for Fun in Learning", Finnish Innovations and Technologies in Schools, Part III, pp. 69-76 , 2014

[13] Angry Birds Playground
www.rovio.com/en/our-work/learning accessed (Feb.2015)

[14] Unity3d www.unity3d.com/ accessed (Feb.2015)

[15] Photoshop http://www.photoshop.com/

[16] 3ds Max www.autodesk.com/products/3ds-max/overview

[17] L.Matthias, A.Mariebeth, K.Katharina , G.Florian, K.Stephanie, P.Reinhold, "Content Creation for a 3D Game with Maya and Unity 3D" CESCG 2011

[18] Gregory Francis, "Game Character Animation Design" CSRS 2009

Semantic Analysis and Data Visualization of Songs from Different Languages

Goran Petkovski, David Zafirovski, Blagoja Evkovski, Usamedin Nuhiju, Dimitrij Mijoski
University Ss. Cyril and Methodius
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia

{petkovski.goran, zafirovski.david, evkovski.blagoja, nuhiju.usamedin, mijoski.dimitrij}@students.finki.ukim.mk

***Abstract*—Accessing the music on the Internet in order to obtain some information for the songs has been a truly challenge in the past decade. Thanks to the new emerging web-technologies, searching and analyzing songs is accessible to the users or developers as never before in the past 50 years. With the vast amount of data available on the Web, this field has initiated the development of data management techniques for distributed data sets over the existing infrastructure of the Web. This leads to new challenges, how to get the precise information about favorite songs or how to compare different songs from many authors on wide time range in real time. In this paper we present a methodology for analyzing the songs based on the lyrics. First, we demonstrate crawling of different web documents containing song information, parsing the HTML code and retrieving the relevant information from the documents. Then after retrieving appropriate information for the songs, we store the collected data in our database. Once finished, the algorithm search for similar songs according on their semantic relevance, finds all the matching keywords with well-known text-mining technique TF-IDF and provides data visualization of the results. In addition, we also evaluate the proposed results through web application interface and some of the use case scenarios.**

Keywords*—semantic analysis, sentiment analysis, information retrieval, web crawler, text mining, unstructured information, TF-IDF, data visualization

I. Introduction

Technology has always been a major tool in improving the quality of life for people. By lowering the barrier for publishing and accessing documents, the Web has been the innovation which changed the way we communicate, as well as the way we gather and share knowledge. Semantic indexing and retrieval has become an important research area, as the available amount of information on the Web is growing more and more. However, the original design of the Web has been intended for human consumption only, so in order to obtain and analyze larger amounts of data, intelligent software tools are needed. This brings the problem of selecting the right information in right time from the massive data sets available on-line and making the information accessible and easy to use for the users [1].

Thanks to the modern advanced web technologies we can handle with large information sets and extract some conclusion about our data with minimal technical experience. The technology we used offer tools for searching and retrieval of music information from the songs, making categorization of different songs into categories based on genre, year, language or country and provide complex analysis and statistics. This can result with intelligent software agents who understand the data and produce practical and precise methods for articulating the meanings of words and sentences, and for revealing connections between language and culture. For the purpose of understanding our data, Cliff Goddard has proposed breadth methods suitable for basic semantic methods extraction [2].

In order to find relevant keywords from our music database, we need to apply techniques for extraction of textual data embedded in our song lyrics, with the best world wide methods for analyzing unstructured information. These methods belongs to one major research field, called Text Mining, in which there are several related topics (such as artificial intelligence, knowledge discovery and data mining). This field emphasize predictive methods, including key areas of text mining, such as: preprocessing, text categorization, information search and retrieval, clustering of documents and information extraction [3]. This is essential in our application, because we want deeply overview of our music textual data, based on different conditions including: year, genre and language.

The main goal we want to accomplished is comparison on the sentiment score from different songs in various time-periods over history and bring some conclusions over the results. This can be: In which period are the most happy songs, What are the most common words in the songs from World War period, Which is the most positive genre or negative. Another interesting part is the part with "Word Cloud", simple overview of the most relevant keywords in every song genre. And at the end we want to develop a user interface presented to the user, based on framework for building rich data visualization frameworks.

In this paper, we present a sustainable system and its methodology for obtaining and analyzing music related data from various public music web-sites, storing the information in our database, and provide example web application with interactive graphs and charts which demonstrate the data mining techniques over the music data.

The paper is organized as follows: in Section 2 we provide detailed overview of our System Architecture, containing pseudo-codes, description, pictures and summary of our implementation. Then in Section 2.1 we specified the implementation of some parts of the program: crawler, semantic analysis and TF-IDF algorithm, and after that the results are presented

in Section 3. And finally in Section 4 we conclude the research of this paper and discuss possible future work in this area.

II. System Architecture

The system is divided into multiple stages of executions:

- Retrieval of song information (genre, year, country of origin, language, album) from AllMusic.com with our custom implementation of web-crawler.
- Generate corresponding classes and store the music related data in our database.
- Analyze the saved songs with text mining methods including TF-IDF and text clustering.
- Provide useful web interface, with interactive graphs and charts for data visualization of results.

These are the main steps of the system methodology, further more in below sections we describe detailed implementation of every applications part.

A. Implementation of web-crawler

In order to have information about our songs (artist name, genre, year, country, album, song lyrics), we implemented our crawler that searches over the Internet for all the song information and retrieve most important parts. The crawler is simply described as a web crawler (because it's searching on the Web for HTML documents) and the steps of execution are shown below:

- First, user enters the artist name in the text field.
- Then, check if the artist already exists in our database, if so reject the request.
- If artist is not found in the database, search for the songs and albums of the artist and lookup up for the HTML documents where that information is containing.
- Parse the HTML code from the web-document, extract the required information and store it to the database.
- Repeat these steps while all songs from the artist are not evaluated.

For this to be done, we used the free open-source Java library available on the Internet, JSoup, for parsing the HTML code and retrieve the most important information from the HTML document. This JSoup Java libary, has very good implementations of finished functions with CSS Selectors, who first search all the HTML tags on the document and compare it against the CSS selector rules provided in the function [4]. If the resulting tag was found, we extract the information from it, including: text, number, date, hyperlink or image. If the extracted information was hyperlink to another web-site URL, we follow that link to find additional information, if not we store the given result to the database.

B. TF-IDF Implementation

After we find all the songs from the specified artist, we start with text-mining of the music textual data (song lyrics). For this to accomplished we used the wide know algorithm for text-mining, TF-IDF, also called as term frequency-inverse document frequency. This algorithm has been proved as very reliable and extremely precise in finding the relevant keywords in one text document, since the famous search engine Google used as top in the first implementation of their search engine. In short terms, it presents a numerical value of how important is word to text document in collection or text document corpus [5]. It is often used as a weighting factor in information retrieval and text mining. For our implementation in the application, we provide algorithm with these steps of execution:

- First we count the number of times a term occurs in the document (song lyrics), this is called *term frequency (TF)*.
- Then, we define *document frequency (DF)* which defines the number of documents in which a term appears (song lyrics from the same genre).
- However, because some words are common, so we must give the term statistical weight for measuring the importance of a term in a text document collection. This is called *inverse document frequency (IDF)*.

Because our musical data was big enough to analyze the term and document frequency on each song, we only analyzed the songs from the same genre. We were able to determine what are the most relevant keywords in genre "pop/rock" or "classic" and make some comparison on the keywords.

```
float relevanceOfWord(String term, String
    document, List<String> documents)
{
/* calculate TF  */
int occurence= term_occurence(term, document);
/* calculate IDF */
int relevantDocuments = 1 + relevantDocuments(
    term, documents);
/* calculate TF-IDF      */
float relevenceTerm = tf * idf;
relevanceTerm = relevanceOfTerm / 100000.0;
return relevanceTerm;
}
```

Fig. 1. Java code for the TF-IDF implementation.

C. Semantic and Sentiment analysis of songs

This is the part where we examine the semantic score of our songs. For that goal we used AlchemyAPI, popular data mining service that use natural language processing technology and machine learning algorithms to extract semantic meta-data from song content [6]. In our case we used mechanism to identify positive/negative sentiment score from the song lyrics. The procedure is simple, we call the appropriate API with the correspondent song text and the service gives JSON responce format. From the responce output, we extracted following fields:

TABLE I. SENTIMENT ANALYSIS OF DIFFERENT SONGS.

Song name	Artist	Positive words	Negative words	Sentiment score
Always	Bon Jovi	romeo, love, baby, shine , heaven	die, blood	0.2344
Viva Las Vegas	Elvis Presley	light city, pretty woman, soul, dreams, money, fortune	dime, devil, night	0.4856
Let it be	The Beatles	wisdom, whisper, mother mary	trouble, broken-hearted	0.3073
Don't Cry	Guns N' Roses	baby, heaven, honey, sugar,	cry, sorrow, goodbye	-0.2689
Poker Face	Lady Gaga	heart, spades, cards, hug	bank, poker, Texas	-0.03481

- Sentiment score- describe the sentiment strength (0.0 = neutral)
- Sentiment type- sentiment polarity: "positive", "negative", or "neutral"
- Sentiment mixed- whether sentiment is mixed (both positive and negative) (1 = mixed)

Calls to the API were made using HTTP POST, so the posted text document is the song lyrics in English language, and the service calculate positive/negative sentiment score from the posted document. This means, if have more "positive" than "negative" words in our text, the service gives us positive sentiment score (>0), in the other cases it gives us negative sentiment score (<0).

As we can see from Table I, for example the song *"Don't Cry"* from the artist *"Guns N' Roses"*, has more positive than negative words, but it has total negative sentiment score. This is because the relevance factor of these negative words is higher than positive words, thus increasing overall negative score.

D. Lyrics Database

As we are not aware of any publicly available popular lyrics database we created our own by mining public websites. Specifically, we collected as many lyrics as we could for the artist in our database. Most of the data came from AZLyrics.com, but a number of other sites (MetroLyrics.com, LyricsFreak.com) were mined. We collected lyrics for 8,200 songs in total.

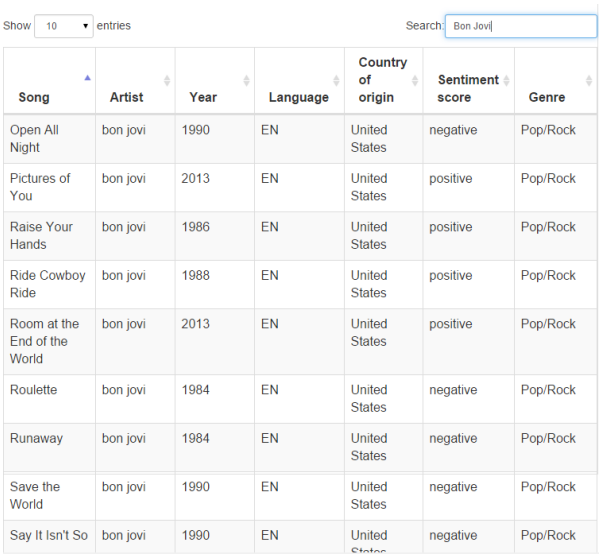


Fig. 2. Illustration of searching the song dataset.

III. RESULTS

After the part with extraction, storing and analyzing the songs in our database, we provide interactive web-based user interface developed with JavaScript and D3.js visualization libraries. This web-application has these parts: searching for the songs, semantic analysis, determining most frequent words in the genres and calculating sentiment scores. Also, we made visualization of the data for both data exploration and explanation, in order to give faster access to actionable insights and allow access to huge amounts of data. In particular, we use the best step-by-step practices for building powerful visualization framework, and interlinking with our important data [7].

When the user want to retrieve information about some song, he enters the text in input text field. In the Figure 2, is given example when the user search for the artist *"Bon Jovi"*, and application gives all the songs from the author with that name. Beside artist name, user can enter name of song, genre, year, country, language and can sort the results in ascending or descending order.

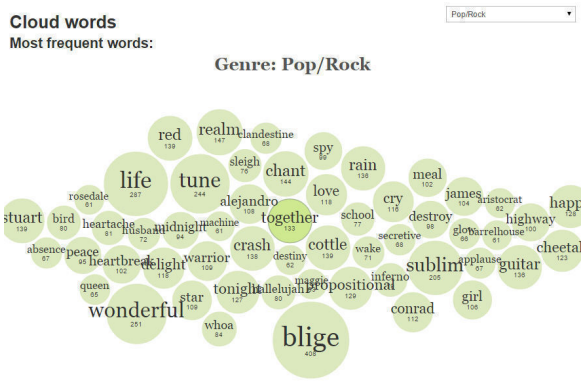


Fig. 3. Most relevant keywords in the genre: Pop/Rock

On the Figure 3 is given example with "Word Cloud" part of the application, where the user selects the genre in which he want to see the results and most important keywords are shown in the main screen.

Also, on the chart given in the Figure 4 is presented time-based comparison on different genres with correspondent sentiment score. The total score is calculated as average sentiment score of all the songs from that genre in the selected year or decade.

IV. CONCLUSION

The concept of sentimental analysis holds the key for organizing and collecting information from the World Wide

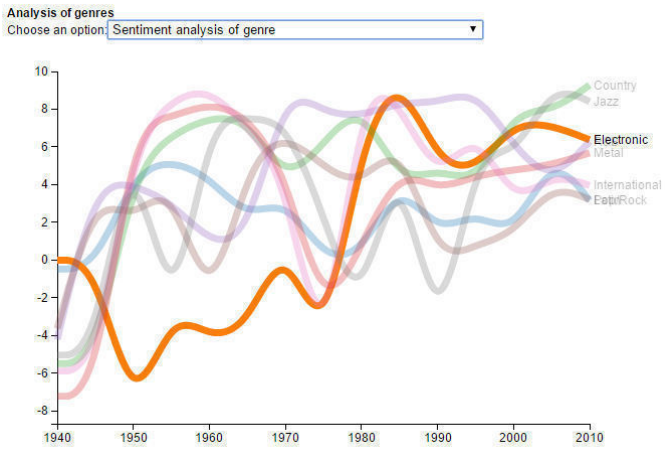


Fig. 4. Graph with sentiment analysis of genres on different time periods.

Web, and using it in various ways to make richer user experience and powerful web-application. Thanks to the modern technology, availability and transparency of the data, we can gain more useful information which do not have even exist in our database, but can carry information of enormous relevance. Today, the concept of publicly available open data represents new fuel for the industry [8].

In this paper we gave an overview of the process of gathering and extraction of songs from different languages (British, French, Spanish) with methodology for transforming and storing them in one database. We worked with big data set, created list of nearly 8.200 songs from 89 different artists, and 22 different genres.

We also provided advanced text mining methods in hope to give creative ways to use available data. For this, we used the popular text-mining technique in order to extract the most important keyword from the songs, for further detailed abstraction on whole new level.

In the future, we would continue our work in this sector and analyze more and more songs from different countries and languages. We want to make complete analyze of the songs including powerful intelligent queries: What are the most popular words in the French Pop/Rock, Who is the most popular artist from Spanish in the 60-th, Which is the most ”positive” and ”negative” song from Spain in the period 1960-1970. For this to accomplished we must collect more musical data, to increase our existing data set, and improve our text mining technique for analyzing big corpus of documents.

On the other hand, we also hope to encourage more music sites to publish song information from different time periods, which will allow leading to better usage of the published music data and allowing development of even more powerful solutions and useful applications.

REFERENCES

[1] T. O’Reilly, ”What is web 2.0”, *OReilly Media*, September, 2009.

[2] C. Goddard, ”Semantic Analysis: A Practical Introduction”, *Oxford University Press*, September, 2011, pp.289-310.

[3] S. M. Weiss, N. Indurkhya, T. Zhang and F. Damerau, ”Text Mining: Predictive Methods for Analyzing Unstructured Information”, *Springer*, October, 2004, pp.138-159.

[4] JSoup API: http://jsoup.org/apidocs/ (Accessed 28 January 2015).

[5] J. Ramos, ”Using tf-idf to determine word relevance in document queries”, In *Proceedings of the First Instructional Conference on Machine Learning*; December, 2003.

[6] Alchemy API: http://www.alchemyapi.com/api (Accessed 28 January 2015).

[7] N. Iliinsky and J. Steele, ”Designing Data Visualizations: Representing Informational Relationships”, *O’Reilly Media*, September, 2011.

[8] T. Berners-Lee and N. Shadbolt, ”There’s gold to be mined from all our data”, *The Times*, December, 2011.

Experimental comparison of several classification algorithms and discussion

Milan Petrov, Dimitrij Mijoski, Duko Serafimovski, Filip Panchevski
Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering
Skopje, Macedonia
{petrov.milan, mijoski.dimitrij, serafimovski.duko, panchevski.filip}@students.finki.ukim.mk

***Abstract*—We present a comparison of several widely used classification algorithms. The comparison is centered on the correctness of the classification models for different types of data sets. As a less important feature, we will consider the time taken for building the model by each of the algorithms. The paper includes experimental check and discussion against actual data sets. The goal of this paper is for a given data set, just by seeing it, the data analyst to be able more easily to select and prioritize the algorithms that he/she is going to run over the data set.**

***Keywords*—classification, data mining, machine learning, evaluating models.**

I. Introduction

"Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications. Classification is considered an instance of supervised learning, which task is learning a **target function *f*** that maps each attribute set x to one of the predefined class labels y. The target function is also known informally as a classification model, which can be descriptive modeling or predictive modeling" [1]. For a data set, descriptive modeling builds a model on the whole data set. The purpose of descriptive modeling is discovering knowledge in the current data. On the other hand, predictive modeling builds a model which most often is used to predict future behaviour. The model is built with part of the data, while the other part is used for testing the predictive capabilities of the model – for every testing instance we compare the model class and actual class. There are many techniques that employ a learning algorithm to identify a model that best fits the relationship between the attribute set and the class label of the input data. A key objective of the learning algorithm is to build a model with good generalization capability; model that accurately predicts the class labels of previously unknown records [1]. More about classification can be read in Introduction to machine learning [2], Pattern classification [3], statistical pattern recognition [4] etc.

Because there are many classification algorithms, it becomes important to ask the question 'What algorithm should we use for a given classification problem?' In our paper we help the reader to be able to decide which models are best for specific types of data sets and to help him choose a model for the data set he will be working on. We will compare some of the most used algorithms for building models with metrics like: accuracy, time needed for building the model, ROC area, F-measure, confusion matrix etc. However we will only present the percent of correctly/incorrectly classified instances (accuracy) and the time required for building the model.

On our data sets we will compare the following algorithms K-nearest neighbors (KNN), Bayes network (BN), Naïve Bayes (NB), ID3 (Iterative Dichotomiser 3), C4.5 decision tree algorithm (J48), Repeated incremental pruning to produce error reduction (RIPPER), Partial decision tree algorithm (PART) and Artificial neural networks (ANN). We will present only the best results obtained by applying the above mentioned algorithms on the following data sets: **Credit approval** [5], **Bank Marketing Data Set** [6], **Forest Cover Type Prediction** [7] and the **Poker Hand Data Set** [8]. Respectively we do preprocessing of the data sets, where we fill the missing values, remove the outliers, normalize the continuous attributes if needed, remove unnecessary correlated attributes, discretization etc. We would not show the steps of preprocessing and the parameters selected for the algorithms accordingly to get the best results compared by the metrics we mentioned before. The preprocessing procedure is not relevant to this paper. The idea is to compare models on already given accuracies, so our focus is only on comparing the algorithms and showing which algorithms are best to start with when solving classification problems on the different types of data sets.

For this research we used the software package Weka 3.6 [9] which is one of the best and most used tools in the data mining area. For the Forest cover type data set and the Poker hand data set we needed to increase the ram memory limit in Weka [10], or we would not be able to run the algorithms nor load the data set.

A. Related Work

There have been many similar researches on this topic. The most worth mentioning is Wolpert with his paper "The Lack of A Priori Distinctions Between Learning Algorithms". Wolpert's paper is one of the first papers that use OST error to check the free relationship between the learning algorithms [11]. Other researches worth mentioning are Thomas G. Dietterich's "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms", whose paper tests if one learning algorithm preforms better than the others on a particular classification problem [12], and Metalearning: Applications to Data Mining, which is the study of principled methods that exploit metaknowledge to obtain efficient models and solutions by adapting machine learning and data mining processes [13].

II. Discussion of the Data Sets

Here we present the data sets and the results from running classification on them. All charts for accuracy (for each of the data sets) are **in percents**.

A. Credit approval

Credit approval data set contains data about clients usage of credit cards, and the class label is whether the client is eligible for credit. The attribute names are replaced with meaningless symbols. It contains 690 instances, 16 attributes (10 nominal and 6 continuous). Figure 1 shows the results from the algorithms applied on the credit approval data set where we can see that the best results were given by the KNN algorithm followed by the Naïve Bayes and the ANN algorithm. All the algorithms showed good classification capabilities and provided good results. The only exception were the results we got from the ID3 algorithm which had the lowest percent of correctly classified instances and did not success to classify some of the instances, which is an often behavior of the ID3 algorithm. We would like to mention that for this problem it would be best to compare the algorithms by the false negative rate, because it will cost the bank more if we give credit to someone who will not return it afterwards than not to give credit to someone who can return it because then the bank only loses the income interest. However we only focus on the accuracy of the algorithms.

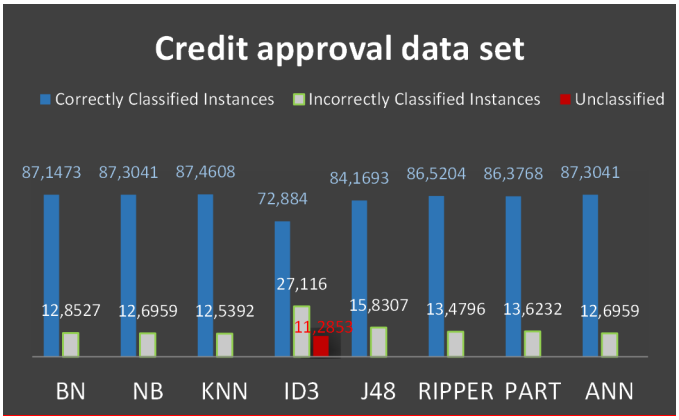


Fig. 1: Credit approval data set results

Figure 2 shows the times required to build the model for each of the algorithms. We can see that it took most time for the ANN, more than 5 minutes. RIPPER algorithm took second longest time for building (0.13) and the other algorithms were very fast with time from 0.01 to 0.03 as showed on the chart.

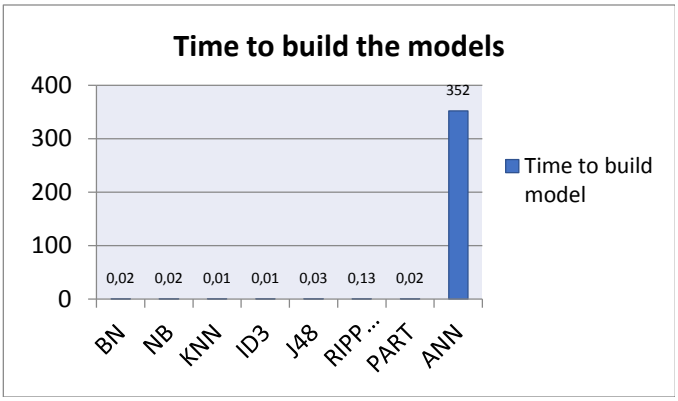


Fig. 2: Time to build the models

B. Forest cover type

Forest cover type data set as input variables has geographic data as elevation, slope, nearest distance to water etc. and output class label describes the cover type of the observed forest. It contains 581012 observations with 54 attributes, of which 44 attributes are binary (nominal) and 10 are continuous. We had to manually change the labels of the class attribute from digits to letters so we can use the data set in Weka, because Weka, by default, does not accept digit only labels as nominal attribute values (e.g. classes 1, 2, 3 versus one, two, three). This can also be done with writing a script in any programming language that can do the work for us. On Figure 3 are shown the results from the algorithms applied on this data set. The best results were given by the J48 and KNN algorithms and worst results were shown by the Naïve Bayes and the Bayes Net algorithm. KNN showed good results because we have a lot of instances and it is highly probable that we will find an already classified instance which attributes are similar to the currently observed instance and their class will be same. NB and BN do not work well with data sets that have many class labels.

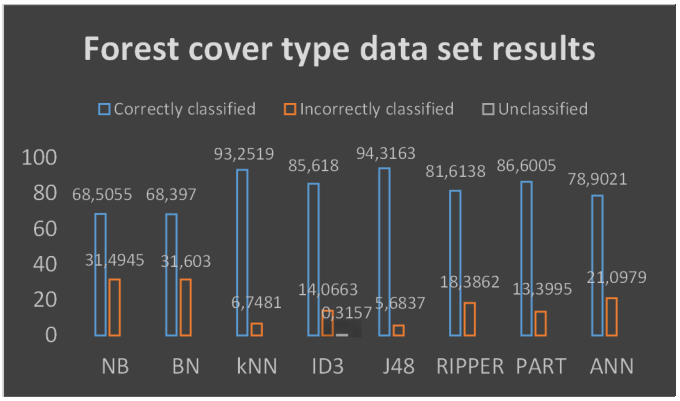


Fig. 3: Forest cover type data set

The best time to build the model was shown by the KNN algorithm, but this time does not include the time necessary train and test the model. The approximate time taken by the KNN algorithm was about 19 hours. This happens because

KNN is a lazy learning algorithm, so all the computations are done when we classify instances – when we call the model function f. The ANN algorithm took longest time to build the model taking about 3 hours as shown on Figure 4.

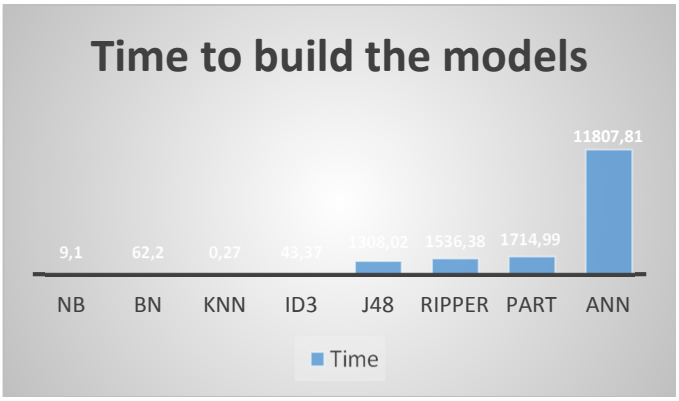


Fig. 4: Time to build the models

C. Bank marketing

Bank marketing data set has data from bank advertising campaign that advertises bank deposit. The campaign was conducted over telephone calls. Input data has client bank data (e.g. balance) and campaign data (e.g. phone call length), and the output class attribute tells if the client will subscribe a term deposit. The data set has 45211 instances with 17 attributes, from which 7 continuous and 10 nominal. For each of the algorithms we get an approximately equal accuracy of 89 percent. The only exception is the ID3 algorithm which gave the lowest percent of correctly classified instances and also was not successful in classifying 7.29 percent of the instances. The best in successfully classifying the instances was the RIPPER algorithm with 90.35 percent accuracy as shown on Figure 5 bellow.

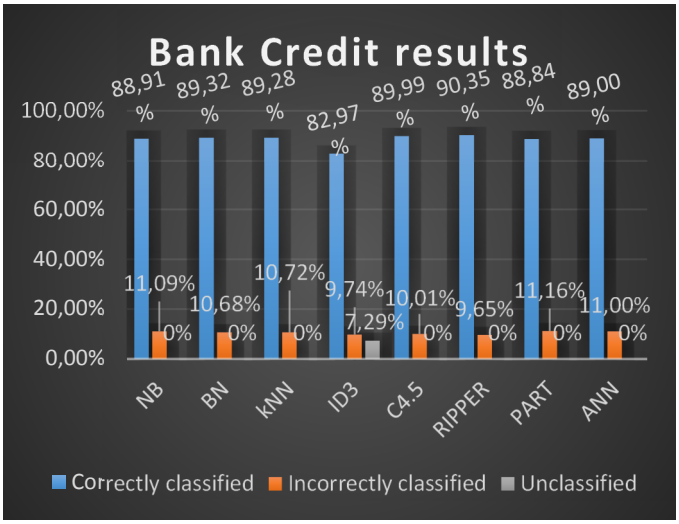


Fig. 5: Bank Marketing Data Set results

In addition to the time required for building the model the best performance showed the KNN algorithm. As expected the ANN algorithm took longest time to build the model and it took 121.41 seconds as shown on Figure 6.

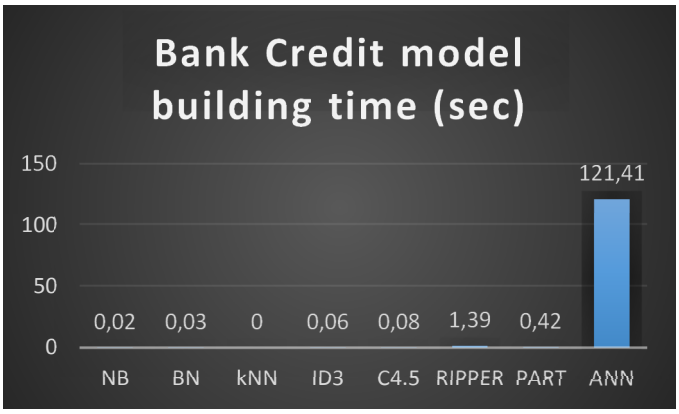


Fig. 6: Time to build the models

D. Poker Hand

In *Poker Hand data set* one instance represents 5 cards dealt from a standard deck (a poker hand) – two attributes, suit and rank, describe one card, so the set has 10 input attributes. Output class is the type of hand, e.g. full house, pair, two pairs etc. For this set we will use a training set of 25010 representatively selected instances and a testing set of 1 000 000 independent instances. We can notice that the ratio between the training set and the test set is very big, the number of instances for training is very small compared to the number of instances for testing the model. We decided on this distribution because our target was to compare the accuracy for classification of the algorithms when we have little information for training, i.e. how will the algorithms work when they do not have a good model built from the training set.

Compared by accuracy, the best results were given by the ANN model with 94.2615% of the instances classified correctly followed by the PART algorithm, and the worst by the ID3 algorithm with only 40.6881% accuracy and also big percent unclassified instances as shown on the Figure 7.

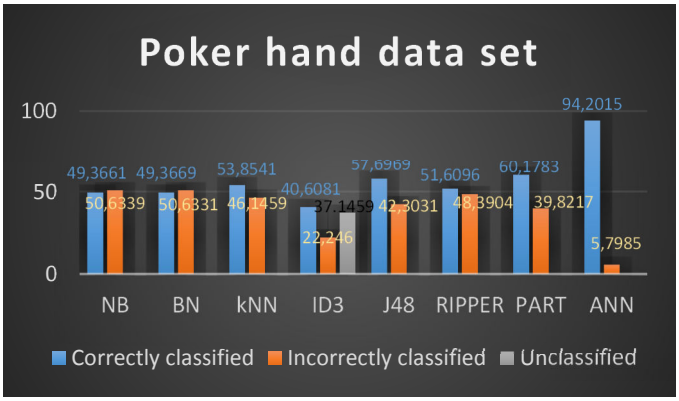


Fig. 7: Poker hand data set results

For the neural network (ANN), the time needed for computation the model was longer than the other algorithms. The

best computational time was shown by the KNN algorithm taking only 0.05 seconds to build the model and as always longest time was taken to build the model by the ANN. The results are shown on Figure 8.

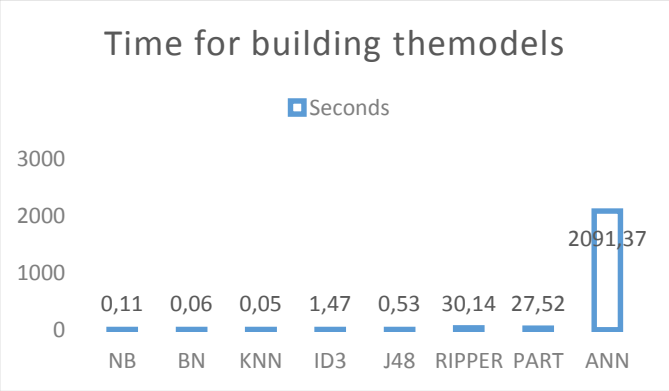


Fig. 8: Time to build the models

III. Conclusion

This paper describes the performance of the most used algorithms in the data mining, machine learning and pattern recognition areas. We only showed the accuracy and time performance of all the algorithms. More information can be found in the papers from the references section.

First we will conclude with overview *centered around the algorithms*. We can conclude that none of the algorithms is a lot better than the others, because for different types of problems one of the algorithms might perform better than the others. Maybe it is **safest to start with the J48 algorithm** because it performed best by average both by time and accuracy. For sure we can tell that **worst accuracy performance is given by the ID3** algorithm which showed the least correct classified instances and also was unable to classify some of the instances for all of the data sets. We can also conclude that **on sets with only few class labels it would be wise to start with the Naïve Bayes and Bayes Net**. Such sets are the Credit Approval and Bank Marketing set (both have binary classes). Also **good performance showed the KNN and ANN algorithms, but their disadvantage is that they can be slow on large data sets**. The ANN algorithms time to build the model and its total computation time was way larger than all the rest of the algorithms and the KNN algorithm took long time when *testing* bigger data sets. The PART algorithm also preformed good on all of the data sets, except on the Poker hand data set, but compared to the other algorithms on that set with the exception of ANN it showed best performance.

Now we will conclude with overview *centred around the data sets*. The *Bank marketing* data set itself did not help us on retrieving any information about comparing the algorithms because they all showed good performance except as always the ID3 algorithm. From the *Forest cover type* data set and the *Poker hand* data set we can see that training and testing on the whole set shows better accuracy than training only on representatively chosen subset and testing on the rest of the data, but that way it takes a lot more time in building the model and also the total computation time is way larger. We can notice that KNN works well on big data sets like forest cover type data set, but as we mentioned before the computation time took about 19 hours which is pretty long.

Having this info, we should have a clearer picture how to practically analyze a given data set.

References

[1] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[2] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. The MIT Press, 2010.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[4] A. Webb and K. Copsey, *Statistical Pattern Recognition*. Wiley, 2011.

[5] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Credit+Approval

[6] ——, "UCI machine learning repository," 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[7] ——, "UCI machine learning repository," 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Covertype

[8] ——, "UCI machine learning repository," 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Poker+Hand

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[10] "Increase the memory limit in weka," 2014, [Online; accessed 17-December-2014]. [Online]. Available: http://infocenter.pentaho.com/help/index.jsp?topic=%2Fperformance_tuning_guide%2Ftask_weka_memory_limit.html

[11] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[12] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[13] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta, *Metalearning: Applications to Data Mining*, 1st ed. Springer Publishing Company, Incorporated, 2008.

High-Level Synthesis, an Introduction and Overview of Present-day Tools

Dimitrij Mijoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
mijoski.dimitrij@students.finki.ukim.mk

***Abstract*—High-level synthesis is a process where a given system whose behavior is described in algorithmic level (instruction level) is translated to register-transfer level which is more appropriate for hardware description. The goal of the process is to ease the hardware designing by allowing hardware to be described with more understandable higher level abstractions, i.e., imperative programming languages such as C. In this paper we will portray the process, give an overview of few modern tools and show an example of compiling with some of the tools.**

Keywords*—High-level synthesis, electronic design automation, compilers.

I. Introduction

HLS is, with simpler words, a way to design hardware with programming languages meant for writing software. In software languages we see data as numbers, and the operations over data are just standard number operations. The programmer does not care how the machine stores and moves the data. When we get down to the hardware, that is an aspect that should be taken care. The hardware designer must care how the number are represented and stored in the machine, how and when they move through the wires, registers, gates etc. Few benefits of designing hardware with software programming languages are shorter design time and increased productivity, easier debugging and lowering the entry barrier for programmers to hardware design.

In Section I and in Section II we will give a theoretical introduction to the process. For a greater theoretical yet still introductory insight it is advisable to look at the following papers [1], [2], [3]. In Section III we will give a short overview of present-day tools.

First we will define a **system** and how it can be described at different levels. System is a device, logical or physical, that at one end has some inputs (motion, sensors, input data) and at other end has outputs (reaction to the input motion, output data). In fact many systems have both physical and logical representations of which one is of greater importance. For example, given a system consisted of gears, motors and levers we are mostly interested in the actual physical construction. Yet we still need the logical representation given with blueprints and mathematical expressions for easier understanding of the system and eventually for rebuilding it. The situation for computer systems is opposite. The concept of data and information is logical, so computer systems are mainly logical systems, but we still need physical machines that can represent and process data. Therefore, we can differentiate different levels for description of systems:

1) Descriptive level – where the system is usually described on declarative way, with mathematical expressions (numbers, variables, functions, sets, logic, etc.)
2) Algorithmic (instruction) level – where we represent the system on an imperative way like a series of steps that the computer has to execute. The steps include operation with variables (assignment, additions, subtraction) and operations that affect the order of execution (if-then, loops, go to etc.).
3) Register-transfer level (RTL) – This is the first ”hardware” level. Every operation from the upper level here is represented as movement of numbers from some register through different paths (buses, multiplexers) and functional units (adders, ALUs) to, again, some register.
4) Logic level – On this level the elements from RTL are mapped as connected logic gates.
5) Circuit level – Here the logic gates and connections are represented with elements from electronics (transistors, wires).

Today the task of transition from the first to the second level and inside the second level from high-level programming language to low-level language is well done by a range of compilers and interpreters. However the transition from algorithmic level to RTL level is mostly done by hand. General-purpose CPUs are designed in a way that first we design an instruction set and then we create a hardware that satisfies the instruction set. As for application specific hardware we look at the whole program and then we write a according RTL design. The transition from RTL to logic level is again well done by hardware description languages and the design tools.

High-level synthesis is exactly the automated transition from algorithmic to RTL level. In other words it solves the problem, to smaller or larger extent, for a written program how to generate hardware satisfying constraints such as maximal number of functional units, critical path length, power consumption, chip area and others.

II. High-Level Synthesis

The first task that has to be done is to actually write source code, to program the wanted algorithm. Usually we write in an imperative (procedural, structural, object-oriented) language. Such languages are aware of the of the concept of

state and the order of execution is explicitly stated. In the paper cited above [1] is mentioned that there were attempts for HLS from functional languages (LISP) and declarative languages (Prolog). For those, additional steps would be required by the compilers to transform the code to imperative list of instructions. The system must allow the program to contain data manipulation expressions (assignment, operators) and control manipulation expressions (conditional branching, loops). It would be convenient if the system allows partitioning in modules in form of functions and procedures. Having that, a simple algorithm may be implemented with one function which contains data instructions and some loops. On the other hand, implementing an instruction set for general purpose CPU may look like a big list of if-else blocks where each block maps to one or a group of instructions. In some blocks we would call different functions.

Once the algorithm is written, the next step is compiling. The compiler should translate the program from the used high level language to an intermediate form usually represented as dependency graphs. The textbook intermediate representation consists of two graphs, data-flow graph and control-flow graph. In the data-flow graph, every instruction is represented as node, instruction inputs as input edges and the output as output edges. That would be the case for data instructions. For control instructions the input is data (values that should be compared, or a result from previous test) and there is no data output, but there is in the other graph. The other graph is the control-flow graph. Here the nodes are same (instructions) and the edges represent the order of execution. Data instructions usually have one input and one output edge, but control flow instructions like conditional branching (if-statements) have two outputs.

Then the compiler may do additional steps to optimize the code such as splitting independent calculations to execute in parallel, removing dependencies, loop unrolling, removing unnecessary code, exchanging instructions with more efficient ones etc. Some optimizations are modifying both graphs, like exchanging and removing instructions, while some are modifying only one graph. For example, if we already have two independent sequences of instructions in the data graphs, those sequences can be parallelized in the control graph using fork-join construct – we put a fork node with two output edges pointing to the start of the instruction sequences, and at the ends of those sequences we put a join node to witch we have two input edges. Some parallelization optimizations modify both graphs, like summing 8 numbers. Instead summing in 7 steps, we can reorder the data-flow tree to look like a balanced tree that can be executed in 3 steps. We modify the control-flow graph accordingly. This parallelization technique is know as reduction and can be done with any binary operator that is associative.

When that is done, the part with classical compiling is done and the process continues with the synthesis. There are two actions that can be described separately, but are interdependent. The first is *time scheduling*, and the second is *spatial allocation*.

With scheduling we map ever instruction to a time step (control step) which is similar to clock cycle from the lower levels. Some of the scheduling algorithms used in high-level synthesis are analogous to Tomasulo's algorithm used in superscalar CPUs. In it functional units are already known and it schedules multiple instructions in one time step to the free units.

The other main action in HLS is (spatial) allocation which maps every instruction and connection between instructions to functional units, memories (registers) and interconnections. This is the part where hardware constraints are of greater importance. Scheduling and allocation can be done in any order, or can be done together. That depends of the algorithms used in the tool. They usually do some form of searching the design space and they return when they find a good enough design that meets the constraints. Once these activities finish, we have an abstract structure of the data path.

The last step including the data path is construction of actual data path. This step is know as *module binding*. Here in place of every abstract functional unit, interconnection and register we put an actual implementation. Implementations are taken from a library with hardware units.

At the end, from the control graph and the scheduling it is constructed a control path. The control path for every control step directs the data through the data path. Since the data path is build to satisfy every instruction in any step, the control path takes care that the right data is taken from the right place and is put in the right place so it satisfies only those instruction that are part of a given single control step. That is done by putting additional multiplexers, coders, decoders and similar units. Those units direct the data and activate only the parts of the data path requested by the current instruction. The control path can be directly constructed using the control graph and the scheduling like a finite state machine, a design usually seen in RISC processors, or it can be more complex programmable control path operated with microcodes, a design seen at some CISC processors.

When all these steps are finished, high-level synthesis is finished. Further steps include testing inside a simulator, synthesis to lower levels, and at the end, programming an FPGA or constructing an integrated circuit.

III. Modern tools

We'll give a short overview of some tools for high-level synthesis.

A. Xilinx Vivado High-Level Synthesis

Most famous Xilinx product is Xilinx ISE which is mainly for RTL design. In 2011 Xilinx acquired the software AutoPilot from AutoESL which is the basis for the Xilinx Vivado High-Level Synthesis [4]. The next year Xilinx released a new product Vivado [5] for hardware design which, according to the manufacturer, has different improvements over ISE and is shipped with the HLS tools.

The HLS tool is integrated in Vivado, but only in the "Vivado System Edition", but can be also obtained separately and used with the older ISE. The tool as input gets a source code and produces a code in hardware description language which is further processed by the ISE or Vivado softwares packages. Supported languages are C, C++ and System-C. The tool is commercial, and can be obtained as 30 day free trial of the whole Vivado package.

B. Two solutions from Cadence

This company calls high-level synthesis with different terms, transaction level modeling. Both offered solutions work with the language System-C. The first is called C-to-Silicon Compiler [6] and is signed directly by them. The second solution was first made by the company Forte, is called Cynthesizer [7], but now is in ownership of Cadence. Both solutions are commercial, there are not any free versions.

C. Synopsis Synphony C Compiler

This company offers a product named Synphony [8]. The compiler works with C and C++ languages. The tool is commercial.

D. Altera OpenCL high level design

The company Altera offers somewhat different solution [9]. Instead programming in C, here we program in Open-CL which is a framework for parallel programming. The used language is extended C. In standard C the programming model matches a random access machine, while Open-CL programming model is most similar to parallel random access machine. In this model in one step we execute the same operation over different data, so this model of programming and execution using the Flynn taxonomy is categorized as single instruction multiple data (SIMD).

The motivation behind this move probably comes from the fact that a usual reason to use an FPGA is to speedup a program that is not fast enough on a general purpose CPU in a way that we use exploit high parallelism on the FPGA by using efficiently many functional units at one single step. Open-CL is an environment where the the parallelism is explicitly stated by the programmer. Thus having an explicitly parallel program should translate better to a highly parallel FPGA. Weakness of this approach is that programming in Open-CL is quite harder than in classic imperative language.

The tool is commercial, but can be evaluated in a 30-day free trial period.

E. Panda Bambu

Panda [10] is a free, open source tool. It was made at the university Politecnico di Milano. It works with C programming language, and outputs in VHDL and Verilog languages. The tool can generate scripts that integrate the hardware description language with familiar software packages for hardware design (Xilinx ISE, Vadio, Altera Quartus, Lattice Diamond) or with HDL simulators (Mentor Modelsim, Xilinx ISIM, XSIM, Verilog Icarus). On Figure 1 is given the software architecture of the Panda-Bambu toolset. As input we give a program in source code form, hardware constraints and a library with pre-built hardware units. Then the HLS tool translates the source code in internal representation which contains few different graphs. The two mentioned in the previous chapter are the most basic, textbook, structures. Then the allocation is done followed by scheduling, binding and final synthesis.

F. LegUP

LegUP [11] is also free and open source tool. It is developed at university of Toronto. The tool accepts source code written in C and outputs to Verilog.

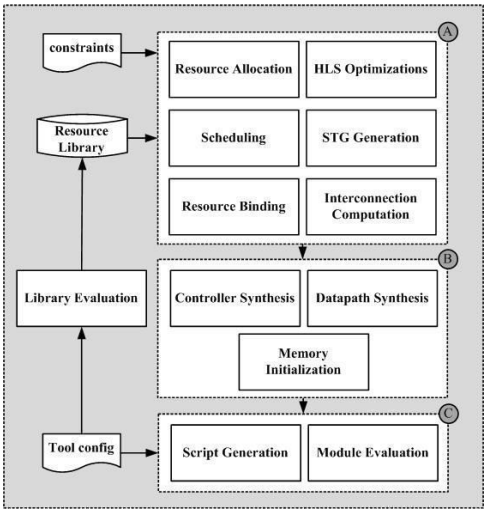


Fig. 1: Architecture of Panda-Bambu

IV. Conclusion and Future Work

From the tested tools we can conclude that although the tools can do the synthesis job to a smaller or larger extent, there is still time to pass to get a greater acceptance from the hardware design community. The future is bright, especially if we take into account the need for faster execution of already written codes.

As future work we will note that a more detailed testing of the tools can be done, for example comparison in the ease of use and benchmarking (execution time of given algorithm).

Appendix A
Example with Panda-Bambu

We will give a short example on how to use Panda. The program is downloaded as source code which needs to be compiled on a Linux operating system. We used the distribution Ubuntu 14.04 installed in a virtual machine. According to the manual (file INSTALL) the following steps were taken: (i) installed the required tools, (ii) called a configuration script with the parameters from Figure 2, (iii) compiling (with make) that took around 3 hours, and (iv) installation (with make install).

```
../configure --prefix=/opt/panda --enable-
    flopoco --enable-icarus --enable-verilator
    --enable-opt --enable-release
```

Fig. 2: Configuration before compilation of Panda.

After the tool was completely installed, an example packaged with the tool was compiled. The example is calculating CRC32 hash. With the commands in Figure 3 we have compiled a single function in the file spec.c. For more details we instruct the reader to look at the compilation scripts inside the CRC32 folder. Different compilation scrips are for different FPGA devices and for their programming software packages.

```
~/panda-0.9.2/examples/crc$ mkdir build
~/panda-0.9.2/examples/crc$ cd build
~/panda-0.9.2/examples/crc$ bambu ../spec.c --
    top-fname=icrc --device-name=EP2C70F896C6-
    DSP --simulator=ICARUS --simulate=../
    test_icrc.xml
```

Fig. 3: Commands for compiling from C to HDL.

ACKNOWLEDGMENT

The author would like to thank Vladimir Zdraveski for assistance in making of this project.

REFERENCES

[1] M. C. McFarland, A. C. Parker, and R. Camposano, "Tutorial on high-level synthesis," in *Proc. of the 25th ACM/IEEE Design Automation Conf.* IEEE Computer Society Press, 1988, pp. 330–336.

[2] ——, "The high-level synthesis of digital systems," *Proc. IEEE*, vol. 78, no. 2, pp. 301–318, 1990.

[3] D. D. Gajski and L. Ramachandran, "Introduction to high-level synthesis," *IEEE Des. Test Comput*, vol. 11, no. 4, pp. 44–54, 1994.

[4] Vivado High-Level Synthesis. Xilinx Inc. San Jose, CA. [Online]. Available: http://www.xilinx.com/products/design-tools/vivado/integration/esl-design.html

[5] "Vivado design suite," White Paper, Xilinx Inc., San Jose, CA, 2012. [Online]. Available: http://www.xilinx.com/support/documentation/white_papers/wp416-Vivado-Design-Suite.pdf

[6] "C-to-Silicon Compiler datasheet," Cadence Design Systems Inc., San Jose, CA, 2013. [Online]. Available: http://www.cadence.com/rl/Resources/datasheets/C2Silicon_ds.pdf

[7] "Cynthesizer Solution datasheet," Cadence Design Systems Inc., San Jose, CA, 2014. [Online]. Available: http://www.cadence.com/rl/Resources/datasheets/cynthesizer_ds.pdf

[8] "Synphony C Compiler datasheet," Synopsys, Inc., Mountain View, CA, 2013. [Online]. Available: http://www.synopsys.com/Tools/Implementation/RTLSynthesis/Documents/synphonyc-compiler-ds.pdf

[9] Altera SDK for OpenCL. Altera Corp. San Jose, CA. [Online]. Available: http://www.altera.com/products/software/opencl/opencl-index.html

[10] Panda bambu. Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano. [Online]. Available: http://panda.dei.polimi.it/?page_id=31

[11] A. Canis, J. Choi, M. Aldham, V. Zhang, A. Kammoona, J. H. Anderson, S. Brown, and T. Czajkowski, "Legup: high-level synthesis for fpga-based processor/accelerator systems," in *Proc. of the 19th ACM/SIGDA international symposium on Field programmable gate arrays*. ACM, 2011, pp. 33–36.

Performance comparison of AES candidates

Ivana Bozhinova*, Eva Tanaskoska†, Sandra Delovska‡ and Andrej Gashteovski§
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Rugjer Boshkovikj 16, Skopje, Macedonia
Email:{*bozhinova.ivana, †tanaskoska.eva, ‡delovska.sandra, §gashteovski.andrej}@students.finki.ukim.mk

***Abstract*—Nowadays multiple different methods and algorithms solve the problem of safe communication and data transfer. Advanced Encryption Standard (AES) is a representative of symmetric-key encryption cipher standard. It was established after a thorough 5-year long testing and evaluation process of the nominated algorithms. By extensive an state of the art research we compare and target the selected five AES candidate algorithms: Twofish, MARS, Serpent, RC6, and Rijndael. Our purpose is to determine the impact of different key lengths on algorithm speed and verify if the winning AES pick has the best execution performances, or it is somewhat slower at the expense of security. The obtained results indicate that all of the candidates have exceptional performance on our testing configuration, but there is not one that can be claimed as absolute speed winner since they are influenced differently by the change of the key length.**

***Index Terms*—AES, performance, encryption, Rijndael, MARS, Serpent, RC6, Twofish.**

I. INTRODUCTION

Cryptography is an important discipline that is not only the subject of an enormous amount of research, but provides the foundation for information security in many applications [1]. Advanced Encryption Standard (AES) is a symmetric-key encryption cipher that was announced as a standard by the National Institute of Standards and Technology (NIST) in 2001. The AES symmetric block cipher is widely used for the U.S. Government needs, as well as by numerous other organizations, institutions and individuals in and outside of the U.S. It consistently shows a very good performance across a wide range of computing environments.

The origins of this cryptographic algorithm date back to 1997 [2] when the National Institute of Standards and Technology (NIST) announced that it needed a successor to the Data Encryption Standard (DES) which could not withstand the current advances in computing technology as it became vulnerable to brute-force attacks due to its too short key and block sizes. The U.S. government required an algorithm "capable of protecting sensitive government information well into the next century." [3]. In order to find the perfect match, NIST organized a cryptographic competition and after extensive analysis conducted by the utmost members of the cryptographic community, the five afore stated algorithms were selected [4]. There were three major categories to the evaluation criteria: security, cost, and implementation characteristics [5]. Taking all these values into consideration, but still concentrating on security, it was determined that the winner should be Rjindael.

This paper is focusing on examination of the cost criteria because with time it has gained equal importance. The topic is challenging since there is still a lot of ambiguity in answering the question which one among the AES candidates has the best performance. We conducted tests of the candidates computational efficiency on a common platform, in order to analyse and compare their performance characteristics. Our hypothesis is that the winning AES does not have the best performance, regardless the key length. We expect that Rijndaels sophistication will impact its performance in favour of security. However, considering the algorithms complexity it is a rather difficult task to conclude which one of them is the fastest. Moreover, we aspire to examine and compare the significance of the key length for the amount of time needed to execute the encryption and decryption for each cipher.

It is structured as follows. In Section II we present the related work in this field and other comparisons and evaluations of these algorithms. Section III describes the final AES candidate algorithms in details, while Sections IV and V present the setup and results of the experimental work comparing the algorithms. Finally, Section VI concludes this paper and points to potential future work.

II. RELATED WORK

Part of NISTs competition for the Advanced Encryption Standard (AES) process included a speed based evaluation on all the candidate block ciphers. Furthermore, this is a popular issue subjected to numerous other researches.

Many researchers have made inapplicable implementations or used incompatible API's, hardware (processors) or software (compiler), what makes the results they obtained questionable. For example in the paper [6] the author implemented self-modifying code that cannot be used in a number of applications. Furthermore, Brian Gladman's implementation of AES candidates [7] uses key-specific static variables rather than to allocate a register to address them, making more efficient performance by freeing some of the registers. In addition, based on the fact that a different API can influence the speed of an implementation, all AES candidate implementations reported in [8] differ in their API's which makes them unsuitable for comparison. Another approach can be seen in a paper written by Andreas Sterbenz and Peter Lip [9]. It describes that all AES ciphers offer reasonable performance in Java, compared to C and assembler implementations. Although they used an Intel Pentium Pro 200 which is far slower than the machine

we tested on, their speeds are two times faster due to the high estimation of the code.

However, in comparison to our results in Section V, the algorithms are ranked in a similar order for speed of encryption and decryption. The first three, RC6, MARS and Rijndael are ordered the same, and then Twofish and Serpent are reversed.

III. ALGORITHMS

NIST specifies a few requirements for the Advanced Encryption standard, some of them being a certain block size of 128 bits and supported key lengths of 128, 192 and 256 bits. The proposed algorithms utilize design components such as Substitution-boxes (S-boxes), used as lookup tables to replace runtime computations thus instating confusion, and Feistel networks, structures consisting of a set of functions which are organized in rounds (iterations), each round providing increased cryptographic security. The main differences accompanying the five finalists are the Feistel functions, number of rounds, S-box generation methods, key schedule algorithms and so forth, implying that every candidate has a rather unique and intriguing working mechanism.

A. Rijndael

Rijndael [10] is a symmetric block cipher with 128, 192 and 256-bit key lengths, supporting equal sized block lengths. The Rijndael algorithm specifies a few linear transformations which operate on intermediate results called states. A state is a rectangular array of bytes consisting of 4 rows and a column number derived from the block length divided by 32. The key size specifies the number of transformation rounds as follows: 10 rounds for 128-bit keys, 12 rounds for 192-bit keys and 14 rounds for 256-bit keys. A single round is composed of four different transformations: SubBytes, ShiftRows, MixColumns and AddRoundKey. The cipher key is a rectangular array with four rows and the number of columns is equal to the key length divided by 32. In the encryption process, the SubBytes step is performed by substituting each byte in the state matrix with a replacement from a 8-bit substitution box (S-box) constructed by combining the inverse function with an affine transformation. Decryption does the opposite, first the affine transformation is taken and then the multiplicative inverse is found. The ShiftRows encryption step shifts the bytes in each row to the left by a certain offset depending on the block length, which increases for every row, starting from the first. The decryption step shifts the bytes to the right. In the MixColumns step, each column is multiplied with a fixed polynomial. AddRoundKey performs a bitwise XOR with the cipher key. Rijndael uses the key schedule algorithm consisting of two components: key expansion and round key selection, in order to expand a key into multiple round keys which are derived from the initial key.

B. Serpent

Serpent [11] is a block cipher with a block size of 128-bits supporting 128, 192 and 256-bit key lengths. The cipher is a 32-round substitution-permutation network which splits the 128-bit block into 32-bit words. Keys shorter than specified are padded by appending 1 to the most significant bit end, followed by zeroes. The cipher consists of three basic functions: an initial permutation, 32 rounds of key mixing operations generated from the key schedule algorithm, passing through S-boxes and linear transformations in all rounds but the last one, where the linear transformation is replaced by an additional key mixing operation. Upon completion of the rounds, the final permutation places the bits into their correct positions. In the initial phrase of the encryption, the initial permutation is applied to the block and the input is then proceeded into the first round. The outputs of the rounds are cascaded into the following rounds, until the final permutation is applied which outputs the final cipher text. Four input bits are mapped into four output bit with the help of 8 different S-boxes. A single S-box is used 32 times in four rounds. This 32 round design is meant to strengthen the algorithm, thus making it usable for quite a longer time compared to ciphers which utilize less rounds. The decryption process uses the S-boxes in a reverse order, applies the inverse linear transformation and a reverse order of the sub keys.

C. RC6

RC6 [12] is a block cipher built as an improved version of RC5, with a block size of 128-bits and 128, 192 and 256-bit keys making use of data-dependant rotations. RC6 can be parametrized with variable word and key lengths, and number of rounds, which makes it a rather versatile cipher. The design consists of six basic operations which are performed over four 32-bit words by default, although the word size can be modified. These words contain the input, and at the end of the encryption process, the resulting cipher text. Each round rotates the positions of the four words. The operations include a quadratic transformation, integer addition, integer subtraction, bitwise XOR, integer multiplication and bitwise rotations to the left and to the right. Before the first and after the last round an additional key addition is performed. The decryption routine is analogous to encryption with inverted steps.

D. Twofish

Twofish [13] is a symmetric block cipher with a 16-round Feistel structure and block size of 128 bits supporting variable-length keys up to 256 bits. Its' main features include the use of pre-computed S-boxes and a complex key schedule. The key has two uses: the first half is used for encryption and the other half is used to enhance the S-boxes' security. The Feistel network is utilizing four key-dependant 8x8 S-boxes using two fixed 8x8 permutations and key material. Twofish's design also includes a 32-bit Pseudo-Hadamard Transform on the output of MDS matrices as methods for cryptographic diffusion and mixing the outputs from two parallel 32-bit functions, 1-bit rotation, as well as the whitening technique by performing a XOR operation over 128 bits of sub key which are XORed with four 32-bit words derived from the plain text before the first and after the last Feistel round. Due to the emphasized

key-dependant nature of Twofish's components, the cipher has a rather complex key-scheduling algorithm which has to provide 40 words of expanded key and 4 key-dependant S-boxes. Insufficient length keys are padded with zeroes to the next specified key-length of the algorithm. The encryption and decryption functions are slightly different, but are based on the same building blocks, therefore it is possible to build a decryption module without duplicating much functionality.

E. MARS

MARS [14] is an iterated 32-round block cipher with varying types of rounds, unlike other block ciphers, for enhanced security and increased attack difficulty, which supports 128-bit blocks, variable length keys ranging from 128 to 1248 bits and 32-bit words. Its' unusual characteristics were designed to remain resilient and fend off attackers long term by making use of strong cryptographic techniques. MARS is a type-3 Feistel network which provides cryptographic diffusion. Every round combines fast and simple operations such as additions, subtraction and XOR, S-box lookups, as well as fixed and data-dependent rotations in conjunction with native multiplication. In MARS, the middle rounds, also known as cryptographic core, differ from the top and bottom rounds known as wrapper layers. The first step of the wrapper layers is adding key words and afterwards, eight iterations of S-box forward mixing and eight iterations of keyed forward transformation. The bottom layer does eight rounds of keyed backwards mixing, eight rounds of unkeyed backward mixing and key subtraction. The core layer performs several rounds of keyed transformations along with S-box lookups, multiplication and rotations, which is a key expansion function. The decryption process slightly differs from the encryption process.

IV. Test procedures

Execution time is the time the algorithm takes to encrypt or decrypt a block of data. It may vary depending on the key size and the data size. Because of this reason, we have tested the algorithms using 128, 192 and 256 bit keys and 512, 1024 and 2048 byte messages and then calculated the speed. Since the dependency of the message size was not the major goal of this experiment, we have taken the average speed in calculation, therefore the results are showing the dependency of the key size only.

For testing purposes, an open-source Java implementation of the algorithms was used. The code was compiled using javac (JDK 1.7, major version 51) and the tests were executed on a machine with Intel(R) Core(TM) i5 2.27Ghz CPU with 4.00GB RAM running Windows 7 Professional 64-bit OS with Service Pack 1. The tests were repeated 1000 times and an average speed was calculated. In every test, a message was encrypted 500 times in order to get a significant time value. Time was measured in nanoseconds and the speed was calculated by dividing the amount of bits encrypted by the amount of time needed for encryption. Encryption speed was measured following the pseudo-code in Algorithm 1.

Algorithm 1 Encryption

```
1:  procedure TestEncryptionSpeed
2:      totalTime ← 0
3:      for i ← 1, 1000 do
4:          set cipher key
5:          set initial vector
6:          start timer
7:          for j ← 1, 500 do
8:              encrypt message
9:          stop timer
10:         totalTime ← totalTime + timerDifferance
11:     time ← totalTime/1000
```

V. Results and evaluation

Speed performances obtained by measuring the encryption and decryption times of each of the algorithms by previously specified procedures are summarized in Table I and Table II accordingly.

Encryption	Twofish	RC6	MARS	Rijndael	Serpent
128 bit key	109.07	273.88	216.36	275.87	193.03
192 bit key	109.90	251.70	211.54	262.91	187.25
256 bit key	107.63	261.49	205.17	232.55	184.71

TABLE I
Encryption Performances

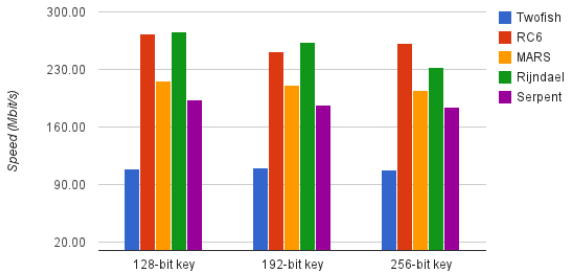


Fig. 1. Encryption performances

The collected results verify that all of the chosen AES finalists work extremely fast both for encryption and decryption. However, from Table I and II it is obvious that the performance of the Twofish algorithm is about two to three times worse than the others. This is caused by the usage of key dependent S-Boxes[9] in its implementation which can decrease the performance on certain JVMs, in particular when using the Just-in-time Compiler (JIT) compiler that comes with the JDK on the Windows platform. This means that Twofish

Decryption	Twofish	RC6	MARS	Rijndael	Serpent
128 bit key	98.76	288.44	216.27	211.29	163.47
192 bit key	97.22	281.72	222.33	218.59	157.09
256 bit key	97.85	269.30	225.06	197.69	153.07d

TABLE II
Decryption Performances

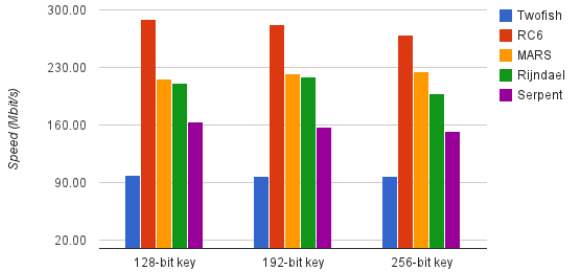


Fig. 2. Decryption performances

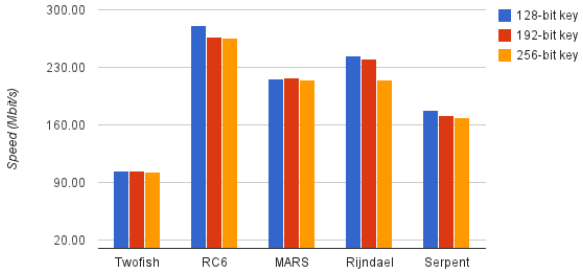


Fig. 3. Average speed for performing encryption and decryption

may be somewhat faster compared to the other algorithms if tested on other platforms by using different Java compilers.

As shown in Figure 1, when encrypting the message using 128 or 192-bit key the fastest algorithm is Rijndael, but for 256-bit keys RC6 has better performance. Furthermore, it can be clearly noted from Figure 2 that both RC6 and MARS decrypt faster than the Rijndael regardless the key length. Hence, as we hypothesized, we can not strictly conclude that Rijndael or any one of the ciphers is absolutely the fastest.

As illustrated in Figure 3, the measurements seem to indicate that for most of the algorithms the key length implies directly on the encryption execution time, resulting in worse performance for longer keys. The reason for this is quite straightforward and it lies in the implementation of the algorithms itself, considering that longer key requires more computations for the encryption to be carried out. This difference is more remarkable for the RC6 and Rijndael ciphers, while for the rest it is not that significant. However, we can not claim that any of the algorithms is completely independent on the key size when it comes to encryption. Interestingly, larger key size does not always lower the performance when decrypting the message. This happens due to the fact that some algorithms have constant or even faster speed and key setup for large keys. In MARS decryption particularly, the performance gets slightly better every time the key length increases. Also, there is a minor improvement in the decryption execution time of Rijndael when working with 192 bits long keys, compared to the time of decrypting using 128 bits long keys. In contrast, the 256 bit key slows the decryption rate down.

VI. Conclusion and Future Work

In this paper we have tested and analysed the performance of the AES candidates using unified hardware and software. Encryption and decryption times obtained show that for most of the defined key lengths RC6 is faster than the other algorithms, followed by Rijndael, MARS and Serpent. However, Twofish runs significantly slower on our configuration. This implies that the winning AES is not the fastest among all of the candidates, in accordance with the hypothesis. Nonetheless, its performance is still satisfactory for its purpose of usage. The key length reflects on the encryption time for most of the algorithms, but not for all of them, unlike our expectations. While its change causes worse performance for MARS, Serpent and Rijndael, it does not make a big impact for RC6 and Twofish. On the other hand, when it comes to decryption, an execution speed-up was denoted using MARS. This means that when it comes to performance over security, using RC6 over Rijndael would be a better choice for most key lengths.

We performed the experiment on the configuration explained in IV. By virtue of the the specific implementation of the algorithms, as it was stated in V, if we carry out the tests on other platforms the performance of some ciphers may differ to some extent. For this reason, in the future we intend to repeat the experiment using the GNU Compiler for Java (GCJ) and the Eclipse Compiler for Java (ECJ). Also, if we had the resources required for testing the security aspect of the candidates, we would broaden our research on this criteria.

References

[1] A. Menezes, P. van Oorschot, S, Vanstone, *Handbook of applied cryptography*, CRC press, 2010
[2] Margaret Rouse. (2014, November).*Advanced Encryption Standard* [Online]. Available: http://searchsecurity.techtarget.com
[3] Samuel Kramer. (1997, January 2). *Announcing development of a Federal information processing standard for Advanced Encryption Standard* [Online]. Available: http://csrc.nist.gov
[4] Elaine Bunten-Mines.(1997,September 12).*Announcing request for candidate algorithm nominations for the Advanced Encryption Standard (AES)* [Online]. Available:http://csrc.nist.gov
[5] Edward Roback and Morris Dworkin.(1998,August 20-22). *First Advanced Encryption Standard (AES) candidate conference* [Online]. Available:http://www.ieee-security.org
[6] B. Schneier, D. Whiting, J. Kelsey, C. Hall, D. Wagner, and N. Ferguson. *The Twofish Encryption Algorithm: A 128-Bit Block Cipher*, Wiley, April 1999
[7] B. Gladman. (January 1999). *AES algorithm efficiency* [Online]. vailable: http://csrc.nist.gov
[8] H. Lipmaa. (January 1999). *A survey for implementations of AES candidates* [Online]. Available from http://home.cyber.ee
[9] A. Sterbenz and P. Lipp. *Performance of the AES Candidate Algorithms in Java*, Institute for Applied Information Processing and Communications Graz, University of Technology, Graz, Austria
[10] J. Daemen, V. Rijmen. *The Rijndael Block Cipher*, Springer Verlag, September 1999.
[11] R. Anderson, E. Biham, L. Knudsen. *Serpent: A Proposal for the, Advanced Encryption Standard* , Cambridge University, England; Technion, Haifa, Israel; University of Bergen, Norway.
[12] R.L. Rivest, M.J.B. Robshaw, R. Sidney, Y.L. Yin. *The RC6 Block Cipher*, M.I.T. Laboratory of Computer Science; RSA Laboratories, August 1998.
[13] B. Schneier, J. Kelsey, D. Whiting, D. Wagner, C. Hall, N. Ferguson, *Twofish: A 128-Bit Block Cipher*, Counterpane Systems, June 1998.
[14] C. Burwick, D. Coppersmith, E. D'Avignon, R. Gennaro, S. Halevi, C. Jutla, S.M. Matyas Jr., L. O'Connor, M. Peyravian, D. Safford, N. Zunic. *MARS - a candidate cipher for AES*, IBM Corporation, June 1998.

Parallel Sorting of Arrays with OpenMP and the Quick-Sort Algorithm

Goran Petkovski, Darko Tosev
University Ss. Cyril and Methodius
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
{petkovski.goran, tosev.darko}@students.finki.ukim.mk

***Abstract*—Parallel computing is nothing new to computer science, it has been evolved now a day because of the availability of fast and more inexpensive processors and advancements in the field of communication technologies. We will concentrate on the modest parallelism of computers, the multithreading programming model. This model provides executing of computer programs on multiple separate processors in shared-memory architecture. In this paper we evaluate the performance of parallel quick sort algorithm on multithreading programming environments such as OpenMP. We review our proposed model for parallel array sorting, and explain the need for providing efficient execution in multithreading environment and use of Quick Sort as primary choice. Then, we perform an experiment on an array with random generated numbers, thus analyzing the performance of the read/write in memory, time of execution and total speedup. Afterwards, we discuss the results and made comparative analysis based on average time of execution and speedup of the parallel over sequential algorithm. This is provided with detailed graphs and tables of the executed results. At the end, we present the benefits of the use of the implemented solution, by comparison to the other already existing parallel sorting algorithms.**

Keywords—Parallel computing, OpenMP, shared-memory model, performance analysis, quick sort, SMP-systems

I. Introduction

Sorting is a fundamental operation that is performed by most computers. Theory and study of different approaches of sorting data has been major research topic over the last 50 years. This is very interesting research area, because sorted data are easy to manipulate and are used frequently in many different applications (spreadsheet programs, database applications, unstructured generated data, etc.) On the other hand, today modern computer society is faced with the challenge of very enormous amounts of data coming from different heterogeneous sources, so the reason for efficient sorting algorithm is getting more and more meanings. This area of sorting algorithms has been deeply studied from many computer scientist, and they developed various approaches for constructing sorting algorithms. However, depending on the nature of the situation and reason of the problem, these approaches come with diverse complexity and usage, according of the system architecture [1].

Traditionally, computer software has been written for serial computation. To solve a problem, an algorithm is constructed and implemented as a serial stream of instructions. These instructions are executed on a central processing unit (CPU) on one computer. Only one instruction may execute at a time, and after that instruction is finished, the next is executed. Parallel computing, on the other hand, uses multiple processing elements simultaneously to solve a problem. Thanks to the advance of computer architecture, today we get multiple processing elements on single chip (multi-core) or on different chips (multi-processing). This architectures can be classified according to the parallelism level of the hardware, so we have cluster and grid computing (multiple computers executing same task) and centralized server machines with multi-core processors (single computer executing the task) [2]. For implementing sorting solution in such environment, we must take some form of breaking the problem into multiple parts, so every processing unit has its own independent part. These processing units can work simultaneously and can be synchronized with the others, for avoiding race-condition of memory resources and dead-lock situations [3].

This paper will present an implementation of the well known algorithm for sorting, Quick Sort, done in the parallel programming environment OpenMP. Even though other efficient algorithms for sorting, such as merge sort, radix sort or bubble sort exist, performing an optimized ordering of arrays on the so-called symmetric multiprocessing architecture [1]it's been a challenge in the domain of parallel computing.

We first start in Section 3 with a brief overview of our methodology for parallel sorting algorithms in general. Then in Section 3.1 we give a high-level overview of our algorithm, followed by a detailed description of the individual phases and a thorough analysis of their performance. Section 3.2 gives us detailed information about our implementation of the algorithm, followed by detailed graphics, pseudo-codes and explanations. Practical performance results on the computer cluster are presented in Section 4 with tables and graphics. Finally, Section 5 concludes the paper and discuss potential optimizations.

II. Related Work

Since this scientific field which we explore contains exhaustive material to process and a lot of information to collect, we will present only the most important findings of this subject. There are some fundamental concepts we will use in our paper, then we will discuss the possible improvements in

[1]Symmetric multiprocessing systems (SMP) systems work with shared memory that is centralized for all processing units and have single operating system with many homogeneous processors.

the experiment and finally we will compare the results with complete experiment.

There are some very good implementations of parallel sorting algorithms optimized for high performance enviroments. The one commonly used for parallel execution is the Merge Sort algorithm (also defined as *divide-and-conquer* algorithm as Quick Sort, who uses master-slave model in the form of tree for parallel sorting). The technique is that, we first divide the list of elements into two halves which each process receives from its precedor, process them, keeps one half for it and the other second half send to its successor [4]. This methodology makes use of all active processors at the same time and this approach we used for implementation of our parallel quick sort.

Another good implementation is Hyper Quick Sort. Initially hyper quick sort is developed for hypercube interconnection networks, but it can be used on any message-passing system having number of processing elements in power of 2 [5]. The main difference between hyper quick sort and parallel quick sort consists in the method of choosing the pivot element. The average element of some block is chosen as the pivot element (generally, on the first processor of the computer system). The average element of some block is chosen as the pivot element (generally, on the first processor of the computer system). The pivot element is selected in such a way that it appears to be closer to the real mean value of the sorted sub sequence than any other arbitrarily chosen value. This concept is very close to our implementation as we use different processors for distribution of data blocks of array.

III. System Architecture

The main goal we want to achieve is to prove that the execution time of parallel sorting algorithm is better than serial algorithm. We will conduct that by an experiment, executing the program with linear increasing of the array's size, while comparing parallel with serial execution time, as well as calculating the linear speed up. Afterwards we will conclude that for different array sizes, we get a different speed up, which of course depends on the size of the cache-memory, the speed of the bus and the time needed for reading/writing data in memory. The main goal is using the parallel execution nature of the enviroment to provide simultaneously execution of the Quick Sort algorithm.

Here we are using the well known parallel enviroment for executing programs on multiple threads, OpenMP. OpenMP was chosen for measurement of the performances and execution of the program, as opposed to other parallel enviroments such as MPI and Posix (pthreads programming), because it represents standard for the implementation of shared-memory programming model, standardized with many finished APIs which can be used, also it uses high abstraction level during calls of functions and it can easily work on many computers with a shared-memory architecture [6].

1) Parallel Sorting Algorithms: With the advent of parallel processing, parallel and optimized sorting has become an important area for algorithm research. Most parallel sorting algorithms can be placed into one of two main categories: merge and partition based sorts [7]. Merge-based sorts consisted of many stages with merge operations on single processor, and perform well only when the number of processors is small. When the number of processors is getting larger, the overhead of scheduling and synchronization also increase, which reduces the speed up. On the other hand, partition-based sorts generally consists of these phases:

- Take a basic element (also called a *pivot*) from the array.
- Sort the array in a way that all the elements whose value is less than the *pivot*'s value come before it, while the elements whose value is greater than the pivot's value come after it (the elements with the same value as the pivot's can go either way). In the end, the pivot's position is the final one.
- This operation is also called a *partition operation*. Recursively repeat the previous steps with the sub-arrays with elements whose value is smaller and respectively with the sub-arrays with elements whose value is bigger.

2) Parallel quick sort: Not only quick sort is considered to be a better performing sorting algorithm but it is also considered to be one of reliable algorithm which can be parallelized. As it shown in Figure 1 the division of the array is done in a way that all the elements with a value smaller or equal to *pivot* are moved before the *pivot*, while the elements with a bigger value are moved after it. This process also finds the final position of the pivot element and returns it as an index.

It can be noticed that an element can be moved several times before the right position is found. Furthermore, in the case where the pivot is a duplicate in the array, the elements can be arranged in a different schedule. This isn't a mistake in the division of the array, as later the sorting will unite the elements in a final sorted array [8].

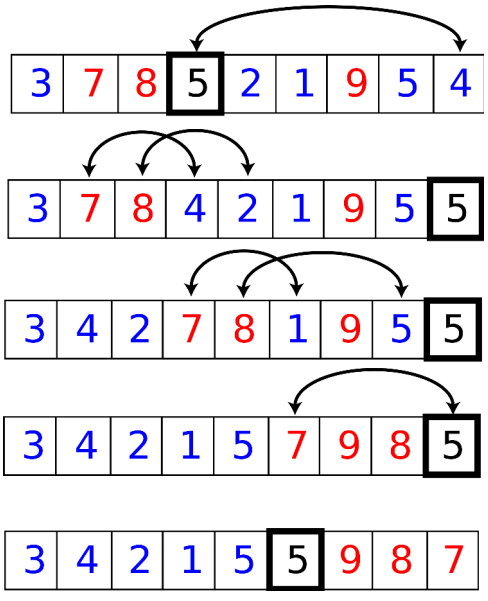


Fig. 1. Illustration of steps of executing quick sort algorithm realized with partition function.

Key feature of Parallel Quick Sort is parallel partitioning of

the data across different processing units. This may be obtained in the simplest way for a network of processing elements. A possible method to execute the first iteration of the parallel method is as follows:

- Select the pivot element from the subsequence and broadcast it to all the processors.
- Subdivide the sub arrays available on each processor into two parts using the pivot element.
- After this parts are formed, exchange of the data among the processors takes place. As a result of data transmissions, the parts of data blocks with data values smaller than pivot element must appear on the processor having position 0. The processors with position 1 must collect all data values exceeding the value of the pivot element.

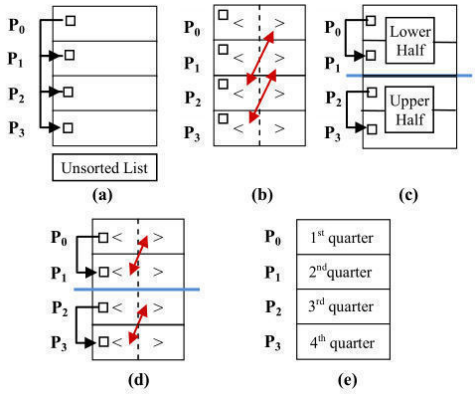


Fig. 2. Figure shows the steps of executing the parallel quick sort

3) Pseudo-code of Parallel Quick Sort Algorithm: As it is shown in the pseudo-code in Figure 3, every recursive call of the quick sort function reduces the size of the array because at least an element will be sorted. This ensures that algorithm completes it's execution in at least *n* recursive calls.

Quick-Sort has some shortcomings in comparison with other comparative algorithms, such as merge sort, which further complicates the efficiency of the parallelization. The depth to which the quick sort tree will go (this refers to the divide-and-conquer tree of branching), directly impacts the stability of the algorithm and the efficiency and size of the memory which the program will require [8].

IV. Results

Measuring of the performance were made on the cluster owned by FINKI (os.finki.ukim.mk), while the program was executed with 4 threads (although the number of the threads can be changed before the execution of the program occurs). The size of the sequence was set up to change from 0 to 1.000.000 elements, in order to show us the better image of the execution time. Also comparison was made between serial and parallel execution.

First we present the average times of execution of the program (parallel and serial execution), taking a fixed size of array, while the number of threads changes exponentially.

```
/* Quick Sort parallel function*/
void qsort_parallel(int *a, int l,int r)
{
int pivot=a[r],tmp;
int less=l-1,i;
for(i=l;i<=r;i++)
{
  if(a[i]<=pivot)
   {
         less++;
         tmp=a[less]; /* swap the elements */
         a[less]=a[i];
         a[i]=tmp;
   }
}
#pragma omp task
qsort_parallel(l,less-1);
#pragma omp task
qsort_parallel(less+1,r);
#pragma omp taskwait
}
```

Fig. 3. Function Parallel Quick Sort implemented in C programming language, which takes three arguments: the array, lower index and higher index.

TABLE I. Average execution of the program with 1.000.000 array elements (in s.)

Threads	Parallel exec. time	Serial exec. time	Speed Up
4	9.909	28.526	2.87
8	10.253	28.526	2.78
16	12.709	28.526	2.24

As we can see from the table I, the relation between number of threads and total speed up is inverse. Namely at 4 threads we get speed up of 2.87, than at 8 threads we get speed up of 2.78 and at 16 threads we get speed up of 2.24. These could be related to the famous Amdahl's law [9], that the speedup of a program using multiple processors in parallel computing is limited by the sequential fraction of the program. However the speed up that we reached is over 2, that says we have a good implementation on our parallelized algorithm.

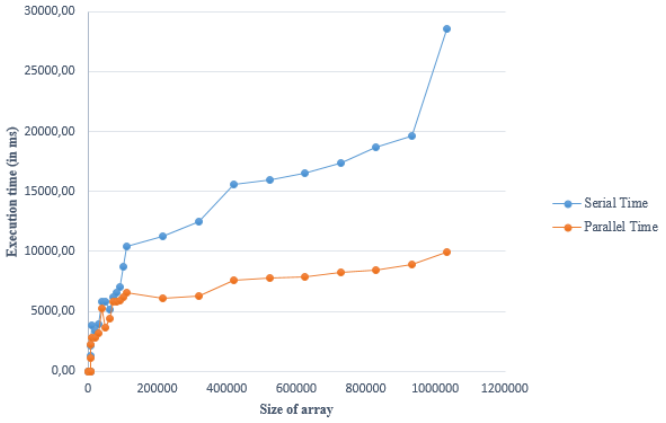


Fig. 4. Comparison of parallel and serial execution times.

We can see from the Figure 4, that the times required for parallel and serial execution of the Quick Sort algorithm are changed linear according to the size of array elements. Another

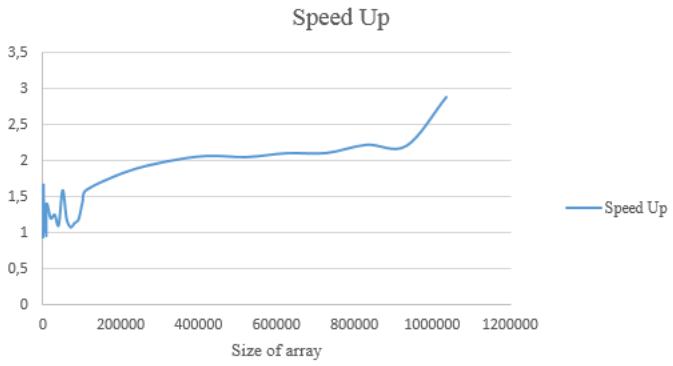


Fig. 5. Linear speed up of the algorithm.

thing that we can conclude is the rapid difference between these times, after the approximately the 200.000 element in the array. This could be reason because of the nature of parallel execution on multiple threads and size of array.

In the Figure 5 we can see the total speed up of the serial versus parallel execution. As mentioned earlier, from the figure it's shown that the speed up is growing faster after the array has gone larger.

V. Conclusion

In this paper we presented algorithm for array sorting (Quick Sort) that is implemented in parallel environment (OpenMP). We were using the approach to sort the sequence using *divide-and-conquer* concept for recursive programming model, each step of the sequence is divided into smaller parts, so these steps are repeated sequentially until initial case of the algorithm is reached.

We tested performance of parallel versus serial implementation of Quick Sort Algorithm. The input data elements was huge array with over 500.000 elements so it was required because this approach will efficient work with huge amounts of data. It is shown that Quick Sort, easily parallizable, has overall of 2x speed-up against serial sorting technique on CPU, so this approach works excellent with large size of data.

However, selecting the appropriate programming paradigm for implementing of the algorithm, still depends on the nature of the problem, and some other parts: hardware and software that will be used in servers, network communication, computer performance metrics. For example, if we have the opportunity for distributed processing and it takes less time for allocation and manipulation of memory, then for sure it's better using that approach.

A future step in our research regarding the parallel sorting of arrays will be creating more general sorting algorithm that will offer opportunities for execution on heterogeneous environment and faster optimal execution time. This includes linear speed up (relation between serial and parallel execution time), dynamic load balance, task granularity, and total efficiency.

References

[1] D. E. Knuth, "The Art of Computer Programming: Sorting and Searching.", *Addison-Wesley*, May, 1998, vol. 3, pp.145-162.

[2] J. L. Hennessy and D. A. Patterson, "Computer architecture: a quantitative approach", *Elsevir*, September, 2011, pp.246-389.

[3] G. Hager and G. Wellein, "Introduction to High Performance Computing for Scientists and Engineers", *CRC Press*, July, 2010, pp. 115-143.

[4] P. Pacheco, "An Introduction to Parallel Programming", *Morgan Kaufmann*, January, 2011, pp. 125-139.

[5] M. Quinn, "Parallel Programming in C with MPI and OpenMP", *McGraw-Hill Science/Engineering/Math*, June, 2003, pp.289-319.

[6] B. Chapman, G. Jost and R. Van Der Pas, "Using OpenMP: Portable Shared Memory Parallel Programming", *MIT press*, October, 2007, pp.124-148.

[7] R. Miller and L. Boxer, "Algorithms sequential and parallel: A unified approach. Cengage Learning.", *Cengage Learning (ICCSN)*, 2012, pp.178-190.

[8] R. Sedgewick and K. Wayne, "Algorithms (4th Edition)", *Addison-Wesley Professional*, March, 2011, pp.90-124.

[9] M. D. Hill and M. R. Marty, "Amdahl's Law in the Multicore Era", *IEEE Computer*, 2008, vol. 7, pp.140-168.

Benchmarking And Comparison Between Node.js Real Time Engines

Irina Arsovska*, Kristijan Trajkovski†, Monika Jandrevska‡, Vlado Velickovski§
*Faculty of Computer Science and Engineering
FINKI
Skopje, Macedonia
Email: arsovska.irina@students.finki.ukim.mk
†Faculty of Computer Science and Engineering
FINKI
Skopje, Macedonia
Email: trajkovski.kristijan@students.finki.ukim.mk
‡Faculty of Computer Science and Engineering
FINKI
Skopje, Macedonia
Email: jandrevska.monika@students.finki.ukim.mk
§ Faculty of Computer Science and Engineering
FINKI
Skopje, Macedonia
Email: velickovski.vlado@students.finki.ukim.mk

***Abstract*—Many of the applications which are widely used today use real time communication. Commonly used technologies for achieving that kind of communication are real time engines. There are many different implementations of this technology. Because of their wide range, developers who develop real time applications are faced with the problem of choosing the right technology which will provide best performance. For the experiments, we have developed a framework for benchmarking performance of Node.js real time engines. Based on the results of our experiments, we have concluded that ws and faye are the fastest, but they still have some drawbacks, outlined in this paper.**

***Index Terms*—Real-time engines benchmark, node.js, real-time communication benchmark, socket benchmark, socket.io, socket.js, primus, faye, websocket, rfc-6455**

I. Introduction

Since the nature of the web applications evolved through time, the performance of real time communication has become a problem for the developers. In order to solve this problem, many developers have worked towards different ways of achieving duplex communication between the server and the browser [1]. Real time engines are a new and revolutionary feature that provides a full-duplex communication channel operating over the Web through a single socket [1] Another definition for real-time engines is that they are libraries which allow developers to develop and maintain real-time applications. Common examples for real time applications are Chat systems, Notification systems, Multiplayer video games, Real-time charts, Air Traffic Control systems ,etc.

The aim of this scientific paper is to measure the performance and to make a comparison between different Node.js real time engines. Our hypothesis is that implementation which will show best performance on our experiments is **ws**. because it is low-level and only provides basic communication. Our hypothesis is supported in [2]. There are a few distinct implementations of Node.js real time engines. The most important difference between them is in the response time due to different times needed for establishing and terminating a connection and times needed for processing the user requests. Another very important attribute that may negatively impact performance is memory usage. The following Node.js real time engines are examined in this scientific paper:

- Socket.IO
  Socket.io is an abstraction layer for WebSockets. This real-time engine provides an easy server and client library for making real-time, streaming updates between a web server and a browser client [1]. Socket.io enables real-time bidirectional event-based communication. It works on every platform, browser or device, focusing equally on reliability and speed [3]. Socket.io is built so that both the sides can send messages or attach handlers to process the incoming messages [1].

- Sock.js
  Sock.js is a JavaScript library(for browser) that provides a WebSocket-like object. Sock.js gives a coherent, cross-browser, JavaScript API which creates a low latency, full duplex, cross-domain communication channel between the browser and the web server, with WebSockets or without [4].

- WebSocket
  The WebSocket JavaScript interface defines a full-duplex single socket connection over which messages can be sent between client and server [5]. The WebSocket simplifies much of the complexity around bidirectional web communication and connection management [5]. Ws is a simple to use WebSocket implementation and probably the fastest WebSocket library for Node.js [2].

- Primus
  Primus provides a common low level interface to communicate in real-time using various real-time frameworks [6]

- Faye
  Faye is a publish-subscribe messaging system based on the Bayeux protocol. It provides message servers for Node.js and Ruby, and clients for use on the server and in all major web browsers [7].

Socket communication. The WebSocket Protocol is designed to supersede existing bidirectional communication technologies that use HTTP as a transport layer to benefit from existing infrastructure (proxies, filtering,authentication). Such technologies were implemented as trade-offs between efficiency and reliability because HTTP was not initially meant to be used for bidirectional communication (see [RFC6202] for further discussion). The WebSocket Protocol attempts to address the goals of existing bidirectional HTTP technologies in the context of the existing HTTP infrastructure; as such, it is designed to work over HTTP ports 80 and 443 as well as to support HTTP proxies and intermediaries, even if this implies some complexity specific to the current environment. However, the design does not limit WebSocket to HTTP, and future implementations could use a simpler handshake over a dedicated port without reinventing the entire protocol. This last point is important because the traffic patterns of interactive messaging do not closely match standard HTTP traffic and can induce unusual loads on some components. [8]

Heartbeat messages. Heartbeat messages are important attribute when using Node.js real time engines. Heartbeat messages are short messages between the client and the server. They are sent on short, periodic intervals. This mechanism is used because the server needs to know if the client is still active in order to continue the communication.

II. Related Work

Because the real time engines we are writing about are relatively new, we faced lack of similar benchmark tests in order to make a comparison between the results.

Similar tests for Websocket, Socket.IO, Engine.IO and Primus have been performed. According to the author of the performance tests, which can be accessed at the following link:, https://medium.com/@denizozger/finding-the-right-node-js-websocket-implementation-b63bfca0539, Engine.IO has a significantly better memory usage compared to others. His tests have shown that for a constant throughput, Engine.IO's number of connections have the biggest impact on performance, then comes reducing size of data, and lastly reducing frequency of data sent.

Benchmark tests for Sock.js have been made, but an older version of Node (version 0.6.5) has been usedhttp://mrjoes.github.io/2011/12/15/sockjs-bench.html. According to the graphic representation of the results, sock.js starts to slow down around 45,000 messages. The approximate memory usage, with 2000 clients and reasonable rates for Sock.js was around 36 MB. Due to the older version of Node, as well as library versions, a new benchmark was needed to analyse and document the current performance, using new versions of the libraries as well as Node.js.

Another benchmark test can be accessed at the following link: http://drewww.github.io/socket.io-benchmarking/. This is benchmark test for Socket.IO only. According to the results, between 9,000 and 10,000 messages per second, the round trip time starts to grow.

III. The Experiments

Our framework. In our experiments, we measure the time needed for the client's request to be sent to the server, the time needed for the server to process the client's request and the time needed for the server to send back a response to the client and the client to receive the response. The overall client-server-client response depends on the software and hardware specifications. The experiments are performed

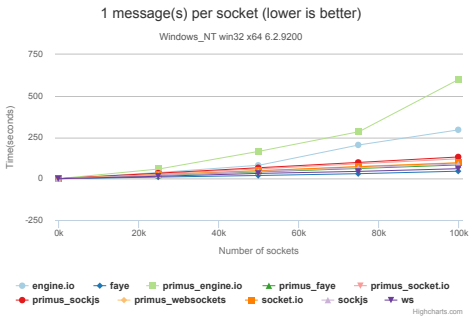


Fig. 1: Benchmark results - 1 message per socket

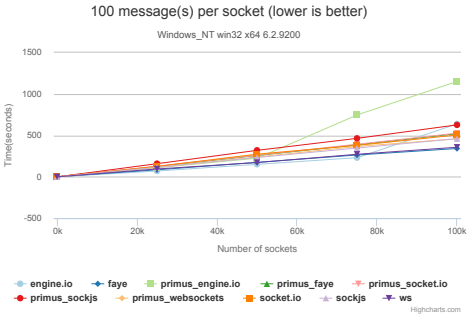


Fig. 2: Benchmark results - 100 messages per socket

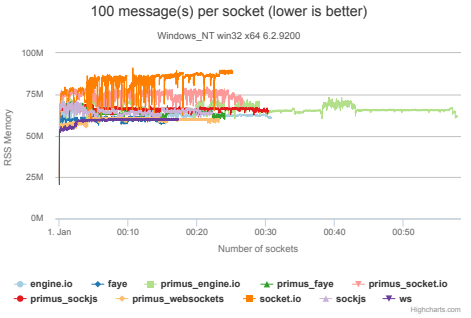


Fig. 3: Benchmark results - memory consumption for 100 messages per socket

on the same system that has the hardware and software specifications shown in Table 1. They were repeated 3 times and each of them needed 2 hours to complete. We developed a Node.js socket benchmarking framework for the purpose of providing experiments' results for this paper. The project is available on GitHub and can be accessed on the following link: https://github.com/freezx/sockbench. It is an open source project.

How our tests work: First of all, the main file, **server.js** requires the needed backend middleware based on the program arguments which are located in the **backends** folder. Each backend middleware provides functions for creating a server, creating a client and running a test with the specified real time engine. Using the provided function, our software creates a server and starts listening for connections on the specified port. Then, it executes several instances of **client.js**, each running on one processor core in the host system. This is done so that the running server can get a bigger number of client connections that are responsive, thus measuring performance under heavy loads. Each client connects to the server and sends a number of messages, and then expects a response for each sent message. If it loses the connection, it reconnects and resends the messages for which it hasn't received a response for. After the required number of sockets finish with these operations, the time needed for their completion is recorded to a file.

How our charts are generated: We have developed a special script for the purpose of showing the results of our experiments. The steps fot generating the charts are explained below:

1) The script collects all recorded data from the benchmarks ever performed
2) It calculates the average values for every output
3) It renders a HTML document containing a charting library which displays the output

TABLE I: Hardware and software characteristics of the system on which the experiments are performed:

Processing Power (CPU)	i5 4670k
Random Access Memory (RAM)	16GB
Operating System (OS)	Windows 8.1

IV. Results

The results are presented in Figure 1 and Figure 2. The X-axis represents the number of sockets tested, and the Y-axis is the round trip time for all sockets. The first experiment, represented in Figure 1, was measuring the needed time to send and receive one message for each real time engine with 1, 25,000, 50,000 and 75,000 sockets. The second experiment, represented in Figure 2, was the same as the first, but this time, instead of 1 message, we measured the needed time for sending and receiving 100 messages. According to the results from the previous two experiments, when the number of sockets exceeds approximately 50,000 sockets, **primus with the engine.io transformer**'s response time grows rapidly. But when the number of sockets exceeds 75,000, **engine.io**'s

response time also starts growing rapidly. Also, in both experiments **ws** and **faye** were the fastest Node.js real time engines. We also measured the memory consumption of the server and the results are shown in Figure 3. According to these results, there is no significant difference between the servers' memory footprint or evident memory leaks.

V. Conclusions and Future Work

While real time applications have become a routine today, the demands for better, faster and more reliable real time communication are growing strong. Our hypothesis, that **ws** and **faye** are the fastest Node.js real time engines, was proven correct by our experiments. According to the evaluation of the experiments' results, **ws** and **faye** have shown best performance due to their simplicity. However, they are less flexible than the other solutions, because they don't provide heartbeats and message encoding by default, so additional work needs to be done to make it more usable. On the other hand, the other Node.js real time engines, Socket.io, Engine.io, Sockjs and Primus, support other features that makes them slower. One of features that costs the most is encoding which is supported only by Socket.io and Primus. Another feature that has impact on the performance are heartbeat messages which are supported only by Socket.io and Primus. We are currently extending our experiments and performing the benchmark tests on Linux. We also plan to test the performance of other advanced features such as broadcasting and rooms.

References

[1] Rai, R., *Socket. IO Real-Time Web Application Development*, Packt Publishing, Limited, 2013
[2] *ws: a node.js websocket library*, https://github.com/einaros/ws, Last access: December 15, 2014
[3] *Socket.io*, www.socket.io,Last access: December 15, 2014
[4] *WebSocket emulation - Node.js server*, https://github.com/sockjs/sockjs-node.git, Last access: December 15, 2014
[5] *Websocket.org*, http://www.websocket.org/index.html, Last access: December 15, 2014
[6] *Primus*, https://github.com/primus/primus, Last access: December 15, 2014
[7] *Faye*, http://faye.jcoglan.com/, Last access: December 15, 2014
[8] *I.Fette & A. Melnikov, The Websocket Protocol, December 2011* https://tools.ietf.org/html/rfc6455#section-2.1

Twitter sentiment analysis: generating automatically labeled sets

Stevica Bozhinoski* and Ana Tanevska†
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
Email: bozhinoski.stevica@students.finki.ukim.mk* and tanevska.ana@students.finki.ukim.mk†

***Abstract*—Social networks provide us with an abundance of free-form data. From this data we can extract a variety of information, one of which is the overall public opinion in certain topics of interest, i.e. their sentiment expression. However, in order to do this we need a manually labeled training set which takes time and resources to make it. In this paper we attempted a different approach for the problem of sentiment analysis of tweets. We trained a Naive Bayes classifier based on manually labeled tweets. We analyzed the classifier and extracted the most informative features. We then used these features to download new sets of tweets and automatically labeled them. These tweets are tested on our previously trained classifier and evaluated in order to determine the importance of each set for future usage.**

***Index Terms*—Natural Language Processing, Sentiment Analysis, Twitter, Classification, Naive Bayes, Social Networks**

I. Introduction

Soon after the Internet became available for a lot of households, the social networks emerged. The social networks made it possible for people from all around the world to connect based on their similar interests. Examples like Facebook[1], Tumblr[2] and Twitter[3] generate an enormous amount of data every day. Twitter is especially famous for its option that allows users to write short messages(up to 140 characters long) called tweets in order to express an opinion about a topic. Many businesses and agencies can use this information to determine the public opinion about their product or service. Also, Twitter can be used for predicting political elections [9] [10] and the situation on the stock market. [3] To train a supervised system that can classify each tweet based on its sentiment, we needed a manually-labeled training set which is hard to obtain, takes a lot of time and is rarely objective. This is why we looked more into the problem of automatically labeling the tweets. In this paper we present our system in a step-by-step procedure as follows: In Section II we're going to take a brief look at the work made by others in this area. In Section III we're going to present the overview of our solution. In Section IV we're going to describe our dataset, show how each of our components work and analyze the performances of our trained binary and 3-way classifier. We're going to show how we obtained the automatically labeled sets of tweets and later based on the binary classifier we are going to test these sets of tweets. In Section V we're going to present the results from the tests with our automatically labeled sets of tweets. Finally, in Section VI we're going to summarize the results of our experiment and present the conclusion.

[1]https://www.facebook.com/
[2]https://www.tumblr.com/
[3]https://twitter.com/

II. Related work

A system for automatically labeling of tweets was presented before by Alexander Pak and Patrick Paroubek. [8] However in their attempt they did not give a starting manually labeled set that can be used as a baseline. There is a big progress made with different classifiers: Support Vector Machines, Naive Bayes and Maximum Entropy. [4] Most of the features are obtained by word unigram [1], bigram and N-gram models. [8] Some classifiers additionally use twitter specific features like number of hashtags, emoticons, abbreviations, capital words [6] and others use convolutional neural networks for feature selection. [5].

III. Overview of our Solution

We started with a set of tweets obtained from SemEval[4] and given to us by our mentor and professor Ph.D. Gjorgji Madjarov. We first used this set for our project "Real-time Sentiment Analysis of Tweets" as a part of the course Pattern Recognition. We chose Python[5] as the programming language mainly for its easy manipulation with strings and its vast libraries for scientific calculations, string processing and natural language processing. The tweets first are preprocessed and the features relevant for this task are extracted. After that we trained the binary and 3-way classifiers and performed 10-fold cross validation for computing the performances of each classifier. You can refer to the architecture in Figure 1. Next, we extracted the most informative features and used most of them to download new sets of tweets and manually label them. Finally, after we labeled them they were tested and evaluated with the same metrics.

IV. Experiment

A. Our dataset

Our dataset is composed of a set of tweets manually labeled as 'positive', 'negative', 'neutral', 'objective' and 'neutral-

[4]http://www.senseval.org/
[5]https://www.python.org/

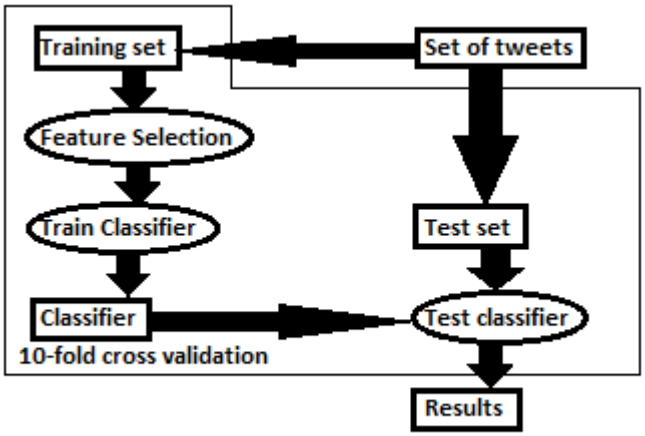


Fig. 1: In this figure we show the architecture of our classifier and the 10-fold cross validation used to calculate the performances.

OR-objective' based on the sentiment of the tweet. For the purpose of this task we removed the tweets labeled as either 'objective' and 'neutral-OR-objective'. After this removal we were left with a set of 9085 tweets that we used to train our 3-way classifier. We further removed the tweets labeled as 'neutral' and this resulted with 6202 tweets labeled as either 'positive' or 'negative' used for our binary classifier. On the table you can see detailed composition of the sets used in our experiment.

TABLE I: On this table we display the composition of the sets used by our classifiers.

Set	positive	negative	neutral	total
3-class	4421	1781	2883	9085
Binary	4421	1781	0	6202

B. Preprocessing

Natural languages used in social networks such as Twitter constantly evolve. We encounter slang, neologisms, emoticons and hashtags. Emoticons are sequences of characters used to specify an emotion. For the needs of our project we used the emoticons from Wikipedia[6] labeled as smiley, laughing and frown. Hashtags are phrases that begin with # character and are used to specify the topic of a tweet. Hashtags also help the tweet to become more visible inside the Twitter community.

In the preprocessing part of our system we tokenized the tweets into words. Interpunction signs were then removed from the beginning and the end of each word. Then we selected the following features:

- Boolean values for each of the words present in the training sets. Every word represents a separate feature at unigram level. Because of the frequency of different URLs and usernames, we converted every URL into http://someurl and every username into @someusername. We additionally looked if a negation was present before a word. If it was, we added the _NEG suffix to the word.
- Number of positive emoticons. In this category we counted the emoticons labeled by Wikipedia as 'smiley' and 'laughing'.
- Number of negative emoticons. In this category we counted the emoticons labeled by Wikipedia as 'frown'.
- Number of hashtags.
- Number of negations. In this category we counted the words containing the sequence of characters 'n't' at the end of the word and the word 'not'.

C. Training the classifier

Our choice for a classifier was the Naive Bayes classifier from the Natural Languages Toolkit (NLTK)[7] Python library. [2] We chose this classifier because of its good performance when working with a big feature space. For the evaluation of the classifier, we used 10-fold cross validation. We used precision, recall, F1 score and accuracy as metrics for evaluation of both the binary and multi-class classifier. They are explained as follows:

- Precision is determined by the number od true positive cases divided by all of the cases predicted as positive.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

- Recall is determined as the number of true positive cases divided by all of the cases labeled as positive.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

- F1 measure is a harmonic mean of the precision and recall. [7]

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

- Accuracy is the number of correctly classified cases divided by the number of elements.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

TN, TP, FN and FP are values from the confusion matrix. TP denote the number of tweets correctly identified, TN is the number of tweets correctly rejected, FP is the number of tweets incorrectly identified and FN is the number of tweets incorrectly rejected.

The results from the evaluation of the binary classifier are given in table II, while the results from the evaluation of the 3-way classifier are given in table III.

[6]http://en.wikipedia.org/wiki/List_of_emoticons

[7]http://www.nltk.org/

TABLE II: Here we show the precision, recall, F1 score and accuracy for the binary classifier.

Classifier	Precision	Recall	F1	Accuracy
Binary	0.9007	0.7018	0.7889	0.7149

TABLE III: Here we show the precision, recall, F1 score and accuracy for each of the classes: positive, negative and neutral trained on our 3-way classifier.

Class	Precision	Recall	F1	Accuracy
Positive	0.6293	0.4473	0.5229	0.5992
Negative	0.3483	0.3536	0.3509	0.3733
Neutral	0.5687	0.2144	0.3114	0.5157

D. Most informative features

After we have trained the classifier we used the function show_most_informative_features()[8] from the NLTK library. This function retrieves the features that have the biggest role in determining if a tweet is positive, negative or neutral. This function is pretty useful because it also displays the strength of the word sentiment. The most informative features from the binary classifier are present in table IV, while the ones from the 3-way classifier are present in table V.

TABLE IV: In this table we present the most informative features gained from the training of the binary classifier. Here we also show the strength of the respective feature. For example: the word 'lose' is 20.7 times more probable to be in a tweet labeled as negative than in a positive one.

Feature	Ratio	Values
lose	negative : positive	20.7 : 1.0
sad	negative : positive	20.3 : 1.0
make_NEG	negative : positive	19.0 : 1.0
tired	negative : positive	19.0 : 1.0
injury	negative : positive	19.0 : 1.0
countNEG	negative : positive	18.9 : 1.0
worst	negative : positive	17.4 : 1.0
dont	negative : positive	15.2 : 1.0
hate	negative : positive	14.6 : 1.0
excited	negative : positive	14.4 : 1.0

E. Extraction of tweets

Based on the most informative features received in the previous step, we downloaded sets of 1000 tweets(12000 in total) that contain one of the words: 'worst', 'tired', 'lose', 'injury', 'hate', ':(', 'sad', 'don't', 'love', ':)', 'fun' and 'excited'. From now on in this paper we are going to use the word seed meaning a sequence of characters that was used for downloading the sets of automatically labeled tweets.

We automatically labeled the tweets that contained the seeds 'worst', 'tired', 'lose', 'injury', 'hate', ':(', 'sad' and 'don't' as negative and tweets that contained the seeds 'love', ':)', 'fun' and 'excited' as positive. In the download process we checked if more than 70% of the words belonged to Standard US English using the Python Enchant[9] library. We determined that when we raised the bar higher than 70%, the system didn't accept tweets with few abbreviations, emoticons and slang written in English, whereas when we lowered this bar, we encountered tweets written in foreign languages that used only a few words and phrases written in English. Also, our automatically labeled tweets are unique i.e. only one tweet with the same sequence of characters is present in its respective set.

After the process of downloading a total of 9000 tweets and automatically labeling them, we used these as our test set on our already trained binary classifier separately.

TABLE V: In this table we present the most informative features gained from the training of the 3-way classifier. Here we also show the strength of the respective feature. For example: the emoticon ':(' is 54.5 times more probable to be in a tweet labeled as negative than in a positive one. *Because of indecent language, some of the letters are replaced by special characters.

Feature	Ratio	Values
:(	negative : neutral	54.5 : 1.0
f@#k*	negative : neutral	26.4 : 1.0
a$$*	negative : neutral	24.3 : 1.0
excited	positive : neutral	23.3 : 1.0
lose	negative : positive	20.7 : 1.0
fun	positive : neutral	20.7 : 1.0
sad	negative : positive	20.3 : 1.0
tired	negative : positive	19.0 : 1.0
make_NEG	negative : positive	19.0 : 1.0
injury	negative : positive	19.0 : 1.0

TABLE VI: In this table we show the sets of tweets containing their respective seeds, F1 score and accuracy.

seed	F1	Accuracy
worst	0.8931	0.807
tired	0.9047	0.826
lose	0.8998	0.818
injury	0.8130	0.685
hate	0.9384	0.884
:(	0.9954	0.991
sad	0.8851	0.794
dont	0.8331	0.714
love	0.9004	0.8189
:)	0.8532	0.744
fun	0.8888	0.8
excited	0.9642	0.931

V. Results and evaluation

Finally, we tested the automatically labeled tweets on our binary classifier and calculated the F1 score and accuracy for each of the sets. The results are given in table VI. We can see that the most informative set of tweets we got is the one containing the seed 'frown' with accuracy of 99.1%, while

[8]http://www.nltk.org/howto/chunk.html

[9]http://www.abisource.com/projects/enchant/

the least informative is the set of tweets containing the seed 'injury' with accuracy of 68.5%.

VI. Conclusion

From the results of testing our binary classifier we can see that we achieved an accuracy of 71%. This result is pretty good since in real life, around 79 %[10] of people agree on the sentiment of a topic. Working with social networks is somewhat hard because of the sentences present in natural language. We can see a lot of abbreviations, slang, foreign words, emoticons, neologisms, ambiguity and other non-standard words. Also sarcasm is an important issue - in most of the cases where sarcasm was present, it was classified as false.
From the automatic labeling of tweets we can conclude that using the right seeds makes a big difference in the quality of the labeled tweets. If we use the ':(' and 'excited' as the negative and positive seeds respectively, we get set of automatically labeled tweets which is really close to the manually labeled tweets. On the other side, 'injury' is the least informative seed from our experiment when it comes to binary classification with an accuracy close to that of randomly choosing a tweet.

VII. Future Work

In the future we would like to extend the automatic labeling system to more seeds. We would also like to expand our work with the new concept of convolutional neural networks (CNN) for feature selection that has given good results in image-processing tasks. There are also some experiments made on text sentiment analysis tasks that show promising results in the area of Natural Language Processing. The smaller feature space will enable us to test this algorithm on a bigger number of classifiers.

Acknowledgment

We would like to thank professor Ph.D. Gjorgji Madjarov for his advices and support during the process of writing our paper and correcting us where needed.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[5] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[6] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR13)*, 2013.

[7] Arzucan Özgür, Levent Özgür, and Tunga Güngör. Text categorization with class-based and corpus-based keyword selection. In *Computer and Information Sciences-ISCIS 2005*, pages 606–615. Springer, 2005.

[8] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.

[9] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[10] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

[10] http://mashable.com/2010/04/19/sentiment-analysis/

PageRank Approach to Ranking National Football Teams

Verica Lazova *

Lasko Basnarkov

Faculty of Computer Science and Engineering
Cyril and Methodius University
Skopje, R.Macedonia

E-mail: *lazova992@gmail.com

***Abstract*—The Football World Cup as world's favorite sporting event is a source of both entertainment and overwhelming amount of data about the games played. In this paper we analyse the available data on football world championships since 1930 until today. Our goal is to rank the national teams based on all matches during the championships. For this purpose, we apply the PageRank with restarts algorithm to a graph built from the games played during the tournaments. Several statistics such as matches won and goals scored are combined in different metrics that assign weights to the links in the graph. Finally, our results indicate that the Random walk approach with the use of right metrics can indeed produce relevant rankings comparable to the FIFA official all-time ranking board.**

***Keywords*—football, random walk, ranking, network analysis**

I. Introduction

Football, being the world's most favored sport, draws people's attention in every field, from the simple means of entertainment to more complex objectives of statistics, research and data analysis. Since the FIFA world cup first took place in 1930 until this day, there have been around 20 tournaments held, each comprising of about 64 matches, not counting the qualification rounds [1], [2]. Therefore, there is significant amount of data that one could inspect, analyse and draw conclusions from.

Having that in mind researchers are tackling problems regarding playing strategy, ranking of teams or performance analysis from different aspects including economic, demographic, cultural and climatic factors [3]. A team's game strategy for example can be observed from graph theory perspective by constructing a network of passes between players. In this context different centrality measures can be used to determine the importance of particular players [4]–[6]. Other subject of interest might be modelling football matches in terms of scores during the game. For example, in [7] the authors discuss a statistical model for scoring times in a match.

Here we address the problem of ranking national football teams. Our main task is to use the available statistics, in order to come up with an alternative ranking method for the football teams based on their achievements at the world cups. There are different rating methods currently in use and they produce relevant results. FIFA have their own 4-year points based FIFA/Coca-Cola rating system [8] and world cup all-time ratings [9] that includes all championships since their origin. There are also the World Football Elo Ratings based on the rating system FIDE uses to rate chess players [10].

A good ranking method should not only take into account how many times a team has won, but also consider how strong an opponent they have defeated. Victory against stronger opponent is preferable and thus more significant than victory against weaker opponent. One method that incorporates such logic is the PageRank (Random walk) method, which is applicable to vast varieties of network based problems that require ranking in some way. Other than the well known problem of rating web-pages [11] it is also utilized in social network analysis, in tasks such as link prediction, information diffusion and communities detection [12]–[14]. Also it is used in NLP for the purpose of text summarization and word sense disambiguation [15], [16]. For previous attempts of employing PageRank mechanism in sporting events we refer the reader to [17]–[19].

The rest of the paper is organized as follows. In Section II we present the ranking problem and the PageRank based method for solving it. We also give description and statistics of the data that was available to us. The obtained results are presented in Section III including a discussion and comparison to the official rankings and then we conclude the paper in Section IV.

II. Materials and Methods

A. Data

The data we used was obtained from 11v11, web-site for football statistics that contains all time figures about the matches of the world cup, qualification games inclusive [20]. For each national team there is information on which country they have played against, the number of matches won, drawn and lost, as well as the number of scored and conceded goals during all match-ups. Throughout this paper we use the term *match-up* in context of *a single game played between two teams*. And a *match-up pair* are *every two teams that have played against each other*. The dataset, contains 210 countries and statistics on 2335 match-up pairs that have played against one another, or 7141 games in total, during which 20298 goals were scored. The average number of games per match-up pair

Table I
SET OF TESTED WEIGHTING FUNCTION AND THEIR SCORE IN NORMALIZED NUMBER OF INVERSIONS AS SIMILARITY METRICS TO THE OFFICIAL RANKINGS. LESS IS BETTER.

#	WEIGHTING FUNCTION	INVERSIONS
1	$f_{i,j} = \frac{l_{i,j}}{g_{i,j}} \cdot \frac{1}{G-g_{i,j}+1}$	0.032
2	$f_{i,j} = \frac{l_{i,j}}{g_{i,j}}$	0.038
3	$f_{i,j} = \frac{l_{i,j}}{g_{i,j}} + \frac{c_{i,j}}{c_{i,j}+s_{i,j}}$	0.040
4	$f_{i,j} = l_{i,j}$	0.040
5	$f_{i,j} = \frac{c_{i,j}}{s_{i,j}}$	0.041
6	$f_{i,j} = \frac{l_{i,j}}{w_{i,j}}$	0.043
7	$f_{i,j} = \frac{l_{i,j}}{g_{i,j}} + 0.5 \cdot \frac{d_{i,j}}{g_{i,j}}$	0.044
8	$f_{i,j} = \frac{c_{i,j}}{c_{i,j}+s_{i,j}}$	0.044
9	$f_{i,j} = \frac{c_{i,j}}{g_{i,j}}$	0.046
10	$f_{i,j} = c_{i,j}$	0.050

is 3.0582, and the average number of goals scored per match-up pair is 4.3465. Mexico versus USA is the pair with the largest number of games played against one another. About 28 games were played during which around 100 goals were scored, 15 of which were won by the US, 6 were drawn and the other 7 resulted in a victory for Mexico. The country with the most games played is Brazil with about 200 matches and also is the country with most games won and most goals scored as expected.

B. Method

The ranking method explored throughout this paper is the PageRank with restarts algorithm applied to a graph build around the supplied data [11]. Each national team is a single node in the graph and two nodes are linked if the two teams (the match-up pair) have ever competed against each other in a world cup tournament. The weight of the link is determined by a weighting function that involves one or more metrics such as number of games played between a match-up pair, the number of won, lost and drawn games, or the number of scored and conceded goals. The various weighting functions we have tested are given in Table I.

Within the functions we use the following notation:

- $f_{i,j}$ weight of the link from node i to node j;
- $g_{i,j}$ number of games played between the two teams;
- $l_{i,j}$ number of games lost by team i amongst all the games i and j played;
- $w_{i,j}$ number of games won by team i amongst all the games i and j played;
- $c_{i,j}$ number of goals conceded by team i during all the games i and j played;
- $s_{i,j}$ number of goals scored by team i during all the games i and j played;
- $d_{i,j}$ number of games drawn between the two teams;
- G maximum number of games played between any match-up pair;

Another factor that affects the PageRank is the damping factor. The damping factor corresponds to the probability that a random walker would discontinue the walk and jump to a random node [21]. The damping factor other than being necessary as assurance that the random walk would converge to a stationary distribution, it is also intuitive. The intuition behind the use of damping factor within our match-ups network is the following: although the graph is dense not every team have played against every other. So when using weighting metrics such as the loss ratio (function 1 in Table I) the damping factor would mean adding some wining chances to all the teams that have never been played against. It also adds some wining chances to a team that has never won a game within a match-up.

The PageRank is calculated using the power method [22]. This method is an iterative algorithm (eq. 2) that finds the dominant eigenvector π, which corresponds to the invariant distribution of the time a random walker spends at a certain node - the PageRank. By normalizing the adjacency matrix A we get the transition probability matrix Q with elements as given in eq. 1, where N is the number of nodes.

$$Q_{i,j} = (1-d) \cdot \frac{A_{i,j}}{\sum_{k=1}^{N} A_{i,k}} + \frac{d}{N} \tag{1}$$

$$\pi^T = \pi^T Q. \tag{2}$$

Note that Q is guaranteed to be irreducible and aperiodic as a consequence of the nonzero damping factor d.

Table II
NUMBER OF GAMES PLAYED AND RESULTS FOR EACH MATCH-UP PAIR

PAIR	GAMES	RESULTS
A-B	3	A wins 2, B wins 1
A-C	3	A wins 2, C wins 1
A-D	3	A wins 3, D wins 0
B-C	3	C wins 3, B wins 0
B-D	3	D wins 3, B wins 0
C-D	3	C wins 1, D wins 2

Table III
THE PAGERANK OF EACH TEAM IN DESCENDING ORDER

TEAM	GAMES	WIN	PAGERANK
A	9	7	0.333
C	9	5	0.281
D	9	5	0.211
B	9	1	0.175

C. Example

For the sake of demonstration, let's consider a toy example that illustrates our goal. Suppose there are 4 teams and the given statistics for each pair are shown in Table II. The graph (Fig. 1) is built using loss ratio as metric (function 2 at Table I). Therefore the weight of a given link from i to j is the part of the games that i has lost to j. For instance there is a link from A to C with weight of $\frac{1}{3}$ and also a link from C to A with weight of $\frac{2}{3}$. That means out of 3 matches A and C have

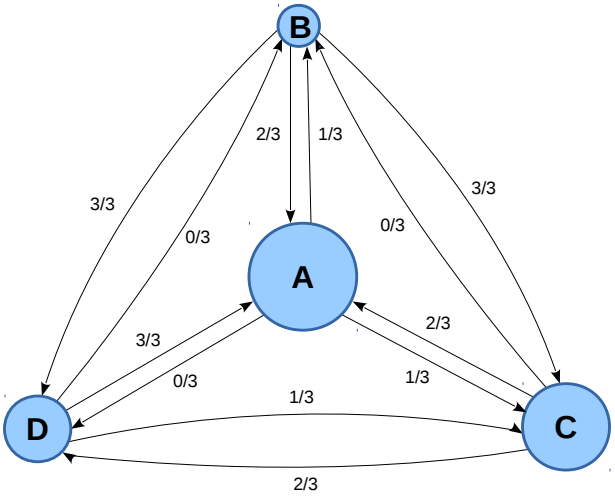


Figure 1. Graph representation og the games played, the size of each node is proportional to it's PageRank

played against each other A has won 2 matches, C has won 1 and no matches were drawn. The next step is calculation of the PageRank. Therefore we need transition probability matrix which is calculated according to eq. 1 with a common damping factor value of 0.15.

Finally the results are shown at Table III. A is pointed as highest ranked and B is lowest ranked team as expected. On the other hand, team C and team D both have won 5 games as shown in Table III. However, PageRank takes into account the strength of the defeated opponent not only the number of winnings. As a result, team C is ranked higher since they have won a game against A, considered as strong opponent, in contrast to team D who have winnings only against weaker opponents.

III. Results and Discussion

In order to find the most precise ranking several different weighting functions have been tried and almost all of them delivered similar results. The results were evaluated by comparing the PageRank to the official world cup ranking. We have used normalized number of inversions as evaluation metric [23], taking the official FIFA all-time rankings as referent ordering. The tested weighting functions and their scores are listed at Table I. Lower score means the results generated using the corresponding metric are more similar to the official ranking. We only used the top 30 highest ranked teams in the comparison because we wanted to give them higher priority and get their ordering right at the cost of misplacing some of the lower rated teams. The error of the weighting functions also depends on the damping factor. The minimum is achieved when the damping factor value is very small, around 0.05. That is the value we used in the evaluations of the metrics shown in Table I. Fig 3 shows errors (in normalized inversions count) for the top 5 metrics as functions of the damping factor. As expected the error increases with the growth of the damping factor. Table IV

Table IV
Top 20 highest ranked national teams using combination of loss ratio and number of games the two teams played as weighting function (function 1 in Table I) and 0.05 damping factor. The 4-th column gives their position in the official ranking

#	COUNTRY	PAGERANK	OFFICIAL
1	Brazil	0.040375	1
2	Italy	0.037992	3
3	Germany	0.033801	2
4	Netherlands	0.031052	8
5	Argentina	0.029159	4
6	England	0.029100	6
7	Spain	0.027904	5
8	France	0.025670	7
9	Czechoslovakia	0.025155	NA
10	Sweden	0.022882	10
11	Mexico	0.022034	13
12	Hungary	0.022014	16
13	Uruguay	0.020660	9
14	Belgium	0.020255	14
15	Portugal	0.020211	17
16	Poland	0.019528	15
17	Denmark	0.019206	25
18	Croatia	0.018993	27
19	Switzerland	0.016650	21
20	Yugoslavia	0.016466	NA

shows the top 20 teams (for brevity), according to our best weighting function. The 4-th column contains the positions for each team at the official rankings board. The position is marked green if the team holds the same place in both ours and the official rankings. The position is marked with red if there is a large displacement (Denmark and Croatia). If a team is not found in the official ranking (Czechoslovakia and Yugoslavia in our case) their position is marked with NA. Fig 2 shows the match-ups graph. Each team is a node in the graph represented by their national flag and the size of each node is proportional to it's PageRank. In the figure a portion of the links are omitted for the sake of clarity, thus the real graph is much denser than it appears.

Possible issue when using PageRank as ranking method might be the following: A node can obtain a high PageRank score if it has a high ranked neighbour from which it can receive significant amount of votes or if it has many low ranked neighbours. In our example, if a national team is high ranked then they must have either defeated many low ranked teams or achieved remarkable results against a highly ranked opponent. This property of the Random Walk affects our results especially since we treat all matches equally, without taking into account whether it is qualification round or final game. As a result there might be teams that have received high ranking only because they have played and won against many low ranked opponents in less significant qualification matches.

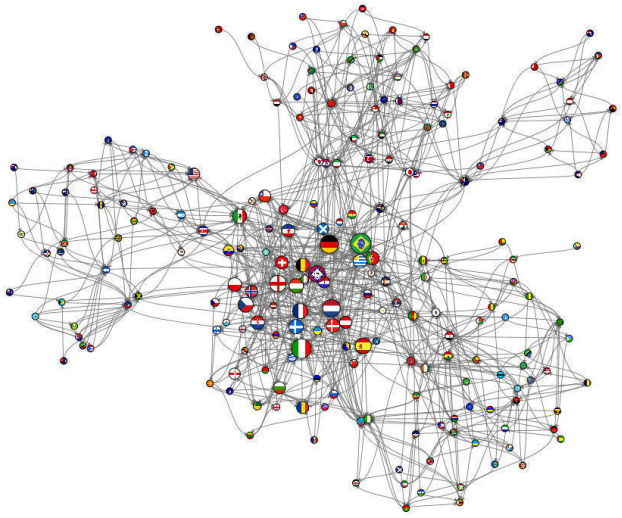


Figure 2. The graph of the match-ups built with the combination of loss ratio and number of games the two teams played as weighting function (function 1 in Table I). The size of each node corresponds to their PageRank (damping factor of 0.05 used). For the sake of clarity only the strongest links coming out of each node are shown.

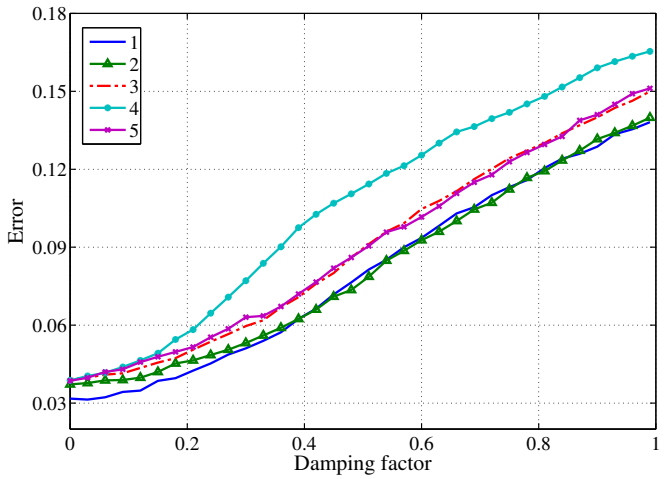


Figure 3. The error in normalized number of inversions of the first 5 weighting functions in Table I as function of the damping factor

IV. Conclusion

Throughout this paper we explored the PageRank method for ranking national football teams. Our results showed that even with simple weighting functions such as ratio of the goals scored or matches won, the PageRank algorithm derives promising results. The rankings this method produced are similar to the official FIFA all-time rankings. However, it is difficult to evaluate whether the PageRank with use of more sophisticated weighting function and more features within the dataset could lead to a better rating scheme than the official. Anyway, under the assumption that the FIFA ranking system is proper and accurate, RandomWalk despite the simple dataset and weighting metrics can replicate it's results in a great deal.

Acknowledgement

We would like to thank Andrej Gajduk and Igor Trpevski for fruitful discussions and comments. VL also thanks TAPAN MNG D.O.O.E.L. Negotino for the internship opportunity during which the presented work was completed.

References

[1] F. I. de Football Association *et al.*, "Fifa competitions and olympic football tournaments 1908-2017," 2014. [Online]. Available: http://www.fifa.com/worldcup/organisation/documents/index.html

[2] ——, "Fifa world cup comparative statistics 1982-2014," 2014. [Online]. Available: http://www.fifa.com/worldcup/organisation/documents/index.html

[3] R. Hoffmann, L. C. Ging, and B. Ramasamy, "The socio-economic determinants of international soccer performance," *Journal of Applied Economics*, vol. 5, no. 2, pp. 253–272, 2002.

[4] J. L. Peña and H. Touchette, "A network theory analysis of football strategies," *arXiv preprint arXiv:1206.6904*, 2012.

[5] J. Duch, J. S. Waitzman, and L. A. N. Amaral, "Quantifying the performance of individual players in a team activity," *PloS one*, vol. 5, no. 6, p. e10937, 2010.

[6] M. Hughes and I. Franks, "Analysis of passing sequences, shots and goals in soccer," *Journal of Sports Sciences*, vol. 23, no. 5, pp. 509–514, 2005.

[7] M. Dixon and M. Robinson, "A birth process model for association football matches," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 3, pp. 523–538, 1998.

[8] "Fifa/coca-cola world ranking," http://www.fifa.com/fifa-world-ranking/ranking-table/men/, accessed: 2015-01-25.

[9] F. I. de Football Association *et al.*, "Fifa world cup all-time ranking," 2014. [Online]. Available: http://www.fifa.com/worldcup/organisation/documents/index.html

[10] "World football elo ratings," http://www.eloratings.net, accessed: 2015-02-11.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." 1999.

[12] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 635–644.

[13] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *Knowledge Discovery in Databases: PKDD 2006*. Springer, 2006, pp. 259–271.

[14] A. Stanoev, D. Smilkov, and L. Kocarev, "Identifying communities by influence dynamics in social networks," *Physical Review E*, vol. 84, no. 4, p. 046102, 2011.

[15] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.(JAIR)*, vol. 22, no. 1, pp. 457–479, 2004.

[16] E. Agirre and A. Soroa, "Personalizing pagerank for word sense disambiguation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 33–41.

[17] J. P. Keener, "The perron-frobenius theorem and the ranking of football teams," *SIAM review*, vol. 35, no. 1, pp. 80–93, 1993.

[18] S. Mukherjee, "Identifying the greatest team and captain—a complex network approach to cricket matches," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 23, pp. 6066–6076, 2012.

[19] F. Radicchi, "Who is the best player ever? a complex network analysis of the history of professional tennis," *PloS one*, vol. 6, no. 2, p. e17249, 2011.

[20] "11v11 - home of football statistics and history," http://www.11v11.com, accessed: 2015-01-25.

[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[22] A. N. Langville and C. D. Meyer, "Deeper inside pagerank," *Internet Mathematics*, vol. 1, no. 3, pp. 335–380, 2004.

[23] D. E. Knuth, *The art of computer programming: sorting and searching*. Pearson Education, 1998, vol. 3.

Information retrieval over Macedonian test collection using word forms and transliteration

Ivana Bozhinova*, Stevica Bozhinoski† and Jasmina Armenska‡
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
Email:{bozhinova.ivana*, bozhinoski.stevica†}@students.finki.ukim.mk, jasmina.armenska@finki.ukim.mk‡

***Abstract*—The extensive amount of information available today increases the need for effective and efficient information retrieval (IR). With the existing IR systems users can pose questions using natural language, but they are usually presented with full documents instead of the precise answers. Question answering (QA) is a technology that aims to provide the users with an accurate answer to questions posed in a natural language, instead of presenting a list of ranked documents which are presumed to contain the answer. In this paper we present a Natural Language Question Answering System designed to work with multiple-choice questions posed in Macedonian language. As a baseline we use the well-known IR model, the Vector Space Model. We then implemented and examined various techniques in order to overcome the obstacles we faced and improve the results. One of the most sensitive problems concerning Macedonian language is the possibility that a word appears in many different forms. To group the word forms derived from the same word, we used single-link clustering, where the similarity between the words is estimated using the Dice coefficient. In order to make achievement in finding the correct answer to the questions, we have also implemented an algorithm for transliteration, so that we can cope with the numerous English words that appear in the questions. The empirical results show that the two approaches we implemented (the transliteration and the word forms) improve our systems effectiveness over the baseline we used.**

***Index Terms*—Information Retrieval, Question Answering, Transliteration, Word Forms**

I. Introduction

Search engines have become a standard for information retrieving in our daily life. They can successfully manage with the enormous document collections available today. The advances made in modern IR systems have even motivated nowadays users to pose more complex and complete questions, instead of a sequence of a few keywords. But the task for finding precise answer to natural language questions can be time-consuming and quite boring. Usually, IR systems retrieve a list of ranked documents that might have the associated information but they leave to the user to extract the required answer [9]. Today we need a way to simply retrieve facts, which on the other hand, gives us an opportunity to focus on more mind challenging and creative tasks.

Since these kinds of information needs are difficult to fulfill by the traditional IR techniques, users have started to take the advantages of QA systems. QA systems enable users to access the knowledge resources in natural way (i.e. by asking questions) and get the proper response in concise words [4]. Automatic question answering will definitely be a significant advance in the state-of-art IR technology in the future. Nevertheless, this is a quite challenging task with many issues that have to be resolved. One of the issues for existing QA systems is to understand the natural language questions correctly and deduce the precise meaning in order to extract the correct answer. The system should be adapted to language specification, because every natural language has its own characteristic morphology and syntax.

Until now, the research concerning the development of QA system capable to answer questions posed in Macedonian language is in its infancy [2]. Therefore, in this project we focus on developing a QA system capable for answering questions using a collection written in Macedonian language. We address the specifications of Macedonian grammar, especially putting emphasis on the different grammatical forms that a word might have. We use the Dice measure for finding word forms in order to locate the correct answers easily. On the other hand, since Macedonian language uses the Cyrillic Alphabet we frequently encounter words and names from different languages and hence different scripts. In order to improve our results we implemented a module for transliteration. We have tested numerous techniques to incorporate these modules into the system.

The structure of this paper is organized as follows: section II presents the most relevant related work, section III details our systems architecture and section IV deals with our experiments. The last section presents the conclusions drawn from the results and a brief overview of our future steps.

II. Related work

Most of the modern QA systems achieve satisfactory results for questions posed in English, since there are some powerful NLP tools for that language. One of the famous ones is the Princeton Universitys WordNet[1], which is a large lexical database of English, containing interlinked synsets by means of conceptual-semantic and lexical relations. It labels the semantic relations among words, so for each word we not only have the Part Of Speech (POS) tag [12], but also the word forms and synonyms. Another useful tool is the CMU Link Parser[2] that builds relations between pairs of

[1]http://wordnet.princeton.edu/
[2]http://www.link.cs.cmu.edu/link/

words. NLP tasks also include stemming to manage with the vocabulary mismatch problem. Stemmers generally make a little or no difference in accuracy. However, when applied to short documents they prove to be quite effective [7]. Although there are stemmers developed for some languages, such as English, Malay, Dutch, German and French, we lack one for the characteristics of the Macedonian language.

Cooper and Ruger developed a system for answering questions without using predetermined answer possibilities [3]. They first classify the questions in a few groups according to the type of the answer, then retrieve the passage using the Managing Gigabytes engine and add all of the hyponyms for the answer type concept using WordNet. By taking the disjunction of those hyponyms they build a regular expression and every region of text that matches it is considered as a candidate answer. Each answer is scored and ranked using various heuristics mainly based on the punctuation. The IBMs Statistical QAS [6] applies the maximum entropy classification [11] for question/answer type prediction and named entity marking. It works by first looking for the most relevant passage in an encyclopedia database. Relevance scoring is based on morph unigram and bigram features extracted using a decision tree based tokenizer, POS tagger and morphological analyzer. Then they calculate the tf-idf weights for the matched words, their word forms, synonyms and clusters.

It is quite challenging for us to develop a QA system with a satisfactory accuracy because, as we mentioned before, the Macedonian language has its own specifics and there are no open-source NLP tools. We do not have POS tagging information for a big part of the words used in our collection, nor information about synonyms, collocations and word forms. Our approach is a continuation of the work already done in the research Information Retrieval using a Macedonian Test Collection for Question Answering [1].

III. Our solution

Currently, there are several textual QA systems that include different techniques and architectures. Most of them have a number of components in common, and these are: question analysis, retrieval of relevant documents, document analysis (or passage analysis) and answer selection [8]. Creating our system, we focus on the implementation of two of those components, namely the process for locating documents that contain the correct answer to the questions, as well as choosing the right answer among given possibilities. The system is developed using Python.

A. Our collection

Our collection consists of four documents written in Macedonian language, as well as a set of multiple-choice questions [1]. The documents topics are: History of computers, Introductory concepts, Hardware and Software, and this corpus contains totally 12412 words. For each question there are four possible answers, from which only one is correct and it can be found in one of the documents.

B. Preprocessing

In the preprocessing step, we transform the documents and the questions into a format suitable for analyzing. We remove the punctuation marks and the special characters from the text and also exclude the stop words. After that we represent each document by a vector whose components are the unique terms from all of the documents in the collection. Each term is associated with a weight representing its importance in the document, as well as in the whole collection. Thus, the set of the documents in the collection can be viewed as a set of vectors in a vector space. On the other hand, every question (query) can be treated as a short document and represented by a vector in the same vector space.

C. Computing the weights

The weight of a term in a document can be determined in many ways. A common approach uses the so-called *tf-idf* method, in which the weight of the term is determined by using only two factors:

- term frequency

$$tf = 1 + log(term_count_in_the_document) \quad (1)$$

- inverse document frequency

$$idf = log(1 + \frac{1 + no_of_documents}{no_documents_containing_term}) \quad (2)$$

In other words, the *tf-idf* weighting scheme assigns a weight to each term which is highest when the term occurs many times in a small number of documents, lower when the term occurs fewer times in a document, or occurs in many documents, and lowest when the term occurs in virtually all documents [5].

D. Measuring the similarity between a document and a query

A standard way to quantify the similarity between a document d and a query q is to apply the cosine measure on their vector representations.

$$sim(d, q) = \frac{d * q}{|d| * |q|} \quad (3)$$

The numerator represents the dot product, also known as the inner product of the two vectors, while the denominator is the product of their Euclidean lengths that represent their *cosine document length normalizations* [9]. The cosine normalization is required to lessen the impact of longer documents, since they have higher term frequencies and more different terms, which increases the number of matches between a document and a query and makes these documents much more likely to be retrieved.

E. Transliterating foreign words

Some of the questions from our collection contain words (usually names) written in the original language, other than Macedonian. Since they appear in both forms throughout the documents, the questions and the answers, we needed a way to connect them so our system could treat them equally. For this reason, we have implemented an algorithm that maps each

Latin letter or a group of letters into its Cyrillic equivalent and using it we created a transliteration table containing the named entities written in a foreign language and the corresponding transliterations in Macedonian. However, this leaded to a new problem - for one foreign word given as an input, the algorithm usually outputs multiple possibilities for transliteration, from which most of the time only one is correct. In order to find the correct one, we check for a match in the words in the documents. If there is none, we then look in the words extracted from the Macedonian Wikipedia[3]. If there is a foreign named entity in the question or answer that is not in the table, it is transliterated at runtime.

F. Clustering for word forms

The Macedonian language has a specific morphology, which means that one word can appear in different word forms. Since the words can appear in various forms in the documents, the questions and the answers, we needed a way to group the different forms of the same word. Therefore, we have used the Dice measure in order to estimate how similar two terms are. We have implemented this by first dividing each word into bigrams. Then, we have calculated the number of common bigrams of the two words (i) and the number of bigrams each word contains (w_1 and w_2):

$$sim(w_1, w_2) = \frac{2 * i}{w_1 * w_2} \tag{4}$$

Afterwards, we have used hierarchical clustering in order to create the groups. More precisely, we have implemented the single-link clustering, starting with each word as a separate cluster and then combining the two clusters with the maximum Dice coefficient in each step. The clustering stops when the highest coefficient has a value smaller than 0.7. However, Macedonian language contains words whose Dice coefficient is very high, but they are carrying different meanings. Since most of the word forms of the same word start with the same bigram, we have extended the similarity measure by assigning a value of 0 if the two terms do not satisfy this condition. After we grouped the word forms, we tested various techniques on how and when to use them in the question's or answer's query in order to achieve higher accuracy.

IV. Experiment and evaluation

The process of choosing the correct answer consists of two subtasks that can be implemented and improved in various ways. The system has to find the right document that contains the relevant information (IR), and then looks for the correct answer in it (QA). We first executed the tests separately, so that the results of the QA part could be independent from the accuracy of the IR part. Then we performed tests on the whole system.

[3]www.mk.wikipedia.org

A. Information retrieval

1) Baseline: The first step in finding the correct answer of a question is to retrieve the document which contains the relevant information. In our solution we choose the document which has the highest cosine similarity with the query (words from the question) and we use this method as a baseline. We tested the performance using the baseline and then measured the improvements gained using transliteration, different approaches for adding word forms and a combination of both. The overall IR results are presented in Table I.

TABLE I
IR accuracy

Baseline	Clustering	Transliteration	Both
83.9744%	92.3077%	83.9744%	92.3077%

2) Word forms: We add a word form to the query only for the words that are missing in the document under consideration. We used two approaches: the first one is to add the most similar word form (using the Dice measure) to the missing word from our dictionary. The other solution is to add the most similar word form which is appearing in every document separately, that would result in a different query for each document. Both methods were very beneficial, since the accuracy increased from 83.9744% to 87.1795% and 91.0256% accordingly.

However, we noticed that for some longer questions (which result in longer queries) our system started to choose wrong answers, even if it was guessing them correctly before. This motivated us to perform more testing with restrictions on the length of the query. So, we decided to add word forms only if the query consists of less then four words. For both of the aforementioned approaches the accuracy was significantly increased (to 92.3077% and 90.3846%, respectively).

3) Transliteration: In this step we only use the words from the questions which contain very few foreign named entities, so the transliteration did not impact the accuracy at all.

B. Question answering if the right document is known

Once we know the document that contains the relevant information, we then determine the right answer. In this phase documents paragraphs are used in order to achieve better results. The unit paragraph is identified by the way MS Word defines paragraphs namely, each section that ends with pressed Enter (new line) is treated as a retrieval paragraph. We examined two different strategies listed below, combined with two different weighting schemes for the query terms (tf-idf from the collection and tf in the query). All the results from these measurements are depicted in Figure 1.

1) Comparing each paragraph with each answer: The first strategy we tried was to calculate the cosine similarity between each paragraph in the right document and the four possible queries composed of one of the answers, plus the words of the question. The combination that would have the maximum similarity is supposed to be consisted of the right answer and

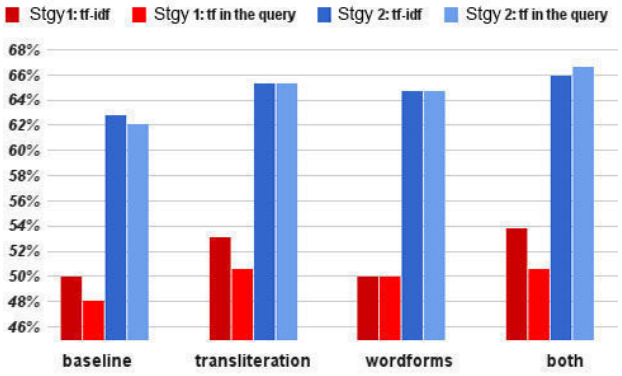


Fig. 1. Accuracy of the QA Module

the paragraph that contains it. This approach resulted with an accuracy of 50%. The usage of the word forms did not result with an improvement, while the transliteration slightly increased the accuracy to 53.8461%. We also tried different term weighting for the queries: each word weights as the number of times it appears in the query. But this approach gave us even worse results, as shown in Figure 1.

2) Finding the relevant paragraph and comparing it with each answer: This approach first finds the most relevant paragraph in the right document, using the same baseline technique explained in IV-A1 enhanced with transliterations and a word form for each word in the query that does not appear in the relevant document. Then, it chooses as a correct answer the one that has the highest score with this paragraph. The results from this method are significantly better, with an accuracy of 62.82% when using only the baseline. This is the highest baseline accuracy we got. Applying the word forms and transliteration techniques increased this value to 64.74% and 65.38%, respectively. Furthermore, we tested the other weighting metric for the query words: each word weighs as the number of times it appears in the query. Using this alternative metric, the system performed slightly worse for the baseline (62.18%), made no change compared to the tf-idf metric when using word forms and transliteration, but gave us the highest QA results when using both of them (66.67%).

C. Question Answering with Document Retrieval

We combined the IR and QA modules explained foregoing with the techniques that gave the highest accuracy. The overall results are shown in Table II. They prove that the two components we implemented improve the systems effectiveness.

TABLE II
TOTAL ACCURACY

Baseline	Clustering	Transliteration	Both
60.8974%	62.1795%	64.1026%	65.3846%

V. CONCLUSION AND FUTURE WORK

The main goal of the research presented in this paper was creating a QA system capable for answering questions posed in Macedonian language. We used the vector space model as a baseline in order to conceive the benefits from implementation of the transliteration component, as well as the process of expanding the search queries in various ways. We got the best results for document retrieval when we expanded only the queries that contain less then four words with the most similar word forms from the dictionary. When looking for the answer, the transliteration component resulted with a substantial improvement. Therefore, we can conclude that the two approaches we implemented (the transliteration and the word forms) are key factors in determining the correct answer to a particular question.

Our future steps are intended towards developing more language-specific techniques that would improve the accuracy of our system. One possible approach is to define new metrics that would take into consideration the distance between the query terms in the text. Through the usage of statistical methods, such as the t-test or the chi-square test for collocation discovery and synonym extraction [10], we could draw the correlated words and word sequences that would be an enrichment of our vocabulary. We also plan to expand the Macedonian test collection with additional documents and questions.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to professor Katerina Zdravkova for her enthusiastic guidance and suggestions throughout this research.

REFERENCES

[1] Jasmina Armenska, Aleksandar Tomovski, Katerina Zdravkova, and Jovan Pehcevski. Information retrieval using a macedonian test collection for question answering. In *ICT Innovations 2010*, pages 205–214. Springer, 2011.

[2] Jasmina Armenska and Katerina Zdravkova. Comparison of information retrieval models for question answering. In *Proceedings of the Fifth Balkan Conference in Informatics*, pages 162–167. ACM, 2012.

[3] Richard J Cooper and Stefan M Rüger. A simple question answering system. In *TREC*, 2000.

[4] Sanjay K Dwivedi and Vaishali Singh. Research and reviews in question answering system. *Procedia Technology*, 10:417–424, 2013.

[5] Amarnath Gupta and Ramesh Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, 1997.

[6] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J Mammone. Ibm's statistical question answering system. In *TREC*, 2000.

[7] Leah S Larkey, Lisa Ballesteros, and Margaret E Connell. Improving stemming for arabic information retrieval: light stemming and co-occurrence analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–282. ACM, 2002.

[8] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Mining the web to validate answers to natural language questions. In *Proc. of the 3rd Intl Conf. on Data Mining*, 2002.

[9] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[10] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[11] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, volume 17, pages 591–598, 2000.

[12] Fei Song and W Bruce Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM, 1999.

Isolating Steps in Periodic Movement

Viktor Janevski, Aleksandar Kuzmanoski, Dimitar Spasovski, Marija Mihova, Mile Jovanov

Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering
Skopje, R. Macedonia

***Abstract*—The main focus of this paper is developing an algorithm for isolating individual footsteps made during the abovementioned type of movement, based on the data gathered using a Microsoft Kinect sensor. A footstep is an act of movement characterized by the lifting of one's foot and bringing it down elsewhere or by moving each foot in succession, as done in walking or running. Such algorithm will be useful for modeling different types of periodic human movements and their recognition. Therefore, sequences of steps following a repetitive pattern will be of special interest in this paper.**

Keywords—Kinect sensor; footstep recognition; feet movement analysis

I. INTRODUCTION

Periodic movement footsteps are usually done with a certain pace, which means the speed of performing the single steps is nearly constant. Another trait is the amplitude of the steps, i.e. the length of individual treads also remains approximately equal. The goal is to develop means of isolating every single step during such a movement.

There are many ways to track the human motion. In [1] a Microsoft Kinect sensor [2] was used for the purpose of continuous, in-home gait measurement. In [3] two cameras, a static and a moving one, were used to gather data for an image processing algorithm in order to detect periodic motion. In [4], a use of robust camera estimation and a localization method were proposed in order to track the human movement during some sport activities.

All of the abovementioned papers are focused on analyzing the movement process for a specific purpose. Our objective is to develop a more generic technique that will divide the movement to its smallest constituents, i.e. steps and analyze each one separately.

In other words, our technique offers a very detailed analysis of the human feet movement. This is the reason why this algorithm can be included in a variety of movement and gesture recognition problems that include the legs, such as dancing choreographies recognition. Most of the Macedonian folk dances have a strict sequence of moves, done mostly by using the legs. Analyzing each move separately and then analyzing the whole sequence of moves can help us determine whether that particular data represent some dance or not.

Our primary motivation is to contribute to the efforts done in [5] and [6] by attempting to improve the process of traditional dance recognition. Knowing the number of steps made by the dancer during one period of the dance is useful in order to clarify potential uncertainties in the process of classification.

The algorithm can also find its application in the field of medicine, in particular physiotherapy and orthopedics. It can detect people with certain musculoskeletal issues and disorders by analyzing the steps and speed. It is also possible to apply the algorithm in any area which requires an estimation of the quality of human motion or an optimization of movements. Furthermore, it can be used in sports and fitness to approximate the distance travelled and the number of calories burned during walking or running inside a closed space, as on a treadmill, but certainly using a more affordable device which requires less space. In the end, an entertaining or educative application of the proposed algorithm would be navigating a virtual environment.

The following subsection explains the technology being used for the process of data gathering and experimenting. After that are given some definitions and explanations of the terms needed for better and unambiguous understanding of the algorithm which is explained in great detail in the subsequent section of the paper. In the end we present the results of the test phase the algorithm underwent.

A. Kinect sensor and Kinect SDK

The data needed for this paper was collected with a Microsoft Kinect [2]. It is a motion sensing and speech recognition input device developed primarily for the owners of Xbox video game consoles to be able to interact with them using body gestures and spoken commands. With the release of the software development kit for Windows 7, the Kinect sensor obtained a new dimension, which permitted numerous new applications in different study areas, similar to the one being analyzed in the following pages. Among the several benefits of using the SDK that are crucial for the positive results of this research is the access to the data streams from the depth sensor and the color camera sensor. In addition, the skeletal tracking capability provides detailed data in the form of Cartesian coordinates for twenty joints of the detected human's body including both ankles, thus greatly simplifying the examination of the positions of the feet during the periodic movement.

II. PRELIMINARIES

The following parts of this paper establish a common basis for further introduction of the algorithm and techniques used for the isolation of footsteps. First, it was necessary to decide what should be considered a step and what should be rejected.

After a wide-ranging discussion, we agreed to define a step as given in the subsection that follows.

A. *Definition of a step*

Every movement involving the raising of one's foot, traveling some distance from the starting point with respect to the floor, lowering it until it stands firmly on the ground and then moving the other foot towards the first in order to join them is considered to be a step. Furthermore, every action of bringing one foot in front of the other, again placing it on the ground, is also considered a step.

B. *Attributes for modeling steps*

It is easily concluded that the choice of a period that represents a single step is done entirely based on the movement done with the legs. The position of the feet and its change provide sufficient information when and whether a step has been made, which leads to the fact that the algorithm for footstep isolation should analyze the position of both feet and ankles, particularly the position of one relative to the other. Therefore, the two important factors that need to be considered in order to conclude whether a step has occurred are:

- The period between the raising and the lowering of one foot, which indicates a potential step
- The distance traversed with the elevated foot during the aforementioned period, which is a crucial element in the process of decision-making

In order to analyze the former it is necessary to use the Vertical Feet Distance (FVD) and for the latter we introduce the Horizontal Feet Distance (HFD) attribute.

The attributes introduced in [5] do not offer the necessary thoroughness required to obtain detailed information about the distance of the feet in distinct moments of time, since we are interested in detecting the changes in any direction, not only along the axes. Since the implementation of the footstep isolation algorithm requires the computation of the horizontal distance between the feet, in this part we introduce new attributes that will be more useful in such situation: Feet Horizontal Forward Distance (FHFD) and the Feet Horizontal Side Distance (FHSD).

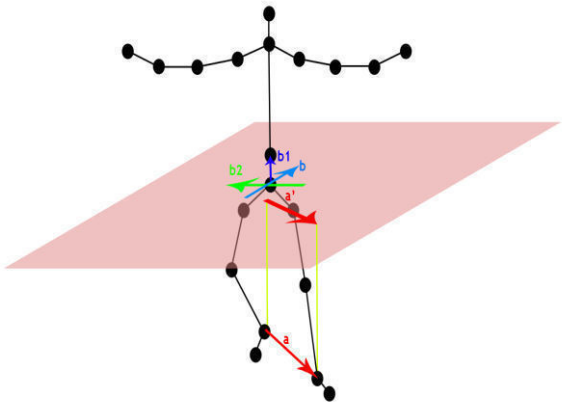


Fig. 1. The vector representing the HFD attribute

1) *Horizontal Feet Distance (HFD) attribute*

The Horizontal Feet Distance (HFD) is the module of the vector $\vec{a'}$, which is a normal projection of the vector $\vec{a}$, formed by the skeleton ankles, on the plane defined by the two vectors $\vec{b}$ and $\vec{b_2}$, where $\vec{b}$ is the vector product of the vectors $\vec{b_1}$, whose endpoints are HIP_CENTER and SPINE, and $\vec{b_2}$, whose endpoints are the points HIP_LEFT and HIP_RIGHT of the Kinect skeleton.

2) *Feet Vertical Distance (FVD) attribute*

The reason for revisiting the definition of the Feet Vertical Distance described in [5] is to mention that its values are taken by absolute value, since we are not interested in which foot was raised, but rather whether such action has occurred. Therefore, the meaning of the attribute is somewhat altered.

C. *Definitions of used terms*

1) *Extremes*

By definition, the value $x(t^*)$ at a given time period t^* is said to be a local maximum if and only if there exists $\varepsilon > 0$ such that $x(t^*) \geq x(t^* + \varepsilon)$ and $x(t^*) \geq x(t^* - \varepsilon)$. Similarly, the value $x(t^*)$ at a given time period t^* is said to be a local minimum if and only if there exists $\varepsilon > 0$ such that $x(t^*) \leq x(t^* + \varepsilon)$ and $x(t^*) \leq x(t^* - \varepsilon)$. Stated differently, $x(t^*)$ is a point where the monotonicity of the function changes. The specific criteria under which a local extreme is considered to be global differ for each attribute and are explained into more detail in the respective subsections.

2) *Tolerance rate*

In order to successfully isolate the extreme points, we introduce an additional parameter $tRate$ - maximal tolerance rate, which indicates the minimal number of successive values that violate the current monotonicity of the function. The value of this parameter is a small positive integer and it depends completely on the attribute.

A point is considered a local extreme if there are exactly $tRate$ successive values which violate the current monotonicity compared to the point being examined. By creating a list in the preceding manner, each pair of neighboring local extremes will be of different type (a minimum will always be followed by a maximum and vice versa).

The starting monotonicity is decided by democratic majority.

III. THE ALGORITHM FOR EXTRACTING INDIVIDUAL STEPS

Depending on the complexity of the periodic movements, it is possible that one period consists of more than one single step and it may be in our interest to count them. We find several applications where the number of steps in a period is significant or it is necessary to separate one step from the others. In addition, each such step might possess different characteristics that can be analyzed in order to obtain more information about the problem being solved, such as the distance traversed by each foot, the length or the duration of the step, etc. Furthermore, it can be used to examine more thoroughly whether the sequence of steps is similar and to what

extent, thus providing an additional tool to either confirm that the period was successfully isolated or to discard it.

The problem of extracting the individual steps constituting the periodic movement is reduced to the problem of choosing the appropriate subsequence of data from the entire sequence obtained during the process of data collection, for which it would be possible to claim it represents a single footstep. The process of isolating such subsequences should be done cautiously, since the beginning and the end of the corresponding step can usually be ambiguous when regarded as numeric values. Each chosen subsequence should represent nothing more and nothing less than a single footstep, without taking data from the preceding or succeeding movement units.

A major issue which must be addressed when using the Microsoft Kinect is that the sensor is completely unaware of the floor. Unlike sensing floors, the Kinect does not provide information whether both feet are on the ground, but only that they are on the same height, which does not necessarily imply the same thing. The raising and the lowering of each foot is examined relatively to the other, meaning that if both ankles are in the air during a jump, the vertical distance between them will be equal to zero.

Observing the human walking, it can be perceived that every step requires lifting one of the feet while the other remains on the ground, which causes the value of the FVD attribute to increase. When the foot is being lowered in order to make the tread, the FVD decreases and tends to zero. Significant information about the time periods, when the foot is being raised or lowered, can be obtained by studying the values of this attribute and their tendency.

It is preferable to inspect whether the values of the FVD belong to some closed interval containing zero in order to determine whether both feet are placed firmly on the ground, rather than simply comparing them to zero, since the former contributes to the elimination of noise and sensor inaccuracies during data collection. If the attribute values are between zero and the statistically determined upper-bound of the interval, it will be concluded that both feet are on the ground, otherwise one of them is raised. In addition, the analysis of local extremes would enable us to find the exact moments in time when a change has occurred. Differently stated, local extremes help in the process of determining the beginning and end of steps.

After the isolation of data for a potential footstep, one needs to use a different attribute which will show whether some change in position has occurred in between. The existence of motion shall be determined by the use of the HFD attribute. Using the HFD attribute we are able to check if the distance between the ankles has changed between the raising and the lowering of a foot, which represents an actual tread. More precisely, between the beginning and the end of the footstep, a global maximum in the HFD function must exist. It should be noted that the global extremes are a subset of the previously computed local extremes which satisfy a certain condition.

A. *Finding local extremes*

The process of obtaining the local extremes is done by observing the monotonicity of the HFD (Horizontal Feet Distance) and the FVD (Vertical Feet Distance) attribute values. The chosen values of $tRate$ are three and two, for the HFD and the FVD attribute, respectively.

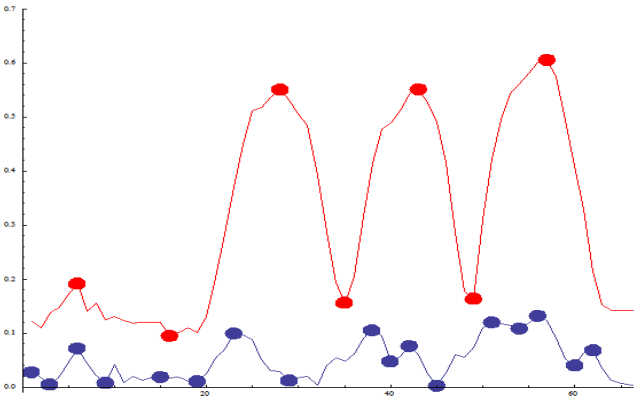


Fig. 2. Local extremes found from the collected data, where the red curve represents the HFD and the blue curve represents the FVD values

B. *Determining the potential endpoints of a footstep*

In the process of collecting data it was noticed that the values of the FVD in different moments t and $t + k$, where $k > 0$, were distinct, although it was expected for them to be equal since the feet had the same respective positions as in time t. These alterations are caused by noise and sensor inaccuracies, and had to be annulled. Therefore, a bound must be set, such that every value of the FVD lower than the determined bound would be considered equal to zero, which represents the state when both feet are placed on the ground. The values found in the interval are potential points in time that mark the beginning or end of a single step. The upper-bound is statistically determined to be 0.025, based on a previously collected data sample where we searched for the maximal value representing a state of both feet down, which in theory should have been equal to zero.

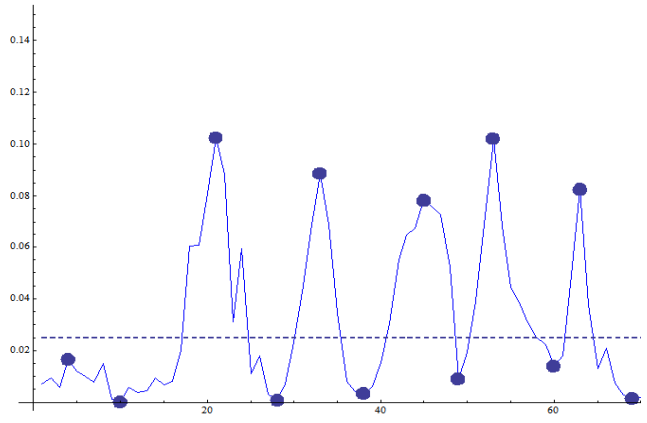


Fig. 3. Local extremes on the FVD curve and the statistically determined upper-bound y=0.025, represented by the dotted line

C. *Determining the occurrence of a single step*

It is necessary to separate the global maxima of the HFD attribute from the set of local ones. As it was mentioned in the sections above, the Horizontal Feet Distance shows whether one of the feet has been moved from its original position after raising it from the ground, potentially to make the step. The main problem with the mentioned attribute is how to determine the bound above which the extremes would be considered global, since the footsteps may vary greatly in length between different data collection processes. From the previously stated, it is more than clear that the bound chosen cannot be constant, i.e. it must depend on the obtained data. Therefore, the bound introduced is the line of regression obtained with the method of least squares, since it best fits the collected data and is not easily influenced by outliers.

IV. Experimental Results

The process of evaluation of the algorithm for isolating steps in a periodic movement included an experiment in which three people were involved. During the experiment, the people whose movement was recorded walked on a flat surface, with no obstacles on the way, facing the Kinect sensor, from a distance of 3.5 to 4 meters up to a distance of 1 to 1.5 meters from the sensor itself. The experiment attempted to test the algorithm's ability to successfully separate single steps which varied in magnitude and which were done with a different walking pace. The number of steps done during the recordings ranged between three and eight, but we also tested the extreme cases when not a single step occurs, such as situations where the person stands still or only raises his foot, which do not satisfy the previously given definition of a step.

The overall success rate of the algorithm, based on 26 data sequences, is 94.77%. The percentage was calculated as an arithmetic mean of the individual precision for each test case. The obtained number of steps from each recording was compared to an expected value, determined by hand according to the given definition of a step. The equation used for the estimation of correctness is the following:

$$successRate = \frac{\sum_{i=1}^{n} \frac{\min(expected_i, obtained_i)}{\max(expected_i, obtained_i)} \cdot 100}{n} \quad (1)$$

In (1) n is the number of data sequences collected.

In 20 of the total 26 cases, the expected and the obtained number of steps were equal, which is 76.92%. In the other 6 tests, the obtained number differed from the expected in one or two steps, which is mostly due to noise and hardware imprecision. Naturally, the steps made were not perfectly equal. Another issue was the data collection rate, since the Kinect sensor did not manage to gather enough data, as a result of the fast pace with which the person moved. Last, the great difference in the magnitude of the steps can potentially lead to problems, since the linear regression tends to best fit all the given data, thus eliminating some extremes on the curve that were valuable.

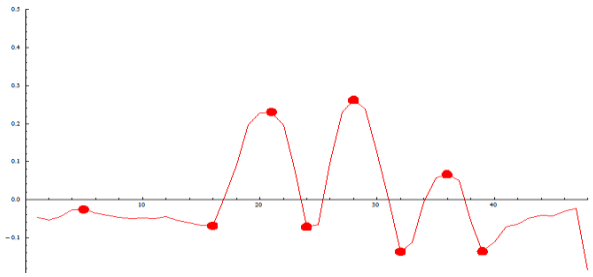


Fig. 4. Local extremes on the fitted HFD attribute

V. Conclusion

In this paper we described a new technique for analyzing the human leg movement. Using the data from the Microsoft Kinect sensor, two attributes for step modeling were developed. By inspecting the values of these two attributes, we can analyze the movement of the feet. The analysis includes the number of steps, the size and the speed of each step, their starting and ending points.

The main conclusion is that we developed a technique which is more generic than other movement detection algorithms. With minor modifications, the algorithm can help solve problems from different fields that have no correlation between each other.

Further work might include research and consultations to experts on the definition of a step in Macedonian traditional dances, as well as introducing certain modifications to the algorithm in compliance to those definitions. Moreover, we will concentrate on designing an algorithm for recognition of dances which will consider the number of steps in each dancing period, counted by using this new technique.

Acknowledgement

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje. We also express our gratitude to Microsoft Macedonia for supporting the test phase of the presented algorithm by lending us the needed hardware.

References

[1] E. E. Stone, M. Skubic, "Unobtrusive, continuous, in-home gait measurement using the Microsoft Kinect", IEEE Transactions on Biomedical Engineering, vol. 60, no. 10, pp. 2925-2932, October 2013

[2] Z. Zhang, "Microsoft Kinect sensor and its effect", IEEE MultiMedia, vol. 19, no. 2, pp. 4-10, 2012

[3] R. Cutler, L. Davis, "Real-time periodic motion detection, analysis and applications", Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 326-332, 1999

[4] C. Panagiotakis, I. Grinias, and G. Tziritas, "Automatic human motion analysis and action recognition in athletics videos", 14th European Signal Processing Conference, pp. 1-5, 2006

[5] D. Martinovikj, M. Mihova, M. Jovanov, "Analysis of the problem of Macedonian folk dance recognition", 10th Conference on Informatics and Information Technologies, pp. 165-168, 2013

[6] D. Martinovikj, M. Mihova, M. Jovanov, "Characteristic attributes for recognition of traditional Macedonian folk dances", Review of the National Center for Digitization, no. 25, pp. 81-90, 2014

Generating Boolean-unlinear quasigroups of order 8 and their cryptographic properties

Bojan Petrovski Atanas Todorovski Tome Stojkovski Marija Mihova
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, R. Macedonia
petrovski.bojan@students.finki.ukim.mk; todorovski.atanas@students.finki.ukim.mk
stojkovski.tome@students.finki.ukim.mk; marija.mihova@finki.ukim.mk

***Abstract*— based on the theoretical results for the properties of Boolean-Linear to Constant Operands Quasigroups we developed an algorithm that will generate all such Quasigroups. Furthermore, we organized them in classes of isotope Quasigroups. We further examined the cryptographic properties of Quasigroups between classes and within the same class using linear and nonlinear permutations.**

Keywords—Quasigroups; BLCO; Matrix; Cryptography;

I. INTRODUCTION

Quasigroups are algebraic structure that have many applications in mathematics and computer science, mainly used in coding and cryptography. A class of codes Random codes based on quasigroups, is proposed by Gligorovski in [4], while J. Denes and A. D. Keedwell are the main cryptologists that apply quasigroups [6, 7, 8]. Some specific Quasigroups are used for construction of block ciphers [3, 9] and hash functions [5]. Since the quality of a crypto-product depends of its resistance on different types of attacks, the differential and statistical crypto-analysis is integral part in designing such product.

The focus of our research is placed on special type of Quasigroups of order 8, we named it as BLCO Quasigroups. Since these Quasigroups have a simple Boolean representation, they are easy usable in practice. In this paper, we define BLCO quasigroups, finding that some of them can be regarded as isotopic quasigroups, which allow us to classify them into isotopic classes. One of the main goal of this paper is to analyze some of the cryptographic properties of such classes of quasigroups. More specifically, we use quasigroup string transformations for encryption and decryption based on binary quasigroups known as e-transformation and d-transformation, defined in [11]. We analyze how a change of one bit in a given input binary string affects the binary strings obtained by applying E-transformation as a multilevel encryptor, known as the avalanche effect.

II. BLCO-QUASIGROUPS

Definition. 1 Given Quasigroup $(\{0,1\}^n, *)$, its representation as vector valued Boolean polynomials is the function $f: \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^n$ defined by $f(\vec{x},\vec{y}) = \left(f_1(\vec{x},\vec{y}), \dots, f_n(\vec{x},\vec{y})\right) = \vec{x} * \vec{y}$, where $\vec{x} = (x_1, \dots, x_n)$, $\vec{y} = (y_1, \dots, y_n)$ and each $f_i(\vec{x},\vec{y})$ is a Boolean polynomial function of Boolean variables $x_1, \dots, x_n, y_1, \dots, y_n$.

Normalized quasigroup is a quasigroup for which for all $\vec{x}$, $\vec{x} * 0 = 0 * \vec{x} = \vec{x}$.

The most of the theory about BLCO Quasigroups of order 8 is given in [4]. Here we give the definition and some of the properties that

Definition. **2** The quasigroup $(\{0,1\}^n, *)$ is called a *Boolean-Linear to Constant Operands* (*BLCO-quasigroup*) if the quasigroup operation can be represented by a vector valued Boolean function $f(\vec{x},\vec{y}) = \vec{z} = (z_1, \dots, z_n)$, where for some constants $c_k, a_{ki}, b_{ki} \in \{0,1\}, k, i = \overline{1,n}$

$$z_k = c_k + \sum_{i=1}^{n} a_{ki}x_i + \sum_{i=1}^{n} b_{ki}y_i + \sum_{i,j=1}^{n} d_{kij}x_iy_j\,, \qquad (1)$$

where addition and multiplication are the binary operations OR and AND respectively.

Using the notation: $A = [a_{ki}]$, $B = [b_{ki}]$, $\vec{c} = [c_k]$, $D'_i = \left[d'^{i}_{kj}\right]$, where $d'^{i}_{kj} = d_{kij}$, and $D''_j = \left[d''^{j}_{ki}\right]$ where $d''^{j}_{ki} = d_{kij}$, $\vec{x} * \vec{y}$ can be represented in matrix form as

$$\vec{x} * \vec{y} = \vec{z} = \vec{c} + A\vec{x} + B\vec{y} + \sum_{i=1}^{n} x_i D'_i\vec{y} = \vec{c} + A\vec{x} + B\vec{y} + \sum_{j=1}^{n} y_j D''_j\vec{x}\,. \qquad (2)$$

Example 1. The matrix representation of the Quasigroup defined by

$$\vec{x} * \vec{y} = \big(x_1 + x_3 + y_2 + x_3y_3,\ x_2 + y_1 + y_3 + x_3y_3 + 1, x_3 + y_3 + (x_1 + x_2 + x_3)y_2\big)$$

is

$$A.\ \vec{x} * \vec{y} = \begin{bmatrix}1&0&1\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} + \begin{bmatrix}0&1&0\\1&0&1\\0&0&1\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix} + x_1\begin{bmatrix}0&0&0\\0&0&0\\0&1&0\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix} + x_2\begin{bmatrix}0&0&0\\0&0&0\\0&1&0\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix} + x_3\begin{bmatrix}0&0&1\\0&0&1\\0&1&0\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}$$

Normalized BLCO-Quasigroup is a BLCO-Quasigroup in which A and B are identical matrices, and $\vec{c}$ is a zero vector, i.e. these quasigroups have the form

$$\vec{x} * \vec{y} = \vec{x} + \vec{y} + \sum_{i=1}^{n} x_i D'_i \vec{y} = \vec{x} + \vec{y} + \sum_{j=1}^{n} y_j D''_j \vec{x}.$$

Example 2. One normalized BLCO-Quasigroup (NBLCO-Quasigroup) is the quasigroup defined with

$$\vec{x} * \vec{y} = \begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} + \begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix} + x_1\begin{bmatrix}0&0&0\\0&0&0\\1&0&0\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix} + x_2\begin{bmatrix}0&0&0\\0&0&0\\1&0&0\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix} + x_3\begin{bmatrix}1&1&0\\1&1&0\\0&0&0\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}.$$

The next Theorem is proven in [3]:

Theorem 1. For each BLCO-quasigroup $(\{0,1\}^n, \hat{*})$ there is NBLCO-quasigroup $(\{0,1\}^n, *)$ and nonsingular matrices A and B such that

$$\vec{x} \hat{*} \vec{y} = A\vec{x} * B\vec{y}.$$

We will call the quasigroup $(\{0,1\}^n, \hat{*})$ **linear transformation** of the quasigroup $(\{0,1\}^n, *)$.

III. Isotopism and Isomorfism of Quasigroups relative to matrix representation

Given a quasigroup of order n and a permutation π on the set $\{0, 1, 2, \ldots, n-1\}$, by applying π on the rows, on the columns or on the elements of that quasigroup we will obtain a quasigroup again. More specifically, if the operation $\vec{x} * \vec{y}$ defines a quasigroup, then $\pi(\vec{x}) * \vec{y}$, $\vec{x} * \pi(\vec{y})$ and $\pi(\vec{x} * \vec{y})$ also define quasigroups. The next Corollary is obvious:

Corollary 1. Let $(\{0,1\}^n, *)$ be NBLCO-quasigroup. Given a permutation π on the set $\{0,1\}^n$, the quasigroup $(\{0,1\}^n, \hat{*})$ defined by

$$\vec{x} \hat{*} \vec{y} = \pi^{-1}(\pi(\vec{x}) * \pi(\vec{y}))$$

is a normalized quasigroup.

Definition. **3** We will say that the quasigroups $(\{0,1\}^n, *)$ and $(\{0,1\}^n, \hat{*})$ are **isomorphic relative to matrix representation** iff there is a permutation π on the set $\{0,1\}^n$ such that

$$\vec{x} \hat{*} \vec{y} = \pi^{-1}\big(\pi(\vec{x}) * \pi(\vec{y})\big).$$

If π is a permutation defined as $\pi(\vec{x}) = A\vec{x}$, where A is a nonsingular binary matrix then we will say that the quasigroups $(\{0,1\}^n, *)$ and $(\{0,1\}^n, \hat{*})$ are **isotopic relative to matrix representation.**

Definition 4. We will say that the quasigroups $(\{0,1\}^n, *)$ and $(\{0,1\}^n, \hat{*})$ are **commutatively relative to matrix representation (CRM)** iff

$$\vec{x} \hat{*} \vec{y} = \vec{y} * \vec{x}.$$

The next Theorem is proven in [3]:

Theorem 2. Let $(\{0,1\}^n, *)$ be NBLCO-quasigroup and let the permutation π on the set $\{0,1\}^n$ is defined as $\pi(\vec{x}) = A\vec{x}$, where A is a nonsingular binary matrix, then the quasigroup $(\{0,1\}^n, \hat{*})$ defined by

$$\vec{x} \hat{*} \vec{y} = \pi^{-1}(\pi(\vec{x}) * \pi(\vec{y}))$$

is a NBLCO-quasigroup.

From the Theorem 1 we can conclude that in order to generate all BLCO quasigroups it is sufficient to generate all NLBCO quasigroups. Therefore, for quasigroups of order 8 we generated all NBLCO. In the generating algorithm, we used some properties of the matrices D'_i given in [10] and we obtained 2192 BLCO-Quasigroups of order 8. .

Eliminating all isotopic Quasigroups it is obtained that there are in total 30 different isotopic classes of BLCO-Quasigroups. Moreover, we were checking for CRM quasigroups, and we found that there are classes consisting of CRM quasigroups of some other class. In fact, there are 6 pairs of such classes, so there is in fact only 24 quasigroups that cannot be obtained one from another.

An interesting result is that there are quasigroups isomorphic on some BLCO-Quasigroup that are not BLCO-quasigroups. We are expecting that such quasigroups will have better cryptographic properties then BLCO-quasigroups. That prompted us to generate such quasigroups. For that reason, we analyzed the form of the nonlinear permutations and we found that some of them can be represented as

$$\pi(\vec{x}) = A\left(\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} + (I_p + I)\begin{bmatrix}x_1x_3\\x_1x_3\\x_1x_2\end{bmatrix}\right), \qquad (3)$$

where A is a nonsingular binary matrix, I_p is a permutation of the identity matrix and I is the identity matrix. The inverse permutation have the form

$$\pi^{-1}(\vec{x}) = \left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + (I_p{}^T + I) \begin{bmatrix} x_1x_3 \\ x_1x_3 \\ x_1x_2 \end{bmatrix} \right) (A^{-1}\vec{x}).$$

IV. Analyses of the Avalanche effect

We analyze the avalanche effect of the BLCO-quasigroups of order 8 and their isomorphic quasigroups trough a modification of the E-transformations with a given key (*a,b,c*) defined in [f]. The *e*-transformation is defined as follows:

Let $x_1x_2x_3x_4...x_{3n-1}x_{3n}$ be bit string, then

$$e*,(a,b,c)(x1x2x3x4\dots x3n-1x3n) = x1'x2'x3'x4'\dots x3n-1'x3n, \quad (4)$$

where,

$$(x1'x2'\,x3') = (a,b) * (x1x2\,x3)$$

$$(x'i, x'i+1, x'i+2) = (x'i-3, x'i-2, x'i-1) * (xi, xi+1\; xi+2), i = 4,\dots,3n-2.$$

Consecutive e-transformations based on $*$ can be applied on a given binary string, as a composition of e-transformations using the same or different leaders for each transformation. This composition of *k* mappings is said to be E-transformation. In some of the experiments each odd transformation is performed from left to right, and each even transformation is performed from right to left.

A. Experiments and results

For our experiments, we choose 14 of the 24 NBLCO classes, but in this paper we presenting the result of two of them. There are *three groups* of experiments conducted.

Group 1. 500 different 150-bit messages that were coded using five different 96-bit keys ((*a, b, c*) in formula (4)).

Group 2. We used 100 different messages that had a length of 900 bits, using the same five different 96-bit keys

Group 3. We tested 15 messages with a length of 4500 bits. All messages and keys was generated randomly.

In the experiments, *we observed two important properties*: the number of flipped bits between the original and coded messages and the Strict Avalanche effect by introducing a one-bit change in the message in 3 different places.

The results are presented threw two NBLCO-quasigroup classes, the class of the quasigroup K1 defined by

$$\vec{x} * \vec{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + x_1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + x_2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + x_3 \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix},$$

and the class of the quasigroup K2 defined by

$$\vec{x} * \vec{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + x_1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + x_2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + x_3 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

These two classes are chosen because in K1 the matrices D_1, D_2 and D_3 are linearly dependent, while in K2 D_1, D_2 and D_3 are linearly independent.

The results are presented through Table1-Table 4. Table 1 and Table 2 presents the results for K1, while Table 3 and Table 4 presents the results for K2. The rows corresponds to different keys. First column corresponds to a suitable NLCBO-quasigroup, next column correspond to one of its linear transformation, while last columns correspond to one of its isomorphic, but not isotopic, quasigroup.

In order to find if there is a difference in cryptography properties between quasigroups in different classes, first, we compare different NBLCO-quasigroups. The average flipping of bits for the classes K1 and K2 are presented in Table 1 and Table 3 respectively, while the average Avalanche effects for the classes K1 and K2 are presented in in Table 2 and Table 4 respectively. The results of the performed experiments show that on the quasigroups for which the matrix D_1, D_2 and D_3 are linearly independent perform much better on the Avalanche effect and are much more consistent in the flipping of bits.

In order to find if there is differences between isomorphic quasigroups in the same isotopic class, we take NBLCO -quasigroup. We compared some of its linear transformations with some of the linear transformations of quasigroups that are isomorphic, but not isotopic, with it.

The results show that only a linear transformation quasigroup of a NBLCO -quasigroup that performed poorly in the first test its cryptographic properties, significantly improved the cryptographic proprieties, especially the Avalanche effect, Table 1 and Table 2. The NBLCO -quasigroups that had good cryptographic properties in the first test do not change significantly, Table 3 and Table 4.

The isomorphic, but not isotopic, quasigroups of a NBLCO quasigroups have much better cryptographic properties. From the other hand, for NLCBO quasigroup for which the matrices D_1, D_2 and D_3 are linearly independent, we were not able to find significant difference between BLCO quasigroups and the quasigroups that are not isotopic with appropriate NBLCO quasigroup. But, this is expected since such NLCBO quasigroups have good properties their own.

Key	NLCBO	linear transf.	isomorphic
1	0.500787	0.500507	0.501653
2	0.500613	0.499987	0.499613
3	0.499653	0.501520	0.501413
4	0.501053	0.499933	0.500720
5	0.500347	0.501200	0.501920

Figure 1. Representation of average flipping of bits for the Quasigroup K1

Key	NLCBO	linear transf.	isomorphic
1	0.224444	0.499796	0.498818
2	0.224444	0.498631	0.500533
3	0.224444	0.500587	0.501276
4	0.224444	0.500604	0.500813
5	0.224444	0.502453	0.500049

Figure 2. Representation of average Avalanche for the Quasigroup K1

Key	NLCBO	linear transf.	isomorphic
1	0.499853	0.498880	0.499187
2	0.499653	0.502640	0.506067
3	0.500680	0.501120	0.499147
4	0.499493	0.501333	0.499973
5	0.500893	0.501480	0.502973

Figure 3. Representation of average flipping of bits for the Quasigroup K2

Key	NLCBO	linear transf.	isomorphic
1	0.499249	0.500293	0.500022
2	0.499200	0.499916	0.498724
3	0.498964	0.501822	0.498564
4	0.499573	0.501378	0.501867
5	0.499142	0.500316	0.500227

Figure 4. Representation of average Avalanche for the Quasigroup K2

V. Conclusion

Using one represent of the 24 classes of BLCO quasigroups that we obtained with the classification, any BLCO quasigroup of order 8 can be easily generated as isotopic quasigroup or its linear transformation. As shown in our analysis of the cryptographic properties with the use quasigroups obtained from a given NBLCO with different transformation, all BLCO quasigroups produce results that satisfy the criteria of the avalanche tests.

As a future work we plan is to continue with analyzing other cryptographic properties of NBLCO quasigroups and their isomorphic.

Acknowledgment

. This paper is partially supported by Faculty of Computer Science and Engineering at the University St. Cyril and Methodius in Skopje, Macedonia

References

[1] D. Gligoroski and V. Dimitrova and S. Markovski, “Classification of Quasigroups as Boolean Functions, their Algebraic Complexity and Application of Gröbner Bases in Solving Systems of Quasigroup Equations”, Invited short-note for RISC Book Series, Springer, "Groebner, Coding, and Cryptography", Ed. M. Sala, 2007

[2] D. Gligoroski, V. Dimitrova, S. Markovski, Quasigroups as Boolean Functions, Their Equation Systems and Grobner Bases. In: Sala, M.,Mora, T., Perret, L., Sakata, S., Traverso, C. (eds.), Gr¨ obner Bases, Coding, and Cryptography, pp. 415-420. Springer, Heidelberg, 2009.

[3] D. Gligoroski, S. Markovski, and S. J. Knapskog. A public key block cipher based on multivariate quadratic quasigroups, 2008.

[4] D. Gligoroski, S. Markovski, and L. Kocarev. Erro-Correcting Codes Based on Quasigroups, 2007

[5] D. Gligoroski, R.S. Ødegå rd, M. Mihova, S.J. Knapskog, A. Drapal, V. Klima, J. Amundse, M. El-Hadedy, Cryptographic hash function Edon-R, Proceedings on the 1st International Workshop on Security and Communication Networks (IWSCN), 2009, IEEE, (2011), 1-9.

[6] J. Denes and A. D. Keedwell. Latin Squares and their Applications. Academiai Kiado, Budapest, 1974.

[7] J. Denes. Latin squares and non-binary encoding. Proc. conf. information theory, CNRS, pages 215-221, Paris, 1979.

[8] J. Denes. On latin squares and a digital encrypting communication system. P.U.M.A., Pure Math. Appl., 11(4):559-563, 2000.

[9] M. Hassinen and S. Markovski. Differential cryptanalysis of the quasigroup cipher. Definition of the encryption method. In Differential cryptanalysis, Petrozavodsk, June 2004.

[10] M. Mihova, B. Petrovski, A. Todorovski, T. Stojkovski, Classes of Isotopic Normalized BLTM quasigroups of order 8, LOOPS 2015

[11] S. Markovski, Quasigroup String Processing and Applications in Cryptography, In: 1st International Conference of Mathematics and Informatics for Industry, pp. 278-289, Thessaloniki, Greece, 2003.

Graphtec – the interactive research tool for running experiments on graphs

Petar Grchevski, Biljana Zaeva, Elena Golubovska
Faculty of Computer Science and Engineering
University St. Cyril and Methodius
Skopje, Macedonia
petar.numb@gmail.com, biljana.zaeva@yahoo.com, elena.golubovska@gmail.com

Abstract

In this paper we present an interactive tool which provides a visual representation of a directed or undirected graph, designed in C# in the Visual Studio environment. The tool provides the opportunity for drawing graphs, running graph algorithms as well as saving the new state of the graph with the selected preferences. The tool is easily extendable, which means that users would be able to add and test their custom algorithms. As a demonstration we have implemented the popular Ford Fulkerson algorithm for determining the maximum flow between two nodes for a given network.

Keywords— Graphs, Algorithms, Research, Experimetns, Ford-Fulkerson, Flow

I. INTRODUCTION

Graphs are of extreme importance to computer science and the I.T. industry. They have a wide range of uses in computer networks, protein interaction research, social networks such as Facebook, artificial intelligence, intelligent systems, machine learning etc. Hence, most people that are working within the field of computer science have at some point in their career encountered graphs.

Even though graphs are a widely covered topic by scientists and programmers alike, their use is not simple. While doing research and development, graphs usually have graphical representation. However, even though experiments are designed this way, when it comes to running them on a computer, most people use standard input and output to perform them, hindering their results, due to the inability to represent them graphically. The ability to accompany their algorithms with instant graphically representation of the completed task, will increase their productivity. In addition, teachers will have an easier job explaining how graphs work to students.

Because of this, we decided to develop an interactive research tool which would allow users to easily perform experiments and receive a real time graphical representation of the results. We decided to make the tool open for modification so that users can easily add their own algorithms in addition to the previously implemented algorithms. The tool has been specifically designed for the implementation of algorithms concerning reliability of flow networks. However, its design allows for the implementation of any algorithm which works with graphs.

Graphtec can be used for many different purposes, some of which are:

1) applying graphs (and related algorithms) in practice - economics, social sciences, urban operations and others
2) teaching graph theory - especially useful for teachers
3) studying, analyzing and solving problems on graphs - especially useful for teachers, students and even researchers
4) constructing and designing graphs for other purposes

We would also like to note other similar applications exist. In section II of this paper we have covered the most popular ones, what they offer, what they lack and why our tool, Graphtec, is superior. The covered applications are Graphwiz[5], Graph Magics[6] and Graph Creator[7].

II. ON GRAPHS AND SIMILAR TOOLS

In this section we provide a small intro in graphs, namely their brief history and more importantly their representation. Furthermore we provide several examples of other graph applications similar to Graphtec, as well as a comparison of features and usage.

A. What does a graph represent?

A graph is an ordered pair $G = (V, E)$ comprised of a set V of vertices or nodes together with a set E of edges or links. Vertices are also called *nodes* or *points*, and edges are also called *arcs* or *lines*. Each edge has either one or two vertices (v, w) associated with it (where *v* is called the *source* of *e* and *w* is called the *target* of *e)* called its endpoints. Furthermore a graph with empty node list is called *empty. [1]*

B. Brief history

The study of graphs is known as graph theory, and was first systematically investigated by D. Konig in the 1930s. Unfortunately, as Gardner notes, "The confusion of this term [i.e., the term "graph" to describe a network of vertices and edges] with the 'graphs' of analytic geometry [i.e., plots of functions] is regrettable, but the term has stuck." [2]

C. Representation of graphs

In computer science, different data structures are used for the representation of graphs.

1) An adjacency matrix A – represents a two-dimensional n x n matrix (where n is the number of vertices) in which the rows represent the source vertices and colums represent destination vertices. Its (i,j)th entry is equal to one if v_i and v_j are adjacent and zero when they are not adjacent. In other words A=[a_{ij}] where :

$$a_{ij} = \begin{cases} 1 \text{ if } \{vi,vj\} \text{ is an edge in the graph} \\ 0 \text{ otherwise} \end{cases}$$

In addition, the adjacency matrix of an undirected graph is symmetric because $a_{ij} = a_{ji}$, for all i, j. Also if the graph does not contain any loops then each diagonal entry is zero $a_{ii}=0$ for all i. Adjacency matrices can also be used to represent graphs with loops and multi-edges. Then the (i, j)th entry equals the number of edges connecting the pair of vertices. Moreover the adjacency matrix for a directed graph need not be symmetric.

2) An adjacency list – a list of edges associated with every vertice V. In a directed graph two lists are requred:

adj_edges(v) = {e ϵ E | v = source(e)}, i.e., the list of edges starting in v, and

in_edges(v) = {e ϵ E | v = target(e)}, i.e., the list of edges ending in v (where e in an edge from the set of all egdes E). The list adj edges(v) is called the adjacency list of node v and the edges in adj edges(v) are called the edges adjacent to node v. For directed graph we often use out edges(v) as a synonym for adj edges(v). However, in an undirected graph only one adjacency list is requred for every vertice V.

3) An incidence matrix - is a two dimensional boolean matrix in which the rows represent the vertices and the rows represent the edges. The entries indicate whether the vertex at a row is incident to the edge at a column. *[4]*

D. Differences and Graph applications

There are numerous graphing applications made by students and professional companies but they differ significantly from the Graphtec desktop application we created. Namely, we provide several examples in order to reveal those differences:

1) Graphwiz [5]

Graphviz is graph visualization software. As well as Graphtec it is open source. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains. Unlike Graphtec, the Graphviz layout programs take descriptions of graphs in a simple text language, and make diagrams in useful formats, such as images and SVG for web pages; PDF or Postscript for inclusion in other documents; or display in an interactive graph browser, whereas in the Graphtec users can create the preferred graph interactively with simple instructions. Graphviz has many useful features for concrete diagrams, such as options for colors, fonts, tabular node layouts, line styles, hyperlinks, and custom shapes. In contrast to the Graphtec, it lacks the use of graph algorithms and serves only to provide visual representation.

2) Graph Magics [6]

The ease with which graphs are constructed is very similar in Graphtec: edges and vertices can be added with just a simple click ; movement and resizing of vertices is simply done by dragging and dropping ; one can edit edges weights and capacities by typing them directly into edges representation, or by a using a special table of values. It also offers Execution and step by step visualization of 17 different algorithms on graphs (finding the shortest path, minimal spanning tree, maximum flow, minimum cut, Chinese postman problem, the largest set of independent vertices, maximum clique, eulerian and Hamiltonian circuits and others) are offered. However, unlike the Graphtec, this sophisticated program is not open source and it does not offer the opportunity for students to integrate other useful and unique algorithms themselves. In addition the Graphtec is free of charge and does not require any payment, whereas in order to use the Graph Magics users need to pay some particular amount.

3) Graph Creator [7]

The Graph Creator is an online tool for creating graphs and the use of Graph Explorer is to investigate your graph and the problem it represents. Its main purpose is to investigate ideas such as planar graphs, complete graphs, minimum-cost spanning trees, and Euler and Hamiltonian paths. In contrast, the Graphtec implements more sophisticated algorithms such as Ford-Fulkerson and is available for desktop users without the need of an internet connection. Moreover the Graph Creator does not offer an option to save the created graph nor to reuse it. One of the main features of the Graphtec is indeed the possibility to reuse already saved graphs and interactively modify them and run different algorithms multiple times.

III. FUNCTIONALITY AND USAGE

This section describes the tool we developed in Visual Studio for drawing graphs and running algorithms on them. Although there are data structures in Visual Studio that can be easily used for representing graphs, none of them were fully satisfactory for the options and the possibilities we wanted our application to offer, structurally and graphically.

Our application consists of multiple windows forms along with classes that represent the nodes and the edges of the

graph as well as classes for all the algorithms that can be run on the graph.

Below, we present you with thorough description of each of the windows forms and its functionalities.

A. First window – Graphtec

The first window to open when the application starts running is the window in which a user can create a graph, by adding nodes and edges. It has two modes - Node Mode and Edge Mode, both resulting in different behavior of the same action such as double clicking on the blank surface or right clicking on a certain edge.

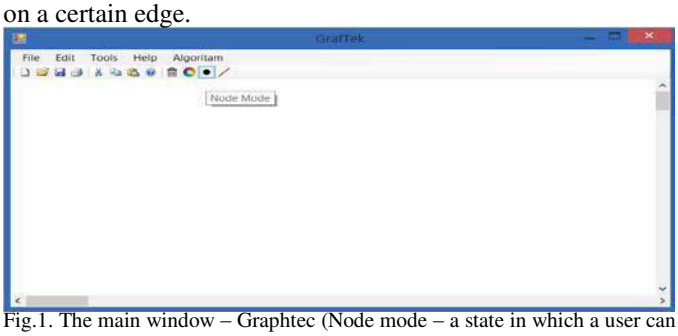


Fig.1. The main window – Graphtec (Node mode – a state in which a user can add nodes)

In the picture above, Figure 1, it is clear that the user has chosen the Node Mode, which is predefined, and it enables him to add a node by a simple double click anywhere on the white surface. When in this mode, the user can select one or multiple nodes just by a single click on them, which he can delete or move by dragging them over.

Apart from the standard tools such as save, open and other similar commands, this window offers tools that allow the user to manipulate with the graphical appearance of the nodes and edges. As shown in Figure 2, a user can pick a color beforehand. After changing the color each node that will be added will have the chosen color.

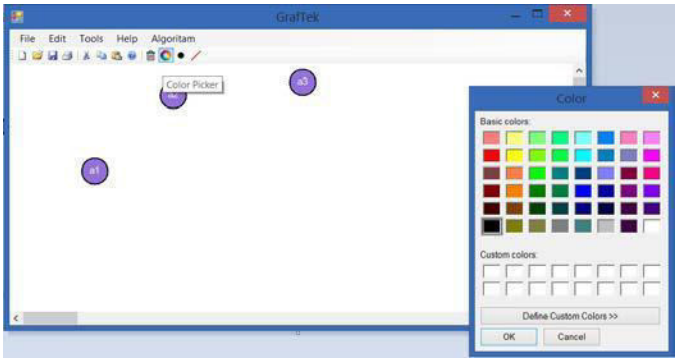


Fig.2. Picking different color for the nodes that will be added next

Switching between modes is enabled by clicking on the appropriate buttons, which are very intuitive and easy to recognize. By clicking the Edge Mode button and selecting two nodes a new form will open asking the user for the weight of the edge he wants to add between the selected nodes respectively. If the user does not state otherwise the predefined weight of the edge is zero.

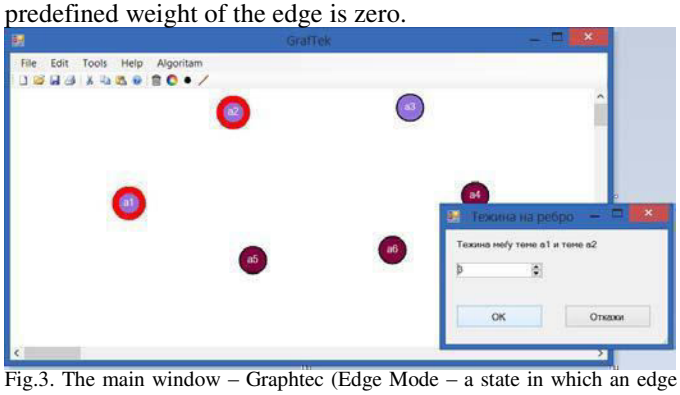


Fig.3. The main window – Graphtec (Edge Mode – a state in which an edge can be added) and the form for assigning value to the weight

By selecting a certain edge and right clicking on it, another form will open asking the user if he wants to change weight of the currently selected edge. By repeating the procedure the user can easily create a graph like in the picture below, Figure 4.

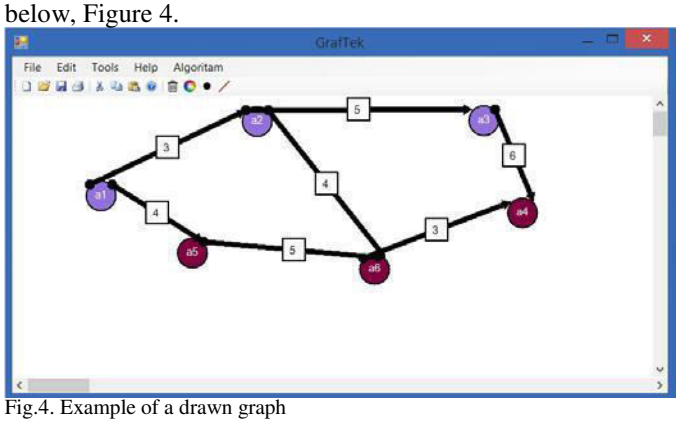


Fig.4. Example of a drawn graph

The Algoritam button leads the user to a new form where he can choose an algorithm to run on the graph he had drawn as well as all the options he needs to analyze.

B. The Algorithm Window - Options

The Options window form is simple and visually clear form. Its usage is very intuitive. It contains a list of all the algorithms that can be run on the graph, as well as a list of all the options the user wishes to see once the algorithm has been run. By choosing a certain algorithm, the user can easily see and choose from the list with options, the ones that are adequate for the chosen algorithm.

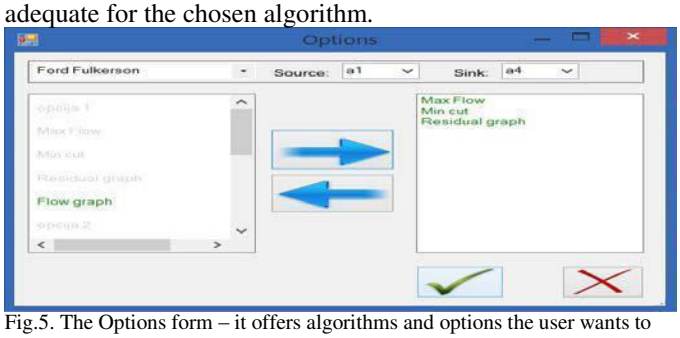


Fig.5. The Options form – it offers algorithms and options the user wants to analyze

After choosing all the options a user wants, he can proceed by clicking the button with a checkmark, which will lead him to the MDIContainer – a form which contains a number of window forms depending on the number of options the user had requested to see.

C. The third window – MDIContainer

This window is based on MDI(Multiple Document Interface) which is an interface that allows multiple windows to reside under a single parent window. It is very practical in scenarios like our application because it can provide the user with clear view of all the windows at the same time, which can be of great importance when it comes to comparing, experimenting and analyzing.

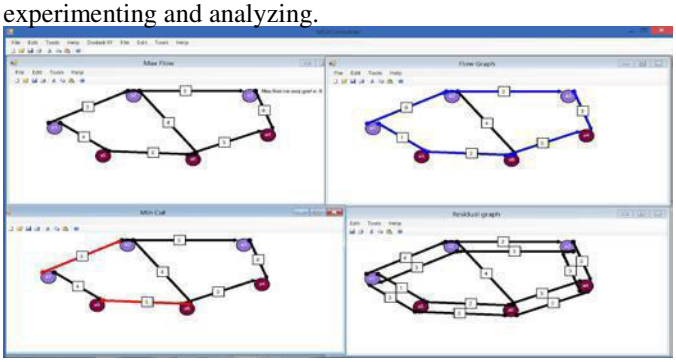


Fig.6. The MDIContainer form – contains windows representing all of the options the user had requested to see. The results shown are from running the already implemented Ford-Fulkerson Max-Flow algorithm.

In order to be able to use the results later, the user can save them by simply using the predefined tool save and open. In Figure 6 it is shown how it can be done.

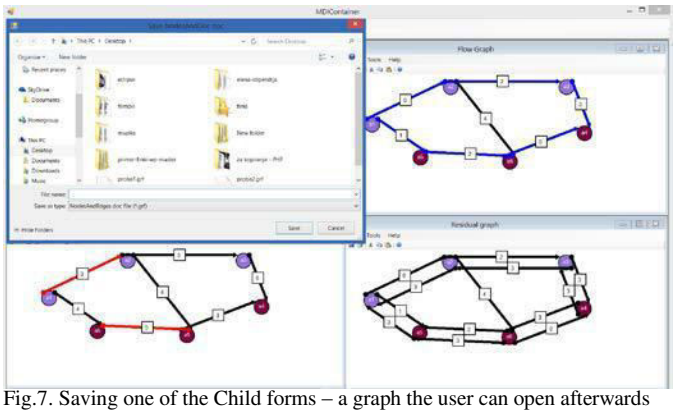


Fig.7. Saving one of the Child forms – a graph the user can open afterwards

D. The representation of the graph

The representation of the graph is done with the native C# structure – Dictionary, which is equivalent to a HashMap in other programming languages.

We will also add a parallel implementation through an adjacency matrix in order to accommodate for algorithms that rely on it for optimal performance.

IV. CONCLUSION AND FUTURE WORK

This paper doesn't cover every feature of the tool. For further information see the documentation and/or code which have been provided online as open software.

This tool will greatly improve experiments conducted on graphs by providing visual output. We see this tool being used at Universities and research labs, as well as for personal use for learning and testing. As future work, we would like to implement more popular algorithms such as Dijkstra's algorithm, Prim's algorithm or similar algorithms as default options. Also, the tool was made in C# in Microsoft Visual Studio because we felt most comfortable. We would rethink rebuilding the application in a different environment using a different programming language that might be more suitable.

ACKNOWLEDGMENT

We would like to give special thanks to Professor Dr. Marija Mihova and Professor Dr. Mile Jovanov for providing us with the required knowledge for development of the tool, giving us the opportunity to develop it under their guidance, as well as their continuous support during the making of this paper. We also like to thank the 'Faculty of Computer Science and Engineering' at the University St. Cyril and Methodius in Skopje, Macedonia, for providing us with the necessary tools and resources for creating this software.

REFERENCES

[1] N.Drankos (2008, January 8). Graphs [Online] Available:http://www.algorithmic-solutions.info/leda_manual/graph.html

[2] L. Lemay (2002, May 14). The history of graphs [Online] Available: http://light.physics.auth.gr/graphs/gr_history_en.html

[3] K.Etessami (2010, December 13). Discrete mathematics and mathematical reasoning [Online] Available: http://www.inf.ed.ac.uk/teaching/courses/dmmr/slides/13-14/Ch10.pdf

[4] E.Charles (2014, June 20). Graphs [Online] Available: http://en.wikipedia.org/wiki/Graph_%28abstract_data_type%29

[5] E.Gansner (2012, September 15). Graphwiz [Online] Available: http://www.graphviz.org/Documentation.php

[6] D.Ciubatti (2004, January 3). Graph Magics[Online] Available: http://www.graph-magics.com/overview.php

[7] A.Kong (2011, March 20). Illuminations resources for teaching math [Online] Available: http://illuminations.nctm.org/Activity.aspx?id=3550

Educational game with Question-Answer method

Angela Arsova Martin Stojanovski Sara Miftar Vladimir Jovanovski

Students at Faculty of Computer Science and Engineering
Cyril and Methodius University
Skopje, R.Macedonia

Bojan Ilijoski Metodija Jancheski Mario Ampov Mile Jovanov

Faculty of Computer Science and Engineering
Cyril and Methodius University
Skopje, R.Macedonia

***Abstract*— In this paper we present the new game "World of IO". It is 2D educational game that helps you to learn C# programming language. The scripts of the game are also written in C#, and the game is developed using Unity 3D[1], as part of the laboratory exercises of the faculty course "IT learning systems". This game is easy and interesting. One gets the points by answering questions, but in order to know the right answers she should have studied first. We present the game as an example of learning game, and also, as an example of a student project that can be developed as a part of a faculty course.**

Keywords—educational, questions, game

I. INTRODUCTION

'IT learning systems' is optional course which is part of the curriculum of our Faculty of Computer Science and Engineering in Skopje. It is available for all the students that previously had the introductory courses in Informatics. During this course you can learn a lot about the learning models, tools that are used in traditional learning environments, integration of technology in the learning process, the ICT systems, e-learning 2.0 and studying through playing games. Students get this knowledge through lectures, projects, discussion and laboratory exercises. When they finish the course they should be able to independently create a learning activity using ICT (for example, 2D game).

As enrolled students, we had an assignment to create a 2D educational game with Unity 3D, a program that we were learning during our laboratory exercises."Unity is a flexible and powerful development platform for creating multiplatform 3D and 2D games and interactive experiences. It was created in 2005 and announced only for Mac OS, at Apple's Worldwide Developers Conference. It is used to develop video games for web sites, desktop platforms, consoles, and mobile devices. It has been extended to target more than fifteen platforms. Unity 3D is a complete ecosystem for anyone who aims to build a business on creating high-end content and connecting to their most loyal and enthusiastic players and customers. The best thing about this program is that any individual can download it for free. It is also free for any company or incorporated entity to use as long as their annual turnover is less than $100K per year."[1]

Playing a game always feels joyful, even when you are losing, you are still playing it to try to master it and win. That is why we figured out that we can make a game using this program that can be used as educational tool in schools. There can be different approaches and methods when building an educational game. In the rest of the paper we will present the game and the method used. The game is an example of learning game, and also, as an example of a student project that can be developed as a part of a faculty course.

II. THE APPROACH AND SETTING OF THE GAME

We implemented the Question-Answer method in our game. "Question-answer is called Socratic method because he developed it – It is a teaching technique in which a teacher does not give information directly but instead asks a series of questions, with the result that the student comes either to the desired knowledge by answering the questions or to a deeper awareness of the limits of knowledge"[2].

The story of the game is as follows. We created faculty environment which includes *laboratory assistants* in areas that are considered to have easy questions, *assistants* area that have hard questions and *professor* area that have very hard questions. When player asks for question, random unanswered question will be given to him. Question types depend on current area where the question is asked, so 3 types are possible:

- 1 correct answer can be selected out of 3 presented optional answers this answers are represented with Radio Buttons, questions can be found at Laboratory Assistant's area
- 3 possible answers, any combination of all three can be checked to complete the answer, this answers are represented with Check Boxes, questions are found in Assistant's area

- Textual answer, player should write the answer in Text Field mostly two to three words will be needed found Professor's area

If player decides that he really don't know the answer of the question, he can ask for hint. Hints aren't based on text that includes the answer, but text which will point the player to right direction to find the right answer, so logical thinking is required. Hints are found in some environment object depending in which area is the player currently.

Every wrong answered question decreases question's points by portion, so player can make decisions (1) to not answer the question now and leave it for later, so no points will be decreased and he can check the hint, however if he decides to do this, next asked question will be again randomly picked, so he needs to remember the hint; (2) or he can try to answer the question with any possible combinations of answers, so in this case he will lose a lot of points, but will still know the answer.

For example if the question says: *How can we load level 3 in Unity? (20 points)* and possible answer is only one and the given answers are:

1. *Application.LoadLevel (3) (Correct Answer)*
2. *Application.SetLevel(3)*
3. *Application.CreateLevel(3)*

If player can try to answer 3rd answer as correct, he will receive "Wrong answer" message and points will be decreased to 12.

Game is repeatable and high scores are maintained so players can play it again and again to win the highest score. This means they will learn everything that is offered.

III. VISUAL SETTING OF THE GAME

A. *Main menu*

In main menu players can choose male or female character and enter their desired name for the scoring system as you can see in Fig. 1. Entering name is required and already used names can be entered again. Key *arrow left* is used to select the male character and key *arrow right* is used for female character. After this selection key *enter* can be used to enter into the game.



Fig. 1. *Main menu*

B. *Map*

Map shown in Fig. 2 is consisted of 3 areas:

1. Laboratory Assistants area
2. Assistants area
3. Professor area

Assistants and Professor area doors are locked until specific amount of points are collected. Player camera sees only portion of the whole map, which is rectangle ¼ of the image(Fig. 2) size.



Fig. 2. *Map*

C. *Player character*

Players can pick between 2 characters. Male at left and female on right shown in Fig. 3



Fig. 3. Male and Female character

D. *Question givers*

Laboratory assistants area has the characters Pinky and Blacky shown in Fig. 4. While in Assistants area live Assistants Nora and Joe shown in Fig. 5. Professor Mstach's character is shown in Fig. 6



Fig. 4. Laboratory Assistants Pinky and Blacky

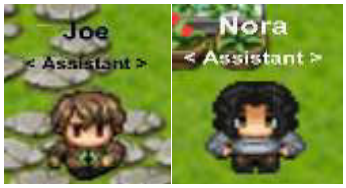


Fig. 5. *Assistants Joe and Nora*



Fig. 6. *Professor Mstach*

E. Hint objects

First 2 rocks are hint objects and are found in Laboratory Assistants area, the plant and the table are located at Assistants area and board is in Proffesors House, hint objects are shown in Fig. 7.

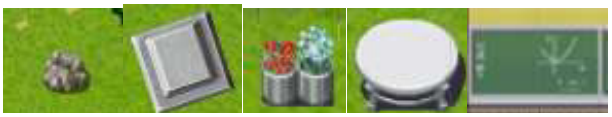


Fig. 7. *Hint objects*

Hints 1 and 2 can be found in Laboratory Assistants area, hints 2 and 3 are located in Assistants area and hint 5 is in professor's house.

IV. GAME PROCESS AND CALCULATIONS

We implemented simple and easy controls for players and we fit every action into the keyboard. Character's movement can be controlled with the keys:

- *'W'* or '*Arrow Up'* moves north/top;
- *'S' or 'Arrow Down'* moves south/down;
- *'A' or 'Arrow Left'* moves west/left;
- *'D' or 'Arrow Right'* moves east/right.

Every question giver (NPC) has its own patrol path, all patrolling paths go in big circle back to the starting position. Whenever player's character collide with some NPC, NPC pauses its route and expects to be asked. Player can ask the collided NPC with clicking *'Enter'* on the keyboard, than on bottom of the screen, centered, will be showed dialog. The dialog on the left side shows the question text and current question's points(decreased points will be also updated), on right side depending on current question type, will be showed 3 radio buttons, 3 checkboxes or 1 text input field.

Selecting radio button or checkbox can be done with keys '1', '2' and '3' respectively by top down order or using the mouse, for the textual answer player can click on input text field box to write answer. Confirming the answer is done with the key '*Enter*' or canceling the question with the key '*Esc*', so no points will be decreased and hint can be viewed for current question.

Whenever the player can't answer any more questions, 'Give up' button is available on top-right corner so current game will finish and score will be saved and new game can be started.

Every player starts with 1 hint, every time a question is answered correctly it gives +1 hint point. Remaining points after answering wrong are calculated depending on question's type. Wrong answer percent value determines how many points will be taken from remaining points. If there are any more wrong answers, points are not decreased. Points are rounded to smallest integer value greater or equal to calculated decimal value.

1. For questions with one possible answer, points are decreased to this percent:

   - 1 wrong answer to 60%;
   - 2 wrong answers to 30%

   …of remaining points.

   Example:
   Question's original points are 10,
   1 wrong answer will decrease points to 6;
   2 wrong answers will decrease points to 2.

2. For questions with one or more answers, points are decreased to this percent:

   - 1 wrong answer to 90%;
   - 2 wrong answer to 80%;
   - 3 wrong answers to 70%;
   - 4 wrong answers to 60%;
   - 5 wrong answers to 50%

   …of remaining points.

3. For questions with textual answer, points are decreased same as one or more possible answers question type[2]

Our score system only tracks tops 3 high scores.

V. FLEXIBILITY OF THE GAME



Fig. 8. XML structure for the questions

In the game questions can be easily changed because they are written in *XML*. Our *XML* structure have root node called *NPCs* which have child nodes with tag names of laboratory assistants, assistants and professor, so questions can be separately created for each question giver. *Questions* node is

child node of each question giver node and *Question* is child node in *Questions*. *Question* node has child nodes like:

- *Text* – questions text,
- *Type* – questions type (1 possible answer, 1 or more possible answers or textual),
- *Points* – starting points,
- *Hint* – logical hint for the question

Also, another child node of *Question* is *Answers* node, which have *Answer* child nodes(1 child node for textual answer or 3 child nodes for radio button and checkbox answers), *Answer* node have 2 child nodes:

- *Text* - answers text
- *Correct* – true or false

One of the best thing about the questions, as mentioned before is the possibility to modify and even change the questions and possible answers at any time. This gives the game a great flexibility and ability to fit in exactly every topic and school subject without any problems. Our XML structure for Questions is shown in Fig. 8.

VI. CONCLUSION

"World of IO" can be used in primary schools as an interesting way of knowledge evaluation. Students will be motivated to learn more, because more knowledge means having higher score. They will compete with each other in a good way. The teacher could give a motivation, like the first one or the first five on the high score list would not make an exam, for example.

We know that there are a lot of educational games out there, but what makes our game special is that the questions are optional and can be easily replaced.

This means that we can adapt the same game for every other subject, not only C#. Second best is that can be played on computers, laptops, tablets and mobile phones. This makes "World of IO" available at any time and every place.

ACKNOWLEDGMENT

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

We would like to mention and give credits to artists DoubleLeggy[1] for character sprites[2] and painhurt[3] for ground tiles[4], professor's house[5], professors chalk board[6], environment objects like trees, doors, tables, etc.. and house roof[7].

[1] http://doubleleggy.deviantart.com/
[2] "Player's female and male characters": http://doubleleggy.deviantart.com/art/RE2-Sprites-v1-0-148599973
"Laboratory Assistants Pinky and Blacky": http://doubleleggy.deviantart.com/art/RE-CV-Sprites-v1-0-151003724
"Assistants Nora and Joe": http://doubleleggy.deviantart.com/art/Visoes-3-Sprites-156625949
"Professor Mstach": http://doubleleggy.deviantart.com/art/RE-Side-Character-Sprites-v1-0-151126630
[3] http://painhurt.deviantart.com/
[4] http://www.deviantart.com/art/RPG-Maker-VX-ModernRTP-TileA2-206955520
[5] http://painhurt.deviantart.com/art/RPGVX-ModernRTP-Indoor-2-207389244
[6] http://painhurt.deviantart.com/art/RPG-Maker-VX-ModernRTP-TileD-206956501
[7] http://painhurt.deviantart.com/art/RPG-Maker-VX-ModernRTP-TileB-206956314
http://painhurt.deviantart.com/art/RPG-Maker-VX-ModernRTP-TileA3-206955900

REFERENCES

[1] "Unity3D" – *http://unity3d.com/*
[2] American Heritage ® Dictionary of the English Language, Fifth Edition. Copyright © 2011 by Houghton Miffilin Harcourt Publishing Company. Published by Houghton Mifflin Harcourt Publishing Company

On the Way to Future's High Energy Particle Physics Transport Code

Gábor Bíró[†‡], Gergely Gábor Barnaföldi[‡], Endre Futó[‡]
[†]Eötvös Loránd University
1/A Pázmány Péter sétány , H-1117 Budapest, Hungary
Email: biro.gabor@wigner.mta.hu
[‡]Wigner Research Centre for Physics of the Hungarian Academy of Sciences
29-33 Konkoly-Thege Miklós Str., H-1121 Budapest, Hungary

***Abstract*—High Energy Physics (HEP) needs a huge amount of computing resources. In addition data acquisition, -transfer, and -analysis require a well-developed infrastructure too. In order to prove new physics disciplines it is required to higher the luminosity of the accelerator facilities, by which we can produce more-and-more data in the future experimental detectors.**

Both testing new theories and detector R&D are based on complex simulations. Today have already reach that level, that the Monte Carlo detector simulation takes much more time than real data collection. This is why speed up of the calculations and simulations became important in the HEP community.

The Geant Vector Prototype (GeantV) project aims to optimize the most-used particle transport code, applying parallel computing and exploit the capabilities of the modern CPU and GPU architectures as well. With the maximized concurrency at multiple levels, the GeantV is intended to be the successor of the Geant4 particle transport code that has been used since two decades successfully. Here we present our latest result on the GeantV tests performances, comparing CPU/GPU based vectorized GeantV geometrical code to the standard Geant4 version.

Keywords*—*High Energy Physics, High Performance Computing, particle tracking, numerical simulation, computer simulation, vectorized Geant, GeantV

I. Introduction

High Energy Physics (HEP) requires a significant amount of computational resources at several layers of the data taking processes. Detector upgrades of the experiments generate further needs of the computational capacity at hardware, software, and middleware levels. Since the increase of the CPU-clock frequency saturated few years ago, developers turned to push more effort to the software-development side in order to exploit all the capabilities exist in the parallel architectures and many-core computing.

Early performance analysis of the detector and theory simulations have been already explored the bottleneck of the numerical calculations, which found to be the particle transport – the most time-consuming part of the simulations. Due to this reason, the investigation of the possibility to rebuild the most widely used particle transport framework, the Geant (GEometry ANd Tracking), in a vectorized/parallelized way has been started [1–16].

Recent version is the Geant4 simulation framework, which is in use by most of the experimental collaborations of the Large Hadron Collider (LHC, CERN) and many other facilities all around the world. Its main purpose is to simulate the passage of particles through matter using Monte Carlo methods. The development of the original simulation package started in 1994. Since 2012 studies for the possible successor, the Geant Vector Prototype (GeantV) project has started [1, 2, 17]. The Geant development was always a worldwide collaboration and it has many application areas outside of particle and nuclear physics, such as: space science, nuclear reactor development, or medical/radiological fields.

The basic idea to build the parallel version of these simulations lays on the vectorized geometry calculations. The vectorized GeantV code executes single instructions on multiple data (SIMD) in parallel. Thus, a relevant speedup has already achieved on CPU architectures which supports vector instructions.

Nowadays, it is a popular trend to transform numerical codes able to run on the fast, General-Purpose computing on Graphics Processing Units (GPGPU) [18]. This fashion fits nicely to the strategy of GeantV development, since geometrical computations with GPUs are pretty effective. In the case of HEP geometric calculations for all particles are independent, thus it is reasonable to prepare the particle transport code for such cases and implement support for GPUs or other many-core architectures.

The VecGeom library (Vectorized Geometry) is the geometry package which was started to develop in the context of the Geant Vector Prototype project in 2013 [2, 3]. By the time it will be finished, this would become a standalone library that provides a GPU support via the Computer Unified Device Architecture (CUDA) compiler as well. This CUDA backend already exists in VecGeom, which supports only NVIDIA devices at the moment. On the other hand we investigated how could we increase the portability and the performance on multiple architectures, with the usage of the Open Computing Language (OpenCL) framework [19].

In this paper first we give a general overview of the Geant framework, then we introduce the generalization of the particle transport simulation. Finally we show the speedtest results of VecGeom on CPU and GPU architectures and present the comparison between them.

II. Project goals

Here, we introduce the general aims of the GeantV project focusing on the parallelization and vectorization.

A. New generation of particle transport simulations

The famous Moore's law stated in the '60s says that the number of transistors integrated on one single chip grows exponentially with time [20]. The scaling law presented on Fig. 1, is still valid in hardware sense. Although relevant speedup has not been achieved recently, because the clock frequency of the CPUs has reached the currently maximal value of 3-4 GHz since the middle of the last decade. Due to the increasing number of CPU cores per chip, and other advanced architecture technologies, it is still possible to increase the computing performance. The exploitation of this increasing performance needs new, so far not widespread programming techniques. Since the popularity of video games increased recently, the graphical hardwares (GPUs) were greatly improved. Today these hardwares become suitable and used even for scientific calculations [3, 17, 18, 21–23].

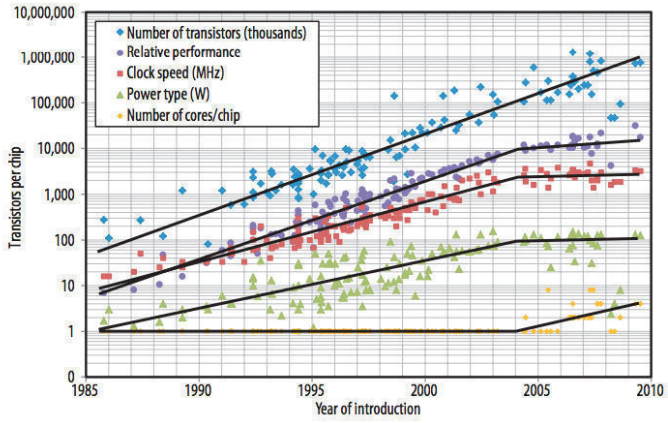


Fig. 1. The famous Moore's law says that the number of transistors integrated on one single chip grows exponentially with time [20]. However, it is still valid in hardware sense, but significant computational power has not been achieved so far due to the CPU-clock speed saturation on 3-4 GHz [24].

The power of the computing performance may not appear by increasing the clock frequency, but in the utilization of parallelization and vectorization. The parallel computing with CPUs has long history. The first SIMD instruction set appeared at late '90s, however in the last few years the new generations of CPUs have more-and-more advanced SIMD support (SSE4, AVX-512) and therefore they are becoming more effective with a vectorized software. The outline of the operating principle of the SIMD instructions is seen on Figure 2.

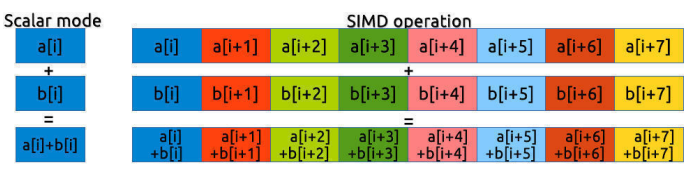


Fig. 2. *Left panel*: A scalar code performs one operation per CPU-tick, while a vectorized, SIMD-instruction based operation is capable to make the same operation on a data vector at the same time (*right panel*).

The main goal of the Geant Vector Prototype Project (GeantV) is to exploit the benefits of the modern CPU, GPU and many-core architectures at the same time. The project started in 2012 with preliminary performance analysis which showed that in a typical detector simulation the geometry calculations itself using major of the computational resources, $40 - 50\%$ in time [1]. On the other hand, geometry calculations are independent of each other, which gives hand-on opportunity for parallelization. The layout is seen on Figure 3, where scheduler manages geometrical algorithms and physical processes [3, 5, 17].

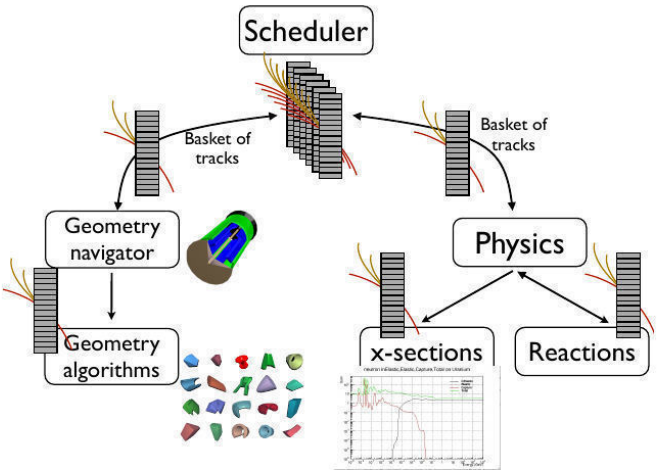


Fig. 3. The structure of the GeantV prototype: scheduler manages geometrical algorithms and physical processes [5, 17].

Development of this code aimed to get maximal performance beside maintainability and hardware independence. On the top of this, the predecessor, Geant4 has been developed and used since two decades and the successor has to be suitable for long term as well. Thus the software has to be compatible not just with the older hardwares but with future ones too. Following Fig. 3 the development of GeantV has three different levels:

1) Scheduler: While in Geant4 the particle transportation happened serially, GeantV will manage multiple particles at the same time. Because the transportation is a local process and most of the simulation steps occur in a small part relative to the whole detector geometry, the main idea was to collect the particles which are in the same type of volume into vectors (baskets) and perform computations parallel using SIMD instructions. In this way both the necessary memory containing the different parts of the detector, and the necessary computational time can be strongly reduced. The scheduler manages the baskets of particles as an interface between the physics and geometry.
2) Physics: At present the physics of GeantV is tabulated, which means that instead of theoretical calculations the physical processes are pre-calculated and ordered in multidimensional tables. Because in this way all the possible final states are given in advance, the necessary computational time is much less. On the other hand, because of the finite number

of possible final states, the precision will be less as well. For higher precision one should pre-calculate and load more final states, requiring more memory. The physics of the final version of GeantV is still in the design phase.

3) Geometry: In a particle transport simulation a very significant part of the runtime (40-50%) is devoted to the geometrical calculations. Therefore it is important to take advantage most of the computational capacities of the hardware in the implementation of the geometry code. The main purpose of the VecGeom (Vectorized Geometry) package is to use more efficiently the different hardware architectures via vectorization.

III. High Performance Computing in HEP

In parallel to the development of detector- and data-acquisition technologies, computing architectures went through a major evolution during the last years. In order to keep the scientific advancement, the necessary infrastructure needed to improve following the construction of the huge particle accelerators and giant detector systems. This trend is still ongoing: the hardware manufacturers release the newer and better hardwares regularly, sometimes motivated by the needs of the HEP community.

As we have mentioned earlier, in Geant4 in high-energy physics simulations the detector-particle interaction uses the main computational-resource part. Moreover any detector-simulation and -design require particle transport simulations especially one needs to calculate the detector efficiency or perform predictions by theoretical calculations. We note, besides HEP applications several other disciplines were also served such as medical applications or space sciences, where there is a strong need to simulate the transport of particles through matter.

In a transport simulation the most significant quantity regarding the computational time is the number of steps. While the particle passing the matter, it will interact with its environment many times, and the distance between two interactions called step length. For a given process, i this length is determined by the mean free path, λ_i, which is calculated by the following formula:

$$\lambda_i = \frac{1}{n\sigma_i} \quad , \tag{1}$$

where σ is the total cross section of the given process and n is the density of the medium. The total cross section is the sum of the cross sections for all processes, i,

$$\sigma_{total} = \sum_i \sigma_i \quad . \tag{2}$$

If the size of mean free path is large (e.g. in vacuum), the step length can be large as well, so one may needs only a few steps until one can eliminate the particle (when it leaves the range of interest). On the other hand, if the mean free path is small, it means that interactions take place even in short distance. In this case one needs more steps with smaller λ_i which require more computational time to simulate the same distance at the same precision. The inverse of the mean free path equals to the inverse sum of the mean free path of all the possible interactions:

$$\frac{1}{\lambda_{total}} = \sum_i \frac{1}{\lambda_i} \quad . \tag{3}$$

The two main parts of a Geant4 particle transport simulation are physics and geometry.

A. Physics

For a given particle type at fixed energy it may happen different kinds of interactions. One can distinguish the following three types:

- At rest: It occurs when the kinetic energy of the particle is zero. The selection criteria for 'at rest processes' is the lifetime of the particle (e.g. radioactive decay).
- Along step: Processes which happen during the transportation (e.g. ionization).
- Post step: The processes which take place after one step, after the transportation has been fully completed. This depends on the interaction length (e.g. elastic scattering).

The process which is being implemented in the actual step is selected by the lowest mean free path. In the first step the code calculates all the possible mean free paths via Monte Carlo method for all the possible interactions at a given energy. In the next step the shortest mean free path will be implemented. If the lowest mean free path is larger than the distance from the geometric border (which is the boundary layer of two volume with different materials), then the particle is being transported to the border. In the next step the mean free paths will be calculated within the material of the new volume.

These two methods (the calculation of the physical interaction length (*GetPhysicalInteractionLength*) and the execution of the step (*DoIt*)) are implemented in every steps. During a simulation the tracking of a particle is ended when its energy becomes lower than a given threshold or it decays, or it leaves the investigated geometry. The execution of a step (transportation and the implementation of the processes) is calculated by Monte Carlo methods as well.

B. Detector geometry

Complex detector systems are built with basic three dimensional geometric units like rectangular solids and tubes. The fundamental geometric calculations, like coordinate transformations and distance calculations are implemented in the source code of these elementary bodies. Using boolean operations one can construct complex, composite elements as well. Finally, using hierarchical connections, one can build the whole detector geometry as an ordered system of the basic building blocks. The geometric calculations, like e.g. distance calculations do not depend on the particle type. This is one of the most computational intensive tasks, since it has to be carried out in every step. Thus it is crucial to implement them in an efficient and fast way.

The time is determined by the number of steps. In Geant4 the simulation is serial, which means that it tracks all the particles individually, one after the other. This results that the running time of the simulation will be proportional to the number of the particles. Since the simulation of the large detector systems needs a huge amount of resources, the next generation of particle accelerators and detectors require to develop the new generation of particle transport simulations. For example: the simulation of one single particle in the ALICE detector system needs a few ∼ms, while in the LHC there are particle collisions and particles to be detected in every ∼ns.

C. VecGeom

The VecGeom library is the vectorized geometry package of the GeantV which is devoted to speed up the code of the elementary three dimensional solids and the basic calculations: rotations, translations, logical operations, and distance calculations [2, 3]. During a simulation there are three questions to be answered for all the particles:

- Is the particle inside or outside the solid?
- What is the minimal distance between the particle and the border of the solid?
- What is the maximal step length for a particle with a given velocity and direction?

The purpose of VecGeom is to answer these questions as fast as possible with parallelized calculations. The main idea is to collect all the particles which are located in the same type of volume and perform the calculations for all of them simultaneously with SIMD operations. In addition, the same operations have to be executed with different architectures, while minimizing code duplication. Using of template classes the VecGeom library is able to run the same code on either CPUs or GPUs even with or without vector instructions.

For vectorization on CPU architecture one may have multiple choices. The VecGeom uses the Vc SIMD library to perform explicit vectorization [25]. Thereby the code will be always vectorized regardless of the compiler. In order to use the same code on GPU architectures the VecGeom uses CUDA platform which works very nicely with the CPU vectorized C++ code too. A further option can be the OpenCL (and SYCL) version, which is under construction [19].

In order to present the power of the VecGeom library, we performed a speed test of two bodies, a Box and a Tube, presented on Figure 4. During the test the speed of the DistanceToOut method were measured, which gives the distance of randomly placed particles with random speed escaping from the bodies. Fig. 4 shows the relative speedup to the Geant4 calculations as a function of the number of particles and the vertical axis shows . In order to reduce the statistical fluctuations the calculations at a fixed particle number was repeated 5000 times.

The tendency of the tests is clearly visible: in case of CPU calculations a relevant speedup were measured even with low particle number, while in case of GPU to higher the particle number increase the speed of the calculations.

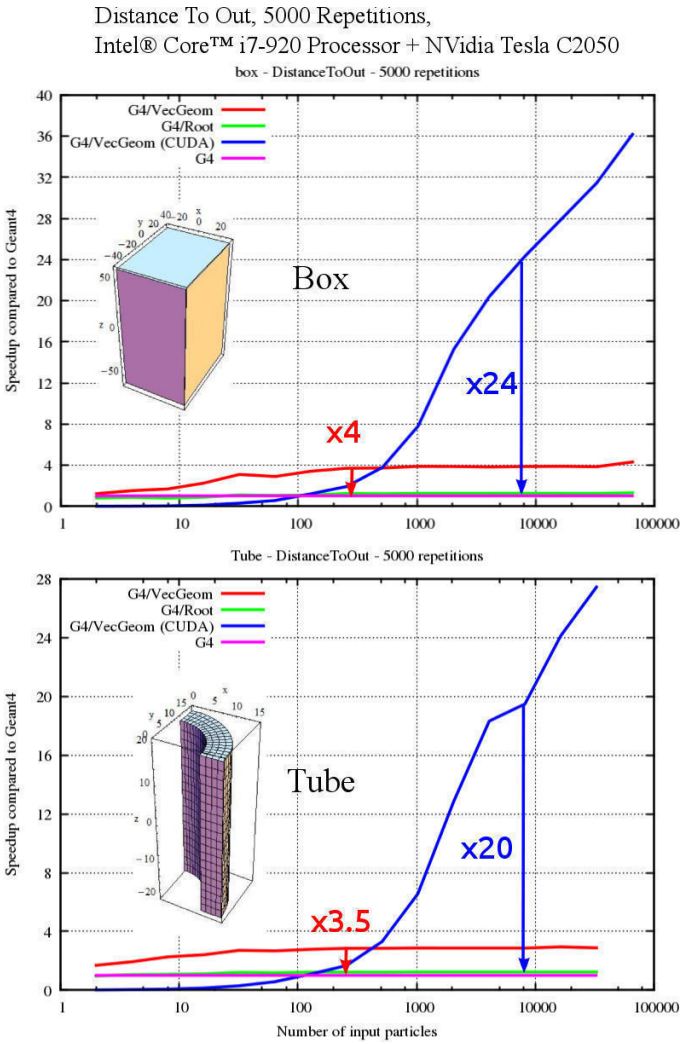


Fig. 4. Speedtest of the VecGeom *Box* and *Tube* bodies - the used hardwares are an Intel Core i7-920 CPU (SSE4.2 instruction set) and a NVIDIA Tesla C2050 computing processor [26].

IV. Conclusion

Since future HEP facilities require more-and-more speedup both by hardware and software way, the development of the next generation of GEometry ANd Tracking (Geant) code for many-core architectures was started. The aim is to build the vectorized GeantV code able to manage several times speedup calculations either on CPU and on GPU architectures. Here, we presented the first promising results with the VecGeom library for simple body (Box and Tube) cases which were done in collaboration with the CERN PH-SFT group during a CERN Summer internship in 2014.

Acknowledgment

This work was supported by Hungarian OTKA grants, NK106119, K104260, and TET 12 CN-1-2012-0016. Author G.G. Barnaföldi also thanks the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Author G. Bíró also acknowledge the support by the Wigner Research Centre of the H.A.S. and the support of the CERN-PH-SFT group.

REFERENCES

[1] J. Apostolakis, R. Brun, F. Carminati, and A. Gheata, "Rethinking particle transport in the many-core era towards GEANT 5," *Journal of Physics: Conference Series*, vol. 396, no. 022014, 2012.

[2] M. Gayer *et al.*, "New software library of geometrical primitives for modeling of solids used in Monte Carlo detector simulations," *Journal of Physics: Conference Series*, vol. 396, no. 052035, 2012.

[3] J. Apostolakis, R. Brun, F. Carminati, A. Gheata, and S. Wenzel, "Vectorising the detector geometry to optimize particle transport," *Journal of Physics: Conference Series*, vol. 513, no. 052038, 2014.

[4] P. Canal *et al.*, "High energy electromagnetic particle transportation on the GPU," *Journal of Physics: Conference Series*, vol. 513, no. 052013, 2014.

[5] J. Apostolakis, R. Brun, F. Carminati, A. Gheata, and S. Wenzel, "The path toward HEP High Performance Computing," *Journal of Physics: Conference Series*, vol. 513, no. 052006, 2014.

[6] J. Apostolakis *et al.*, "Parallelization of particle transport using Intel® TBB," *Journal of Physics: Conference Series*, vol. 513, no. 052025, 2014.

[7] F. Carminati, "Geant 5 - an update," Annual Concurrency Forum Meeting 2013 (CERN).

[8] S. Wenzel, "Towards a high performance geometry library for particle-detector simulation," ACAT 2014 (Prague, Czech Republic).

[9] J. de Fine Licht, "First experience with portable high-performance geometry code on gpu," GPU Computing in High Energy Physics 2014 (Pisa, Italy).

[10] J. de Fine Licht and S. Wenzel, "GeantV Geometry: SIMD abstraction and interfacing with CUDA," Annual Concurrency Forum Meeting 2014 (CERN).

[11] S. Wenzel, "Status of Geometry developments for SIMD and GPUs," Annual Concurrency Forum Meeting 2014 (CERN).

[12] S. Wenzel, "Vectorizing the geometry library for simulation – experience and results from a prototype and future directions," 5th International Workshop on Future Challenges in Tracking and Trigger.

[13] A. Gheata, "Fast Detector Simulation And The Geant-V Project," ACAT 2014 (Prague, Czech Republic).

[14] F. Carminati, "Status and Plans for Geant V prototype," Annual Concurrency Forum Meeting 2014 (CERN).

[15] S. Y. Jun *et al.*, "High Energy Electromagnetic Particle Transportation on the GPU," Annual Concurrency Forum Meeting 2014 (CERN).

[16] S. Y. Jun, "Sampling secondary particles in high energy physics simulation on the GPU," GPU Computing in High Energy Physics 2014 (Pisa, Italy).

[17] J. Apostolakis, R. Brun, F. Carminati, A. Gheata, and S. Wenzel, "A concurrent vector-based steering framework for particle transport," *Journal of Physics: Conference Series*, vol. 523, no. 012004, 2013.

[18] B. Oancea and A. Tudorel and R.M. Dragoescu, "GPGPU computing," *Challenges of the Knowledge Society*, p. 2026, 2012.

[19] G. Biro, "Investigation of OpenCL support in the VecGeom geometry package," *CERN Summer Student Report*, 2014, report number: CERN-STUDENTS-Note-2014-234.

[20] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.

[21] G. Moore, "Parallelization of the PENELOPE Monte Carlo particle transpot simulation package," *Nuclear Mathematical and Computational Sciences: A Century in Review, A Century Anew*, 2003.

[22] S. Jarp and A. Lazzaro and A. Nowak, "The future of commodity computing and many-core versus the interests of hep software," *Journal of Physics: Conference Series*, p. 052058, 2012.

[23] Sz. Borsanyi and Z. Fodor and C. Hoelbling and S.D. Katz and K.K. Szabo, "Full result for the QCD equation of state with 2+1 flavors," *Phys.Lett.*, vol. B730, pp. 99–104, 2014.

[24] http://cr4.globalspec.com/blogentry/25398, downloaded: 2014.11.10.

[25] http://code.compeng.uni-frankfurt.de/projects/vc.

[26] http://gpu.wigner.mta.hu/index.php?id=labor_hu.

MICROWAVE IMAGING: CURRENT TRENDS AND FUTURE PERSPECTIVES

Danail Tavchioski
Faculty of Computer Science and Engineering, University "Ss. Cyril and Methodius", Skopje, Macedonia

Microwave imaging has gained a great deal of attention among researchers over the past decade. It has evolved from older detecting/locating technologies like Radar. Today, it is used to evaluate and discover hidden objects in a structure or media using electromagnetic waves in the microwave spectrum (300Mhz-300Ghz). State-of-the-art applications that use this technology include medical imaging, concealed weapon detection, structural health monitoring and through-the-wall imaging.

This paper provides an overview of microwave imaging and its applications. We will present the current trends and future perspectives of this technology, as it evolves at a very fast pace and improves other disciplines and techniques.

Keywords—microwaves, medicine, gpr, radar, cwd

I. INTRODUCTION

Microwave imaging as the name suggests is an imaging technique used in places where embedded or hidden objects need to be evaluated and afterwards examined. Microwave imaging uses extend to variety of areas and technologies, like telecommunications, radar, medicine and even spacecraft communication. It evolves from the early radar implementations used to reveal objects where other techniques fell short. By using microwave transmitters on the one end and microwave receivers on the other, we have a system that detects the reflection of electromagnetic waves (EM) and can therefore discern the object that needs further evaluation. Possible medical uses made this technology extensively investigated in the past few years. The results are mostly positive and encouraging. The breast cancer diagnosys systems that are in use today, like X-ray mammography and Ultrasound imaging although satisfying, they are nowhere near perfect. Microwave imaging helps determine cancer better and at the same time has less (very little) harmful effects on patients. The results it yields have bigger resolution and more contrast, which helps breast cancer specialists in their analysis.

This paper is organized as follows. In part two we will discuss the technology behind microwave imaging. The procedure it follows, the technical devices and their practical configuration. In part three we will elaborate the different applications of microwave imaging techniques. Medical applications like cancer detection, GPR (or ground penetrating radar), concealed weapon detection and NDT&E (non-destructive testing and evaluation) uses. In part four we will have a glimpse at the future perspectives of microwave imaging, its possible implementations and further research needed to make it widely applicable. Finally we will conclude this paper in part five.

II. TECHNOLOGY BEHIND MICROWAVE IMAGING

The basic microwave imaging system is made of two parts, the hardware and its supporting software for further processing and evaluating. The principal on which microwave imaging works is electromagnetic wave reflection. The system includes two types of antennas, transmitters and receivers. Any n number of transmitters send electromagnetic waves with frequency in the microwave range (300 MHz - 300 GHz) [1], and any m number of antennas receive the waves (receivers) that pass through a test object. By theory, if the waves pass through an object made of homogeneous materials, no reflection will be detected. If the object has some irregularity with different electromagnetic properties compared with the homogeneous medium, it will reflect a portion of the electromagnetic waves. The more contrast between the homogeneous material and the irregular material, considering the electromagnetic properties of the both, more EM waves will be reflected.

The configuration of the system will implement either mono-static or bi-static radar (radar in sense of detector) [2]. The mono-static implementation has an antenna that is responsible for both sending EM waves and receiving the reflecting EM waves. A bi-static implementation makes use of more antennas (or an array of antennas), some responsible only for transmitting and others only for receiving reflected waves (Figure 1). In any implementation, the distance between the antennas should be carefully managed, meaning that any two antennas should have a distance smaller than the wavelength they operate on [1]. This being best practice, allows less degradation of the received (reflected) signal and therefore provides raw data that can be later more accurately (and more easily) analysed. An array of antennas can also be used. Here the space will be more complex to manage but bigger area will be covered more efficiently by gathering separate samples by each antenna. Furthermore, a flexible system can be used to simplify the process; namely, a moving antenna (by some kind of aperture) collects samples as it "scans" the test object [3]. The samples need to be processed simultaneously and then passed on to post-processing.

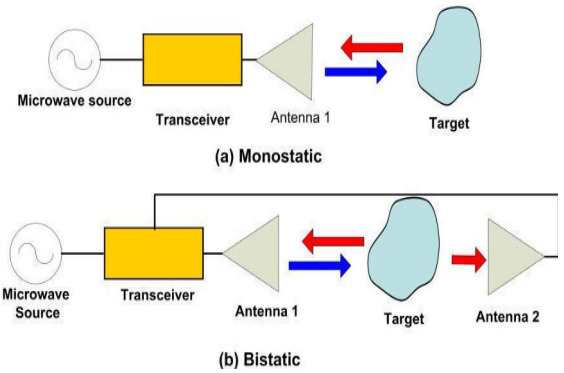


Fig. 1. Mono and Bi-static radar [2].

The post-processing part uses algorithms to make the raw data received from the hardware, human understandable. According to the algorithm, the microwave imaging technique is either quantitative or qualitative [4]. Active microwave imaging methods rely on recovering the dielectric properties of an object under test from the scattered fields measured from a number of antennas distributed around it. To do so, a microwave signal is transmitted from one antenna and the scattered signals are received at the remaining antennas of the array. This process is successively repeated until all the antennas have been used as transmitters.

Among the existing imaging algorithms, one can distinguish between radar-based or tomographic methods. The radar-based method obtains a qualitative image of the most significant scatters in the reconstruction area. Tomographic methods on the other hand can provide a quantitative complex permittivity image by solving the corresponding inverse scattering problem. The most often used algorithms include "Chalmers" or Time Domain Inversion algorithm (TDIA), and UWB (Ultra Wide Band) Magnitude Combined Tomographic algorithm (UPC). The first one is an iterative technique, while the second falls partially in the area of diffraction tomography [5].

III. Applications

The properties of microwaves make microwave imaging a desirable method in many technologies. Its uses extend even further when combined with previous (already in wide use) methods like Ultrasound Imaging. Therefore we will elaborate few of the most widely researched and anticipated applications of microwave imaging.

A. *Medical uses (Breast cancer detection)*

1) Microwave Imaging

Microwave signals provide a good compromise between resolution and penetration of body tissues, and hereby offer a good opportunity to replace (or at least improve) the results offered by other techniques like Ultrasound imaging or X-ray. Microwave imaging is found to be an effective diagnostic tool for breast cancer [6] patients and its techniques are elaborated further on. Other lesser uses include brain stroke diagnostics or cardiovascular diseases detection [7].

Breast cancer is one of the most common cancers diagnosed in women of any age. Studies made on larger scale over the years show that more than 8% of women will suffer from this disease during their lifetime [8]. Nearly 1.7 million new cases of breast cancer were diagnosed around the world in 2012 (that takes 25% of all cancers) and an estimated 522,000 breast cancer deaths occur in the world every year. Since 2008 breast cancer incidence around the world has increased by more than 20%, while mortality has increased by 14%. By these facts, at the current rate, 13 million breast cancer deaths will occur in the next 25 years around the world [9].

Current methods for breast screening are X-ray mammography and in some cases Magnetic Resonance Imaging (MRI) or Ultrasound Imaging. X-ray mammography is a fairly good tool for diagnostics but falls somewhat short in some important elements. Firstly, the relatively high false-negative rate, which can sometimes be 30% [10]. Secondly, its high false-positive rate. Namely, according to statistics, on average 75% of breast biopsies prompted by a suspicious mammography abnormality prove to be benign [11]. Thirdly, screening mammography shows less accurate images, failing to detect up to 30% of cancers greater than 5mm in diameter [12] in women with radio-graphically dense breast tissue prevalent in younger women which is due to the relatively poor soft-tissue contrast. Finally, X-rays pose a small risk of inducing cancer due to the ionizing radiation exposure [13] and should not be used very often.

MRI on the other hand has a moderately higher success rate of correct diagnosis but also has high operational costs. The only sure diagnosis of all types of breast cancers depends on a biopsy, which is an invasive procedure to remove and examine cells or tissue for the presence of cancer. However, a biopsy returns results to be benign in 80% of breast lesions detected by mammography findings [14].

The situation calls for alternative, new and improved diagnostic tools to detect cancer more precisely and reduce physical and mental suffering of the patients. The development of low-cost diagnostic method with better chance at successfully detecting breast cancer, and which can be more easily accessed by as much women in the world as possible, may be microwave imaging. It has been widely assumed that normal breast tissue is largely transparent to microwaves because they are featured with a low relative permittivity and conductivity at the microwave frequency bands, whereas lesions, which contain more water and blood are characterized by a high relative permittivity and conductivity at the microwave frequencies and hence they cause a significant backscatter [15]. Upon this assumption, which is supported by tests and measurements, microwave imaging systems are being designed to detect the presence of a small object inside a breast causing a considerably larger backscatter than the surrounding medium. Normal breast tissue is largely transparent to microwave radiation, whereas malignant tissues, which contain more water and blood, cause microwave signal backscattering. This scattered signal can be picked by an array of microwave antennas and analysed using a computer.

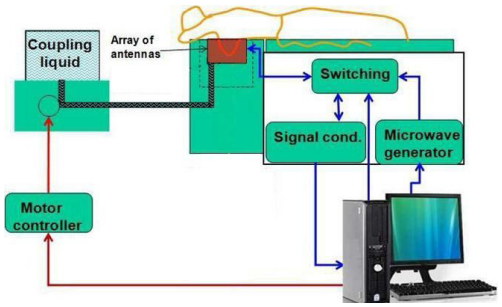


Fig. 2. Prototype microwave imaging radar system [2].

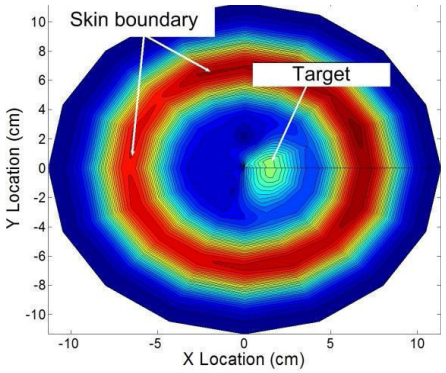


Fig. 3. Microwave image produced by the system [2].

The configuration of a prototype radar system is shown in the Figure. 2. The system consists of a circular cylindrical scanning platform with a resolution of 1° to support a breast phantom, and a mechanical scanning platform with resolution of 0.1 mm in the vertical axis. The scanning platform supports an array of wideband antennas. The antenna is connected to a microwave Vector Network Analyser that measures the scattered signals. The collected scattered signals are then processed in a personal computer to get an image of the breast. The imaging capabilities of the aforementioned radar system are carried out for an artificial breast phantom. Sample of the obtained images is shown in Figure. 3. The figure clearly shows the boundaries of the layer representing the skin, size and shape of the object that represents the tumor [16, 17].

2) Hybrid Imaging

In order to remove the obstacles facing the success of the microwave-based imaging methods, the hybrid technique is proposed and tested [18]. The hybrid technique utilizes the dielectric (electrical properties) and elasticity (mechanical properties) contrast between tumors and healthy tissue in order to produce a three dimensional image of the breast. The information from the hybrid image significantly enhances early diagnostic accuracy.

A system has been devised to show how hybrid imaging can be performed. A very short frequency-modulated pulse is applied from an Ultra Wideband (UWB) antenna beside the breast of the patient in the face-down position via a switching unit. At the same time, an acoustic signal is sent from a transducer below the breast. The proposed system utilizes the benefits of few techniques to produce a relatively accurate three-dimensional image. Microwaves are used to give a full view of the dielectric contrasts, while the acoustic wave shows the elasticity properties and distributions within the breast. The final image will have higher resolution and contrast than any of the techniques used separately. Cylindrical array of UWB antennas encircling the breast collect the scattered signals that are consequence of the difference in electrical and mechanical properties of different types of breast tissues. Important parameters of the signals are processed according to their correlation with the breast tissue characteristics. Then they are processed and digitized on a computer before images are produced. The operator is then able to distinguish tumors from the healthy tissue by the high backscatter (reflection of waves) and high stiffness [19].

In order to remove undesired reflections due to the mismatch between the antenna, the skin layer and the space separating them, the breast under test is immersed in a coupling liquid which has material properties that reduce the backward scattering at the skin layer, and thus increasing the dynamic range of the system [20].

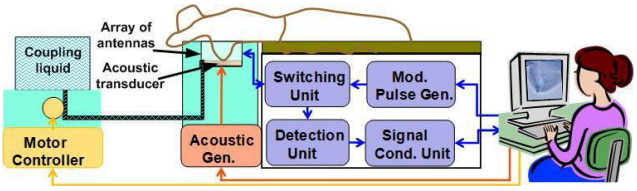


Fig. 4. Hybrid imaging system (schema) [19].

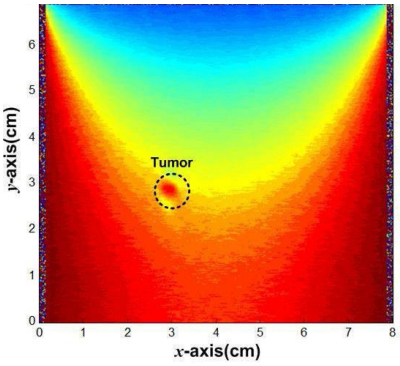


Fig. 5. Successfull detection of small tumor using hybrid imaging [2].

B. Ground penetrating radar (GPR)

A ground penetrating radar is a non-destructive sensor technology used to investigate and evaluate objects underground or inspect structures of buildings. As a diagnostic instrument it finds grand application because it's nature is microwave-based. It uses electromagnetic radiation in the microwave band (UHF/VHF frequencies) of the radio spectrum and detects the reflected signals from subsurface structures. The high-frequency radio waves that range from 10MHz to 1GHz are usually polarized. The objects buried under the media surface (the media can be soil, rocks, ice, water,

pavement and similar) reflect a portion of the sent radio signal depending on their properties [29]. Objects with materials having different dielectric constants reflect, refract or scatter signals back to the surface to the GPR receiving antenna. The variations in the received signal are then processed and analysed [28]. An example is given in Figure 6.

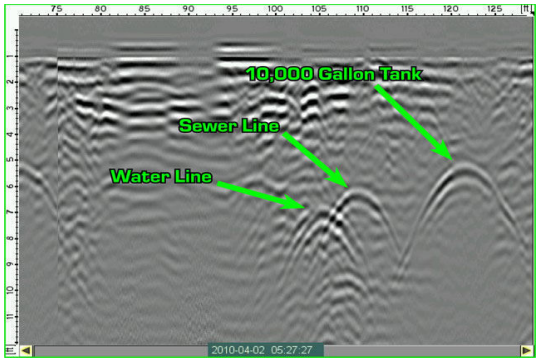


Fig. 6. GPR underground scan [30].

Depending on the ground on which the GPR works, some variables as the frequency of the signal, the radiated power and the electrical conductivity of the ground itself have specific influence on the results. The higher the frequency of the microwave the lower the depth that can be imaged and vice versa. However, higher frequency wave tends to have better resolution. Differences in electrical conductivity attenuate the wave and thus decreases or increases the depth that can be reached. Fine tuning the GPR is needed depending on the material being tested. The GPR can (according to the operator) lay on the ground or work from a certain distance but with lesser quality results.

The shortcomings of the technology and its weakest points are the high-conductivity materials like clay, rock or soils that are salt contaminated. Interpretation of the radar-grams (images captured with the radar) is not a simple job, and considerable expertise is required to conduct GPR surveys. GPR also has high energy consumption and remote field surveys are harder to conduct.

Engineering applications like non-destructive testing (NDT) are done on building structures and pavements, including locating of utility lines, studying soils and bedrock, defining landfills, contaminant plumes and water leakages in reservoirs [21]. Military uses radar imaging reconnaissance satellites like the Lacrosse to discover underground bunkers and similar facilities [22].

C. Concealed weapon detection using Phased Antenna Array

A lot of research has gone into concealed weapon detection techniques, especially after 9.11 in the United States. The technique of phased array imaging consists of illuminating the target with microwave radiation from a certain distance. This excites the low order complex natural resonance (CNR) frequency and allows for a late time response (LTR) of the concealed item to be spatially located on the subject [23]. The post-processing of the images will be able to discern between objects of importance (weapons) and benign items like mobile phones or keys for example.

IV. FUTURE PERSPECTIVES

Sparse microwave imaging radar is a novel strategy of microwave imaging radar. It introduces sparse signal processing technology to microwave imaging and combines them to form a new theory and system that can be used to procure better results and lower the complexity of the first. [24] The sparse microwave signal is acquired by sparse sampling in space or time domain. After the signal processing and information extraction, the geometrical and physical characteristics of the observed object can be acquired with lower data rate and less hardware complexity of the imaging radar.

Compared with the traditional radar imaging methods based on matched filter theory, the sparse microwave imaging algorithms could take advantage of obviously fewer measurements far below the Nyquist Sampling rate to reconstruct the target scene by solving an lq ($0 < q \leq 1$) norm optimization problem [25]. It is convenient for data acquisition, storage, transmission and processing, and it also could reduce the system complexity and achieve high resolution microwave imaging.

The medical use of microwave imaging techniques takes time to become a reality. A large scale study showed that the dielectric properties of a variety of normal, malignant and benign breast tissues can vary from individual to individual and also depend on age [26, 27]. In some cases, the difference in dielectric property between healthy normal tissue and malignant tissue is too small (less than 10%) to get a good image with enough contrast. Another challenge comes from the fact that in some cases when a microwave penetrates in a heterogeneous medium, the wave reflects, refracts and scatters too much and it is very difficult to make a three-dimensional image. The trend here is a transition to developing a hybrid system that overcomes some shortfalls of microwave-only imaging for detection of breast cancer.

Computational acceleration for image reconstruction is an important part of the post-processing stage. Researchers are turning to MATLAB language for its high performance, ability and ease of use. MATLAB Graphical User Interface (GUI) is a powerful tool to exploit advanced computational and graphical functions of MATLAB, though not the most efficient. When it comes to computationally complex problems as 3D microwave image reconstruction, computer scientists turn to other prominent programming languages, such as FORTRAN, C or C++. Parallel computing has been in the spotlight for some time now, and MATLAB makes use of it. The Parallel Computing Toolbox implemented in the MATLAB software, parallelizes high-level constructs such as for-loops, using multi-core processors or computer clusters. [31, 32]

V. CONCLUSION

Microwave imaging has been around for decades but its potential not very widely recognized. Many of today's and tomorrow's technologies can grow and evolve thanks to

microwave imaging techniques, and fields like Medicine will have a better chance at helping people live longer and better lives. There are challenges of great variety and problems occur that have to be addressed in the future. Researchers put their efforts in recognizing the full arrange of possibilities gained by using microwave imaging and results are everything but fruitless. People are collaborating in many universities and institutes around the world to study the uses of microwave imaging and we have yet to see the full potential of this somewhat old but nevertheless modern technology.

ACKNOWLEDGMENT

I would like to thank my professor and mentor Biljana Stojkoska, PhD, for her expert opinion and guidance in writing this paper.

REFERENCES

[1] Mojtaba Fallahpour, "Synthetic aperture radar-based techniques and reconfigurable antenna design for microwave imaging of layered strutures". Missouri University of Science and Technology, 2013.

[2] Younis M. Abbosh, "Breast Cancer Diagnosis Using Microwave and Hybrid Imaging Methods", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.5, No.3, June 2014.

[3] M. Soumekh, Synthetic Aperture Radar Signal Processing, 1st ed. New York, NY, USA: Wiley, 1999.

[4] L. Lianlin, W. Zhang, and F. Li, "Derivation and Discussion of the SAR Migration Algorithm Within Inverse Scattering Problem: Theoretical Analysis," IEEE Trans. Geosci. Remote Sens. E, vol. 48, no. 1, pp. 415–422, Jan. 2010.

[5] Guardiola, M. ; Fhager, A. ; Jofre, L. , "Circular microwave tomographic imaging. Experimental comparison between quantitative and qualitative algorithms", Proceedings of the 5th European Conference on Antennas and Propagation, EUCAP 2011. Rome, 11-15 April pp. 2800-2804.

[6] P. Meaney, M. Fanning, D. Li, S. Poplack, and K. Paulsen, "A clinical prototype for active microwave imaging of the breast," Microwave Theory and Techniques, IEEE Transactions on, vol. 48, no. 11, pp. 1841 – 1853, Nov. 2000.

[7] S. Semenov, R. Svenson, V. Posukh, A. Nazarov, Y. Sizov, A. Bulyshev, A. Souvorov, W. Chen, J. Kasell, and G. Tatsis, "Dielectrical spectroscopy of canine myocardium during acute ischemia and hypoxia at frequency spectrum from 100 khz to 6 ghz," Medical Imaging, IEEE Transactions on, vol. 21, no. 6, pp. 703 –707, 2002.

[8] K. Subashini, K. Jeyanthi "Breast Cancer Detection And Classification In High Resolution Ultrasound Images" International journal of innovative research in electrical, electronics, instrumentation and control engineering Vol. 2, Issue 5, May 2014.

[9] American Cancer Society, Global Cancer Facts & Figures 2nd edition, Globocan 2012- http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

[10] P. Huynh, A. Jarolimek, and S. Daye, "The false-negative mammogram," Radiographics, vol. 18, no. 5, pp. 1137–1154, 1998.

[11] S. Nass, I. Henderson, and J. Lashof, Mammography and beyond: Developing technologies for the early detection of breast cancer, National Academy Press, Washington D.C., 2001.

[12] E. A. Sickles, "Nonpalpable, circumscribed, noncalcified, solid breast masses: Likelihood of malignancy based on lesion size and age of patient," Radiology, vol. 192, pp. 439–442, 1994.

[13] E. Fear, P. Meaney, and M. Stuchly, "Microwaves for breast cancer detection" IEEE Potentials, vol. 22, no. 1, pp. 12-18, 2003.

[14] Younis M. Abbosh, "Breast Cancer Diagnosis Using Microwave and Hybrid Imaging Methods", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.5, No.3, June 2014

[15] Y. Xie, B. Guo, L. Xu, J. Li, and P. Stoica, "Multistatic adaptive microwave imaging for early breast cancer detection," IEEE Trans. Biomed. Eng., vol. 53, pp. 1647–57, 2006.

[16] W.C. Khor and M.E. Bialkowski, "Investigations into Cylindrical and Planar Configurations of a Microwave Imaging System for Breast Cancer Detection" Proc. IEEE Antennas and Propagation Society International Symposium, pp. 263-266, Albuquerque, New Mexico, USA, Jul 2006.

[17] A. Abbosh, H.K. Kan and M.E. Bialkowski "Design of compact directive ultra wideband antipodal antenna" Microwave and Optical Techn. Letters. vol. 48, No. 12, pp. 2448-2450, 2006.

[18] A. Abbosh, and S. Crozier, "Hybrid Imaging Method for Early Breast Cancer Detection," Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International , vol., no., pp.1-4, 18-20 Dec. 2008.

[19] W. Khor, H. Wang, M. Bialkowski, A. Abbosh, and N. Seman, "An experimental and theoretical investigation into capabilities of a UWB microwave imaging radar system to detect breast cancer," European Microwave Conference, 2007.

[20] A. Abbosh, M. Bialkowski, and S. Crozier, "Investigations into optimum characteristics for the coupling medium in UWB breast cancer imaging systems", IEEE-Antennas and Propagation Symposium, USA, 2008.

[21] F. Soldovieri, L. Crocco, A. Brancaccio, R. Solimene and R. Persico, "Applications Of Ground Penetrating Radar And Microwave Tomography In Water Monitoring And Management", International Water Technology Journal, IWTJ, Vol. I - Issue 1, 2001.

[22] Lacrosse Sattelite, http://en.wikipedia.org/wiki/Lacrosse_(satellite)

[23] S. W. Harmer, S. E. Cole, N. J. Bowring, N. D. Rezgui, and D. Andrews, "On Body Concealed Weapon Detection Using a Phased Antenna Array", Progress In Electromagnetics Research, Vol. 124, 187-210, 2012.

[24] "Statement tasks and project plan of 973 program: Studies on Theory, System, and Methodology of Sparse Microwave Imaging," 2009.

[25] B. Zhang, H. Jiang, W. Hong, and Y. Wu, "Synthetic aperture radar imaging of sparse targets via compressed sensing," in Synthetic Aperture Radar (EUSAR), 2010 8th European Conference on. VDE, 2010, pp. 1–4.

[26] M. Lazebnik, et al, "A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries," Physics in Medicine and Biology, vol. 52, pp. 2637–2656, 2007.

[27] M. Lazebnik, et al, "A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries," Physics in Medicine and Biology, vol. 52, pp. 6093–6115, 2007.

[28] I. Catapano, F. Soldovieri, A. Affinito, J. Hugenschmidt, "Full 3D Microwave Tomography enhanced GPR surveys: a case study.", Geophysical Research Abstracts, Vol.16, EGU2014-13874, EGU General Assembly 2014.

[29] Daniels DJ (ed.) (2004). Ground Penetrating Radar (2nd ed.). Knoval (Institution of Engineering and Technology). pp. 1–4. ISBN 978-0-86341-360-5.

[30] B. Rushing, "GPR Problems and a GPR Alternative" Blog Post, http://www.berntsen.com/Blog/entryid/26/gpr-problems-and-a-gpr-alternative, Sep 10, 2013.

[31] A. Golnabi, P. Meaney, N. Epstein, K. Paulsen, IEEE Members, "Microwave Imaging for Breast Cancer Detection: Advances in Three-Dimensional Image Reconstruction". NIH Public Access, Conf Proc IEEE Eng, Med, Biol, Soc. Sep 16, 2013.

[32] Q. Fang, "Computational Methods for Microwave Medical Imaging", Doctor of Philosophy thesis, Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire. Dec 2004.

Defining Virtual Networks using SDN

Swtich virtuelizations with "OpenDaylight"

Bojan Petrovski
Ivona Najdenkoska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, R. Macedonia
petrovski.bojan@students.finki.ukim.mk
najdenkoska.ivona@students.finki.ukim.mk

***Abstract*— SDN is a new technology that breaks the bundled hardware and software in current networking equipment in two separate planes. In this paper, we demonstrated that with the use of an open source controller we could program specific behavior in network components, like Virtual Networks, relatively easy in a high level programing language.**

Keywords—SDN; Virtual; Network; Swtich; OpenDaylight;

I. INTRODUCTION

The explosion of mobile devices and content, server virtualization, and advent of cloud services are among the trends driving the networking industry to reexamine traditional network architectures. Many conventional networks are hierarchical, built with tiers of Ethernet switches arranged in a tree structure. This design made sense when client-server computing was dominant, but such a static architecture is ill-suited to the dynamic computing and storage needs of the enterprise data centers, campuses, and carrier environments. [1] In today's technological society the amount of data produced grows continuously, and with that grows the need for an effective way of transporting this data around the world. The Company Cisco estimated the data traffic on the Internet in 2000 at 82 PB per month; in 2014, this estimate was 62475 PB pet month, which suggest an average increase in data traffic of 60% per year. [2] The increase it traffic directly increased the need for better networking equipment that is replaced within approximately 18 months. The huge cost of the new equipment stimulated the large Internet and Telecommunications companies to invest in alternative technologies for networking. One such alternative are SDN's or Software Defined Networks. SDN is a relatively young technology that breaks down the traditional networking technologies in two separate parts Hardware and Centralized Software that enable better control over the Network. In this paper, we look at the current possibilities of SDN and propose a simple way to create Virtual Networks.

II. NETWORK ARCHITECTURE

A. *Traditional Networks*

Most current networks are build using different network component: routers, switches, bridges, hubs, repeaters that are interconnected and enable routing of the data. The routers and switches work with complex, distributed control software that is proprietary and specific to the hardware manufacturer. The software implements and uses different network protocols that have been tested for interoperability. The network is managed by managing the different network components that in turn affects the network as a whole. This individual management is a large contributing factor in the complexity of the networks. Most Networking components today are built with three major segment: Specialized hardware, which receives and forwards packets with great speed, Operating System that works on top of the hardware and high-level applications that use the network protocols.



Fig 1. Hierarchy of a standard Network component.

B. *SDN Netrowks*

SDN or Software defined networking is a relatively modern technology that implements two basic principles:

- The separation of the control layer that makes the decisions how to handle the traffic (Software) and the Data layer that receives and forwards the traffic according to the control decisions.

- Consolidation of the control layer such that one software control program, controls multiple data layer components.

The separation and centralization of the control software enables simple and efficient management of the network and utilization of its hardware potential. This also enables us to alter the behavior of all the hardware components so a switch that supports SND can be configured to behave as a: router, switch, firewall, network address translator or something in between.

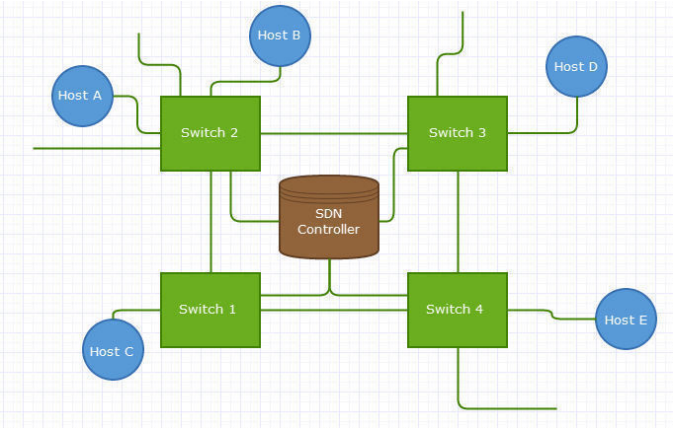


Fig 2. Representation of an SDN Network, with four switches, multiple hosts and a central SDN Controller that is usually a server.

Today there are a couple of SDN Controllers that implement different APIs, however the most widely used is the "OpenFlow" controller, which was adopted early on by some large companies like HP and NEC. OpenFlow is a protocol that allows a server to tell network switches where to send packets. In a conventional network, each switch has proprietary software that tells it what to do. With OpenFlow, the packet-moving decisions are centralized, so that the network can be programmed independently of the individual switches and data center gear. [3] In the past year, many other network equipment manufacturers have added some support for "OpenFlow".

III. "OPENDAYLIGHT" VIRTUAL NETOWK

For the creation of the virtual network we chosed to use the tools provided by the „OpenDaylight" project for two reasons: its large manufactures support[2], and its open source code that is free. This project is a combined effort supported by the Linux foundation in corporation with some of the largest technology companies to facilitate the wide spread adoption of SDN. „OpenDaylight" offers a prepackaged version of Ubuntu that contains a network simulator and all the necessary libraries for SDN software development.[4]

For the purposes of this experiment, we used a single switch network with four hosts. At the beginning of the simulation since there is only hardware with no software the network is not functional and can be checked with a simple ping command.

For the network to be functional, we need to install the appropriate software for the switch, which can be divided in three main components:

- Incoming packets handling
- Packets parsing
- Command messaging

Since our goal was to create multiple virtual networks using a switch, the primary concern is the Incoming packets handling method. This is the most important part of the software and can logically be divided in 4 parts.

- Invalid packet handling is the first part of the method that deals with invalid incoming packets that are immediately dropped.
- Hub functionality is the second part of the method that enables the switch to function as a hub when there is a demand for that.
- MAC learning is the third part of the method that first learns the incoming MAC address, and then checks for the destination MAC address. If it knows the destination, address it sets the appropriate ports, if not it uses the hub functionality to flood the ports.
- Data transmission part is the last part of the method that uses the ports set by the previous part to transmit the packet data.

This part of the software is shown in the pseudo - code below:

Any packet that arrives at a node without a pre-cached rule will result in a callback of *receiveDataPacket*. So overriding that method allows the application to get a copy of the packet.

```
inPkt: incoming data packet
formatted data packet: decoded incoming packet

function recieveDataPacket(RawPacket inPkt){
    if  inPkt is equal with null
            then return inPkt is ignored;
            extract the incoming node (i.e., switch) and node-
            connector (i.e., switchport)
            //NodeConnector incoming_connector =
            //inPkt.getIncomingNodeConnector();
    if the network device is hub
            then flood all ports with the incoming packet inPkt;
    else
            decode the incoming packet using
            the dataPacketService;
            if the formatted data packet isn't instance of
                    Ethernet
            then return inPkt is ignored;
```

```
        learn the MAC address of the switch-port which
        received the formatted packet;
        check if the outcoming_connector knows the
        destination MAC address
        //NodeConnector outcoming_connector =
        knowDestinationMAC(formattedPak);
    if the outcoming_connector doesn't know which is
the dest. MAC address
        then all ports are flooded with the incoming packet;
        else
            if the formatted packet doesn't flow from the
                incoming_connerctor to the
                outcoming_connector;
            then return the packet is ignored;
            use the dataPacketService to send a packet out
            of a node-connector (switchport) - transmit the
            data packet;
return the incoming data packet is consumed;
end;
```

To implement our virtual networks we first define lists of ports that correspond to a certain network. Then we make changes to the third part of the handling method. After the MAC learning has finished but before the ports are flooded, we check if the incoming and outgoing ports are in the same virtual network and if they are the method continues if not the packet is dropped.

Changes to the flooding method need to ensure that if a flood occurs only the ports that are in the same virtual network with the incoming port are flooded. To achieve this we pass the incoming port as a parameter to the flooding method and then iterate over the lists with ports to find the virtual network that needs to be flooded. Once the network is found, we iterate over all the ports and forward the packets.

The changed code from the third part of the handling method is shown in the pseudo - code below:

First we define two lists of node connectors:

```
VN1: List of Node Connectors;
VN2: List of Node Connectors;

if inPkt is equal with null
    then return inPkt is ignored;
else
    if (VN1 contains the incoming_connector and VN2
    contains the outcoming_connector) or (VN1 doesn't
    contain the incoming_connector and VN2 doesn't
    contain the outcoming_connector)
        if the formatted packet doesn't flow from the
        incoming_connerctor to the
        outcoming_connector
            then return the packet is ignored;

        use the dataPacketService to send a packet
        out of a node-connector (switchport);
    return the data packet is consumed;
    else
        return the data packet is ignored;
```

After the software was deployed, we successfully created two virtual networks in the SDN network simulator.

IV. Conclusion

SND is a powerful new technology that enables network administrators to modify and adapt their network software. This small experiment shows that by using open source controllers even inexperienced administrators can configure and manage the software using simple high-level methods and create a switch or other networking component with a specific behavior. In near future it is expected that the SDN will lead to designing a fully programmable networks, that will be adaptive, flexible, agile and fast. That will contribute to the development of many data centers and enterprises, their connection and faster sharing of information.

Acknowledgment

We thank professor Marjan Gushev for the proposed idea for this paper.

References

[1] Software-Defined Networking: The New Norm for Networks; ONF White Paper

[2] Cisco Systems, “Visual Networking Index”.

[3] Open daylight, “Project Members”.

[4] Open Network Connection: "OpenFlow"

[5] SDNHub

Author Index

Abazi Chaushi, Blerta	135
Ackovska, Nevena	154, 168, 172, 225
Aksoski, Stefce	123
Ampov, Mario	330
Andreevski, Aleksandar	131
Armenska, Jasmina	314
Armenski, Goce	67
Arsova, Angela	330
Arsovska, Irina	302
Baran, Virgil	45
Barnaföldi, Gergely Gábor	334
Bashova, Katerina	144
Basnarkov, Lasko	310
Berényi, Dániel	35
Bidikov, Vladislav	154
Bozhinoski, Stevica	306, 314
Bozhinova, Ivana	294, 314
Bozinovski, Adrijan	58, 87
Božinovski, Adrijan	83, 144, 268
Burjakoski, Dejan	123
Bíró, Gábor	334
Chaushi, Agron	135
Chorbev, Ivan	114, 178
Cincaroski, Kosta	168
Dedinec, Aleksandra	40
Delipetrev, Blagoj	63
Delovska, Sandra	294
Dika, Zamir	135
Dimitrievska Ristovska, Vesna	212
Dimitrievski, Filip	67
Dimitrov, Georgi	352
Dimitrova, Tamara	238
Dimitrovski, Ivica	127
Djordjevic, Milan	231
Donchev, Ivaylo	5

Evkovski, Blagoja	282
Filiposka, Sonja	40, 92
Futó, Endre	334
Gashteovski, Andrej	294
Gjorgjevikj, Dejan	110
Gjorgjiev, Kjiro	25
Golubovska, Elena	326
Grchevski, Petar	326
Grozdanov, Vassil	212
Grubi, Arbnor	203
Gushev, Marjan	67, 73, 77
Hasani, Zirije	49, 54
Hrisafov, Kocho	83
Ilijoski, Bojan	119, 330
Ilioski, Ivan	168
Ivanoska, Ilinka	156
Jakimovski, Boro	253
Jandrevska, Monika	302
Janevski, Viktor	318
Jovanov, Dimitar	98
Jovanov, Mile	25, 318, 330
Jovanovik, Milos	131, 273
Jovanovska, Elena	225
Jovanovski, Gorjan	98
Jovanovski, Vladimir	330
Jovkovska, Jelena	58
Kalajdziski, Slobodan	156
Kalemdzhievski, Nikola	273
Kapralov, Stoyan	1
Kirandziska, Vesna	168
Kitanovski, Ivan	127
Kjosevski, Angjel	273
Kon-Popovska, Margita	197
Koteli, Nikola	123, 273
Koteska, Bojana	216
Krasniqi, Venera	21
Kukenska, Valentina	13, 17
Kulakov, Andrea	92, 110, 150

Kuzmanoski, Aleksandar 318

Lameski, Petre 92, 110
Lazova, Verica 310
Loskovska, Suzana 127
Lévai, Péter 35

Madevska Bogdanova, Ana 25, 106
Madjarov, Gjorgji 114
Mahoski, Kliment 102
Manev, Mladen 1
Markoska, Elena 73, 77
Miftar, Sara 330
Mihova, Marija 21, 318, 322
Mijoski, Dimitrij 282, 286, 290
Minev, Petar 13, 17
Mirceva, Georgina 150
Mishev, Anastas 216
Mishev, Kostadin 273
Mishkovski, Igor 114, 182
Mitreski, Kosta 123, 273
Mitrevski, Pece 186
Mitrovska, Ristana 268

Najdenkoska, Ivona 344
Najdova, Natalija 178
Nakevski, Dragan 168
Naumoski, Andreja 123, 257
Nenovski, Blagoj 186
Nicolin, Alexandru 31, 45, 141, 221
Nikolov, Goran 197
Nikolovski, Vlatko 114
Nizetic Kosovic, Ivana 162
Nuhiju, Usamedin 282

Pachovski, Veno 87, 144, 268
Pacovski, Veno 58
Pancevski, Filip 286
Pejov, Ljupco 216
Petkovski, Goran 282, 298
Petrov, Milan 286
Petrovski, Bojan 322, 344
Piperkoski, Mile 182
Popeska, Zaneta 119
Popeska, Zhaneta 102

Radeva, Paolina	13, 17
Raportaru, Mihaela Carina	31, 221
Risteska Stojkoska, Biljana	156, 207
Ristov, Sashko	67
Ristov, Sasko	73, 77
Sazdovski, Vasko	253
Sekuloski, Igor	172
Serafimovski, Duko	286
Serban, Laurentiu	141
Shalamanoski, Jovan	278
Simjanoska, Monika	106
Skokov, Vladimir V.	35
Spasov, Dejan	235
Spasovski, Dimitar	318
Stankov, Emil	25
Stojanov, Riste	114, 131, 243
Stojanovska, Ivana	83, 192, 203
Stojanovski, Martin	330
Stojcevska, Biljana	83, 203
Stojchevska, Biljana	192
Stojkoska, Biljana	278
Stojkovski, Tome	322
Tagasi, Petar	87
Tanaskoska, Eva	247, 294
Tanevska, Ana	306
Tavchioski, Danail	339
Todorova, Emilia	5, 9
Todorovski, Atanas	322
Tosev, Darko	298
Trajanov, Dimitar	40, 92, 131, 243, 273
Trajkova, Slobodanka	243
Trajkovski, Igor	262
Trajkovski, Kristijan	302
Trajkovski, Nenad	192
Trivodaliev, Kire	156
Trojacanec, Katarina	127
Varró, Sándor	35
Velichkovski, Vlado	302
Velinov, Goran	197
Venov, Dimitar	262

Zaeva, Biljana 326
Zafirovski, David 282
Zdravevski, Eftim 92, 110
Zdravkova, Katerina 73

ISBN 978-608-4699-05-7

