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

Proceedings of the 14th International Conference for Informatics and Information Technology

Held at Hotel Bistra, Mavrovo, Macedonia
07-09th April, 2017

Editors:
Aleksandra Popovska Mitrovikj
Biljana Tojtovska
Kire Trivodaliev

ISBN 978-608-4699-07-1

Conference on Informatics and Information Technology 2017
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:
Aleksandra Popovska Mitrovikj
Biljana Tojtovska
Kire Trivodaliev

Technical editing: Ilinka Ivanoska and Vesna Kirandziska
Cover page: Vangel Ajanovski

Total print run:150
Printed in Skopje, Macedonia, 2017

ISBN: **978-608-4699-07-1**

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

004.7:621.39(062)
004(062)

PROCEEDINGS of the 14th Conference on Informatics and Information Technology (14; 2017; Mavrovo) Proceedings of the 14th Conference on Informatics and Information Technology: CIIT 2017, April, 7-9 Mavrovo, Macedonia / editors Aleksandra Popovska Mitrovikj, Biljana Tojtovska and Kire Trivodaliev. - Skopje : Faculty of Computer Science and Engineering, 2017. - 229 стр. : граф. прикази ; 30 см

Библиографија кон трудовите

ISBN 978-608-4699-07-1
1. Popovska Mitrovikj, Aleksandra [уредник] 2. Tojtovska, Biljana [уредник] 3. Trivodaliev, Kire [уредник]

Preface

This volume contains the papers presented at CIIT 2017: the 14th International Conference on Informatics and Information Technologies held on April 07-09, 2017 in Mavrovo, 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.

In the fourteenth edition, the key CIIT conference mission remained to provide an opportunity for young researchers to present their work to a wider research community, but also facilitate multidisciplinary and regional colaboration. Building on the success of the past thirteen conferences, this year conference attracted a large number of submissions resulting in presentations of 48 short and full papers. The conference was comprised of nine sessions. Traditionally, the conference included two student sessions presenting the work of the best undergraduate students, selected on the basis of their submitted projects, prepared during the previous year. The format of the conference allowed the participants to attend most of the talks that covered a diverse spectrum of research areas.

Three distinguished key note lecturers gave plenary sessions covering the different areas of the conference. Prof. Smile Markovski, retired professor at the Faculty of Computer Science and Engineering, UKIM, Skopje, gave a talk on Probabilistic Quasigroups, Vesna Prchkovska, PhD, co-founder & CSO of Mint Labs, Barcelona, Spain, gave a talk on Seeing the brain: How neuroimaging transforms the diagnosis and treatment of patients with brain disorders, and Ognjen eki, PhD, postdoc researcher at the Distributed Systems Group, Institute of Information Systems, TU Wien, gave a talk on Cyber-Human Smart Cities: The Internet of Things, People and Systems. The conference also welcomed a guest student speaker, Ana Tanevska, MSc, PhD student at the Robotics, Brain and Cognitive Sciences Unit, Istituto Italiano di Tecnologia (IIT), Italy, on the topic of Autonomous and cognitive human-robot interaction.

Part of the conference success is owed to the support received from partners and sponsors: Ss. Cyril and Methodius University, Makedonski Telekom, Nextsense, Sorsix, Software4Insurance, Macedonian Winemakers and Pivara Skopje.

All in all, this year the CIIT conference has outgrown the role of being an excellent opportunity for young researchers to present their scientific growth, to a more premier role, that is to bring researchers together for establishing collaborative links between disciplines, for testing the ground for innovative ideas and for engaging the wider academic community.

September, 2017
Skopje

Aleksandra Popovska Mitrovikj
Biljana Tojtovska
Kire Trivodaliev

Organization

Conference chairs

Aleksandra Popovska Mitrovikj	Assistant professor - Faculty of Computer Science and Engineering, Skopje
Biljana Tojtovska	Assistant professor - Faculty of Computer Science and Engineering, Skopje
Kire Trivodaliev	Assistant professor - Faculty of Computer Science and Engineering, Skopje

Organizing Committee

Ilinka Ivanoska	Teaching/research assistant - Faculty of Computer Science and Engineering, Skopje
Pance Ribarski	Assistant - Faculty of Computer Science and Engineering, Skopje
Petre Lameski	Assistant - Faculty of Computer Science and Engineering, Skopje
Vesna Dimitrievska Ristevska	Assistant professor - Faculty of Computer Science and Engineering, Skopje
Vesna Kirandziska	Assistant doctorand - Faculty of Computer Science and Engineering, Skopje

Program Committee

Adrijan Boinovski	School of Computer Science and Information Technology, University American College Skopje, Macedonia
Aleksandar Shurbevski	Kyoto University, Kyoto, Japan
Aleksandra Popovska-Mitrovikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Aleksandra Mileva	Faculty of Computer Science, University Goce Delcev, Macedonia
Alexandru Nicolin	Horia Hulubei National Institute for Physics and Nuclear Engineering and University of Bucharest, Faculty of Physics, Centre for Theoretical Physics, Romania
Ana Madevska-Bogdanova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia

Andreja Naumoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Antun Balaz	Institute of Physics, University of Belgrade, Serbia
Biljana Tojtovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Biljana Stojkoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Bojan Marinkovic	Mathematical Institute of the Serbian Academy of Sciences and Arts, Serbia
Danco Davcev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Dejan Gjorgjevikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Dimitar Trajanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Eugenia Stoimenova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Galina Bogdanova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Georgina Mirceva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Gjorgji Madjarov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Goce Armenski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Goce Ristanoski	Data61, Commonwealth Scientific and Industrial Research Organisation, Australia
Haris Gavranovic	Faculty of Engineering and Sciences, International University of Sarajevo, Bosnia and Herzegovina
Hristijan Gjoreski	University of Sussex, United Kingdom
Hristina Mihajlovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Hristo Kostadinov	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Igor Mishkovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Ivan Chorbev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Ivaylo Donchev	St Cyril and St Methodius University of Veliko Turnovo, Bulgaria
Ivica Dimitrovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Katerina Zdravkova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Kire Trivodaliev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia

Konstantin Delchev	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Lasko Basnarkov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Magdalena Kostoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Marcin Paprzycki	Systems Research Institute, Polish Academy of Sciences, Poland
Margita Kon-Popovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Marija Mihova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Marija Slavkovik	University of Bergen, Norway
Marjan Gusev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Mile Jovanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Milos Jovanovik	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Miroslav Mirchev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Natasha Stojkovikj	Faculty of Computer Science, University Goce Delcev, Macedonia
Natasha Ilievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Nevena Ackovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Ognjen Scekic	Distributed Systems Group, Vienna University of Technology, Austria
Sahra Sedigh-Sarvestani	Missouri University of Science and Technology, USA
Sasho Gramatikov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Sasko Ristov	University of Innsbruck, Austria
Simona Samardjiska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Slavcho Shtrakov	Neofit Rilsky South-West University, Bulgaria
Slobodan Kalajdziski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Smile Markovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Smilka Janeska-Sarkanjac	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Sonja Gievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Stela Zhelezova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Stoyan Kapralov	Technical University - Gabrovo, Bulgaria
Suzana Loshkovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Tsonka Baicheva	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Vangel Ajanovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Vasil Grozdanov	Neofit Rilsky South-West University, Bulgaria
Veno Pachovski	School of Computer Science and Information Technology, University American College Skopje, Macedonia
Verica Bakeva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Vesna Dimitrova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Vesna Dimitrievska Ristovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Vesna Prckovska	MintLabs - University of Barcelona, Spain
Vladimir Trajkovikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Yuri Borissov	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Zaneta Popeska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Macedonia
Zlatko Varbanov	St Cyril and St Methodius University of Veliko Turnovo, Bulgaria

Table of Contents

Education and eLearning

Multimedia and Signal Processing

Theoretical Foundations of Informatics and Applied Mathematics

Parallel Processing, Cloud Computing and Computer Networks

Assisting children with Down syndrome in ICT era

Sofija Kitanovska
sofija.kitanovska@gmail.com

Nevena Ackovska
nevena.ackovska@finki.ukim.mk

Faculty of Computer Science and Engineering,
Skopje, Macedonia

Abstract— **The purpose of this study is to show that technology has a big influence on children wih disabilities and can improve their quality of life. This research takes into consideration children with Down syndrome. Each child's characteristics is reviewed, which enables us to show how a computer application can improve their knowledge. The paper elaborates on each part of the application named "Shareni chorapi", and shows its importance for the child's development and improvement of their intellectual and motor skills.**

Keywords – ICT; Assistive technologies; Down syndrome; Children

I. INTRODUCTION

There are many organizations dedicated to people with special needs. However, having a computer application that looks after their needs is an excellent way for people with special needs to improve their abilities. This is especially true for children with special needs, since they can learn to cope with their environment much better in their young age. In this paper, we provide proof that technology can significantly improve the development, education and integration process for children who have Down syndrome.

The paper provides an overview of basic information about the disease, typical characteristics of a child who has Down syndrome. Afterwards we provide the specific goals one should have in mind when creating software for children that have Down syndrome, in order to help and improve their intellectual and motor skills. At the end, we will explain the application "Shareni chorapi" created by Sofija Kitanovska and what is the feedback of using it.

II. DOWN SYNDROME AND ITS CHARACTERISTICS

Down syndrome is a condition to a newborn, which occurs because of the presence of an extra chromosome in the nucleus [1]. Each person who has Down syndrome is unique and possesses unique characteristics, but there are some specific characteristics that represent the disease. Children with Down syndrome are usually quiet, have low muscle tone, flat facial features, small nose, upward slant to the eyes, small, abnormally shaped ears. Apart of all that, they are capable of any activity that other children can perform. These children learn best from what they see, so involving them in everyday activities may be a good progress for them to properly develop and understand the environment [2], [3].

During their early years, young children are developing a sense for imagination and creativity, they are curious and eager to learn everything in their environment. Considering that the early years are the best years for a child to evolve, grow and learn from the environment, and having in mind the possible delay in the child's development, it will further be discussed how a computer application may help children under the age of 6 to develop basic knowledge or to improve their already acquired knowledge [4].

III. MARKET RESEARCH : SIMILAR APPLICATIONS

The rapid growth of technology plays a major role in helping people with disabilities, especially children. For people with Down syndrome to be fully included in the society, they must have the opportunity to use and extract knowledge through technology. Therefore, technology can be used as an educational tool, which may encourage the development of the child's reading, writing, language, motor, memory and social skills [5], [6].

There are many applications for that purpose, that facilitate learning for children with Down syndrome. It allows children to gain knowledge, through visual presentation, motivational graphics, words also represented as spoken, because in this way, the children strenghten their capability of word recognition. The most used applications are: ProloQuo2Go and TalkTablet – dedicated to people who have speaking difficulty, ChoiceWorks helps children perform their daily routines, to understand and control their emotions and increase their patience and Injini: Child Development Game Suite, dedicated to children with cognitive, language and motor delays [7].

IV. USED TECHNOLOGIES

The application "Shareni chorapi" was built on two core technologies. To achieve the purpose and build an application available for both mobile and web version, the first technology used was Apache Cordova.

Apache Cordova represents a new way of developing applications. It is used for creating hybrid applications, with reusing the code from web to native applications, available for mobile devices [8]. The second technology aimed at testing the application was PHP through the XAMPP server. The XAMPP(X-Cross Platform, A-Apache, M-MySQL, P-PHP, P-Perl) server provides everything an application could be build upon. X-Cross Platform means that it can be used on different platforms, Apache represents a web server, MySQL is the database used for retreiving data, PHP is widespread server-side programming language on the web and Perl is dynamic programming language. Fig. 1 features an architecture

diagram that outlies how these technologies are combined.

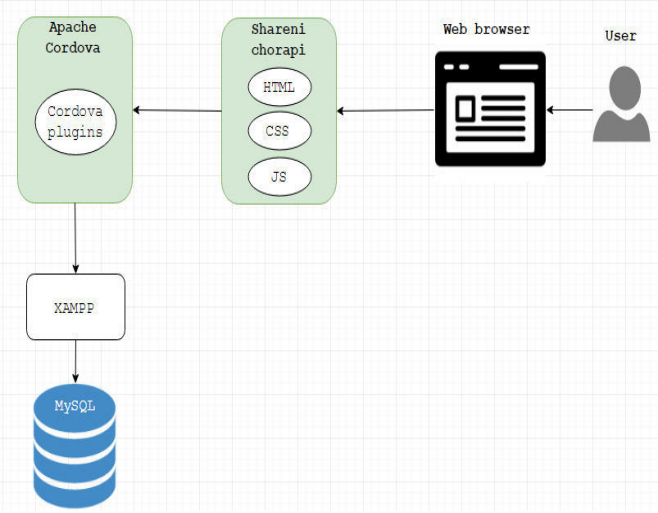


Figure1. Architecture diagram

V. THE APPLICATION

The functionality and features of the already created application and how it will help the children with Down syndrome to evolve will be explained in this section.

When building an application for children with Down syndrome, the most important things to take into consideration are how to develop motor skills, self-help skills and academic skills [9].

It is shown that using a computer is motivating for the children with diffucilties in learning, in a way that they are more focused on the program they are using, especially if the program consists of images, sounds, animations, interactive tools and music [10].

The main focus on creating applications for children with Down syndrome is including audio and visual representation on anything that appears on the screen.

The application consists of two important sections. The first section represents a new way of communication between the children with Down syndrome and the people in their environment. The main focus in this section is learning and spelling the words, with the final outcome of generating a sentence. Every word on the display is accompanied with a corresponding image which best explains the shown word.

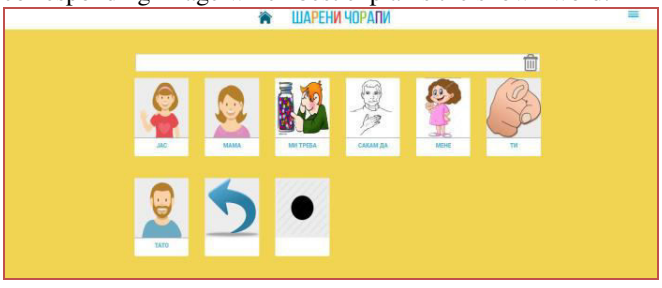


Figure2. Communication section helps the child with Down syndrome communicate with its environment

The second section represents a novel way of education, which helps children test and improve their current knowledge, accompanied through games, each one of them focused on a different topic, whether that is color, figures, animals, fruit, as well as a memory game.

Apart of these sections, within the application there exists an administrative panel. This panel is meant for the parent or caregiver to control the settings in the communication section such as:

- Choice of voice that will be used when pronouncing the words
- Image or description change
- Adding a new image

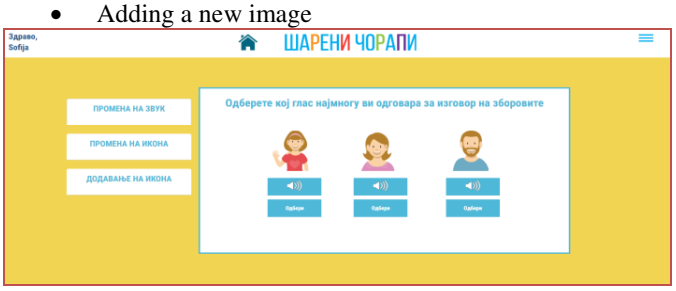


Figure3 . Administrative panel

VI. RESULTS AND DISCUSSION

In this part of the paper, a brief overview about the development process will be elaborated. The process was extensive and it included cooperation with experts in this field. The results are very positive, since feedback was overly positive and the experts and the children with Down syndrome embraced the application and found it useful.

The testing phase was done on real subjects. The application was reviewed by the organization “Open the windows”, where we got approval for future work, especially because an application of this type, also written in macedonian language does not exist. The application was assessed by a group of children with Down syndrome, who found the application very interesting and amusing.

This application can also be used by people who do not have Down syndrome, but suffer by some of the symptoms that are the focus of this application. Another key strength of the application is the possibility to be used by different users who are unable to talk, as a way of expressing their thoughts.

The application represents a new way for children to express themselves, to build or excel their skills.

In the future, we hope to improve the vocabulary by adding more words, classify the words into categories for better navigation, and add more games in the second section which would help the children to learn in a fun way.

REFERENCES

[1] http://www.encyclopedia.com/doc/1G2-3451600527.html
[2] http://www.ndss.org/Down-Syndrome/What-Is-Down-Syndrome/
[3] Down syndrome 21+ no. 2/2016 Skopje, Macedonia
[4]http://www.inclusive.co.uk/articles/downs-syndrome-computers-and-the-curriculum-a286

[5] https://downsyndrome.ie/computers-technology/

[6] https://www.down-syndrome.org/information/education/technology/

[7]https://www.care.com/c/stories/6621/22-best-mobile-apps-for-kids-with-special-nee/

[8] https://msopentech.com/opentech-projects/apache-cordova/#

[9]https://www.time4learning.com/teaching_your_down_syndrome_child.shtml

[10] https://www.down-syndrome.org/information/education/technology/

Different Statistical Methods in Predicting Student Course Enrollment

Ljupcho Rechkoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Rugjer Boshkovikj 16, PO Box 393, 1000 Skopje, Macedonia
reckoskiljupco1@gmail.com

Zhanko Mitreski
University of Information Science and Technology
St. Paul the Apostle
Partizanska bb, 6000 Ohrid, Macedonia
zankoohrid@gmail.com

***Abstract*— Estimating the number of students at each university course is a difficult problem, but somehow it is obvious that one student takes the course acording to his or hers characteristics and capabilities. There are groups of undergraduates whose course enrollment rates are significantly different than the rest of the university population. In this study we analayze the historical enrollment data so we can make some predicion about the future enrollment. The prediction models will be analysed on real data from the last 5 years at FCSE.**

Keywords*— *Course enrollment; Statistical methods; Data mining; Prediction methods.

I. INTRODUCTION

Predicting university course enrollment is an important feature in managing faculty resources, such as determining the number of classes for each course, allocating seats for multiple course sections or finding instructors for the classes. In order to make prediction more accurate, we may use historical data about popularity of a specific course, similar interest among students, transition rate between semesters and number of finished courses. Using this information we can predict the chosen courses by a student for any given semester, based on his or her course history.

Students have different understandings about certain lectures and courses, different desires and ability to overcome the materials studied in those courses. Some students are good at applying practical actions, others are good mathematicians or programmers, and some are good at theory and research activities. Based on these factors, they choose the elective courses during the academic years. It is expected that students with similar understandings would pick similar set of courses. Furthermore, students who attend same courses, tend to be friendlier to each other, exchange experiences and share information about other courses that they previously attended. Considering these activities, we can expect that if student A and student B, both attend some elective course M, and if student A has previously attended a course N, then student B is likely to choose course N on the next course election period.

In this paper, we want to test the correlation between subjects, i.e. are there conditions that influence the course selection process, or the same are randomly selected by the students. We want to test this correlation by analyzing the student enrollment set. The analysis is based on information obtained from students that are studying at FCSE.

Having awareness of the relation between the students and courses, besides predicting the course enrollment, we can improve the student management in the future, i.e. change elective courses based on number of students interested in such courses, provide rules and order for certain courses or design guidelines for studying based on the set of students in the current academic year.

II. RELATED WORK

Most published work is about enrollment modeling centers on the entire student body, i.e. they analyzed the number of students which will continue their education or transfer from one to another university [6]. There is little published work on course enrollment.

Kraft and Jarvis [1] in their article “An Adaptive Model for Predicting Course enrollment” explain prediction model for the Calculus course within the Department of Mathematical Science. Once the data was collected, they proceeded with building the prediction model. In their first step - identification of significant factors, they identified characteristics of students whose course enrollment rates are different than the rest of the university population by splitting the students into groups: new students, continuing students, and returning students. After that they use historical enrollment rates in which they deduced the minimum, maximum and average enrollment rates for each group in the course in the period of three years in order to predict enrollment for the next semester by recording the conditional probabilities of students in each group. Finally they proved that the model in predicting the total number of students is accurate and robust because it can be extended to predict any course enrollment.

Johnson and Strohkorb [2] used logistic regression, the model that is used in classification or prediction of a binary outcome based on a set of predictors. As an input data they used all the courses taken by each student from the Fall semester 2002 to the Spring 2014 semester. The class in the model is either the student takes a given course in a particular semester or does not take the course. They predict

enrollment for each course individually. After applying the logistic model to the training data they used test data to fit the model.

Gerwing and Balachandran [3] proposed three variable work models for predicting course enrollments: the work model, the eligible-work model accounting for prerequisites, and the eligible-work model with program requirements. Conditional probabilities are used for each group to predict the total number of students who will enroll in the course. The work model uses the conditional probability that a student will take a course, given that he or she has not already done that, to predict how many students will enroll in that course. The eligible work model uses the same logic from work model but takes into consideration prerequisite courses. The eligible-work model with program requirements categorizes eligible and ineligible students based on the work they have completed about degree project works.

Some analyses about recommendation in the course enrollment process for the students at FCSE are made in Ajanovski articles [4, 5]. He used recommended system, the system that seeks to predict rating that a user would give to an item. Using the historical data of the students, Ajanovski gives each student personalized course recommendation. He built a system called Integrated Student Information System in a form of a web page with simple dashboards, showing a map of all the previously enrolled courses. The system also shows other information about courses such as which courses are critical, the ones with a significant amount of failed students and which courses are most popular.

The motivation for these papers is the fact that the students are overestimating themselves and enroll more elective courses than they can attend to, resulting with failure in completing the mandatory courses. The set of students is analyzed based on examination results, number of times of unsuccessful course enrollment, and previously enrolled courses. Based on this, the student is provided a set of courses that he or she is likely to choose and pass in the next semester.

III. Data And Methods For Analysis

Having the historical data of courses that students have chosen in the past will give us certain results about the courses that they might choose in the future. Firstly, we have analyzed the courses and their relationships in order to make a connections between different courses. We have used some statistical and data mining techniques, such as hierarchical clustering, decision trees and Naïve Bayes in order to develop the relationship between the courses and we have compared the obtained results with different approaches.

For that reason, in this part we give a description of the data collection process and statistical methods that we use for our study.

A. *Input data*

In our case study, we used data collected from Moodle software learning management system, which contained all the information needed for analysis: name and surname of the students, year of enrollment, and enrolled courses by study year and semester (winter or summer). Before we started with processing the data by determining course relationships and building appropriate classification model, we first removed the data that is not needed in the analysis, such as professors' courses and mandatory courses. We have processed the data with WEKA, a software tool which provides a number of built in implementations of data mining algorithms and R, free software environment for statistical computing and graphics. We have decided to use the following algorithms: Hierarchical clustering, Decision Trees, Naïve Bayes, Logistic regression and Apriori association rule algorithm. After finding a learning function, $y = f(x)$ we applied it, in order to predict the associated class label y of a given tuple X. In our case, the class label is the course of interests and X tuple contains the courses that are chosen from a particular student.

The data that is used as input for analysis is in the following form:

Table 1. Attributes – courses, Rows – Students; Here 1 means that the student has taken that subject

Student	Course 1	Course 2	Course 3	Course 4
Student-1	1	0	1	0/1
Student-2	0	1	0	0/1
Student-n	0/1	0/1	0/1	0/1

B. *Hierarchical clustering*

Clustering is a method for finding groups of objects, such that a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. There are more types of clustering, hierarchical clustering, partitioned, exclusive, fuzzy and so on. Because we are interested to know which courses are enrolled together by the students we consider hierarchical clustering – a set of nested clusters that are organized as a tree. The method uses a measure of dissimilarity between sets of observations. This is achieved by use of an appropriate measure metric which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Some commonly used metrics for hierarchical clustering are: Euclidean distance, Manhattan distance, Maximum distance, Jaccard distance and others. In our study we used R software and implemented Jaccard distance for hierarchical clustering. At the end the result is shown in a cluster dendrogram which can be interpreted easily.

C. *C4.5 algorithm*

The C4.5 algorithm is a decision tree algorithm and it is the most used algorithm for classification and prediction. Decision tree builds classification models in the form of a tree structure. It divides a dataset into subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The root node in a tree corresponds to the best predictor. We used this algorithm to predict whether or not the student will take given course in a particular semester.

As a class attribute any course of interests can be used. At the end the result is interpreted as if then rules, i.e. if the student has chosen courses A and B and C then the student will take course D. The advantage of decision trees is that they can handle both, categorical and numerical data. The core algorithm for building decision trees called ID3 employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy* and *Information Gain* to construct a decision tree.

$$Entropy(t) = -\sum_{j} p(j|t)\log_2 p(j|t) \quad (1)$$

$$GAIN_{split} = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)) \quad (2)$$

After finding the information gain for all of the attributes, it is necessary to find the optimum attribute that produces the maximum GAIN = maximum GAIN of all attributes. After that, the optimum attribute is inserted into the node of the decision tree. Since it is the first node, it is the root node of the decision tree. Once the optimum attribute is obtained, the data can be split accordingly to the optimum attribute.

D. Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. Bayes classifier is based on the assumption that information about classes in the form of prior probabilities and distributions of patterns in the class are known. It then uses posterior probabilities to assign the class label to a test pattern. Finally, a pattern is assigned the label of the class that has the maximum posterior probability.

$$P(A_1, A_2, \dots, A_n|C) = P(A_1|C_j)P(A_2|C_j) \dots P(A_n|C_j) \quad (3)$$

New point is classified to C_j if following equation is maximal:

$$P(C_j)P(A_1|C_j)P(A_2|C_j) \dots P(A_n|C_j) \quad (4)$$

We use this algorithm to classify whether or not the student will take given course based on courses that the student has taken in the past.

E. Logistic regression

Logistic regression predicts the probability of an outcome that can only have two values, 0 and 1, which represent outcomes such as pass/fail, win/lose or healthy/sick. The algorithm can be used to estimate the probability of a student to take specific course, based on the student previously enrolled courses. In contrast with linear regression, where we assume the data is normally distributed, in logistic regression we first assume that the data follows a binomial distribution. The prediction for a given class is made by substituting the coefficient in the following logistic formula

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n)}} \quad (5)$$

F. Apriori algorithm

All of the algorithms mentioned above are used to build classification or prediction models. In our study we used also an association rule algorithm – Apriori. We use this algorithm to find all the rules that will give information about which courses the student will take in the next semesters based on the previously course enrollments. For example, the students who take course A also tends to take courses B or C at the same time.
Ex: A => B or C [support = 6%, confidence = 70%].
The support of an association pattern is the percentage of task-relevant data transactions for which the pattern is true, $(A => B)$

$$support(A => B) = P(AUB) \quad (6)$$

$$confidence(A => B) = P(A|B) \quad (7)$$

IV. Results

A. Hierarchical clustering

After preparing the input data, we firstly used R software for applying hierarchical clustering because for this kind of analysis it is the best way to find which courses are enrolled together by the students. Appendix A shows the whole dendrogram using binary method for calculating the distance of hierarchical clustering. The clusters in the dendrogram are divided with rectangles. Fig. 1 shows one cluster from the dendrogram. It is reasonable that the courses that are associated with network technologies are in the same clusters.

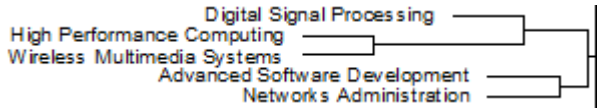


Fig.1. One cluster from dendrogram

Also analyzing the dendrogram, one can conclude that subjects that are connected with web programming, web technologies, mobile programming technologies are in same clusters.

B. Decision trees

We used WEKA software for applying algorithms for classification, prediction and association rules in order to extract some useful information from the data. As example in our analysis for classification, we used the course Next

Generation Network as class attribute. The decision tree is represented below.

```
Digitalization and representation = 0
|   Network software = 0
|   |   Virtual societies = 0
|   |   |   Calculus 3 = 0
|   |   |   |   Introduction to recognizing shapes = 0: 1 (58.0/12.0)
|   |   |   |   Introduction to recognizing shapes = 1
|   |   |   |   |   Natural languages processing = 0: 1 (2.0)
|   |   |   |   |   Natural languages processing = 1: 0 (3.0)
|   |   |   Calculus 3 = 1: 0 (5.0/1.0)
|   |   Virtual societies = 1
|   |   |   Natural languages processing = 0: 1 (3.0/1.0)
|   |   |   Natural languages processing = 1: 0 (5.0)
|   Network software = 1
|   |   Network OS = 0
|   |   |   Contemporary methods of network analysis = 0: 0 (44.0/11.0)
|   |   |   Contemporary methods of network analysis = 1
|   |   |   |   Parallel and Distributing processing = 0: 1 (7.0/2.0)
|   |   |   |   Parallel and Distributing processing = 1: 0 (3.0/1.0)
|   |   Network OS = 1: 1 (3.0)
Digitalization and representation = 1: 0 (11.0)
```

Fig. 2. C4.5 decision tree

The decision tree is read with if / else then rule statements. An example for a rule from Fig. 3: "If the student has chosen the course Digitalization and Representation then she/he will not choose Next Generation Network. But if the student has not chosen Digitalization and representation and have chosen Network Software and Network Operating Systems then she/he will also choose Next Generation Network. The number in parentheses shows how many samples reached that leaf versus how many were incorrectly classified. Any course can be treated as a class attribute in order to predict the enrollment for that course.

In order to see how accurate classification model from training set is, we used WEKA detailed accuracy statistics. The following table gives details for decision tree classification algorithm, but the same can be achieved for other algorithms.

Table 2. Detailed Accuracy by class for decision tree algorithm

TP rate	FP rate	Precision	Recall	F-Meas.	Class
0.740	0.310	0.711	0.740	0.725	0
0.690	0.260	0.721	0.690	0.705	1

From Table 2, TP rates (number of rows predicted positive that are actually positive) for both classes are higher than FP rates (number of examples predicted positive that are actually negative) and also the precision is higher than 60 % which lead to a conclusion for an accurate model.

C. *Naïve Bayes*

The same input data can be used to classify the class attribute course with Naïve Bayes tool in WEKA. The software calculates the number of all the attributes for each class and then classifies the tuple for which the posterior probability is maximal. We used the course Next Generation Network as class attribute.

Fig. 5 represents the first 21 instances of the training set and each instance holds some given probability value. The instance 21 is marked with error + and is classified as incorrectly identified instance.

inst#	actual	predicted	error	prediction
1,	1:0,	1:0		0.954
2,	2:1,	2:1		0.797
3,	1:0,	1:0		0.944
4,	1:0,	1:0		0.96
5,	1:0,	1:0		0.989
6,	1:0,	1:0		0.989
7,	1:0,	1:0		0.964
8,	1:0,	1:0		0.989
9,	1:0,	1:0		0.534
10,	1:0,	1:0		0.55
11,	1:0,	1:0		0.989
12,	2:1,	2:1		0.797
13,	1:0,	1:0		0.657
14,	1:0,	1:0		0.71
15,	1:0,	1:0		0.657
16,	1:0,	1:0		0.528
17,	1:0,	1:0		0.783
18,	2:1,	2:1		0.838
19,	2:1,	2:1		0.695
20,	2:1,	2:1		0.838
21,	1:0,	2:1	+	0.592

Fig. 3. Probability of first 21 instances and error of classification

D. *Apriori algorithm*

Finally, we show how WEKA outputs the rules when applying Apriori algorithm. We want to find the association rules that have minSupport = 50% and confidance = 50 %. After the algorithm is finished, the list of the first best rules found is the following:

1. Machine Learning = 1 32 ==> Network Software=1 31

2. Modern Methods for Network Analysis =1 43 ==> Next-gen Networks = 1 29

3. Network Software = 1 57 ==> Machine Learning=1 31

4. Introduction to Telecommunications=1 28 ==> Next-gen Networks = 1 15

5. Natural Language Processing=1 38 ==> Next-gen Networks = 1 20

6. E-Business=0 Calculus 3=0 Advanced Programming = 0 123 ==> Digitalization and ePresentation =0 123

7. Digitalization and ePresentation = 0 Advanced Programming = 0 130 ==> E-Business=0 129

8. Digitalization and ePresentation=0 E-Business=0 130 ==> Advanced Programming=0 129

9. Digitalization and ePresentation=0 Calculus 3=0 Advanced Programming=0 124 ==> e-Business=0 123

10. Digitalization and ePresentation=0 E-Business=0 Calculus 3=0 124 ==> Advanced Programming = 0 123

One can easily interpret these rules. The first rule is read as following: *If someone chose Machine Learning then he/she will choose Network Software*. The tenth rule: *Those who did not choose Digitalization and Representation, E-Business and Calculus 3 then they will not choose Advance programming.*

V. CONCLUSION AND FUTURE WORK

With all of the analysis, we found some relationship between the courses. We used different algorithms to extract useful information about overall data set and we used it as a training data set. Our future work would include division of the data set into training and test set and then fitting the model when applying data from test set. The idea is to populate the test set with data from students that have enrolled three or four years before the analysis, and then process the data with the mentioned algorithms**.** For example for a given student with appropriate courses, using the decision tree algorithm we would apply the algorithm for each subject that the student has not taken and then see which courses will lead to leafs with values 1(the courses that are connected with class attribute course). The process can be repeated with all the algorithms and at the end the idea is to compare which algorithm gives the most efficient results.

The main idea of our future work is not to make a prediction to only one course, but to select a set of elective courses that are most correlated to each other in order for a student to achieve 40 ECTS.

Future work regarding this study would also include creating our own classifier. To get better results from our data we could include not only the courses but also the professors that teach associated course in particular semester and the grades of the students achieved in that semester. With all that information it is possible to build a model that will give information on how grades are associated with course enrollment and how professors influence students to enroll that course.

To conclude, statistical methods are useful when performing data analysis when deducing rules and classifying data. All algorithms in the study show a good percentage of correctly classified instances. Both, R and WEKA software are a data mining standard tools for performing such data analysis. They provide to the user a readable interface and a great comparison bases when developing new methods for classifying data and data processing techniques in general.

REFERENCES

[1] C. R. Kraft and J. P. Jarvis, An adaptive model for predicting course enrollment

[2] B. Johnson, S. Strohkorb, Predicting course enrollment, Frankly Speaking: Volume 5, Issue 7, 20014

[3] K. R. Balachandran, D. Gerwin, Variable-Work models for predicting course enrollments. Operations Research 21(3), 1973, 823-834.

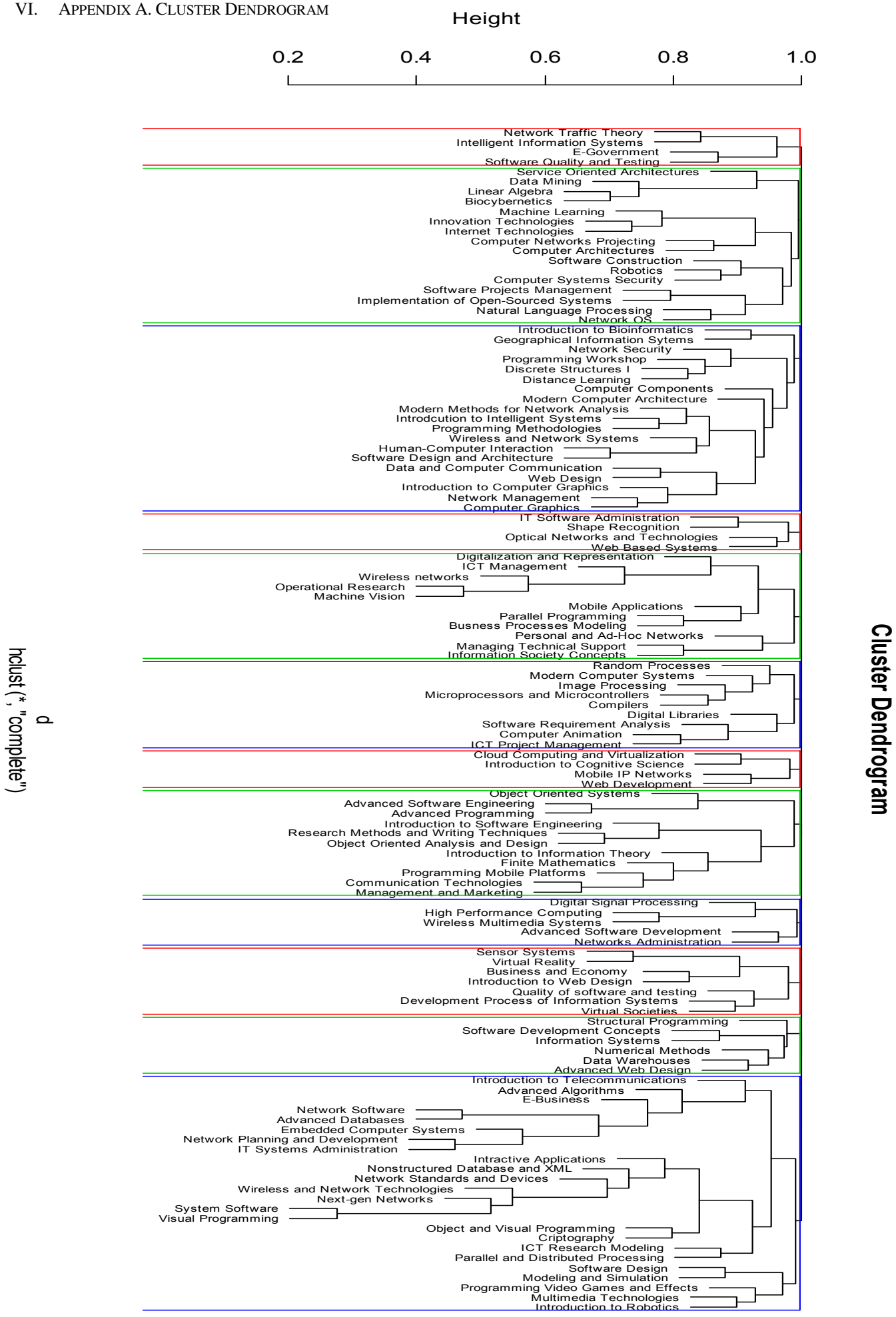
[4] V. Ajanovski, Context aware recommendations in the course enrolment process based on curriculum guidelines, IADIS International Conference e-Learning 2013, 419-422

[5] V. Ajanovski, Integration of a course enrolment and class timetable scheduling in a student information system, International Journal of Database Management Systems; Chennai5.1 (Feb 2013): 85-95.

[6] A. Nandeshwar¸ S. Chaudhari, Enrollment prediction models using data mining, April 2009

[7] Dr. S. Sayad, An introduction to Data Mining, Data Mining map http://www.saedsayad.com/data_mining_map.htm

VI. APPENDIX A. CLUSTER DENDROGRAM



Information System for Mapping the Coverage of Reference Curriculum Guidelines in the Teaching Curricula of a Higher-Education Institution

Vangel V. Ajanovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
E-mail: vangel.ajanovski@finki.ukim.mk

Web: https://ajanovski.info

***Abstract*—This paper presents the results of a project of the introduction of a specialized Information System for mapping the coverage of knowledge topics as given in reference curriculum guidelines established by renowned world organizations (such as ACM and IEEE in the area of computing), as a step to enable the establishment of a qualitative, and not just quantitative process to reforming the existing curricula and introduction of new knowledge areas or topics in order to foster improvement in teaching capacities and capabilities.**

***Keywords*–Curriculum management; curriculum analysis; curriculum mapping; reference guidelines; information systems.**

I. Introduction

During the time I was working at the Faculty of Computer Science and Engineering, and at the Faculty of Natural Science and Mathematics in the past, both at the S-s Cyril and Methodius University in Skopje, Macedonia, I have witnessed several iterations of curriculum design and development processes. These institutions have focused on nurturing computer science since their beginning, but global trends in the past few decades were such that we have created separate study programs for software engineering, information technology, information systems and computer engineering, and respective reconstructions of the curricula. Beside this, since computing overall is a field in continuous flux, to have the benefits of the constant redefinition of technologies, new theories and algorithms, we have also undergone many curriculum redesign efforts, of which I have witnessed the ones that finished with the accredited study programs in 2001, 2005, 2009, 2011, and 2013.

Being in position of constant change, some provoked by new trends in the world, a quite a few mandated by legislative changes it became obvious that a structured solution is required that will enable to keep a record at least, of all the final results in all these curricula reconstructions and also to enable the institution to keep track of student progress through one or another version of a curriculum. According to Macedonian laws a student has the right to remain within the original study plans and programs, that were active at the moment of her admission for a minimum of 8 years. The effect of this is that students admitted in 2006/07 have the right to keep studying according to their study plan (which is the one that was accredited in 2005), until 2014. Although processes of accreditation have occurred in 2009, 2011 and 2013, bringing changes in the curricula, such students have the right to continue studying according to the plans from 2005. The effect is that at the same time we have to track each student according to their original program and if the student decides to switch to one of the newly accredited programs, we have to allow this change and enable the student to keep the past work in terms of credits and grades.

Why would a student want to study 8 years or more, for a 3 or 4-year undergraduate program? Well, our country is plagued with unemployment (30% of the population), and our system is such that allows the students to grasp opportunities as they come, and be able to continue studying with a decreased load (less than 30 ECTS per semester) or retake failed courses several times.

One might look upon all this as just a special case, that is not applicable to any other part of the world and hence not of wider interest. Since unemployment is a worldwide problem and the necessity to study part-time, online, as part of evening schools, or generally in programs with a decreased number of regular classes it is becoming obvious that this discussion is already or can become applicable to any university in the world. Since, these special circumstances ask for a solution that would be a flexible one, a solution that allows many scenarios. And in fact, such a solution has evolved as part of a long-term research project that originated from the 1990s and is still developed. The solution – ISIS (Integrated Study Information System) presented for the first time to a wider

This paper presents research and development that was part of the project ISISng [1] – The project was partially financed by the Faculty of Computer Science and Engineering.

audience in 2009, [2] was built to enable a university to have a record of all the institutions, study plans, programs and curricula from the past and also ones that are active at the moment, enabled to track the students enrollments in each semester (or more flexibly terms if there are more than two semesters) even for students that have been enrolled to two or more study programs at the same time, enabled to differentiate if a student has enrolled a course according to a specific study program and enabled to have all of this differ either at the level of the institution or at the level of a study-program within that institution.

The ISISng further evolved in several directions with some additional capabilities (see [3], [4]), but something that it never had – was the possibility to have a structured record on the specifics of what should be taught as part of curricula. This information was only recorded as a full-text document as part of the external documentation in the accreditation process and as such, it did not give many opportunities to be used for an in-depth curriculum analysis.

The design of a curriculum should involve a systematic process passing through several phases. First, a set of learning goals and objectives should be identified, then the core knowledge topics should be decided on, and subsequently learning activities and assessment activities should be identified and developed [5]. The whole of the process should be streamlined in such a way, that the learner is guided through the educational experience by the curriculum design itself.

In this paper, the results of the effort towards an extension of the ISIS system are presented, with an implementation of a structured process in which the curricula are mapped to the knowledge unit topics proposed as part of internationally known curriculum guidelines (such are the ones produced by ACM, IEEE, and AIS). This is a necessity that allows a more formal analysis of the existing curricula and their success, and should be considered as a prerequisite of a more formal process of curriculum design.

II. Related work

Curriculum mapping has been used and discussed by many authors for at least 3 decades. This section will point-out some references in this area that I consider to be of interest and relevant as comparison and context to the system that I propose.

The notion of curriculum mapping as part of a more formal process for review and analysis of the actual curricula being used at a certain institution has been extensively elaborated by Hayes Jacobs, as devised to be used at K-12 schools, but similar process can be performed also at higher-education institutions [6]. As discussed by Hayes Jacobs the mapping is a procedure for finding out what teachers actually do through the course of the calendar school year and putting it on a map in the form of a calendar. In contrast to this approach, the idea presented in this paper is to perform mapping topic-by-topic (which might be at the same time week-by-week, but it does not have to be so) in order to focus on understanding what is that one wants to teach, as opposed to what one really teaches.

Szabo and Falkner describe where techniques in curriculum mapping can be used to enable a fine-grain analysis in a course level on the timeliness of introduction of certain topics within the course. [5] While this can be of great use to certain teachers in certain courses, and can be used to analyze in much detail the inter-dependencies among courses, and the gain from it would be improvement of the definition of course prerequisites, the idea behind this paper is not to go to such level of depths of analysis, but to only discuss at the level of topics and how good of a coverage for the area or the field do they give.

Romkey and Bradbury discuss the process of curriculum mapping but from the point of view of students and how would they perform it [7]. But, the process that is described in this paper is bound to happen during new curriculum design, period when students can not be involved in wider scale, especially not by teachers and during courses. In the notion of the idea of mapping as proposed in this paper, student involvement might be possible to arrange as part of a collaboration between the higher-education institution and their respective student or alumni organizations, during the period of curriculum reconstruction.

Techniques for semi-automated or fully-automated curriculum analysis, mapping and comparison are also part of this research field. As a notable example, Sekiya, Matsuda and Yamaguchi developed a method of systematic analysis of curriculum syllabi by using LDA (latent Dirichle allocation) and Isomap, which they used as a basis for several research efforts [8] [9].

I believe that only when we are able to explicitly state what we believe to implicitly know, we might be able to claim the knowledge. So, if an institution undergoes an evaluation of the study programs based on automated tools, they will be able to calculate some useful indicators of the trends and situation, but sometimes the journey is equally or even more important than the destination. This is why I propose a tool and a system which main goal is the navigation throughout the curricula and knowledge space, as part of a learning process about the institutional self.

When discussing tools to enable manual curriculum mapping, I would point out the STOPS tool, as a graph based study planning and curriculum development tool [10]. This tool can be used both by students and developers and offers different point of views. While the authors of the STOPS tool focus on mapping the curricula outcomes and try to define the prerequisites from the outcomes, in this paper the idea is to map the curricula to a standard BoK topics (such as ones that are part of curricula guidelines) and the prerequisites are a more generalized notion that should be defined on the basis of the topics that are mapped within the curriculum.

III. Analysis of Curriculum Mapping Extent

The main vision behind the new curriculum mapping extension of the ISIS system that was described in the introduction, is that: the design of new curricula, reconstruction of the past curricula and assessment of the success of the implementation of current curricula – should define, keep and integrate tight associations between the respective curricula proposals and their contents on one hand, and the body of knowledge as defined in reference curriculum guidelines on the other hand.

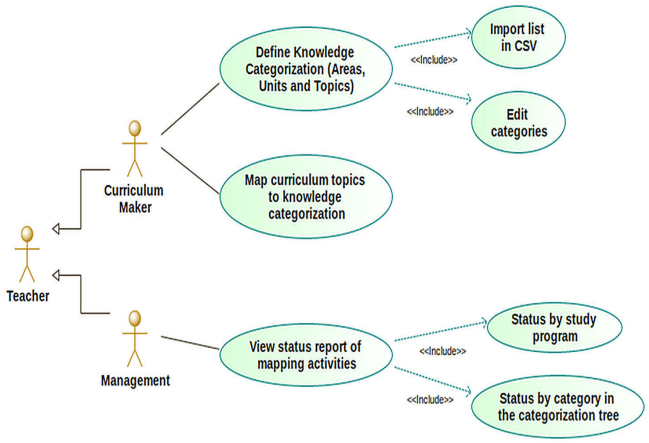


Fig. 1. Priority functional areas of the system: definition of body of knowledge categorical structure, mapping of curriculum topics to the knowledge categories and reporting functionalities.

The system functionalities are grouped in 3 functional areas, as illustrated in the UML use-case diagram in Fig. 1:

- Definition of the Body of Knowledge
- Mapping of curricula to Body of Knowledge
- Reports on the status of the mapping effort

The definition of the Body of Knowledge is a relatively straightforward process since this categorical structure is usually organized as a hierarchical tree having the knowledge areas at the top-level, the knowledge units at the middle and topics at the bottom (leaves) of the tree. As such they are simple to implement, with the only complication is that it should be allowed to have in use several independent categorical structures or alternate definitions of the Body of Knowledge.

The mapping of the implementation of curricula to the relevant Body of Knowledge and the reference guidelines could be done at several levels:

- Mapping at the level of whole study programs to certain specific knowledge areas from the guidelines. This is too broad and would not allow gathering information on more details on how curricula are implemented
- Mapping at the level of whole curricula per subject domain, could be considered as a bare level of mapping, since it gathers info on the knowledge unit that one curriculum implements, but there is no information on how it is implemented.
- Mapping at the level of topic for each curriculum, allows much more details and fine-grained recording of how a certain curriculum is implemented and how does it relate to the other curricula. Such a solution gives the possibility to see if several curricula implement the same parts of the Body of Knowledge, and to investigate if there are curricula that are mere unwanted duplicates of each other or separate instances that in fact offer additional knowledge and better specialization. This will also allow to have a detailed account on the human resources and competencies that will be required to establish such a curriculum and teach it with success and to do it in the most optimal manner by eliminating possible duplicate coverage where it is not needed.

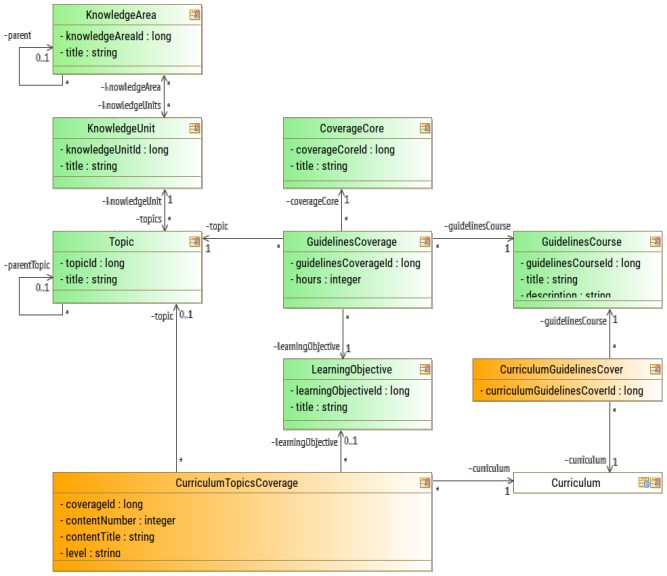


Fig. 2. Class diagram of the coverage mapping with definitions of knowledge areas, knowledge units per area, proposed topics in each unit, and the proposed coverage of each topic.

As an experiment, the analysis and design of this system started in the field of computing, with the investigation of the curriculum guidelines in several fields defined by ACM, IEEE, AIS and AITP (As an example see: [11]). During the design of the experimental system, this level of mapping per curriculum topic was chosen and its implementation is detailed in Figure 2. Note that the Curriculum class references parts of the class diagram in Figure 3.

IV. Software Construction

The new Mapping information system was designed to function as a sub-system of the ISIS (Integrated Study Information System), hence benefiting from the ability to use parts of its code-base and practices established within other projects to introduce curriculum management. This allowed a jump forward the development directly to the construction phase.

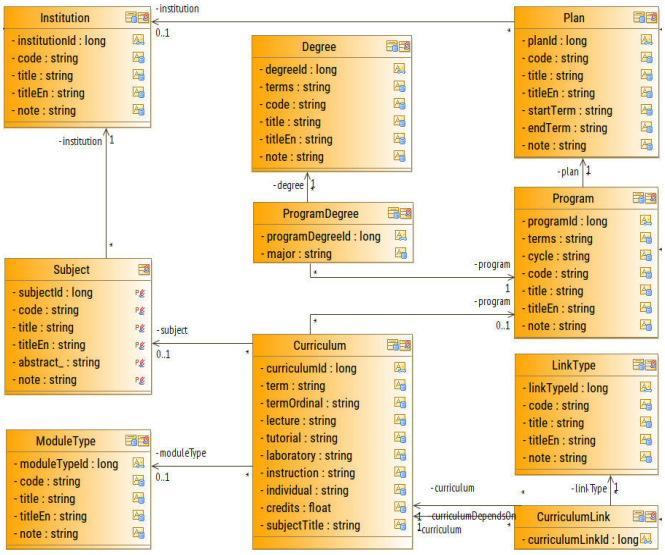


Fig. 3. Class diagram of the structure of study plans in each institution, with study programs and offered degrees, curricula per program for each subject domain, and graph structure showing the full graph of links between curricula allowing many types of links.

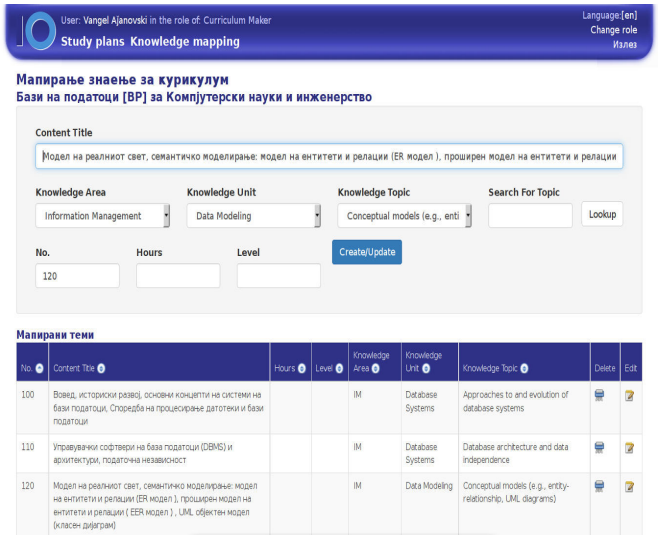


Fig. 4. The web page where all the content mapping occurs. The bottom part shows the list of content topics expected to be taught in a curriculum. In the top part of the screenshot there is a form for definition of new content topics, or editing already defined ones.

The construction phase included: design of new system use-case scenarios, new background and front-end services, new user interface and integration with the existing subsystems of the existing solution.

See Fig. 3 for more detail on the implementation of the structure of the study plans within the ISIS system.

The most important use-case scenario – *A Curriculum Maker maps content topics in a certain curriculum to the Body of Knowledge categorization* is defined as follows:

- The Curriculum Maker opens the list of curricula in a study program
- The Curriculum Makers selects the respective curriculum for editing
- The System displays the list of content topics that are defined for the curriculum
- For each of the content topics, the Curriculum Maker selects a Knowledge Area, Unit and Topic from the Body of Knowledge (or searches for the exact topic that mentions the given search word)
- Alternative: The Curriculum Maker can use specific KA, KU or KT to indicate that she has not yet decided on the most appropriate mapping
- Optionally, the Curriculum Maker specifies how many lecture hours are dedicated to the topic, what is the level of complexity expected for teaching this content topic

The screenshot on Fig. 4 displays an example implementation of the main scenario in a web interface form.

V. Deployment

The success of a project of this caliber at any institution depends on the strong willingness of the management and their concurrence to a single vision that such a system is truly needed and will make benefits to the institution.

Significant investment in both time and effort is required to fully map the entirety of an institutional curricula catalog. The main problem is that this can not be done by delegating the work to administration or technical staff since a sufficient level of competence has to be met in any given knowledge area in order to be able so successfully map the content topics from a certain course curriculum to the most relevant knowledge topics. This can only be done by the teaching or research staff and could take a significant amount of time (especially when designing new curricula).

Because of this, it is recommended to undertake a pilot project of the overall process, but in a real-life scenario at the institution in question, with specific Body of Knowledge per different subject domains. The pilot should include a representative sample of curricula chosen from several curriculum types and different subject domains. In that way, the management can be assured on the feasibility of the process, and a confident decision can be made on the transition of such a system to a production environment.

In order to be able to fully test the system, and assess the success, a complete functional test environment was designed to be performed on a real-life sample from the Curricula at the institution in question. This required that all the proposed categories from the Body of Knowledge tree for the relevant curriculum guidelines were already imported correctly into the system (Fig. 5).

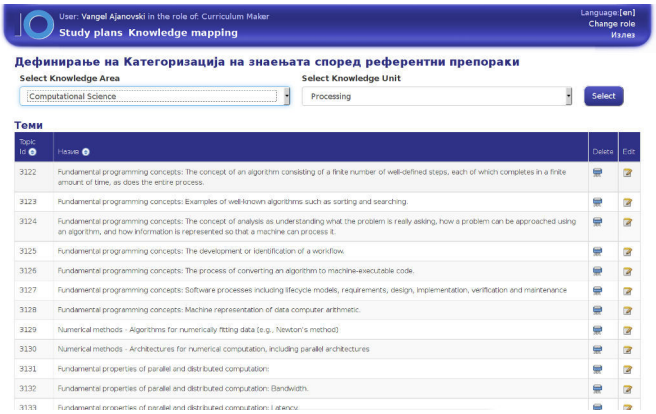


Fig. 5. Imported ACM Body of Knowledge.

VI. Pilot project and test-process results

The test scenario included both editing of the body of knowledge and categorization trees and submitting curriculum mappings, topic-by-topic. This is a highly customized process and depends on the structure of the curriculum materials and how easy it is to turn them into structured list of topics and find the relevant categories from the tree of the body of knowledge.

As an experiment, this process was done by the author, as part of a real-life functional test for the benefit of the author's institution. The test was done by fully using the system and the proposed structure for mapping and as such it lasted a significant amount of time (measured in several hours per course), while requiring many consultations with literature in the relevant areas (even though the authors holds a PhD in the field). To formalize the process, and guarantee a timely finish – the following test-procedure was implemented:

- Open the curriculum for editing
- Find the study program accreditation elaborate and lookup the specific document for the chosen curriculum
- Convert the "description" part of the document to a list of sentenced, in order to convert each sentence into a curriculum content topic
- Each sentence is rephrased so that it can be read as a topic in the curriculum
- Number each of the sentences to enable sorting, but use tens or hundreds in order to allow splitting or joining the topics, and still be able to use the sorting
- Try to find relevant knowledge area, unit and topic from the Body of Knowledge classification tree
- If the content topic can be mapped to several knowledge topics in the Body of Knowledge tree, then split the topic to several sub-topics and repeat the step for each of them
- If several content topics map to a single knowledge topic from the Body of Knowledge tree, join them under a single overall topic
- Decide on the number of lecture hours that should be spent on each of the content topics, and the level of complexity.

Although a test-procedure of such does not prove the true feasibility of the introduction of this system, nor does it prove the effectiveness, it still gives an illustration on the amount of work required and gives hints towards proper planning and management decisions. With this test procedure, applied to a random-chosen curriculum that is not among the competencies of the author, it only took an hour to roughly sketch the content of the curriculum and categorize it to what seemed the most appropriate knowledge topics from the Body of Knowledge, but it would warrant additional effort of checking the mapping and refining it by competent individuals. For a curriculum that is part of the competencies of the author, it took much less time for a rough sketch of the content topics, but it took another 1-2 hours to finalize all the details.

This test, concludes that the average speed of a competent individual in the role of a Curriculum Maker could be up to 3 curricula per working day. The process could be completed in under a month for all curricula under the current accredited study program if their respective teachers are involved in the mapping. If there were many accredited study plan instances in the past, it might take up to that many months to finish the process for all the historical curriculum, but in reality it should take significantly less since more often than not – copies of curricula are created for new accreditation documents and the content is only partly changed, if changed at all.

VII. Reports

Crucial part of the system are several reports that enable to assess the status of the mapping project and analyze the curriculum mapping in order to decide on the quality of the mapping, on the degree of coverage of certain priority knowledge areas and units.

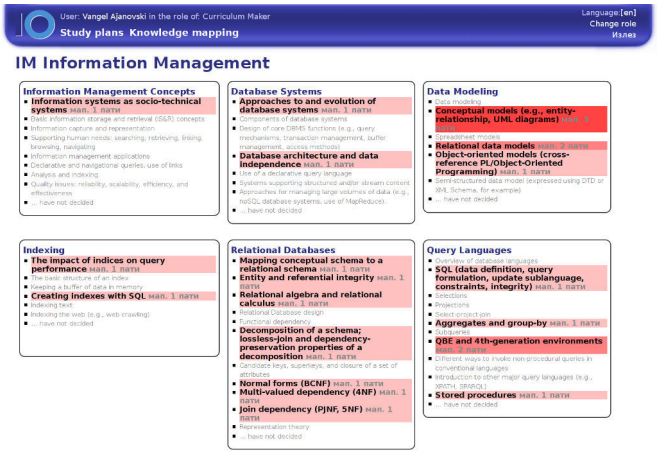


Fig. 6. Report of mapped Body of Knowledge topics.

The screenshot on Fig. 6 shows part of the report on the mapping effort as it goes underway. All Body of Knowledge Areas are displayed, with all Knowledge Units represented as boxes, while inside of them one can see the proposed Knowledge Topics that are part of the Curriculum Guidelines.

The topics are marked in the sense of a heat-map, indicated topics that are references by many curricula with bolder and more intense colors. The total number of lecture hours for each topic is also indicated in the report.

The screenshot on Fig. 7 shows another report displaying a side-by-side comparison of all curricula that map to a certain topic. This report is useful to investigate possible duplicate content across curricula. As an example, if we find that there are many curricula topics that map to a single knowledge topic from the Body of Knowledge, it can serve as an indicator that the same lectures are used over and over again in many courses and this should be investigated in more detail.

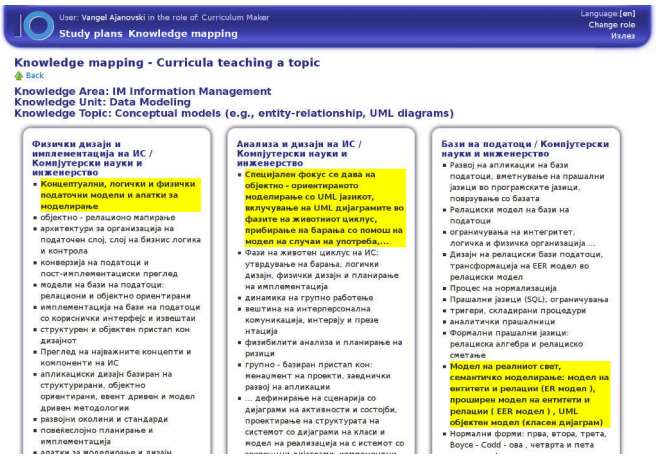


Fig. 7. The screenshot shows an example of a situation when several curricula map learning content to a single topic. The topic from the BoK is "Conceptual Models" and it is mapped to different learning content topics proposed by three different curricula – the image is in original macedonian language as part of the analysis of FCSE study programs.

VIII. Conclusion and Future Work

This paper explains the rationale behind the decision to design and implement an information system for mapping the coverage of reference curricula guidelines in the active study plans and programs of a higher-education institution, and reports on the result of this development. The developed system, includes solutions for the following problems:

- Keeping the structural description of all past, active and future study programs and their differences for a typical HE institution and its many departments, institutes, faculties and centers.
- Structure of the body of knowledge, with areas, units and topics can be defined and redefined and mapping can be established separately for each curriculum, for each subject and for each study program
- Assessment on the current state of curriculum mapping to guidelines, and state of the coverage of study program overall can be performed at any time
- Further analysis can be performed on the similarities and differences on various curricula and on programs overall, by using custom queries within the database.

As ISISng is envisioned as a multi-year project, work has already begun in converging the past efforts to other initiatives at the university related to the improvement of the quality of teaching and research, such as: tracking the implementation of reference curriculum guidelines in teaching week-by-week; monitoring student success and mapping their results to the Body of Knowledge; general human resource management and capacity development about teaching; and mapping and tracking researchers' work to the proposed Body of Knowledge within reference curriculum guidelines.

Further work within the framework of the ISISng project will also continue in the area of curriculum assessment and analysis of the effects that the structure of the study plans and curricula has over learner success (see [12], [13]).

References

[1] ISISng: Integrated Study Information System of the Next Generation Project Web-site – http://develop.finki.ukim.mk/projects/isis

[2] V. V. Ajanovski – "Information System of the Institute Of Informatics: IS for Students by the Students" – Proceedings of the Seventh International Conference on Informatics and Information Technology, Bitola 2010 – pp. 127-129 – ISBN 978-9989-668-88-3

[3] V. V. Ajanovski – "Integration of a Course Enrollment and Class Timetable Scheduling in a Student Information System" – International Journal of Database Management Systems, vol. 5, no. 1, pp. 85-95, Feb. 2013 – DOI: http://dx.doi.org/10.5121/ijdms.2013.5107

[4] V. V. Ajanovski – "A Mobile Virtual Academic Adviser" – 10th International Conference on Mobile Web Information Systems, Paphos, Cyprus, August 2013 – LNCS 8093, pp. 300-303 2013. Springer-Verlag Berlin Heidelberg. 2013 – DOI: http://dx.doi.org/10.1007/978-3-642-40276-0_25

[5] C. Szabo and K. Falkner – "Neo-piagetian theory as a guide to curriculum analysis. In Proceedings of the 45th ACM technical symposium on Computer science education (SIGCSE '14). ACM, New York, NY, USA, 115-120. 2014 – DOI: http://dx.doi.org/10.1145/2538862.2538910

[6] H. Hayes Jacobs, Mapping the Big Picture: Integrating Curriculum and Assessment K-12 (Professional Development). ASCD, 1997.

[7] L. Romkey and L. Bradbury, Student Curriculum Mapping: A More Authentic Way Of Examining And Evaluating Curriculum, in 2007 Annual Conference & Exposition, Honolulu, Hawaii, 2007.

[8] T. Sekiya, Y. Matsuda, and K. Yamaguchi, Analysis of Computer Science Related Curriculum on LDA and Isomap, in Proceedings of the Fifteenth Annual Conference on Innovation and Technology in Computer Science Education, New York, NY, USA, 2010, pp. 4852.

[9] T. Sekiya, Y. Matsuda, and K. Yamaguchi, Curriculum Analysis of CS Departments Based on CS2013 by Simplified, Supervised LDA, in Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, New York, NY, USA, 2015, pp. 330339.

[10] T. Auvinen, J. Paavola, and J. Hartikainen, STOPS: A Graph-based Study Planning and Curriculum Development Tool, in Proceedings of the 14th Koli Calling International Conference on Computing Education Research, New York, NY, USA, 2014, pp. 2534.

[11] H. Topi, J. S. Valacich, R. T. Wright, K. M. Kaiser, J. F. Nunamaker Jr, J. C. Sipior and G. J. De Vreede, (2010). "Curriculum guidelines for undergraduate degree programs in information systems" – Association for computing Machinery (ACM) and the Association for Information Systems (AIS) – Retrieved March, 4, 2013: http://www.acm.org/education/curricula-recommendations

[12] G. Méndez, X. Ochoa, K. Chiluiza – "Techniques for Data-Driven Curriculum Analysis" – LAK 14, March 24 - 28 2014, Indianapolis, IN, USA - ACM 978-1-4503-2664-3/14/03 – DOI: http://dx.doi.org/10.1145/2567574.2567591.

[13] D. C. Rowe, B. M. Lunt, R. G. Helps – "An Assessment Framework for Identifying Information Technology Bachelor Programs" – SIGITE11, October 2022, 2011, West Point, New York, USA. – ACM 978-1-4503-1017-8/11/10

LMS, CMS, LCMS, and VLE in e-Learning- similarities, differences and applications

Metodija Jancheski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
metodija.jancheski@finki.ukim.mk

***Abstract*— This paper discusses the usage of the various managements systems and environments in e-learning, including Learning Management System (LMS), Content Management Systems (CMS), Learning Content Management Systems (LCMS), and Virtual Learning Environment (VLE).**
The first part is dealing with the most common definitions, essential elements, properties, functions and successful examples for each of all these systems.
The next part contains detailed comparison between the systems. The main similarities and differences are explained, as well as their applications. In this comparison, we are considering various aspects of e-Learning, including content (creating, storing, organizing, publishing and sharing), data storage, metadata and learning objects, services and tools, multimedia issues, communication, collaboration, interaction, standards, and the essential components of e-Learning courses.
Finally, the author proposes how to compare existing systems, by using available online tools, and/or system and product analysis. He concluded that serious solutions should satisfy the most requirements of different software quality models. This will be helpful for educational institutions and professionals in the field of e-Learning to select, create or develop the most appropriate system/environment that will meet their requirements and expectations, as well as the educational needs of their students. (*Abstract*)

Keywords— e-Learning; LMS; CMS; LCMS; VLE (key words)

I. INTRODUCTION

After the digital learning material is created, the educator should decide how to make it accessible to students. To select the best method of displaying the resources online, it is necessary first to assess various aspects of the available e-Learning systems and to match them with the strategy, goals and needs of the educational institution. According the results of this assessment, a choice will be made between the following systems: Learning Management System (LMS), Content management system (CMS), Learning Content Management Systems (LCMS), Virtual Learning Environment (VLE). These systems are widely used by the schools, colleges, universities, institutions, organizations and companies. The systems fall in two main categories - proprietary and open source systems.

There are different categories of users of all these systems, including students, employees (trainees), educators, trainers, administrators, technical staff, developers of learning materials, software developers, providers and vendors.

Some organizations develop their own learning management systems and software to address their own unique needs. Others buy off-the-shelf solutions available from a growing number of software vendors. Most organizations use a combination of the two alternatives [1].

II. DEFINITIONS

Follows a review of some common definitions of LMS, CMS, LCMS and VLE.

A. Learning management system definitions

1. A software application used to plan, implement, and assess learning processes. Typically, an LMS provides instructors with ways to create and deliver content, monitor student communication and participation, and assess student performance, and provides students with the ability to use interactive features such as threaded discussions and video conferencing. [1]

2. A high-level, strategic solution for planning, delivering, and managing all learning events within an organization, including online, virtual classroom, and instructor-led courses (Greenberg, 2002) [1].

B. Content management system definitions

1. A Content Management System is a collection of procedures used to describe processes in an environment that requires collaboration between different actors. These procedures are designed to manage: data access, based on user roles; collecting and sharing information; data storage assistance; content redundancy check; and reporting [2].

2. A system used to manage the content of a website [3].

C. Learning content management system definitions

1. A system that enables the creation, storage, management and deployment of learning content in the form of learning objects to serve the needs of individuals.'' (Robert Koolen) [4]

2. Software that allows for both the administrative and content - related functions of e-learning. It combines a learning LMS and CMS in one package. [5]

D. Virtual learning environment definitions

1. Software systems that provide the management of online or elearning [1].

2. An on-line interactive, integrated learning environment supported on the Internet, usually institution based [1].

III. LEARNING MANAGEMENT SYSTEMS

In the literature used, LMS is recognized as complex digital learning platform [1], a high-level, strategic solution [1] software system [6], complex system [7], a large Web-based software application, which comprises a suite of tools [4] or an integrated set of software/programs [2].

Typically, LMS manages the process of formal learning [8]. According Coats (2005), in educational settings, adoption of LMS has been widespread. Using the LMS helps educational institutions and companies to meet their specific needs, adapting the learning (training) to groups of students (trainees) or individuals. This is achieved by programmed learning, i.e. learning through different paths, customized to group of users or individual user. Classically, LMSs offer a collection of courses. They may also include capabilities for assembling individual courses into organized curricula or certificate programs. [7].

As we can see from definitions from previous section, LMS is used for planning, organizing, implementing, delivering, controlling, and assessing learning processes and events. It is also used for housing, organizing, performing and monitoring of various types of e-learning courses (trainings), and their management [1]. That's the reason in many sources (see for example [5, 6]) the terms LMS and CoMS (Course Management System) to be used as synonyms. Conversely, Simonson at all. (2015) stated that CoMS are often erroneously identified as LMS and that it is about entirely different genres of products [6].

The most LMS are Web-based, others should be migrating to fully Web-based implementation [9].

A. Common functions of LMSs

LMSs simplify [7], centralize [2, 4], facilitate and automate [2, 4, 7] the processes of administering education and training, through these functions:

- coordinate registration of learners [1, 4, 9, 10]
- maintain learner profiles [4]
- maintain a catalogue of courses [4, 10]
- store and deliver self-paced e-learning courses [4]
- download e-learning modules and tools [4]
- distribute material [1]
- management of learning materials [10]
- integrating knowledge management resources [10]
- assemble and deliver learning content [2]
- track and record performance and progress of learners [4, 8, 9]
- track training activities [8, 9]
- track the e-Learning program [1, 2, 10]
- test the learners [1]
- assess learners [1, 4, 9, 10],
- track and record assessment results [4]
- provide reports to management [1, 2, 4, 10, 11]
- supporting collaboration and knowledge communities [10]
- include technology that enables chats and discussions [9]
- systems integration [10]

B. Examples of LMS

The list of the most popular LMS includes MOODLE (Modular Object-Oriented Dynamic Learning Environment), Sakai, Blackboard, Elearning manager, Dokeos, Bluevolt, Litmos and Topyx.

IV. CONTENT MANAGEMENT SYSTEMS

A CMS is not a dedicated e-learning application but it can be closely integrated with an LMS and used to support the effective development and delivery of course content—especially if you are working with learning objects. [4] Content can be easily marked with metadata, which enables fast and efficient search and retrieval and use of content. Users with appropriate permissions can edit, add and view the content, while those with lower levels of access can only view the content.

CMS supports an object-oriented approach to e-learning content. [4] CMSs facilitate the integration and automation of the processes of creating, developing, capturing, assembling, publishing and delivering the knowledge content [9]. The other most common features of CMS systems include format management, controlling, revision control, and indexing, search and retrieval [1, 5, 12]. The functions of the CMS allow options about how content should be displayed, in public or private.

With other words, CMS is used for storing, organizing and managing of content (information, data) on Web site located on a central place. It manages various types of content that are stored in the database, including documents, books, images, movies, audio and video files, courses, libraries of learning objects, and simulations.

Typically CMS's comprise of two distinct sections; the content management application (CMA) and the content delivery application (CDA). The CMA allows the content manager or author, who may not know HTML, to manage the creation, modification, and removal of content from a website without needing the expertise of a webmaster. The CDA

element uses and compiles that information to update the website [1, 12].

The principle underlying a CMS is the separation of content and presentation. It is easily visible in online newspaper, which is a classic example of CMS application. The presentation templates are fixed or seldom changed, while content is updated constantly. Journalist don't need professional IT skills to work with CMS, they just save content to the CMS. When a user requests a page, content is automatically poured into the associated template. [4]

A. Examples of CMSs

The most popular CMSs used for creating various types of Web sites are Drupal, Joomla, Dotclear and WordPress. CMS is probably the oldest term used to refer to software as Moodle [13].

V. LEARNING CONTENT MANAGEMENT SYSTEMS

Educational institutions and companies that want to modify individual courses, adapting them to their students or employees (trainees) typically use LCMS. By LCMS, designers of teaching can create small chunks of content, including lessons, topics, modules, tests and other learning objects of individual courses that can be used repeatedly. LCMS manages the integration of these learning objects which allows the e-learning content to be individualized, i.e. to meet the needs of groups of students or individual students. LCMS also simplifies the task of creating, managing, and reusing learning content in multiple courses, lessons, or topics [7, 8].

LCMS usually offers simple opportunities to update and change the content, regardless of time and place.

It is often used in the processes of certification and standardization of different systems in educational institutions and companies.

A. Common functions of LCMSs

As a multiuser system, LCMS facilitate administration and authoring at the course, lesson, and page levels. It manages learning content by maintaining items of content in a central repository. From this database, instructional designers can organize, assemble, approve, publish, and deliver courses and other learning events. An LCMS lets authors create, store, and refine learning objects or other units of content. It helps learners locate and take just the learning they need at the moment. LCMS should simplify and accelerate the content authoring process allowing subject matter experts with the appropriate access rights to self-publish. It is this self-publish function that some people believe gives the LCMS the potential to act as a knowledge sharing and Knowledge Management tool. [4]

B. Examples of LCMSs

Claroline, e-Doceo solutions, Ganesha, Xyleme and Kenexa are some of the most well-known LCMS.

Because Moodle is expandable and allows its separate modules (such as wikis, databases and blogs) to be used by students, to be stored and repeatedly used, it can be considered as LCMS.

VI. VIRTUAL LEARNING ENVIRONMENTS

Virtual Learning Environment is a software system designed to support teaching and learning. It uses Internet browsers to deliver instructional materials. VLE can be a good complement to classical training, within the blended learning. Its primary purpose is to help teachers and tutors in the management of courses designed for their students. Even teachers and tutors with minimal knowledge in the field of information and communication technologies, through VLE can create Web sites for their courses. VLE facilitates communication, testing and sharing of teaching materials, including the textbooks. It allows users to share learning resources with each other and to create them with joint efforts, by using tools like wikis, blogs, RSS and more.

A. Examples of VLEs

The most popular VLE's are Moodle, Blackboard and WebCT.

VII. COMPARISON OF THE SYSTEMS

The systems described in the previous sections are encountered in practice in various variants which differ in terms of different aspects, including design, dimension, scope, tools and functionalities, and costs. The boundaries between them are often invisible or foggy. Each of these categories of systems has its own unique and defining characteristics but for most of them, there are arguably more similarities than differences [14].

There are examples that the differences between systems belonging to the same category can be smaller than between systems belonging to different categories. It is not rare in the literature of the field to find closer determinations, like pure CMS, full-featured LMS, etc.

As for the costs, Dobbs (2003) stated that "Experts agree that an LMS is the most expensive tool in an e-learing initiative [1].

A. LMS vs CoMS

People often get confused regarding the actual functions of a CoMS and an LMS. The source of this confusion lies in the similarities of the two systems. Both perform the functions of enrolling learners, communicating with them, assessing performances, and activating learning materials [2]. For this reason, in this paper we used the terms LMS and CoMS as synonyms.

According Simonson at all. (2015), the primary difference between the two systems is that the focus of a CoMS is on the delivery of courses, while an LMS focuses upon an individual and tracks his learning needs and outcomes achievement over periods of time [6].

B. LMS vs CMS

LMS is a complex digital learning platform, more robust than of CMS [1]. Unlike CMS, which is considered a passive application, LMS gives students the opportunities to see, hear and to interact with each other or to interact with data. The students also can experiment and practice through examples, to conduct assessments, and to give grades or feedback on various aspects of the courses.

C. LMS vs LCMS

LCMS is a software application that has integrated in its structure the essential features of LMS and CMS. Widely speaking, it takes the best-of-breed features and benefits from five major application spaces: knowledge management, content management, document management, online collaboration, and eLearning (including LMS) [9]. Most LCMSs offer some LMS functionality - course administration, course catalogue, learner registration and learner tracking, for example [4].

An LMS and LCMS are not interchangeable nor are they mutually exclusive [4]. Traditionally, an LMS provided management of learning performance, learning requirements, learning programs, and planning, and an LCMS provided for management of learning content (Greenberg, 2002). [1]

An LMS can manage communities of users, allowing each of them to launch the appropriate objects stored and managed by the LCMS. In delivering the content, the LCMS also bookmarks the individual learner's progress, records test scores, and passes them back to the LMS for reporting purposes''. [4]

The next table (Table 1) offers interesting comparison between LMS and LCMS [15].

TABLE I
COMPARISON OF LMS AND LCMS

LMS	LCMS
Offers improved delivery and tracking of content	Offers improved creation and management of content
Used by learners and administrators	Used by content creators and developers
Typically offers courses	Assembles learning objects
Manages people involved in learning (participants)	Manages computer files

D. CMS vs LCMS

Morrison (2003) stated that "an LCMS is a CMS dedicated to learning content and a learning environment" [4], while according Simonson at all "LCMSs are the corporate world's equivalent of CMSs". [6]

One of the main differences between CMS and LCMS is the scope. As we can conclude from the sections described these two types of systems, the reach of an LCMS is more extensive. [1]

An LCMS is a system (primarily Web-based) that is used to author, approve, publish, and manage learning content or learning objects. It combines the administrative and management dimensions of a traditional LMS with the content creation and personalized assembly dimensions of a CMS. (Nichani, 2001) [1].

E. LMS vs VLE

The terms LMSs, and VLEs, are often used interchangeably, although the term LMS is generally used to describe a system of wider scope that includes the ability to perform administrative tasks involved in education such as reporting, documenting, and analyzing. [14]

VIII. ABOUT MOODLE

There are numerous powerful e-Learning software platforms. Some of the successful stories are BlackBoard, WebCT, Moodle, eFront, Dokeos, Claroline, Atutor, ILIAS, OLAT, Sakai, and .LRN. Unlike the first two, the others are open-sourced platforms. Follows brief review of main features and advantages of MOODLE, as a prominent representative of this software category.

Moodle belong to several categories of systems, it is in the same time CoMS, LMS and VLE (e.g. see [14]) or even LCMS (see section V.B.). Moodle's flexibility, in terms of how it can be set up and maintained, is one of the main reasons for its inclusion in virtually every category of online learning software package [14].

It is a Web-based learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalized learning environments [16]. Cost-efficiency, multilingual capabilities, flexibility and scalability are some of the main advantages of Moodle. It is used by universities, community colleges, K-12 schools, business, governments, non-profit organizations, etc. According the newest statistics data from *moodle.org* (March 2017), Moodle is actively used in 234 countries and it is translated into more than 120 languages. It includes more than 11,7 million courses, and about 1,1 million teachers/instructors. The impressive community of more than 100 million users across both academic and enterprise level usage makes Moodle the world's most widely used learning platform. [17]

Moodle.net offers free content and courses shared by Moodle users all over the world. It contains: courses free for downloading and using, courses that everyone can enroll in and participate, other content (such as quizzes, database and glossary entries) that everyone can import into its own courses.

The core theory behind Moodle is social constructivism. It is based on the idea that people learn best when they are engaged in a social process of constructing knowledge. [16]

It is free and open source software which provides an easy way to upload and share materials, holds online discussions and chats, gives quizzes, surveys and dictionaries, gathers and reviews assignments, and record grades. It also can integrate external collaborative tools such as forums, wikis, chats and blogs. Many of these tools are very powerful. For example, quizzes can combine different questions for testing of various types: Calculated, Description, Essay, Matching, Embedded

Answers, Multiple Choice, Short Answer, Numerical, Random Short-Answer Matching, and True/False. [16]

As an open-sourced software, Moodle can be customized and tailored to individual needs. Everyone can adapt, extend or modify Moodle for both commercial and non-commercial projects without any licensing fees. Of course, adapting, extending or modifying Moodle is not an easy task. It required high level of ICT knowledge and experience in educational environments. [16]

IX. CONCLUSION

To make a good choice of e-Learning systems one should be familiarized with their features, functionalities and capabilities. It is equally important if you want to choose any of the existing systems on the market, as well as if you want to develop your own solution, which would include the best features of existing systems and would be tailored to the needs of users and the institution.

Besides, it is useful to research some of the available online tools for comparing systems. One such tool is CMSmatrix (http://www.cmsmatrix.org). It compares the features in over 1300 content management system products. The set of criteria consists of more than 140 criteria, grouped in ten categories: system requirements, security, support, ease to use, performance, management, interoperability, flexibility, built-in applications, and commerce.

There are also available various systems and product analysis, conducted by e-Learning consultancies. One such example is E-learning consultancy Brandon-Hall analyses of the most LMS and LCMS products [4]. There is no doubt that serious solutions should satisfy the most of the requirements in different quality models (QM), including:

1. McCall's QM: a) Product Revision (maintainability, flexibility and testability), b) Product Operation (correctness, reliability, efficiency, integrity and usability), c) Product Transition (portability, reusability and interoperability);

2. Boehm's QM: a) The high-level characteristics (as-is utility, maintainability, portability), b) The intermediate level characteristic (portability, reliability, efficiency, usability, testability, understandability, flexibility), and c) The primitive characteristics;

3. Dromey's QM: a) Correctness (functionality, reliability), b) Internal (maintainability, efficiency, reliability), c) Contextual (maintainability, reliability, portability, usability); d) Descriptive (maintainability, reliability, portability, reliability)

4. FURPS QM: a) functionality, b) usability, c) reliability, d) performance, e) supportability;

5. ISO 9126 QM: a) functionality, b) reliability, c) usability, d) efficiency, e) maintainability, f) portability. [20]

REFERENCES

[1] P. Rogers et al. (editors), "Encyclopedia of distance learning", *Information Science Reference* (an imprint of IGI Global), 2009

[2] S. Ninoriya, P.M. Chawan, B.B. Meshram, "CMS, LMS and LCMS for eLearning", *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 2, March 2011.

[3] M. Burns, "Distance education for teacher training: modes, models, and methods", *Education Development Center*, Washington DC, 2011.

[4] D. Morrison, "E-learning strategies: how to get implementation and delivery right first time", John Wiley & Sons, 2003.

[5] G. M. Piskurich (editor), "The AMA handbook of e-learning: effective design, implemenation, and technology solutions", *AMACOM* (a division of American Management Association), 2003.

[6] M. Simonson, S. Smaldino, S. Zvacek, "Teaching and learning at a distance, Foundations of Distance Education", *Information Age Publishing*, 2015.

[7] W. Horton, K. Horton, "E-learning Tools and Technologies: A consumer's guide for trainers, teachers, educators, and instructional designers", *Wiley Publishing*, 2003.

[8] M. J. Rosenberg, "Beyond e-learning: approaches and technologies to enhance organizational knowledge, learning, and performance", *John Wiley & Sons*, 2006.

[9] L. Bielawski, D. Metcalf, **"**Blended eLearning: Integrating knowledge, performance, support and online learning", *HRD Press*, 2003.

[10] M. J. Rosenberg, "E-Learning strategies for delivering knowledge in the digital age", *McGraw-Hill*, 2001.

[11] "2017 Content Management Fact Sheet" [Online] Available: http://review.content-science.com/2017/01/2017-content-management-fact-sheet/ Last accessed January 2017.

[12] R. Dvorak, "Moodle for dummies", Wiley Publishing, 2001.

[13] A. Giroux, "VLE, CMS, LMS or CLMS?" [Online] Available: http://www.alexandragiroux.net/vle-cms-lms-or-lcms, Last accessed January 2017.

[14] Jason Hollowell**, "**Moodle as a Curriculum and Information Management System", *Packt Publishing*, 2011.

[15] William R. Brandon (editor), "Best of the eLearning guild's learning solutions: top articles from the eMagazine's first five years", *John Wiley & Sons*, 2008.

[16] M. Jancheski, "The importance of e-Learning in the process of life-long education of teachers", 9th International Conference of Education, Research and Innovation (ICERI2016) Proceedings, Seville, Spain, November 2016.

[17] Official MOODLE site [Online] Available: www.moodle.org, Last accessed: January 2017.March 2017.

[18] "Comparison of LMS, LCMS and CMS", [Online] Available: https://florettewilliamsportfolio.files.wordpress.com/2016/02/lms-vs-lcms-vs-cms.pdf

[19] "The classification & evaluation of content management systems", *The Gilbane report*, Vol. 11, No. 2 March, 2003.

[20] H. S. Shukla , D. K. Verma, "Analysis of Software Product Quality Models", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS 15-311), Issue 12, Volume 1, 2015

A New Collection of Educational Scratch Projects Produced by Computer Science Students

Mile Jovanov, Emil Stankov and Bojan Ilijoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
{mile.jovanov, emil.stankov, bojan.ilijoski}@finki.ukim.mk

***Abstract*—Scratch is a free educational programming language specifically designed to be convenient for students of very young age to learn programming. The course "e-Learning Systems" is an elective course for the students at our institution – Faculty of Computer Science and Engineering (FCSE) in Skopje, primarily offered to the students enrolled in the "Computer Science Education" track, in the third semester of the undergraduate studies. Starting from the current academic year (2016/17), students of this course are asked, as a project activity, to implement an educational game or story in Scratch. In this paper we present the new website we have developed for the purposes of the course, which contains the student projects developed throughout the winter semester of 2016/17, as part of the course activities. We describe the form of the student projects and the collection of projects present on the website, as well as their functionalities and application.**

Keywords—development of educational Scratch projects; e-learning systems course; block-based programming.

I. INTRODUCTION

The course "e-Learning Systems" is an elective course for the students at our institution – Faculty of Computer Science and Engineering (FCSE) in Skopje, primarily offered to the students enrolled in the "Computer Science Education" track, and the "Applied e-Technologies" track, in the third semester of the undergraduate studies [1]. According to our implementation of ECTS (European Credit Transfer and Accumulation System), it can also be elected by students from other tracks in a higher semester.

The goals of the course include familiarizing students with the basic concepts of e-learning systems, their usage in education, as well as the functionalities that they offer to their users. Upon successful completion of this course the students will be able: to understand the role and new trends in technology assisted learning environments; to evaluate the learning process in technology assisted learning environments; to evaluate the effectiveness of the courses that are performed with the assistance of systems for technology assisted learning; to identify the needs of students in technology assisted learning environments; to understand the transition from traditional to technology assisted learning environments; and independently – to create an environment for collaborative creation of educational content.

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.

The course contents include: models of learning, traditional learning environments, tools used in traditional learning environments, integration of technology into the learning process, the basic models of using ICT in education, technology aided learning environments, tools for technology assisted learning, comparison between traditional learning environments and technology assisted learning environments, interaction with technology assisted learning, interactive learning through the media, blended learning, designing digital content for learning technology and assisted learning, systems for content management, student oriented learning, collaborative learning and teaching, e-learning 2.0, learning through play, application of ICT in education of: preschool children, children and people with disabilities, gifted children, tools for knowledge management.

The course includes lectures, projects, discussions and laboratory exercises (weekly: 2 hours of lectures, 2 on discussions/exercises and 2 hours in lab). The students pass the course with 50% scored points from all the activities, including work on exercises, short exams and work on a project. Starting from the academic year of 2016/17, students are asked, as a project activity, to implement an educational project in Scratch.

In this paper we present the new website we have developed as e result of the "e-Learning Systems" course held in the winter semester of the current academic year (2016/2017). This website contains all the student projects developed throughout the semester as part of the course activities. Furthermore, we describe the form of the student projects and we go into details for three selected projects from the collection present on the website, with description of their functionalities and application in an educational environment.

II. ORGANIZATION OF THE WORK ON PROJECTS IN THE "E-LEARNING SYSTEMS" COURSE

Student projects can be boring and far from challenging to students. They may need to spend a lot of time on research about topics out of their interest, and still not be required to do any kind of practical work or work that will challenge them adequately on an intellectual level. For the course "e-Learning Systems", we always try to assign students interesting and challenging projects.

The work-on-project in the course is organized throughout the semester. At the beginning of the semester, students are

divided into teams (groups) of four to five members, and they generally work on the project outside the classes – although some of the course's lab classes are specifically dedicated to implementation of projects. Each team must select one member to take the role of team leader – a person responsible for corresponding with the teachers. Since usually there are different project types offered to the students of the course, the team leader also has the responsibility to select an appropriate project type for his/her team – after consulting the team's members. This means that the teams are allowed to "set the bar" for themselves, i.e. to decide on the project difficulty level and the requirements that they will have to meet in order to complete the project.

As a project activity for the course held in the winter semester of the current academic year, students were required to implement an educational game or story in Scratch. Regardless of the students' choice (game or story), as a mandatory part, the project had to include a process of learning and examination of the acquired knowledge on the topic under consideration. The topic for this year's project was "Work in Scratch and basic concepts of programming". The project had to be finished by the end of the semester and presented in public, in front of all the students and teachers on the course.

This year, three types of projects were offered to students. More precisely, it was a single base project (type 1), with added requirements (for type 2 and 3). The project types differed in complexity: type 1 project was the simplest one and type 3 was the most complex one. Of course, they were assigned different amount of points, according to the complexity: type 1 was assigned 7 points, type 2 – 10 points, and type 3 was assigned 13 points.

For the type 1 project, students were expected to implement an e-learning game or story, which had to rely on the basic principles of development and contain the basic elements of a game or a story, as described in [2]. The developed game/story had to look aesthetically satisfying and have at least three different levels. Furthermore, students were required to provide a form of examination (through questions/answers, or some other form that they would find to be appropriate) in the project. The type 2 project had the same basic requirements as type 1. Additionally, students had to create animations, as well as to provide gradation in the examination process (e.g. if they used questions/answers, they had to arrange them so that the easiest question would appear in the first level and the hardest one would appear in the last level). They also had to provide at least two different main characters from which the user would be able to choose at the beginning of the game/story. The type 3 project included everything from the previous type, and plus there had to be a start screen with some animation and a field where the user could write his/her name. Furthermore, students were expected to implement an end screen where the current user's points and the high scores would be displayed. Of course, different types of questions were expected to be assigned different points (e.g. the hardest question might be assigned the most points and the easiest question – the least number of points). Additional important requirement for this type of project was to provide two different storylines (for a story) or gameplays (for a game). In this way, by choosing a different main character from a list of (at least) two characters, the user would have the opportunity to play the game in a different way or to follow a different story, and learn some new concepts on the topic.

The assessment of projects and students was conducted in the following way. Each project, developed by a particular team, won at most the number of points defined as the maximum number of points for the corresponding type of project (7, 10 or 13), depending on how far the requirements for that project type were met. The final grade of the project was based on presentation, gameplay, storytelling, and the learning impact of the game/story. On the other hand, each team member (student) won the points that his/her team got, plus a particular amount of individual points. The individual points were assigned to each student according to the impression he/she had left to the course teachers at the public presentation of projects (at the end of the semester). This impression was formed based on the answers given by the student to 2-3 questions asked by the teachers regarding the implementation of a particular conceptual part of the project. The individual points were also different for the three different project types, and each student could win at most the number of points defined as the maximum number of individual points for the corresponding project type: 5 (for type 1), 6 (for type 2), and 7 (for type 3).

In the following section we explain our decision to use Scratch as a tool, and in Section IV we describe the website we have developed for the purposes of the "e-Learning Systems" course, which contains all the student projects developed throughout the semester. We also describe some examples of successful student projects that are present on the website.

III. WHY SCRATCH?

Scratch ([3], [4]) is a free, very popular, educational programming language specifically designed to be convenient for students of very young age (ages 8 to 16, or from third grade of primary education to high school students) to learn programming. Scratch was developed by the Lifelong Kindergarten Group at the Massachusetts Institute of Technology (MIT). It provides tools for creation of interactive stories, games, art, simulations, and all that using block-based programming [5]. As stated in [4], "Scratch helps young people learn to think creatively, reason systematically, and work collaboratively – essential skills for life in the 21st century". Programming in Scratch is done by dragging blocks from a built-in block palette and attaching them to other blocks to form a program structure called a script – a method known as "drag-and-drop programming" [6]. The Scratch project began in 2003, and the Scratch software and website were publicly launched in 2007 [3]. Currently, there are more than 12 million registered Scratch users, and this statistic grows rapidly each day.

Programming in Scratch introduces students and enables them to understand some of the basic principles of programming in general, such as: sequential execution of statements, conditional execution of statements, repetition of statements, variables and storing values, operators and operands, parallel execution of statements, using and passing messages as a means of communication between objects in a

program, event handling, reacting to input and/or other events, etc. Scratch programming also enables development of some basic IT skills, since creating a typical Scratch application often includes: drawing sprites (characters), inserting photos or other types of files created using a graphics development software as background images, recording sounds and saving them in the MP3 file format, adding different textual fonts to the application, moving sprites to different positions on the screen using graphical commands, manipulation and applying digital effects to graphical objects, etc. Crook [7] argues that besides the fact that Scratch can be used as a core element to teach digital literacy in schools, it also "offers a creative environment to empower children to program computers and could thereby extend the ICT curriculum beyond the teaching of basic office skills". Scratch can also help embed the usage of computers throughout the curriculum and expand their usage beyond the classes in informatics only – which is a common situation in both the primary and the secondary education nowadays.

IV. The Edugames Website: Repository of Succsessfull Student Projects

In order to have a single space where we can gather all the projects developed by students of the "e-Learning Systems" course, and to provide a way for the students' work to be publically available, we have created a website where we have uploaded all the projects implemented in the current academic year. The website's name is Edugames [8]. The home page of the website is shown in Fig. 1. Currently, the structure of the website is very simple – it contains only 4 different main pages which can be reached through the appropriate hyperlinks on the home page, as can be seen in Fig. 1. Clicking on the Games hyperlink opens up the most important one – the page that contains a listing of hyperlinks to each of the projects developed by this year's teams of students. This page is shown in Fig. 2. Clicking on any one of the displayed hyperlinks loads the Scratch project implemented by the corresponding team.

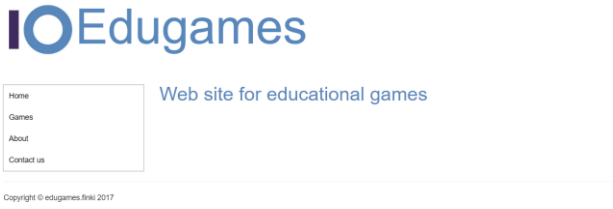


Fig. 1. The home page of the Edugames website.

The students made a serious effort in realizing their projects, and worked continuously throughout the semester – on the course's lab classes dedicated to projects implementation, as well as outside of them. This is also confirmed by the fact that almost 90% of the projects were completed on schedule. Some of the students showed exceptional commitment to the activity and produced educational projects of significantly high quality level. In the following subsections we will describe a few of the most successful projects developed by the students of the course in the current academic year.

Fig. 2. Listing of projects developed by students that are present at the Edugames website (excerpt).

A. *"Coins collector" Game*

The main idea of the "Coins collector" game is to use some of the Scratch elements in order to make a movement of the main character (Fig. 3). In the game, the player has to move the character over the scene using the available Scratch blocks (controls) in order to collect all the coins placed on it. After collecting the coins, the character must reach the goal position, which opens up the next, more complex level of the game. Each new level of the game includes a different, larger set of controls that can be used by the player, so besides the standard controls for moving forward/backward and rotating, there are also controls for door opening, using a piece of weaponry, etc. Additionally, for each different level there is a time constraint for completion, so the player must play quickly and efficiently.

The game offers an interesting way of introduction to the elements of Scratch.

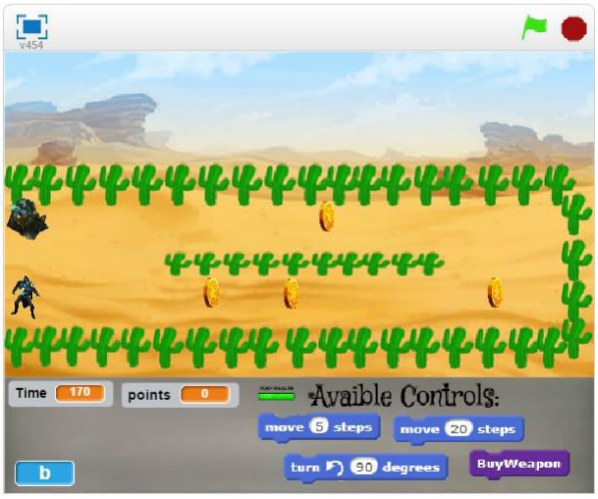


Fig. 3. "Coins collector" game – an example of a successful educational project developed by students of the course "e-Learning Systems", in the winter semester of 2016/2017. The game offers an interesting way of introducing the elements of Scratch.

B. *"The programming course" Game*

The game shown in Fig. 4 represents a classroom simulation. The idea here is that the teacher, through short messages and animation of writing on the blackboard, introduces the student to the contents to be learned. Then, he asks questions to which the student must provide correct answers in order to proceed with the game. The game has many levels, each of which covers different programming concepts, such as variables, conditions, loops, etc. Furthermore, the difficulty of questions rises with each new level.

Fig. 4. "The programming course" game – second example of a successful educational project developed by students.

C. *"Code me" Game*

Another successful educational project, which simulates the process of learning a programming language, is the "Code me" game, shown in Fig. 5. The main scene consists of 4 parts. The first part is the scene where the main character can be moved, as well as some additional objects. The other 3 parts are used to present the description of the story, to present the available controls and to provide a space where the user can write the programming code. First, the story is set in the upper left part named "Instructions" (Fig. 5), and the events that need to happen are described. In the middle part, named "Commands", the controls that are available for the current scene (level) are displayed. There are controls for moving to the left/right, jumping, taking a particular object, leaving an object, using an object, etc. The appropriate controls must be placed in the correct logical order in the upper right part, named "Program", so that executing them will realize the described story successfully. The game has several levels and it becomes more and more complex from level to level: more controls are introduced and the program (sequence of controls) to be executed in order to realize the story becomes larger.

"Coins collector", "The programming course" and "Code me" are three examples of games produced by our students. The benefit from this approach, besides the knowledge gained by the students, is that there now exists a valuable collection of short educational games that may help other students to learn a particular topic through a game. This year's topic was 'Introduction to programming', and the produced games may be beneficial for the younger pupils, having in mind the trend of introduction of computational thinking and programming in the early stages of education in our country [9], as well as in many other countries worldwide.

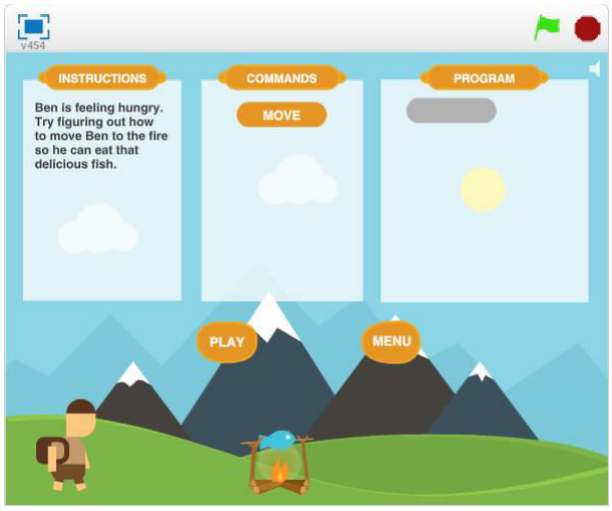


Fig. 5. "Code me" game – third example of a successful educational project developed by students. The appropriate controls (middle section) must be placed in the correct logical order in the right section, named "Program", so that executing them will realize the described story (left section) successfully.

V. Conclusion and Future Work

Our general impression is that students tend to show high interest in Scratch as an e-learning platform. Although some of the students were skeptical at the beginning and shared an opinion that Scratch is a programming environment for children, later they captured our primary idea, which was not to teach them to program in Scratch (students at that age are already capable to program in much more complex environments and languages), but rather to enable them to understand the concept of an environment for learning programming which is intended for children. This is confirmed by the successfully implemented projects, a large part of which were systems for learning programming and/or mathematics suitable for children of very young age. Of course, without doubt, through the process of creation of the project students also learned Scratch as well.

In the future, we plan to continue with this type of project activity for the "e-Learning Systems" course. Our goal is to inspire the future generations of students to create projects of even higher quality, and to publish all these projects and expand the Edugames website. We also plan to group projects by theme and by intended users' age on the website, in order to make them easily accessible for the users for which they were originally intended. This may be done by listing the games by their names and attaching few labels for each game which will allow sorting and fast access to a particular group of games.

References

[1] Website of the Faculty of Computer Science and Engineering (FCSE), http://www.finki.ukim.mk, accessed March 2017.

[2] A. Majumdar, "5 game elements that create effective learning games," eLearning Industry 13 of June 2016 article,

https://elearningindustry.com/5-game-elements-create-effective-learning-games, accessed March 2017.

[3] J. Maloney, M. Resnick, N. Rusk, B. Silverman, and E. Eastmond, "The Scratch programming language and environment," in: ACM Trans. Comput. Educ., vol. 10, no. 4, article 16, pp. 1-15, November 2010.

[4] Website of the Scratch programming language, https://scratch.mit.edu, accessed March 2017.

[5] D. Weintrop, "Minding the gap between blocks-based and text-based programming," in: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 720. New York, NY, USA: ACM.

[6] B. DiSalvo, "Graphical qualities of educational technology: Using drag-and-drop and text-based programs for introductory computer science," in: IEEE Computer Graphics and Applications, vol. 6, pp. 12-15, 2014.

[7] S. Crook, "Embedding Scratch in the classroom," in: International Journal of Learning and Media, vol. 1, no. 4, pp. 17-21, 2009.

[8] The Edugames website, http://edugames.finki.ukim.mk, accessed March 2017.

[9] M. Jovanov, E. Stankov, M. Mihova, S. Ristov, and M. Gusev, "Computing as a new compulsory subject in the Macedonian primary schools curriculum," in: 2016 IEEE Global Engineering Education Conference (EDUCON), pp. 680-685. IEEE, 2016.

Students' attitude towards learning

Mirjana Kocaleva, Aleksandra Stojanova, Natasha Stojkovikj, Biljana Zlatanovska, Blagoj Delipetrev
Faculty of Computer Science
"Goce Delcev" University
Shtip, Macedonia
{mirjana.kocaleva, aleksandra.stojanova, natasa.maksimova, biljana.zlatanovska, blagoj.delipetrev}@ugd.edu.mk

***Abstract*—Education is an important segment in every society. Today we are witness of growing number of students at universities across the country. However, more important than the number of students is the question of how those students are committed and how much time they spend for learning. Sometimes because students have more than one partial exams in one session they are not sufficiently prepared for the exams. So our goal as teaching staff is to discover whether students achieve better results if we allow them to make corrective partial exam or not? For this purpose, we chose the subject Digital logic and we analyzed the results of partial exam and corrective partial exam in the academic year 2016/2017.**

Keywords—education; students' exam; statistical analysis

I. INTRODUCTION

Education is a process of acquiring knowledge and individual development. Education is acquired in educational institutions like schools and colleges. The process of education starts in primary schools then goes on in high schools and ending with colleges. The process of education is arduous and long, but after completing, a particular trade individual is able to find a job and to apply the gained knowledge.

There are two types of education: formal and informal. Formal education is classroom based and is implemented by teachers and this type of education is conducted every day from Monday to Friday. Otherwise, informal education is education outside the classroom without teachers. This kind of education can be provided at every place (home, library…) and can last from one to seven days per week at every time of day. Informal education is also known as electronic learning or e learning.

Education is one of the most important priorities as in every society also in our country. For this purpose the number of universities in the country increase, thereby the number of students who enroll in them also increase. Today education is available for all those who want to gain greater knowledge throughout life. More important than the number of students is the question of how those students are committed and how much time they spend for learning.

II. SUBJECT OF RESEARCH

Sometimes because students have more than one partial exams in one session they are not sufficiently prepared for the exams. So our goal as teaching staff at "Goce Delcev" University (UGD) is to discover whether students achieve better results if we allow them to make corrective partial exam or not? For this purpose, we chose the subject Digital logic and we analyzed the results of partial exam and corrective partial exam in the academic year 2016/2017. The maximum score for partial exam is 25 points [1], [2], [3], [4].

The subject Digital logic is one of the mandatory courses in II semester. The aims of this course are introduction to basic concepts of Boolean algebra and logic circuits that are an integral part of computer systems. This course is also the basis for learning subjects in the coming years.

In this paper, research is aimed at considering the knowledge of students in the subject Digital logic. For statistical analysis and data processing we used MegaStat. MegaStat is Excel add-in that performs statistical analyses with an Excel workbook. It performs basic functions, such as descriptive statistics, frequency distributions, and probability calculations as well as hypothesis testing, ANOVA, regression, and more [5].

The number of tested students is 63 (46 male and 17 female). Corrective partial exam is aimed for students with less than 11 points. On the partial exam from 17 female there are 4 with better results, and after the partial exam and corrective exam the number increase to 7, or from 23.53% the number increase to 41.18%. From the column for male can be seen that also the number for students with better results increase from 4 to 12 or from 8.7% to 26.09%. (Table 1)

TABLE I. OBTAINED RESULTS

Results	Score	Number of male	Percent of male	Number of female	Percent of female
Partial exam (63)	From 0 to 10	42	91.3 %	13	76.47 %
	From 11 to 25	4	8.7 %	4	23.53 %
Corrective partial exam (42)	From 0 to 10	34	80.95 %	10	76.92 %
	From 11 to 25	8	19.05 %	3	23.08 %
After partial exam and corrective partial exam	From 0 to 10	34	73.91 %	10	58.82 %
	From 11 to 25	12	26.09%	7	41.18 %

III. Analysis of results gained of partial exam and corrective partial exam

The results obtained from the partial and corrective partial exams will be processed with statistical data analysis. Statistical data analysis will be performed by

a. Descriptive Statistics,

b. Frequency Distribution and

c. Probability.

A. Descriptive Statistics

Descriptive statistics give us information about count, mean, sample standard deviation, sample variance, minimum and maximum grade and range for population. Also give details about population variance, population standard deviation and standard error of the mean. (Count is the number of students who have this subject. Mean represent the mean score for the students, or average score. Range means distance between the largest and the smallest score.) A description of population is within 1st and 3st quartiles (quartiles split the data into four sections). In addition, information for a median, mode, extremes and outliers are given. (The median is the middle number of a set of data and the mode is the most frequently occurring number, or score in our example. Outlier is a number that is not close to the other numbers in a sample.) Descriptive statistics for our tested students before and after the corrective partial exam is presented in Table 2.

TABLE II. Descriptive Statistics

Descriptive statistics	Partial exam	Corrective partial exam
count	63	63
mean	6,437	8,302
sample standard deviation	4,000	4,466
sample variance	16,004	19,948
minimum	0	0
maximum	16	23
range	16	23
population variance	15,750	19,631
population standard deviation	3,969	4,431
standard error of the mean	0,504	0,563
1st quartile	3,250	6,000
median	6,000	8,500
3rd quartile	8,750	10,750
interquartile range	5,500	4,750
mode	3,000	10,000
low extremes	0	0
low outliers	0	0
high outliers	0	2
high extremes	0	0

The number of tested students is 63. Each student can get the minimum score 0 and the maximum score 25 in depends of their knowledge. The mean of their grades is 6.437 after the partial exam and 8.302 after the corrective partial exam. The range between a maximum and a minimum grade is 16 after the partial exam and 23 after the corrective partial exam. This means that there is a big difference between the students, or there are students who learn a lot, and students who do not know anything. The sample variance as average square deviation from mean is 16.004/19.948 with the sample standard deviation of 4/4.466. That means that the standard deviation is very low and that is good because there are no big deviations from the mean score.

For 1st quartile for partial exam, the grade is 3.250 and for 3rd quartile the grade is 8.750. For partial exam the median is 6 with mode 3. For 1st quartile for corrective partial exam, the grade is 6 and for 3rd quartile the grade is 10.750. The median value is 8.5 and the mode is 10. The interquartile ranges are 5.5 and 4.750 respectively. The interquartile range is distance between first and third quartile. The low and high extremes are zeros. This shows that we did not have extreme. The low and high outliers are zeros when we do partial exam, but when we do corrective partial exam we have high outliers 2.

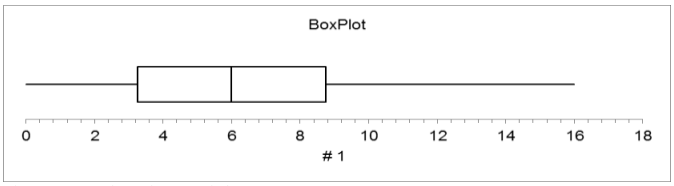


Fig. 1 BoxPlots for partial exam

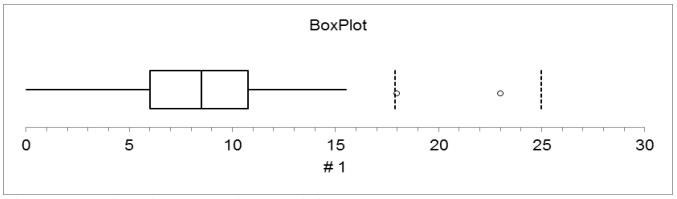


Fig. 2 BoxPlots for corrective partial exam

BoxPlots present a graphical means of summarizing data. In Fig.1 and Fig.2 are presented BoxPlots before and after the corrective exam. It shows the minimum and maximum score, 1st and 3rd quartile points and the median score. In Fig.2 for corrective partial exam besides minimum and maximum score, 1st and 3rd quartile points and the median score are presented also the two outliers in the point 18 and 23 (marked with ° on the graphic).

B. Frequency Distribution

Frequency distribution of a particular observation is the number of times the observation occurs in a set of data. Frequency distribution can be represented in a graphical or tabular format. Distribution displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst and the intervals must be mutually exclusive and exhaustive. Some of the graphs that can be used with frequency distributions are histograms, line charts, bar charts and pie charts. Frequency distributions are used for both qualitative and quantitative data. Frequency distributions are typically used within a statistical context [6], [7].

In the paper, we use histograms (Fig. 3 and Fig. 4) to present frequency distribution for both partial exam and corrective partial exam separately.

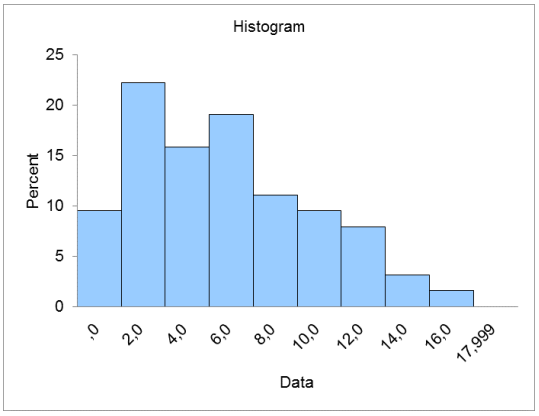


Fig. 3 Histogram for partial exam

The results from Fig. 3 show that the smallest percent we have for the students who obtained more than 16 points. Students have not obtained the score greater than 18. The high percent or over 20% of student have between 2 and 4 points. These results are not good because we have maximum score 25, but we do not have many students with more than 12.5 points (one-half from 25).

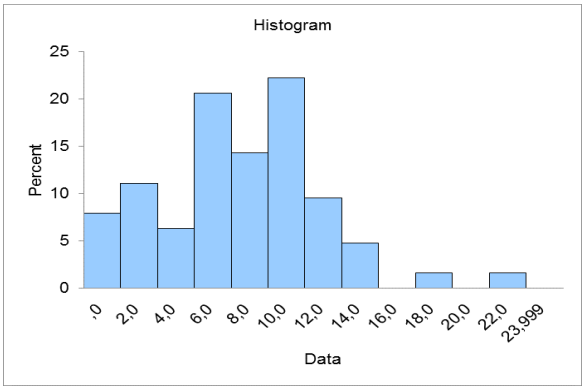


Fig. 4 Histogram for corrective partial exam

From Fig. 4 we can see that the smallest percent we have for the students who obtained between 18 and 20 points and between 22 and 24 points. The high percent or over 20% of student have between 10 and 12 points. There are not students who obtained between 16 and 18 points and between 20 and 22 points. These results are better than the results obtained for partial exam (Fig. 3).

From Table 3 we can conclude that when we talk about partial exam we have a negative growth, or there are larger number of students with fewer points and fewer students with more points. Respectively, we have 51 student with score from 0 to 10, and only 12 with score greater than 11. There is no student that have more than 20 points. In case when we have corrective partial exam we can see that the large number of student have between 6 and 10 points and between 0 to 5 and 11 to 15 we have almost the same number of students. Not so good thing here is the score greater than 16 points. The number of students is too small. We have only two students with score greater than 16, and only one student with score greater than 20 after the corrective exam. This fact is worrisome.

TABLE III. FREQUENCY DISTRIBUTION

		number of students	
group	score	Partial exam	Corrective partial exam
1	from 0 to 5	26	15
2	from 6 to 10	25	29
3	from 11 to 15	11	16
4	from 16 to 20	1	2
5	from 20 to 25	0	1

C. *Probability*

In this section we use probability as part of statistic. Probability is the measure of the likelihood of a given event's occurrence. Probability is always a number between 0 and 1. The higher the probability of an event, the more certain that the event will occur. P (A) represents the probability of event A.

Our goal is to calculate following conditional probability:

1. Probability for student to pass, if the student is female
2. Probability for student to pass, if the student is male

For that purpose we define the events

A – the student is female

B – the student is male

C – the student pass the exam

The probability of events A, B and C are:

$$P(A) = \frac{17}{63} = 0.267$$

$$P(B) = \frac{46}{63} = 0.73$$

AC present the event – student is female and student pass the exam. The probability is

$$P(AC) = \frac{7}{63} = 0.111$$

BC present the event – student is male and student pass the exam. The probability is

$$P(BC) = \frac{12}{63} = 0.19$$

Hence, a conditional probability for student to pass the exam if the student is female, i.e. male are:

$$P(C/A) = \frac{P(AC)}{P(A)} = \frac{0.11}{0.267} = 0.412$$

$$P(C/B) = \frac{P(BC)}{P(B)} = \frac{0.19}{0.73} = 0.26$$

From these can be concluded that the probability to pass the exam is bigger for female.

For the subject Digital logic two partial exams are conducted. Each exam maximum score is 25 points and consist of 25 questions. For the student to pass the exam he or she must to have minimum 10 points from one exam, or 20 from the both. Also after the partial exam, there is a final exam, which defines the grade.

Student does not pass the subject when he/she receive grade 5 (five). If the student receive grade more then 5, then the student passed the exam. The lowest grade for student's to pass the exam is 6 (six) and the highest grade is 10 (ten). The grade depends of students' desire for learning and demonstration of gained knowledge.

For one exam to be passed minimum 10 points are needed. To be able to answer the question whether according to points gained from the first partial exam student has a chance to pass the subject Digital logic, we review the results of the corrective partial exam. From Table 3 we have that the chances for students to pass is 30% (19/63=0.3) or probability of 0.3, and the chances for students not to pass is 70% (44/63=0.7) or 0.7.

IV. CONCLUSION

From data statistical analysis we can conclude that better results are obtained when we have a corrective partial exam. The mean points when we have a corrective partial exam is 8.302 and is bigger than 6.437 gained for partial exam. This mean that corrective exams are good for those students who want to correct their points and to have a high grade for the subject. Basic challenge for teaching staff is and will be finding the new methods and the new ways for better motivation of the students, in order for acquisition of more knowledge and skills.

REFERENCES

[1] M. Kocaleva, B. Zlatanovska, A. Stojanova, A. Krstev, Z. Zdravev, E. Karamazova, "Analysis of Students' Knowledge for the Topic Integral", International Conference on Information Technology and Development of Education – ITRO 2016, Zrenjanin, Republic of Serbia, ISBN: 978-86-7672-285-3, pp.155-158, June, 2016.

[2] B. Zlatanovska, M. Kocaleva, A. Krstev, Z. Zdravev, "E - testing against classical testing in subject Mathematics". Yearbook of the Faculty of Computer Science, 4 (4). pp. 29-32. ISSN 1857- 8691, 2015.

[3] A. Stojanova, B. Zlatanovska, M. Kocaleva, V. Gicev, "On the Use of Mathematica in Engineering Education", International Conference on Information Technology and Development of Education – ITRO 2015, June, Zrenjanin, Republic of Serbia, ISBN: 978-86-7672-258-7 pp.103-108, 2015.

[4] M. Kocaleva, N. Stojkovikj, A. Stojanova, A. Krstev, B. Zlatanovska "Improving on teaching curriculum of Calculus 2 at technical faculties", IEEE EDUCON 2017, April 26-28, 2017. (In press)

[5] J. B. Orris, "MegaStat for Microsoft Excel, Windows and Mac", 2017.

[6] "Frequency Distribution", available online at http://www.investopedia.com/terms/f/frequencydistribution.asp

[7] "Frequency distribution" , available online at https://en.wikipedia.org/wiki/Frequency_distribution

GIS Digitalization of the Infrastructure of Public Buildings: A Case Study of the "Boris Trajkovski" Sports Center"

Angjelkoski Antonio, Andreja Naumoski, Kosta Mitreski, Georgina Mirceva
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
angjelkoski.antonio@students.finki.ukim.mk, {andreja.naumoski, georgina.mirceva, kosta.mitreski}@finki.ukim.mk

***Abstract* — Geographic Information Systems (GIS) is a powerful analytical software for computer mapping and spatial data analysis. This tool allows users to save, analyse and present geographical spatial information's on a map. ArcGIS Desktop software is one of the most popular GIS spatial analytical software and consists of several software packages used in many disciplines. The aim of this work is digitalization of the multi-purpose sport center "Boris Trajkovski" in Skopje, Republic of Macedonia using three ArcGIS Desktop packages: ArcMap, Arc Catalog and Arc Toolbox. GIS digitalization is used not only for the sport object itself, but also for the entire complex and its surroundings. The entire sport center consists of Aqua Park, Sport Hall, Water Poll, Skating, Tennis Courts, Carting Road and Parking space. The result is a GIS map of all these object, where the management team can easily solve and improve any problems that may arise. For future work, we plan to make mobile application for easy public and private access.**

Keywords — ArcGIS, Sport Complex, data management, data analytics, object mapping.

I. INTRODUCTION

GIS software package provides practical tools for visualizing and analysing data like hotel information or bank information, revealing trends, dependencies and inter-relationships. Three major packages are included in the ArcGIS Desktop software: ArcMap – software application used for data manipulation, map analysis and publishing; Arc Catalog – package for storage and data management; Arc Toolbox – consists of tools that helps in data processing, geoprocessing, data conversion, interoperability, managing coordinate system and projections. Using ArcGIS the users can acquire, store, manage, and geographically integrate large amounts of information from different sources of data and other programs from different sectors. Integrated information in this way can provide better inside look of the object itself. Another advantage of the GIS is the multi-platform option, data can be presented locally, in the web or on mobile platforms. The research in the relevant literature shows numerous and important case studies about the applicability of sport related activities [1, 2, 3]. All these research paper research the sport activates, which include the sport complex itself not only in urban areas, but also in areas outside of big cities. Additionally, there is a GIS analytics of sport objects inside of urban areas [4], where researchers take into account the need of access, problems in urban planning and managing big events.

The structure of the paper is organized as follows: Section 2 presents the case study of the digitalization of the entire sport complex, while Section 3 concludes the paper and gives direction for future work.

II. METHODS AND MATERIALS

In order to make digitalization possible, we first create a geodatabase in order to easy manage the geo information data. The creation of the digital data follows several steps using Arc Catalog tool. The first step is to create new geodatabase with custom name. In it, we add the feature class, while we are careful when we create the specific type using the coordinate system that is same as the base map. After adding the feature classes, we create tables for each of these classes. For some of the sport complex objects, no additional information is needed, but for some we add additional information.

The digitalization process begins with marking of the target area borders. Each map in ArcMap editor consist from several layers that links the geographical location with a database. For this purpose, we create feature class with type line in the geodatabase. Then in the editor toolbox we activate the editing process by selecting "Start Editing" and we start editing the map. Using Create feature window, we can choose with what tool we gone draw on the map. After this we add the rest of the future classes in the "Table of Content", so with this we continue to draw on the map for each specific feature on the map. For example, if we want to draw street and roads on the map we create feature class from the type Line, for drawing parcels and objects we use feature class type Polygon and for marking the Trees that surround the sport complex we use feature class type Point. If we want to create constructions that are more complex, we draw more objects and then join them all together into one using tools like "Merge". When drawing more complex object, handy tool is "Snapping", which allows the user more robust manipulation by connecting the lines and polygons drawn on the map. When we encounter overlapping on some of the layers, we use different schedule by using the tool "List by Drawing" in table of content.

One more interesting feature drawn on the map is the street infrastructure. In order to create the streets and roads with particular size we use the "Buffer" tool. In order to use this tool, we first select all the objects from the layer "Ulici" in order to get the entire street infrastructure. Layer "Ulici" consist from information reading the street infrastructure around the sport centre. Using this tool, we can connect all the

lines (streets, roads, pedestrian lines) into one entire street infrastructure. The "Buffer" tool can be easily found using "Search" window, by typing the word buffer in the search widow. More information about how to use this tool, the users can get by selecting this tool from the window and moving the mouse to the word. In the dialog window, we select the layer which we want to use the "Buffer" tool, the output layer is automatically created. This layer is then automatically added on the left side on the main window. In the Linear Unit filed we enter the value which we want to expand the main roads and mark them on the map. We enter 3 meters for the main roads, and 1 meter for the pedestrian road. After this in the field "Dissolve", we select all and we wait the ArcGIS to finish the operations. After finishing the operation, the new layer can be found in the table of content.

When we finish editing and mapping the objects on the map, using the tab "Symbology" from the Layer properties we change each layer. In this way, we change the object marking, symbol, colour, shape and size in order to make the object more noticeable on the map. When we want to add additional data for the additional objects or marked space, we have created tables in which we enter the needed information's (number of parking spaces, the capacity of the sport object or number of visitors in AquaPark and etc.) and then using "Join" and "Relate" we join the tables and the objects which we need. After finishing adding and modifying the map and objects, we need to tell the ArcMap to save the map. This is done using Editor Toolbar "Save Edits" and with this, we finish the editing.

A. Tree maintenance example

Many of the activities around and in the sport complex require maintenance. For record keeping, regarding the trees condition that are planted in the sports centre, ArcMap provide very easy and effective solution. First we made additional table where we log and control and condition of the trees around the sport complex. To address that problem, we created geodatabase and for each tree (point) we add additional information. This makes possible to get information fast by only a click on each tree. When a user clicks on a particular tree, a window pops up with additional information presented. The existing data can be upgraded with new data very easily once we set the geodatabase. Beside the textual information that can be entered, ArcMap allows user to add photographs, hyperlinks or any other additional information for much easier solution of the problem. Also, the users can find fields, where it can perform actions to solve the problem. Additionally, it is possible to apply the same concept of records for other areas of everyday life, such as maintenance of traffic lights, maintenance of the benches in parks, maintenance of street lighting and many others.

To add new records, the user can select "Attributes section" of the menu editor, right-click the field "Odrzuvanje" and then the user adds new one. Once the field is added, the field is selected and insert new data records. We also allow the maintenance team to add date of Control, which made control of the status of the tree and if notes are necessary to be add, and then user can inserted them into the field "Zabeleshki". This completes the recording of new data to the tree and wait to take action by those responsible for the problem. The final map is given in Fig. 1. This map can be exported in any format that well most fit for our needs. While it looks nice and understandable to all, the GIS software allows great range of tools that enables fast processing area and display it in an interesting way.

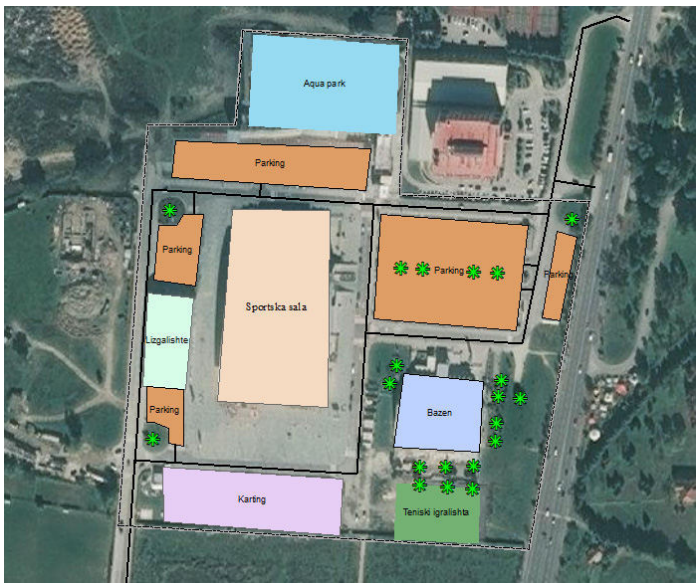


Fig. 1. Digitalized map of sport complex "Boris Trajkovski"

III. CONCLUSION

In this paper, we digitalized the entire sport complex, and not only the main sport hall, but also the additional sport object surroundings. In details, we have described the data collection through the process of making the maps and later through digitalization into final geographical map. We additionally add option to set and response in maintenance need by the management. We plan to further improve the presented system by making mobile application.

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] Y. Jixue, J. Liu, F. Xu, C. Jing. "Management System of Modern Great Sport Events Transportation Based on GIS." In *International Conference on Transportation Engineering 2009*, pp. 2668-2673. 2009.

[2] L. Zhu. "The design of GIS-based information database of sports competition." In *Environmental Science and Information Application Technology (ESIAT), 2010 International Conference on*, vol. 2, pp. 715-718. IEEE, 2010.

[3] G. Yang, X. Xu. "Sports Games Management System Based on GIS." In *Informatics and Management Science VI*, pp. 673-678. Springer London, 2013.

[4] H. Liu, Z. Yangyang, X. Yannan, L. Xinyue, X. Mingrui. "Evaluation of Accessibility to Urban Public Sports Facilities: A GIS Approach Based on Network Analysis Model." In *Information and Computing Science (ICIC), 2012 Fifth International Conference on*, pp. 52-55. IEEE, 2012

Detection of Very Weak Radio Pulsar Signal

Ivan Garvanov
University of Library Studies and Information Technologies
Sofia, Bulgaria
i.garvanov@unibit.bg

Stoyan Vladimirov
University of Library Studies and Information Technologies
Sofia, Bulgaria
stoyanvladimirov@yahoo.com

***Abstract* — Pulsars are rotating neutron stars that emit electromagnetic radiation at regular intervals and can be used for navigation. The detection of a pulsar signal for a short time (in real time) is difficult because the signals are very weak. In the paper we research one detection algorithm which includes three basic stages: epoch folding, moving average filter with a jumping window and CFAR detector. The algorithm proposed in the paper was verified with pulsar signals from Jodrell Bank Centre for Astrophysics.**

Keywords—Pulsar signal detection, navigation

I. INTRODUCTION

Pulsars are fast rotating neutron stars (Fig. 1) that periodically emit broadband electromagnetic pulses [1]. The emission period is thought to be the same as the rotation period. Although individual pulsar pulses vary in strength and shape, the average pulse shape is stable and characterizes each pulsar.

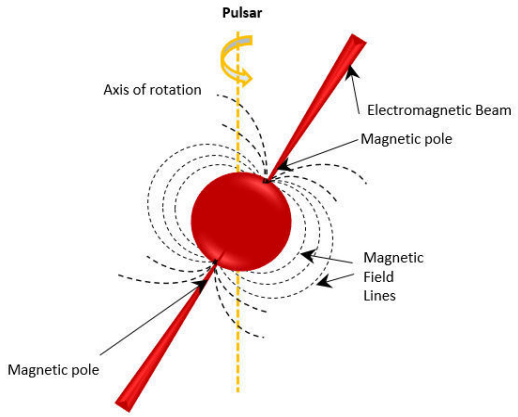


Fig. 1. Pulsar

The idea of using pulsars, rapidly rotating neutron stars, for orientation in space is not new [2], [3], [4]. It is similar to the oldest idea - navigation of ships centuries ago through observation of the visible stars with sextants and using the star and sea charts. This approach is similar, but the difference is in the reception of radio signals instead the light emission of stars.

A similar approach for navigation by satellites is in GNSS systems. This requires to search very specific, fast and effective methods and algorithms for detection and estimation of their parameters.

Earlier, the practical realization of the idea of space navigation was difficult, first, the insufficient number of known pulsars, and secondly - sophisticated technology to detect them. But in recent years the situation has changed significantly. Since the discovery of the first pulsar in 1967, approximately 1900 pulsars have been found.

The main difficulty in detecting a signal from the pulsar by radio telescopes is the low Signal-to-Noise Ratio (SNR) at the receiver input (from -40 dB to -90 dB).

Another difficulty in the study of the pulsars is a great consumption of time needed for detecting the signal from them, about 1-2 hours [5], [7], [8], [9], [10].

Since each pulsar has a unique period, in [4] is applied epoch folding algorithm to shape the pulsar pulse, remove noise, and find the pulsar. Folding is similar to integration except that in folding, the data is broken into a sequence of discrete intervals corresponding to the period of the expected pulsar and then added (or folded) ensuring that the pulsar signal is reinforced with each fold, while the noise approaches a mean zero. The epoch folding method is convenient, but the integration time is too much. It is equal to the number of period of repetition of the signal from the pulsar multiplied by the length of the period.

In the paper, we will discuss the possibility to improve the signal to noise ratio by using Moving Average Filter with a Jumping Window (MAFJW) in time domain signal. As a result of this processing, the number of samples in the record will be reduced in proportion to the number of cells in jumping window. The small number of samples will increase the further signal processing.

II. SIGNAL PROCESSING

The considered algorithm of signal processing of the pulsar signals experimental data is shown in Fig. 2.

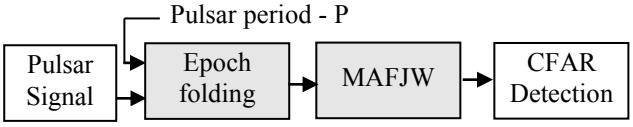


Fig. 2. Block-scheme of signal processing

It includes the following stages: epoch folding of data during N repetition periods of the input data; filtration by the Moving Average Filter with a Jumping Window (MAFJW); estimation of SNR at filter output and, finally CFAR detection.

A. Epoch folding

Most pulsars emit pulses that are too weak to detect individually. Nevertheless, the periodicity of the signals emitted by pulsars makes possible to discover thousands of pulsars, most of which are too weak to yield individually-distinguishable pulses. If the period P of a particular pulsar is known, than the pulsar's average pulse shape (pulse profile) can be determined using the epoch-folding procedure [1], [2], [3], [4]. Epoch-folding is similar to integration except that in folding, the data is broken into a sequence intervals corresponding to the period of expected pulsar and then added. When the number of integrated periods grows, the pulsar signal reinforces with each integrated period while the noise approaches to a zero mean. The standard way implemented in most of radio observatories is to integrate the power of the input signal during K sequential periods. In result of epoch-folding, the output signal y at time discrete n is formed as:

$$y[n] = \frac{1}{K}\sum_{k=1}^{K}\left[I_{s,k} + N_{1,k}(0,\sigma^2)\right]^2 + \\ + \frac{1}{K}\sum_{k=1}^{K}\left[Q_{s,k} + N_{2,k}(0,\sigma^2)\right]^2 \quad (1)$$

In (1), $I_{s,k}$ and $Q_{s,k}$ are the quadrature components of the received pulsar signal in the k-th repetition period, $N_{1,k}$ and $N_{1,k}$ are quadrature components of the receiver zero mean Gaussian noise with variance σ^2 in the k-th period. As follows from (1.8), the output signal $y[n]$ is distributed according the non-central chi-square law with 2K degrees of freedom and the non-centrality parameter:

$$m = \frac{1}{\sigma^2}\sum_{i=1}^{K}(I_{s,i}^2 + Q_{s,i}^2) \quad (2)$$

The relation between the signal –to-noise ratio (SNR) and the integration time $t_{int}=K\cdot P$ during the epoch-folding process can be determined by the following equation given in [1]:

$$SNR = \sqrt{n_p t_{int} B}\left(\frac{A_{peak}}{T_{sys}}\right)\frac{\sqrt{W(P-W)}}{P} \quad (3)$$

Where B is the signal bandwidth, A_{peak} is the pulse peak amplitude, T_{sys} – is the system noise temperature, P is the pulse period, W is the pulse width, and n_p is 1 for single polarization observation or 2 if two orthogonal polarized signals are summed, respectively. The equation (3) can be written as a product of three factors:

$$SNR = \sqrt{K}\cdot\left(\frac{\sqrt{B n_p}}{T_{sys}}\right)\cdot\left(A_{peak}\sqrt{\frac{W(P-W)}{P}}\right) \quad (4)$$

The first factor accounts for the effect of the signal processing (epoch-folding), the second factor accounts for the effect of the receiver parameters, and the third factor accounts for the effect of the pulsar signal parameters.

From (4) can be concluded that using of an antenna with a larger aperture, selecting a frequency band with less background noise, and applying advanced signal processing techniques we will improve the integration time in order to achieve the required SNR for successful detection of pulsar signals.

B. Filtering by the MAFJW

The aim of this study is to examine the possibility to increase the signal to noise ratio of the received pulsar signal by means of one modification of a Moving Average Filter, which uses the Jumping Window (MAFJW). It takes N samples of input at a time and take the average of those N-samples and produces a single output point. It is a very simple structure that comes handy for scientists and engineers to filter an unwanted noisy component from the input data.

$$y[n] = \sum_{k=-N}^{N}\frac{1}{2N+1}x[n-k], \\ n=N+1,\ 2N+1,\ 3N+3,..,\ MN+1 \quad (5)$$

Where n contrary to the traditional Moving Average Filter (MAF) where the number of output samples is equal to the number of input samples, the MAFJW reduces the number of output samples N times, where N is the length of the Jumping Window. The number of input samples is MN, and the number of output samples is M. Therefore in contrary to the traditional Moving Average window, the MAFJW acts not only as a low-pass filter but a decimator as well. When the signal processing is carried out in the time domain, the use of the MAFJW can be very useful in the sense of reducing the processing time.

C. CFAR Detection

The CFAR detection approach is based on the criterion of Neyman – Pearson. According to this criterion, the following algorithm can be used for testing a simple hypothesis H_1 (pulsar signal is present) against a simple alternative H_0 (pulsar signal is absent):

$$H_1:\ if\ \max_n\{P_z[n,K]\} \geq T_{fa}\sum_{l=1}^{L}P'[l,K] \quad (6)$$

$$H_0:\ otherwise$$

It is important to note that the CFAR approach requires to use two zones (for two hypothesis), the noise zone - for setting the threshold, and the signal zone - to be compared with the threshold for signal detection. In radar is always possible to choose these areas because the antenna scans in the space around the target.

In processing systems of pulsar signals for navigation, two variants of the implementation of CFAR detectors are possible: traditional - we have two areas (for noise and for signal) and non-traditional - only one zone is available, which is always a mixture of signal and interference. In the first case,

the antenna must be directed not only at the direction of the pulsar, but also at the direction of interference.

In that case the detection constant T_{fa} is determined in accordance with the probability of false alarm, which should be maintained by the detection algorithm. We assume that the alternative hypothesis H_1 is verified only in one single sample. Using the search strategy "Maximum" the probability of false alarm is defined as:

$$P_{FA} = 1 - \left[1 - P_{fa}\right]^N \tag{7}$$

In (7), N is the total number of samples, and P_{fa} is the probability of false alarm in a single sample. In case of mean zero Gaussian noise [11], [12], the probability of false alarm to be maintained in a single sample is:

$$P_{fa} = \frac{1}{(1+T_{fa})^K} \tag{8}$$

The probability P_{fa} is defined as a solution of the equation (7) for a given the probability false alarm P_{fa}. The solution of (8) gives the detection constant:

$$T_{fa} = P_{fa}^{-1/K} - 1 \tag{9}$$

Because the first variant of the CFAR detector implementation complicates the antenna system for tracking of pulsars at the airplane movement, we choose the second variant of the implementation of CFAR detectors: the antenna is directed only towards the pulsar, i.e we have only one zone, which is a mixture of signals and interference. We provide a variant of the CFAR detector, which is not conventional, as it uses the mixture "signal + interference" for evaluating interference and determining the detection threshold.

In the decision rule (6), $P_z[n,K]$ - is the signal power in the target zone (Fig. 3).

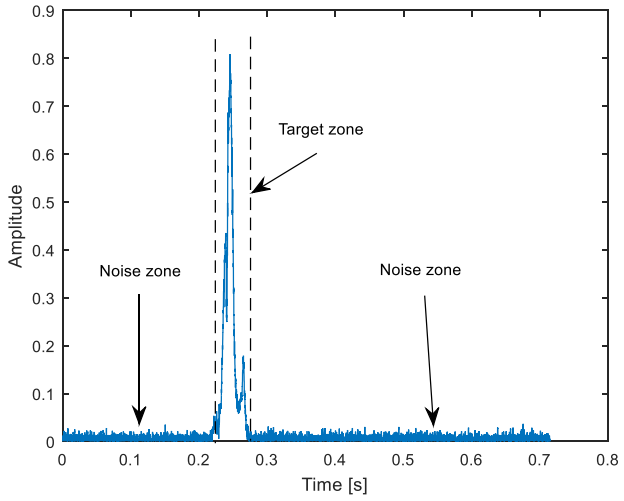


Fig. 3. Pulsar signal

The target zone is determined according to the position of the global maximum of the signal power (*ID)*. In this zone, *n* varies inside the interval [ID-Δ, ID-Δ]. In order to separate the target and noise zones all signal samples are sorted in the ascending order (Fig. 4).

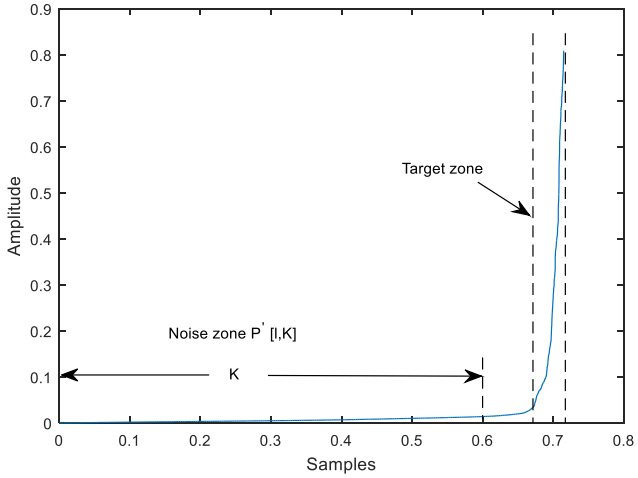


Fig. 4. Sorting of samples in the ascending order for separation of the target and noise zones

In (6), the first K samples P'z[*l,K*] are the signal power samples located in the noise zone, where *l* varies outside the interval [*1, K*], and *K* is the size of the noise zone. The last (2Δ) elements in the assorted samples form the target zone (Fig. 4). The width of the target zone (2Δ) is determined according to the template of the pulsar.

III. Experimental Results

In this study, the experimental records of the signal received from the pulsar B0329+54 provided by Jodrell Bank Centre for Astrophysics is used.

This is the brightest radio pulsar in the northern sky. Otherwise this pulsar is a typical, normal pulsar, rotating with a period of 0.714520 seconds, hence the star makes about one and a half turn in a second, giving it a locomotive kind of sound. We can see in Fig. 5 that each pulse has a different structure, hence the beam of this cosmic lighthouse is constantly changing in shape. This recording has been made with the Lovell telescope in Jodrell Bank.

The signal on Fig. 5 is obtained after signal processing procedure which is about 2 hours. When the SNR is about - 20 dB, the pulsar signal will be as on Fig. 6.

The output of the epoch folding procedure where the received signal is integrated during 84 periods (60 s) and without MAFJW is shown in Fig. 7. When we use moving average filter with jumping window (MAFJW) the output SNR increases (Fig. 8, 9 and 10). The filter output for length of the jumping window N=10 is shown in Fig. 8, for N=40 in Fig. 9, for N=100 in Fig. 10.

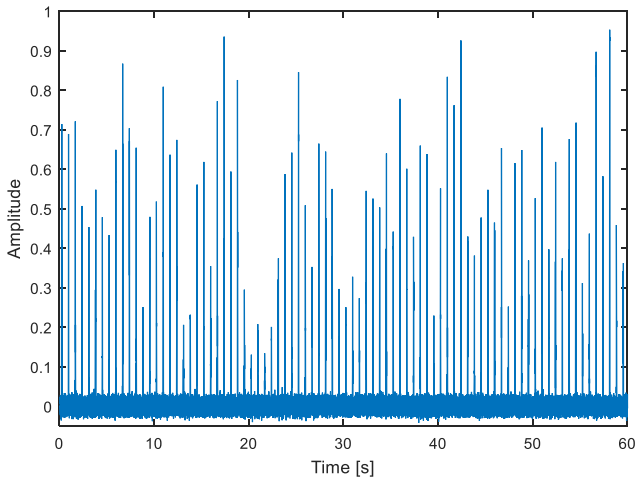


Fig. 5. Pulsar signal (B0329+54)

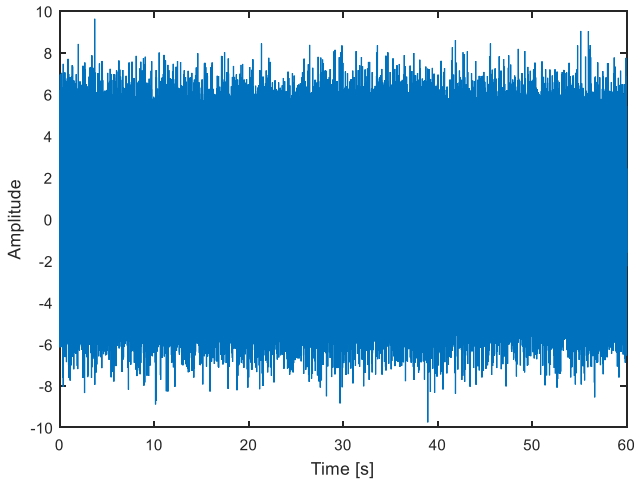


Fig. 6. Pulsar signal B0329+54 and noise (SNR= - 20dB)

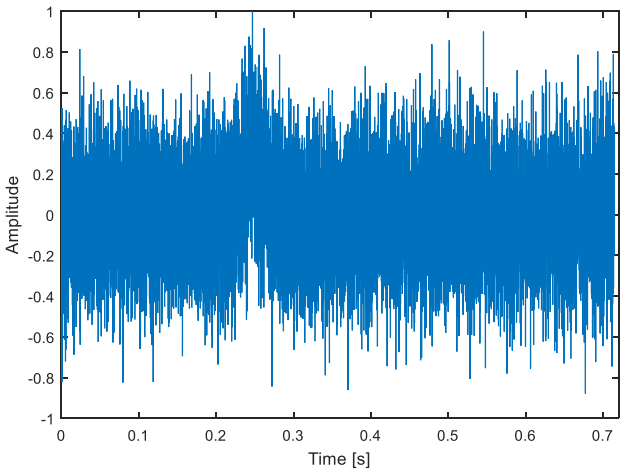


Fig. 7. Pulsar signal after the epoch folding (84 periods)

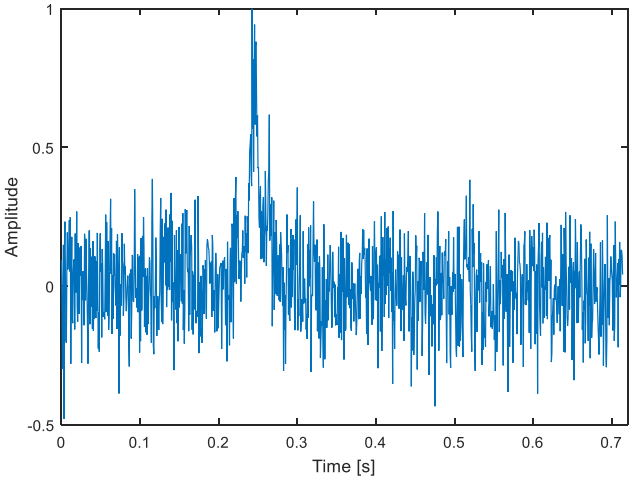


Fig. 8. Pulsar signal after the epoch folding (84 periods) and MAFJW (N=10)

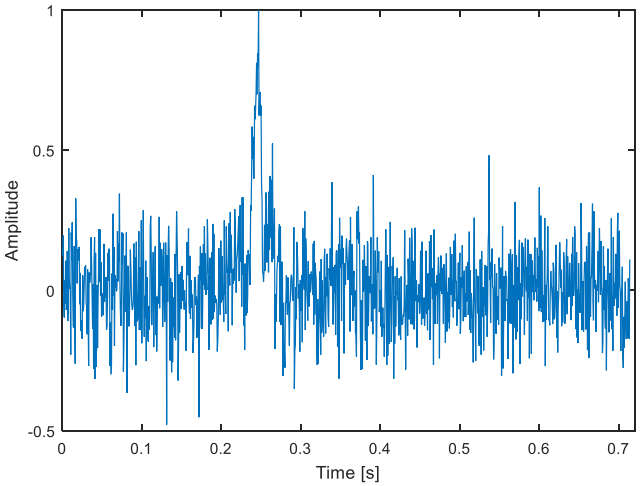


Fig. 9. Pulsar signal after the epoch folding (84 periods) and MAFJW (N=40)

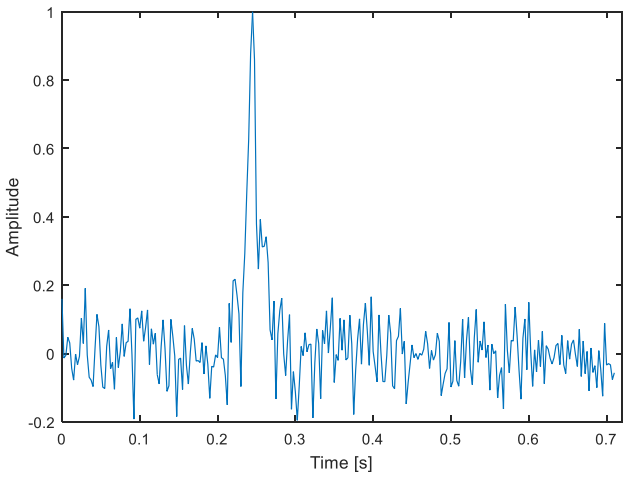


Fig. 10. Pulsar signal after the epoch folding (84 periods) and MAFJW (N=100)

As shown in Fig. 9 and Fig. 10, the key problem concerned with the MAFJW is the optimal choice of the jumping window length (N). When choosing the window length we must take into account not only the level of suppression of the noise variance, but the degree of distortion of the useful signal immersed in noise. The optimal window size of the MAFJW can be determined only in cases where the spectrum of the signal is known Δf and bounded by a certain frequency and the noise power does not exceed a certain level.
The length of a filter jumping window can be defined as:

$$N = f_s / f_{s,new} \tag{10}$$

Where f_S is the sampling frequency at the filter input (in our case f_s= 16 KHz) and $f_{s,new}$ is the sampling frequency at the filter output. Taking into account the Sampling Theorem, the sampling frequency at the filter output must satisfy the following inequality:

$$f_{s,new} \geq \Delta f \tag{11}$$

In (11), Δf is the frequency bandwidth of a pulsar signal. From [6] follows that the frequency bandwidth of the pulsar B0329+54 is about 400 Hz. Therefore the maximal length of the filter jumping window in our case is N_{max} =40.

IV. Conclusions

The obtained results show that the presented algorithm can be successfully used for processing and detection of pulsar signals. The SNR can be improved using the Moving Averaging Filter with a Jumping window. The improvement in SNR depends on the filter window. The length of the jumping window must be chosen carefully satisfying the Sampling Theorem at the filter output. The procedures of epoch folding and filtering can be reversed.

Acknowledgment

This work is supported by the project “Investigation of parameters, properties and phenomena of radio signals from pulsars and their interaction with objects”, DN 07/1 from 14.12.2016

References

[1] Lorimer D., M. Kramer, “Handbook of pulsar astronomy”, Cambridge university press, N.Y.,2005

[2] Sala J., et al., “Feasibility study for a spacecraft navigation system relying on pulsar timing information,” Tech. Rep. 03/4202, ARIADNA Study, June 2004.

[3] Buist, P., S. Engelen, A. Noroozi, P. Sundaramoorthy, S. Verhagen, C. Verhoeven, "Principles and Potential of Pulsar Navigation", 24 International ION Conference, 2011, Portland, USA.

[4] Buist P., S. Engelen A. Noroozi, P Sundaramoorthy, S Verhagen, C Verhoeven, „Overview of Pulsar Navigation: Past, Present and Future Trends“, Journal of the Institute of Navigation, vol.58, № 2, pp.153-164, 2011.

[5] White, N.E. et al., in X-Ray Binaries, Cambridge Astrophysics Series, p.1, 1995.

[6] Garvanov, I., Kabakchiev, Ch., Behar, V., Garvanova, M. The Experimental Study of Possibility for Pulsar Signal Detection. The Second International Conference “Engineering & Telecommunications – En&T 2016”, Noevember 28-30, Moscow-Dolgoprudny, Russia, 2016, pp. 68-72.

[7] Kabakchiev C., V. Behar, P. Buist, I, Garvanov, D. Kabakchieva, N. Gaubich, M. Bentum, “Study of CFAR Algorithms for Signal Acquisition in Radio Pulsar-Based Navigation”, 21st Saint Petersburg International Conference on Integrated Navigation Systems 2014, Saint Petersburg, Russian Federation, pp. 186-194, 2014.

[8] Kabakchiev C., V. Behar, P. Buist, I. Garvanov, D. Kabakchieva, M. Bentum, “Time of Arrival Estimation in Pulsar Based Navigation Systems”, Signal Processing Symposium, Debe, Poland, 2015.

[9] Kabakchiev C., V. Behar, P. Buist, R. Heusdens, I. Garvanov, D. Kabakchieva, “Detection and Estimation of Pulsar Signals for Navigation”, International Radar Symposium, Dresden, Germany, pp. 688-693, 2015.

[10] Garvanov, I., C. Kabakchiev, V. Behar, M. Garvanova, “Target detection using a GPS Forward-Scattering Radar”. IEEE Second International Conference “Engineering & Telecommunications – En&T 2015”, Moscow-Dolgoprudny, Russia, pp. 29-33, 2015.

[11] Izvorska D, “On the something application of axial and central symmetry to work on construct problems”, II International seminar Symmetry: theoretical and methodological aspects, Astrakhan (2007), 194-199.

[12] Slavova S., D.Izvorsk, “Using Maple in Linear Algebra”, International Scientific and Methodical Conference, Penza, (2008), 218-230.

THE INFLUENCE OF QUALITY CONTROL ON THE IMAGE RETRIEVAL: Application to Longitudinal Images for Alzheimer's Disease

Katarina Trojacanec, Ivan Kitanovski, Ivica Dimitrovski, Suzana Loshkovska, for the Alzheimer's Disease Neuroimaging Initiative*
Faculty of Computer Science and Engineering
Skopje, Macedonia
{katarina.trojacanec, ivan.kitanovski, ivica.dimitrovski, suzana.loshkovska}@finki.ukim.mk

***Abstract*—The purpose of the paper is to research the influence of quality control (QC) on image retrieval applied to longitudinal images for Alzheimer's Disease. In fact, structural Magnetic Resonance Images (MRI) used in Alzheimer's Disease studies usually undergo the fully automated image processing pipelines. This process is prone to errors that can bias the results of the subsequent examination. The main goal of the paper is investigated whether the quality control of the automatically processed images can impact the retrieval performance.**

The research includes all subjects from the ADNI database with MRI scans available for four consecutive time points, scans at baseline, and 6, 12, and 24 months later. Hence, 267 in total are selected. All scans are processed using Freesurfer's longitudinal stream. Measurements of cortical and subcortical brain structures such as volumes and cortical thickness of the brain regions were estimated and used as features. The retrieval results were evaluated with and without QC. According to the results, the QC phase significantly improves the retrieval. As a result, it can be concluded that QC influences the overall retrieval performance and is extremely important in this context.

Keywords—quality control; image retrieval; longitudinal images; Alzheimer's Disease

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) is crucial diagnostic technique in neuroimaging nowadays, due to its capability of containing extremely important medical information and noninvasiveness. As such, MRI plays key role in the clinical and research studies for plenty of neurological diseases, such as Alzheimer's, Huntington's disease, schizophrenia, cancer, and stroke. Moreover, performing longitudinal analysis of MRI derived data is crucial in this domain for monitoring progression of the neurological change and/or response to a treatment.

Wide variety of imaging markers are estimated using MRI. For that purpose, in most of the cases automated image analysis is performed using software suits such as Freesurfer [1], [2], [3]. One of the biggest challenges in this context is the level of quality. It can directly affect the derived imaging measurements and reduce reliability and validity of the decision making process or the performed study [4], [5].

Artifacts in MRI and foreign bodies within the patient's body may cause several problems. In fact, they may be confused with a pathology or may reduce the quality of examinations [6]. There exist technical artifacts such as head coverage, radiofrequency noise, signal inhomogeneity, and susceptibility. There are also motion artifacts including blurring and ringing [7], [8]. Motion artifacts are produced by the participant swallowing, blinking, chewing, turning, fidgeting, or repositioning a limb [9]. Hence, it is extremely important to provide quality control (QC) to detect acceptable level of image quality, to define the exclusion criteria and/or to perform correction to some extent [5].

The aim of this article is to analyze the usage of quality control for neurological MRI for Alzheimer's Disease in image retrieval study. In this context, MRI provide valuable and consistent imaging markers for diagnosis, understanding the disease pathology and monitoring the disease progression. They clearly reflect the structural brain changes closely related to the disease progression. Thinning of the cerebral cortex, ventricular enlargement, hippocampus shrinkage etc. are some of the examples [10].

To obtain such measurements, the images are processed using the fully-automated Freesurfer's pipeline. However, the automated procedures can result in variability in the accuracy of the performed processing resulting in failures [4], [5]. The focus of the research is to analyze whether these failures produce difference in the retrieval results. Moreover, the examination is performed in the longitudinal context, using MRI from multiple data point for each patient.

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Several research studies were performed on medical image retrieval for AD [11], [12], [13]. However, they do not address the QC. To overcome this, in this article the retrieval results were evaluated with and without QC. The hypothesis is that the QC will cause improvement in the retrieval performance.

II. MATERIALS AND METHODS

A. Participants and Inclusion Criteria

Data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The main focus of ADNI has been to test whether imaging modalities such as serial magnetic resonance imaging (MRI) and positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be used and/or combined to estimate the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

In this research, a total of 267 subjects from ADNI-1 standardized lists [14] were selected by using the following criteria:

- The patient belongs to AD or Normal Controls (NL) group;
- For each patient scans at baseline, and the 6-month, 12-month, and 24-month follow-ups are available.

B. Image Processing and Retrieval

The image retrieval involves generating feature vector that represents the image given as a query, and then comparing it with other feature vectors previously generated for the other images stored in the system. The feature vector in this research consists of volume and cortical thickness estimates for all four time points. To obtain these measurements, all images are processed with Freesurfer's software suite, which is documented and freely available for download online [15].

For this purpose, MRI MPRAGE T1-weighted 1.5T three-dimensional scans (or equivalent) acquired at regular (6-month or 12-month) intervals are used. The main goal is to get valuable imaging biomarkers reflecting the disease progression. For each subject all scans are performed on the same scanner because an analysis conducted on a subject scanned on different magnetic resonance equipment is likely to be technically inconsistent [14].

For all ADNI data, an initial QC process (which is not subject of investigation in this paper) is performed. Only data that pass the predefined criteria are selected for further analysis. The QC process at this phase includes a comparison of image acquisition parameters in the Digital Imaging and Communications in Medicine (DICOM) header against the expected protocol, a visual check of the image quality by an experienced image analyst, and a quantitative check of the geometric accuracy of the scanner by analyzing data acquired with the ADNI phantom [14], [16].

The data are available with different levels of preprocessing to reduce known image nonidealities, such as gradient nonlinearity, intensity inhomogeneity correction, and phantom-based distortion correction [14].

The Freesurfer's processing pipeline includes removal of non-brain tissue using a hybrid watershed/surface deformation procedure, automated Talairach transformation, segmentation of the subcortical white matter and deep grey matter volumetric structures, intensity normalization, tessellation of the grey matter white matter boundary, automated topology correction, and surface deformation following intensity gradients. After the completion of the cortical models, registration to a spherical atlas follows which utilizes individual cortical folding patterns to match cortical geometry across subjects. Then the parcellation of the cerebral cortex into units based on gyral and sulcal structure is performed.

To extract reliable estimations of volume and thickness measurements using data from multiple time points for each subject, images were automatically processed with the FreeSurfer' longitudinal stream [1]. Specifically an unbiased within-subject template space and image [17] is created using robust, inverse consistent registration [18]. Several processing steps are then initialized with common information from the within-subject template, such as skull stripping, Talairach transforms, atlas registration as well as spherical surface maps and parcellations. This way, reliability and statistical power are significantly increased [1].

In fact, all four available time points for each subject were independently processed using the regular (cross-sectional) FreeSurfer processing stream. Then, a within subject template was built using the previously processed time points (four for all subjects). The subject template is built with the same number of time points for all subjects with similar time spacing (across subjects) for consistency. Finally, the longitudinal runs are created for all available time points. They contain the most reliable and accurate processing results, from which the estimations of volume and thickness measurements were used in this examination. Following this procedure, for each time point 55 volumes and 70 cortical thickness measurements (35 for each hemisphere) were selected, leading to a total of 500 features.

The Freesurfer's automated pipeline can result in variability in the segmentation accuracy for some ROIs and, even more, segmentation failure [4], [5]. The failure might be global, due to extremely poor image quality, registration issues, gross misestimation of the hippocampus, or processing errors. On the other hand, partial failure might occur in one or more regions. The most common cases include these regions: frontal, temporal, insula, parietal, occipital, cerebral white matter, basal ganglia, and ventricle, according to the QC procedures provided by the Center for Imaging of Neurodegenerative Diseases, UCSF [19]. Following this, a total of 114 subjects were detected with such failures. The detected failures might be present in each of the independently processed time points causing subsequent processing error or failure in the final results. Example of such failures for ADNI data are depicted on Fig.1 [19].

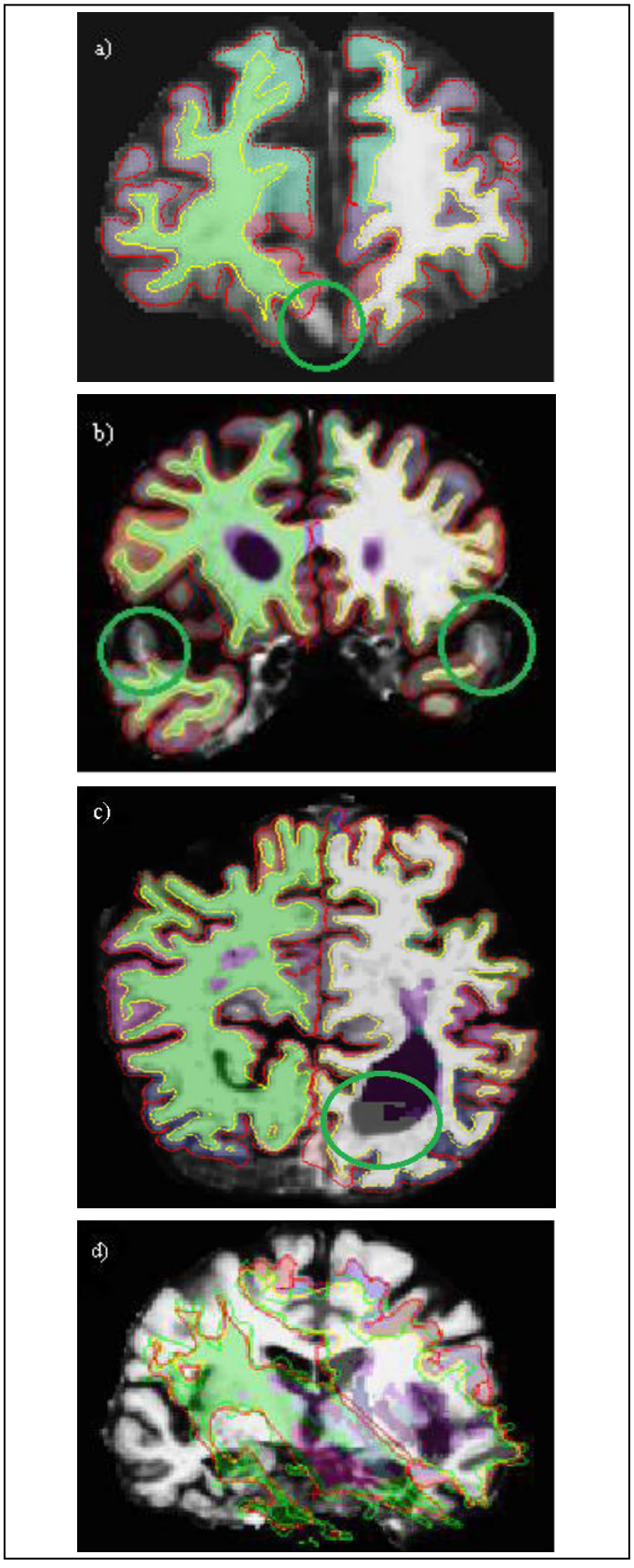


Fig. 1. Processing failures. a) Frontal fail, b) Temporal fail, c) Ventricle fail, d) Overall fail

Due to the lack of medical expert, all subjects with failures in at least one time point were excluded. To be able to evaluate the influence of the QC to the retrieval performance, the retrieval was performed with and without QC. In the first case, all 267 patients were included, while in the second one, 153 subjects undergone the procedure.

According to the previous research [20], feature subset selection significantly improves the retrieval performance. Hence, these phase was also included in the research. It is intended to select the most relevant features, reducing the feature vector dimensionality in the same time. The Correlation-based Feature Selection (CFS) method [21]. It evaluates subsets of features taking into consideration the usefulness of individual features for predicting the class along the degree of intercorrelation among them. This means that valuable feature subsets contain features highly correlated with the class, yet uncorrelated to each other [21].

Considering the number of subjects used in this research, leave-one-out strategy was used, meaning that each subject representation was used as a query against all other representations stored in the database.

Euclidean distance was used as a similarity measurement. To get an unbiased result, the feature selection was performed independently of the query subject. Thus, the specific feature subset for each query subject was obtained.

III. RESULTS

Experimental results regarding the retrieval performance in both cases, with and without QC are presented in this subsection.

Table I contains the results in both cases. Standardized evaluation metrics were used, such as:

- Mean Average Precision (MAP) - the mean of the average precision scores for each query
- Precision at first 1 (P1) - precision of the first (top) returned subject
- Precision at first 10 (P10) - precision of the first (top) 10 returned subject s
- Precision at first 20 (P20) - precision of the first (top) 20 returned subjects
- R-precision (RP) – precision at first (top) X returned subjects, where X is the number of relevant subjects

TABLE I. EXPERIMENTAL RESULTS - EVALUATION

	WITHOUT QC	**WITH QC**
MAP	0.80	0.83
P1	0.85	0.88
P10	0.85	0.86
P20	0.84	0.85
RP	0.76	0.80

From the point of view of the measurements used in the experiments after the feature selection, considering the application domain, features sensitive to the disease are expected to be selected more stable. To be able to examine this, the inclusion rate, i.e. how frequent each feature is selected was also recorded. The feature selected in more than 50% of the cases are depicted in Table II and Table III. Considering the feature vector dimensionality, in most of the cases 30-39 features were selected in the experiments without QC. On the other side, in the experiments with QC, 20-26 were selected in most of the cases.

TABLE II. RATE OF FEATURE INVOLVEMENT IN THE EXPERIMENTS WITHOUT QC

Feature (Followed by the Number of Time Point it Represents)	Involved in the Experiments (%)
Left-Hippocampus_0	100.00
Estimatedtotalintracranialvol_0	72.28
Left-Hippocampus_6	100.00
Left-Amygdala_6	100.00
Left-Inf-Lat-Vent_12	98.50
Left-Putamen_12	99.63
Left-Hippocampus_12	100.00
Right-Hippocampus_24	66.67
Cc_Mid_Posterior_24	99.63
Lh_Entorhinal_Thickness_0	100.00
Lh_Medialorbitofrontal_Thickness_0	100.00
Lh_Bankssts_Thickness_6	100.00
Rh_Isthmuscingulate_Thickness_6	75.28
Rh_Pericalcarine_Thickness_6	71.54
Rh_Precuneus_Thickness_6	64.42
Lh_Bankssts_Thickness_12	100.00
Lh_Entorhinal_Thickness_12	98.50
Lh_Inferiorparietal_Thickness_12	100.00
Lh_Inferiortemporal_Thickness_12	100.00
Lh_Middletemporal_Thickness_12	100.00
Lh_Parahippocampal_Thickness_12	86.89
Rh_Entorhinal_Thickness_12	100.00
Rh_Parahippocampal_Thickness_12	73.41
Lh_Bankssts_Thickness_24	54.68
Lh_Entorhinal_Thickness_24	100.00
Lh_Inferiortemporal_Thickness_24	100.00
Lh_Middletemporal_Thickness_24	100.00
Lh_Parahippocampal_Thickness_24	100.00
Rh_Bankssts_Thickness_24	100.00
Rh_Entorhinal_Thickness_24	100.00
Rh_Inferiortemporal_Thickness_24	98.13
Rh_Middletemporal_Thickness_24	77.90

TABLE III. RATE OF FEATURE INVOLVEMENT IN THE EXPERIMENTS WITH QC

Feature (Followed by the Number of Time Point it Represents)	Involved in the Experiments (%)
Left-Inf-Lat-Vent_6	84.97
Left-Hippocampus_6	60.13
Left-Amygdala_6	80.39
Left-Hippocampus_12	99.35
Left-Hippocampus_24	100.00
Brainsegvol-To-Etiv_24	60.13
Lh_Bankssts_Thickness_6	63.40
Rh_Bankssts_Thickness_6	62.09
Rh_Isthmuscingulate_Thickness_6	99.35
Lh_Bankssts_Thickness_12	99.35
Lh_Inferiortemporal_Thickness_12	100.00
Lh_Middletemporal_Thickness_12	97.39
Lh_Parahippocampal_Thickness_12	98.04
Lh_Superiortemporal_Thickness_12	92.81
Rh_Middletemporal_Thickness_12	64.05
Rh_Pericalcarine_Thickness_12	64.05
Rh_Precuneus_Thickness_12	90.85
Lh_Entorhinal_Thickness_24	100.00
Lh_Fusiform_Thickness_24	98.69
Lh_Inferiortemporal_Thickness_24	58.17
Lh_Medialorbitofrontal_Thickness_24	67.97
Lh_Middletemporal_Thickness_24	100.00
Rh_Entorhinal_Thickness_24	98.69
Rh_Middletemporal_Thickness_24	99.35
Rh_Parahippocampal_Thickness_24	95.42

IV. DISCUSSION

According to the obtained results, it can be clearly concluded that the examined phase of QC after the processing improves the retrieval results. All evaluation metrics leads to the same conclusion. Consequently, it should be emphasized that QC influence on the retrieval performance. This investigation gives results according to which QC procedure is recommended to be included.

From the point of view of the selected features, it should be noticed that in both cases the most frequently selected features are known biomarkers for Alzheimer's Disease such as: including volume of the hippocampus, inferior lateral ventricle, amygdala, cortical thickness of the entorhinal cortex etc. [10] Additionally, most of them are from the last time points (12-month, and 24-month follow-ups). This is reasonable because as the disease progresses the brain changes are prominent.

It is evident that feature selection reduces the feature vector dimensionality significantly in both cases. However, it

should be emphasized that in the case with QC the number of features is smaller than in the case without QC. Thus, even smaller number of features are enough to provide better results.

V. CONCLUSION

In this paper the influence of the quality control after the image processing on the retrieval performance was evaluated. Longitudinal MRIs from ADNI database automatically processed with FreeSurfer were used in the examination. Volume and cortical thickness estimates of the brain structures from four time points were used to generate the feature vector. Feature selection was also applied to select the most valuable and relevant feature while reducing the feature vector dimensionality.

According to the results, the evaluation metrics showed that QC improves the retrieval performance. Regarding the selected features, in most of the cased were automatically selected features that are well known biomarkers for AD. Moreover, although significant dimensionality reduction was present in both cases, with and without QC, in the case with QC this was even more emphasized.

As a result, it can be concluded that QC is extremely important in this context and is highly recommended to be performed. It should be emphasized that failures and/or imprecision in one sub-process of the whole retrieval process, such as the segmentation, can significantly influence on the final retrieval performance and, as a big challenge, should not be avoided in this kind of research.

ACKNOWLEDGMENT

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education. The study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. The Laboratory for Neuro Imaging at the University of Southern California is responsible for the dissemination of the ADNI data.

This work is partially supported by the Faculty of Computer Science and Engineering, Skopje, Macedonia as a part of the project "Deep Learning for Image and Text Analysis".

REFERENCES

[1] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, 2012.

[2] S. M. Smith, N. D. Stefano, M. Jenkinson, and P. M. Matthews, "Normalized Accurate Measurement of Longitudinal Brain Change," *Journal of Computer Assisted Tomography*, vol. 25, no. 3, pp. 466–475, 2001.

[3] Voxel-based Morphometry Extension to SPM8, [http://dbm.neuro.uni-jena.de/vbm/], last visited: 07.03.2017.

[4] C. S. Mccarthy, A. Ramprashad, C. Thompson, J.-A. Botti, I. L. Coman, and W. R. Kates, "A comparison of FreeSurfer-generated data with and without manual intervention," *Frontiers in Neuroscience*, vol. 9, 2015.

[5] L. L. Backhausen, M. M. Herting, J. Buse, V. Roessner, M. N. Smolka, and N. C. Vetter, "Quality Control of Structural MRI Images Applied Using FreeSurfer—A Hands-On Workflow to Rate Motion Artifacts," *Frontiers in Neuroscience*, vol. 10, Jun. 2016.

[6] K. Krupa and M. Bekiesińska-Figatowska, "Artifacts in Magnetic Resonance Imaging," *Polish Journal of Radiology*, vol. 80, pp. 93–106, 2015.

[7] M. L. Wood and R. M. Henkelman, "MR image artifacts from periodic motion," *Medical Physics*, vol. 12, no. 2, pp. 143–151, 1985.

[8] M. Reuter, M. D. Tisdall, A. Qureshi, R. L. Buckner, A. J. V. D. Kouwe, and B. Fischl, "Head motion during MRI acquisition reduces gray matter volume and thickness estimates," *NeuroImage*, vol. 107, pp. 107–115, 2015.

[9] E. Bellon, E. Haacke, P. Coleman, D. Sacco, D. Steiger, and R. Gangarosa, "MR artifacts: a review," *American Journal of Roentgenology*, vol. 147, no. 6, pp. 1271–1281, 1986.

[10] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, J. C. Morris, R. C. Petersen, A. J. Saykin, M. E. Schmidt, L. Shaw, L. Shen, J. A. Siuciak, H. Soares, A. W. Toga, and J. Q. Trojanowski, "The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception," *Alzheimer's & Dementia*, vol. 9, no. 5, 2013.

[11] M. Agarwal and J. Mostafa, "Image Retrieval for Alzheimer's Disease Detection," *Medical Content-Based Retrieval for Clinical Decision Support Lecture Notes in Computer Science*, pp. 49–60, 2010.

[12] M. Agarwal and J. Mostafa, "Content-based image retrieval for Alzheimer's disease detection," *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2011.

[13] M. Mizotin, J. Benois-Pineau, M. Allard, and G. Catheline, "Feature-based brain MRI retrieval for Alzheimer disease diagnosis," *2012 19th IEEE International Conference on Image Processing*, 2012.

[14] B. T. Wyman, D. J. Harvey, K. Crawford, M. A. Bernstein, O. Carmichael, P. E. Cole, P. K. Crane, C. Decarli, N. C. Fox, J. L. Gunter, D. Hill, R. J. Killiany, C. Pachai, A. J. Schwarz, N. Schuff, M. L. Senjem, J. Suhy, P. M. Thompson, M. Weiner, and C. R. Jack, "Standardization of analysis sets for reporting results from ADNI MRI data," *Alzheimer's & Dementia*, vol. 9, no. 3, pp. 332–337, 2013.

[15] Freesurfer [http://surfer.nmr.mgh.harvard.edu/], last visited: 07.03.2017.

[16] C.R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, et al. "The Alzheimer's Disease Neuroimaging Initiative, (ADNI): MRI methods," *J Magn Reson Imaging*, vol. 27, pp. 685–91, 2008.

[17] M. Reuter and B. Fischl, "Avoiding asymmetry-induced bias in longitudinal image processing," *NeuroImage*, vol. 57, no. 1, pp. 19–21, 2011.

[18] M. Reuter, H. D. Rosas, and B. Fischl, "Highly accurate inverse consistent registration: A robust approach," NeuroImage, vol. 53, no. 4, pp. 1181–1196, 2010.

[19] ADNI - Alzheimer's Disease Neuroimaging Initiative, [http://adni.loni.usc.edu], last visited: 09.02.2017.

[20] K. Trojacanec, I. Kitanovski, I. Dimitrovski, and S. Loshkovska, "Medical Image Retrieval for Alzheimer's Disease Using Data from Multiple Time Points," *ICT Innovations 2015 Advances in Intelligent Systems and Computing*, pp. 215–224, 2016.

[21] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1437-1447, 2003.

Using Biomodule for Vital Parameters Measurement in Hospital Environment

Ivana Kozolovska, Bojana Koteska, Monika Simjanoska and Ana Madevska Bogdanova
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Rugjer Boskovikj 16, 1000 Skopje, Macedonia
Email: kozolovska.ivana@students.finki.ukim.mk
{bojana.koteska, monika.simjanoska, ana.madevska.bogdanova}@finki.ukim.mk

***Abstract*—Trauma surgeons at emergency departments in hospitals consider the wireless biomodules as valuable information source about the patients health state. Doctors point of view is very important in designing solution for software support for these biomodules. We have designed and developed a software solution for using Zephyr Bioharness biomodule for patients in hospitals, developing features regarding the vital parameter signal processing and visualization. This solution provides live monitoring of ECG, heart rate, respiratory rate, posture and acceleration. The vital parameters measurements are also stored remotely and used for history and further analysis of the patient health state. As a case study, the solution is set and currently under testing in General Hospital in Celje, Slovenia.**

***Index Terms*—Vital Parameters, Zephyr, Android Application, Hospital.**

I. Introduction

Wearable biosensors can aid the standard procedure of measuring vital parameters in a hospital environment. The data from the biosensors can be collected by a portable electronic device that is located close to the biosensor, for example, a mobile phone or tablet. These biosensors are usually light weighted, long lasting and use low power protocols for sending data. The data received on the electronic device is initially processed and then transferred to the remote server or Cloud for further analysis of the patient health state. The low-cost biosensors have no or low memory capacity and vital parameters streaming starts immediately after the sensors are placed on the patients body. The biosensors data is usually streamed at a frequency of 100 to 1000 Hz. In order to keep the streamed data, we need an external software that will process and store the data. In this paper we present a software solution for using the Zephyr Bioharness biosensor for collecting data from the patients in hospital. Zephyr Bioharness biosensor [1] collects different vital parameters: ECG (electrical activity all over the heart), Heart Rate (number of heartbeats per minute), Respiratory Rate (number of breaths per minute), Temperature (skin temperature), Posture (body position), Activity Level (acceleration), Subject Status as well as the battery level of the device. The data are streamed at a frequency of 250 Hz.

Zephyr BioHarness is widely used for developing eHealth applications. In [2] BioHarness has been tested for validity and has shown to be reliable for determining respiratory rate and respiratory breakpoint during exercise of varying intensity. In [3] the authors present an application developed for usage in fire fighting and sports that includes Zephyr BioHarness sensor. The application has been designed for smart phones, smart watches and tablets. An IoT-enabled mHealth application is presented in [4] to aid the personalized health care services. The application is developed for both Android and Windows phone platforms and the proposed architecture uses Zephyr BioHarness sensor. Another application of Zephyr Bioharness is in a cross-domain application for ambient and health monitoring with the aim to provide a complete picture to the information users (e.g. doctors) [6]. In [7], the authors present a platform that supports a variety of biosensor add-ons and it is used as a pilot project in Washington Hospital .

In this paper we confirm the usefulness of the Zephyr Bioharness in a hospital environment by providing remote patient vital parameter monitoring. Our software solution is set and tested in General Hospital in Celje, Slovenia.

II. Software Solution for Using Zephyr Bioharness Biomodule

The provided software system is intended for remote monitoring of human vital parameters and enables constant monitoring of patients health state in hospitals. The solution is developed under Android platform and supports Android devices with operating system Android 4.4 (and higher). The data gathering from the sensor is performed by using the Bioharness 3 SDK for Android platform. As shown in Fig. 1, the communication between Android device and Zephyr Bioharness 3 sensor is achieved by using the Bluetooth protocol. The application can be also installed on the patient android devices.

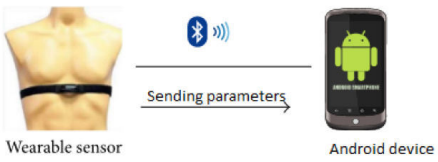


Fig. 1. Patient Monitoring System.

The vital parameter measurements are streamed in real time at a frequency of 250 Hz. Received data are stored locally on

the device as .csv files. If Internet connection is available, the data is sent to the remote SQL database hosted on a Windows server machine by using the android ksoap library and web services developed in C# which provide the communication between the android device and the SQL server.

Our solution enables simultaneous monitoring of five parameters: ECG, heart rate, respiratory rate, posture and peak acceleration. Fig. 2 presents a screen of the proposed solution. Heart rate (HR), respiratory rate (RR), posture and peak acceleration are shown in the upper table. There is also information (MAC Address) for the connected sensor and two buttons for connecting/disconnecting from the sensor.

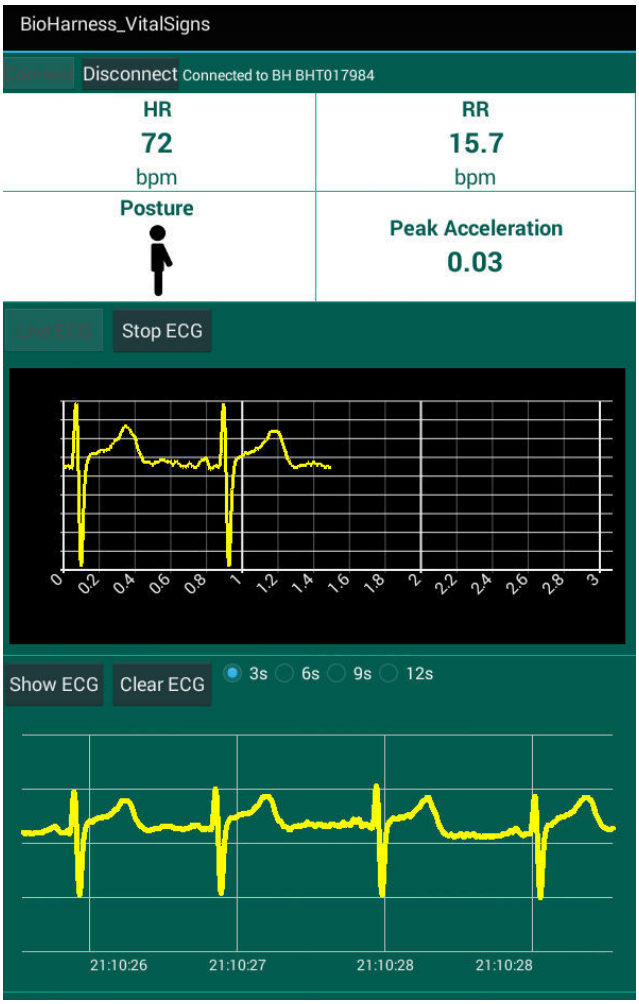


Fig. 2. Monitored Parameters.

As a very important feature requested by the trauma doctors, the application provides live monitoring and history of the electrocardiogram (ECG), and in both cases ECG is graphically represented by using the GraphView component. The live ECG monitoring is shown for three seconds. ECG history is gathered from the .csv files saved in the Android device. The history graph shows ECG data of the paired sensor from the last hour and according to the doctors' requirements there is an option for showing history of 3, 6, 9 and 12 seconds. The time of measurement is presented on the x axis. The graph is also scrollable and zoomable which provides easy detection of ECG signal abnormalities.

The benefits of the application are multi-fold. It provides an alternative opportunity for the doctors to monitor the vital patient parameters at a time without using a few different machines for obtaining each of the parameters. Another advantage is that the solution is wireless. This is very beneficial for the doctors in terms of the space available in the hospital units - too much cables cause lots of problems when treating the patients. The option to zoom and scroll included in the ECG window is also an advantage. Disadvantages considering the reliability of the application are not reported.

III. Conclusion

In this paper we have presented a software for wireless vital parameters monitoring by using the Zephyr Bioharness 3 sensor. The application is developed according to the doctors demands in the General Hospital in Celje, Slovenia where it is tested and confirmed to be reliable. The application provides the ability to monitor the patient's HR, RR, peak acceleration, posture and ECG. The module showing the ECG is more advanced providing the opportunity to show history and also to zoom the signal for making deeper visual analysis by the doctors. The application is modular and easy to use, thus new features can be added easily. The database records created from the patitent's data (ECG, HR, RR) processed by the software solution are very important for further research, especially in biosignal processing domain.

Acknowledgment

This research is supported by SIARS, NATO multi-year project NATO.EAP.SFPP 984753.

References

[1] Zephyr Technology, "Zephyr BioHarness 3.0 User Manual," accessed: 2017-03-07. [Online]. Available: https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf

[2] J. Hailstone and A. E. Kilding, "Reliability and validity of the zephyr bioharness to measure respiratory responses to exercise," *Measurement in Physical Education and Exercise Science*, vol. 15, no. 4, pp. 293–300, 2011.

[3] P. Castillejo, J.-F. Martinez, J. Rodriguez-Molina, and A. Cuerva, "Integration of wearable devices in a wireless sensor network for an e-health application," *IEEE Wireless Communications*, vol. 20, no. 4, pp. 38–49, 2013.

[4] A. Borodin, Y. Zavyalova, A. Zaharov, and I. Yamushev, "Architectural approach to the multisource health monitoring application design," in *Open Innovations Association (FRUCT), 2015 17th Conference of.* IEEE, 2015, pp. 16–21.

[5] V. Gay and P. Leijdekkers, "Design of emotion-aware mobile apps for autistic children," *Health and Technology*, vol. 4, no. 1, pp. 21–26, 2014.

[6] F. Vergari, S. Bartolini, F. Spadini, A. D'Elia, G. Zamagni, L. Roffia, and T. S. Cinotti, "A smart space application to dynamically relate medical and environmental information," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010.* IEEE, 2010, pp. 1542–1547.

[7] T. Gao, C. Pesto, L. Selavo, Y. Chen, J. Ko, J. Lim, A. Terzis, A. Watt, J. Jeng, B.-r. Chen *et al.*, "Wireless medical sensor networks in emergency response: Implementation and pilot results," in *Technologies for Homeland Security, 2008 IEEE Conference on.* IEEE, 2008, pp. 187–192.

Analysis of the urban heat islands effect in Skopje

Kostadin Mishev, Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, R. Macedonia
{kostadin.mishev, dimitar.trajanov}@finki.ukim.mk

***Abstract*— Heat islands is a popular effect which occurs metropolitan or urban areas. Due several physical factors, some areas are becoming significantly warmer than its surrounding. They own higher average temperature than its rural surroundings owing to the greater absorption, retention, and generation of heat by its buildings, pavements, and human activities. [1][2]. The temperature difference usually is larger at night than during the day, and is most apparent when winds are weak. UHI is most noticeable during the summer and winter [3].**

By using a thermal cameras and paraglider, we try to recognize such heat islands in the area of Skopje recording its surface from the top view [5]. We are using Flir Vue Pro camera which detects the infrared portion of the magnetic spectrum presenting the current temperature of the recorded surface [6]. It enables calculation of the relative temperature between two points on image providing an information about the warmness of a single point relatively to another. The final result of each image recorded by such camera is a matrix of recorded temperatures of each point. Analysis of such matrix can provide information about the heated islands that virtually set apart of the other area concerning the temperature difference. Also, we took photos with RGB camera in the same time from the same area which provides additional information about the recorded objects to facilitate their recognition.

In the paper we will provide additional information about the areas that can be concerned as heat islands in Skopje providing brief explanation about the reasons for their appearance by using the analysis of the thermal camera images. Such analysis can explain the reasons of the global warming in single points of view and provides possibilities of their reduction by following the concluded results.

Keywords*— *thermal imaging; heat islands; thermal image analysis; climate changes

I. INTRODUCTION

An urban heat island (UHI) is an urban area or metropolitan area that is significantly warmer than its surrounding rural areas[7]. UHI is most noticeable during the summer and winter. The main cause of the urban heat island effect is from the modification of land surfaces. Waste heat generated by energy usage is a secondary contributor. As a population center grows, it tends to expand its area and increase its average temperature. The less-used term heat island refers to any area, populated or not, which is consistently hotter than the surrounding area. Figure 1 represents typical temperature variance among the urban and rural areas in cities[1].

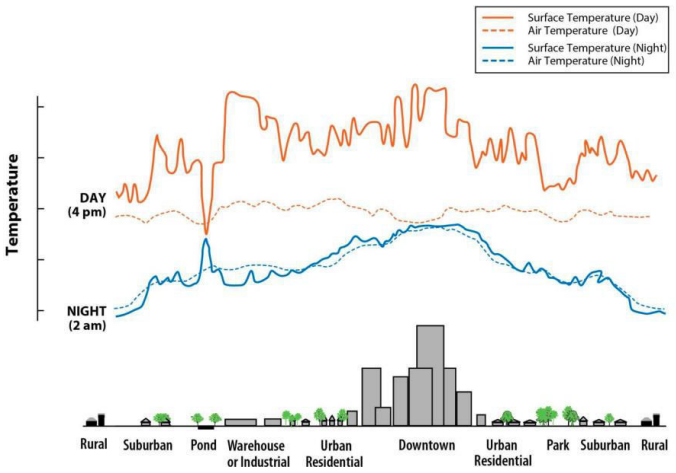


Fig. 1. Typical temperature variance among the urban and rural areas in cities during the night and the day

II. MEASUREMENT OF HEAT ISLAND EFFECT

Due the recommendations of US Environmental Protection Agency, the measurement of the effect of urban heat islands can be performed by measuring the temperature of the urban surface or the air temperature [8]. The surface temperature indirectly influences the temperature of the air. By this way, urban parks and vegetation areas that have lower temperature, contribute for air cooling above them. By other side, densely populated areas contribute higher air temperature. Due the mixture with the atmosphere layers, the ratio between the surface and air temperature is not a constant [9].

A. Measurement of the surface temperature

The measurement of the surface temperature can be done by using thermal images which record the reflected and emitted energy from the object surface like roofs, roads, pavements, vegetation and water surfaces [10]. All surfaces emit infrared

[1] US Environmental Protection Agency, Fact Sheet: Keeping Your Cool: How Communities Can Reduce the Heat Island Effect, November, 2014, Publication Number: 430F14041, https://www.epa.gov/sites/production/files/2016-09/documents/heat_island_4-page_brochure_508_120413.pdf

electromagnetic waves whose length depend on their warmth. By their identification and measurement, it can be determined the temperatures of the surfaces which are recorded.

B. *Measurement of the air temperature*

The second method of heat islands identification can be performed by using air temperature measurement in urban and rural environments. As well as the surface measurement, this method has its own shortcomings like [11]:

- requirement of dense measurement stations all over the urban and rural part of the city;
- it should be considered the type of the measurement station, the different sea level and microclimate of the measured area

C. *Measurement methodology*

During the detection of the effect of heat island in Skopje, it is used the method of the surface temperature measurement. All measurements are done by using two camera, standard GoPro[2] and thermal Flir ProVue[3] which record time-lapse pictures simultaneously. They were set up on a paraglider, which performed a flight from Vodno, as a starting point, to the city park as a final point just to make photos from the above from the covered area. During the flight, there were snapshotted approximately 100 thermal and standard photos which are used in this analysis. We used the GoPro camera just for better coordination of the recorded thermal photos in the analysis. Also, we used a GPS sensor to record the trajectory of the paraglider to facilitate the process of photo merging in generation of 3D map of the recorded area.

D. *Image preprocessing*

After the successful snapshot by using thermal and standard cameras, there were obtained two photo datasets, collections of photos from the same areas but made with different kind of photo sensor making approximately 2000 photos per collection. The standard photo present the visible electromagnetic specter, and the thermal presents the infrared (invisible) specter of colors providing an information about the heat temperature of the recorded objects. First at all, each thermal photo was paired with appropriate RGB photo obtained by GoPro camera by using the time information of the snapshot. Both of cameras were time synchronized, so they were set up to take one photo per seconds and by using the time of snapshot, the appropriate pairs of photos from two collections were aligned. Samples of such pairs are given in Figure 4. Afterwards, the provided GPS coordinates from the GPS device log were used to append georeference stamp of the obtained pairs of photos. Next. Flir software was used for thermal photos analysis. It provides the concept of probes that enable temperature differentiation of two points in the same photo. Such differentiation can provide information about how much one object is more heated than other. Temperature variation can help to discover the heat islands in the observed area. The results of such observation are presented in the section Results.

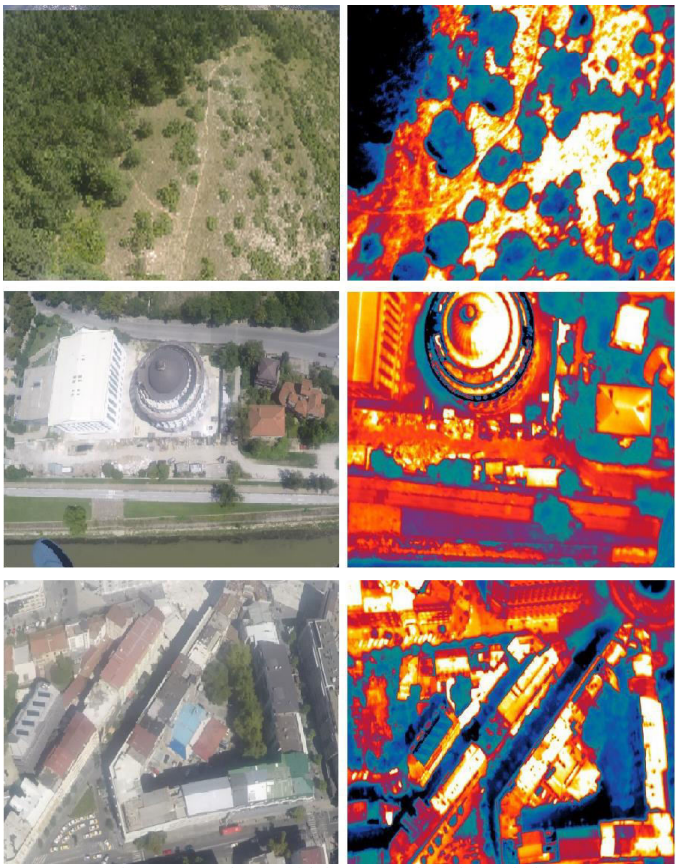


Fig. 2. RGB photos with its appropriate thermal photo

III. RESULTS

Follows the overview and discussion of the obtained results obtained by temperature analysis at each point of variation.

A. *The effects of Urban Heat islands*

The obtained range of temperatures proves the phenomena of existence of the urban heat islands in Skopje. Figure 4 represents map view of selection of probes of thermal information collected from thermal photos. The main considerations are following:

- The range between the maximum surface temperatures in the periphery of the city is approximately 7 degrees, meanwhile the variance regarding Vodno is approximately 12 degrees.
- The river of Vardar has positive impact and influences the temperature, so the temperature of the area around the river is fewer degrees lower that the areas in higher distance from the water.
- The central core, the area of Gradski Zid has the highest temperature which is 1.5 to 2.0 higher than the temperature in the municipality of Kisela Voda. It is important to consider the fact that the City square Macedonia, because of its white color is considerably colder that its environment. For example, the

[2] https://gopro.com/
[3] http://www.flir.com/suas/vuepro/

temperature of the asphalt of the street of Maksim Gorki is 6 degrees higher than the temperature of the City square tiles.

- Generally, by following the tradition of red or dark roof construction, it can be considered a higher temperature on the roofs which are constantly exposed on solar radiation. The buildings that have brighter roofs bring positive impact on the temperature decrease in their environment, but they are too rare in Skopje [13].

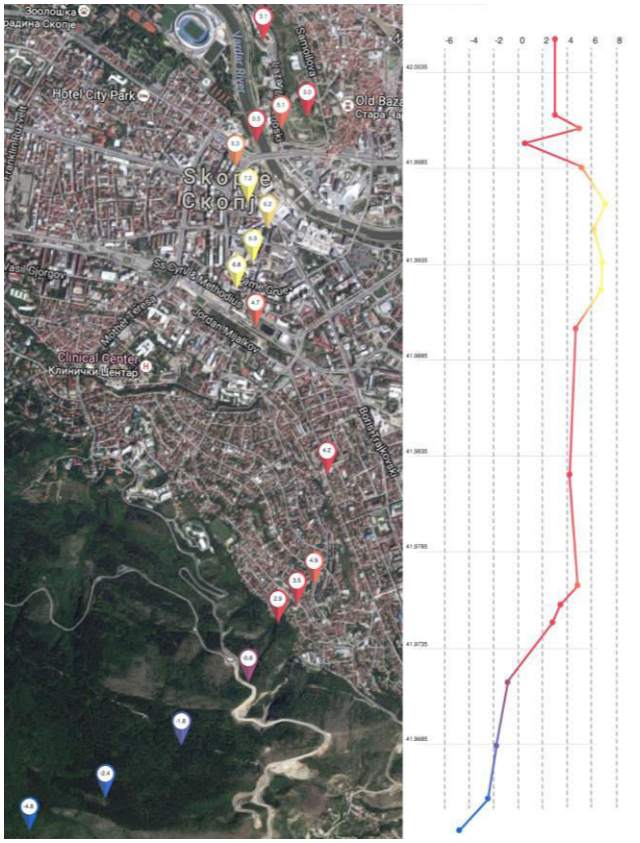


Fig. 4. Overview the relative temperature difference in Skopje in a North-South direction

B. *Micro Analysis*

This analysis covers the central city area around the City square Macedonia and the streets of Maksim Gorki and Nikola Vapcarov. Based on the thermal images which are presented in Figure 5, it can be considered following facts:

- White color of the tiles in the City Square Macedonia has positive impact to the temperature and it is lower that the surrounding asphalt areas, so the central square is cooler than the street of Maksim Gorki (1).
- The dark roofs are greatly warmer than the lighter ones and they have impact to the temperature rise (2).
- Parked vehicles which are directly influenced to a solar radiation are significantly warmed and directly influence for temperature rise of the street of Maksim Gorki (3).
- Due shading, tree existence and vehicle absence, the street of Nikola Vapcarov is significantly cooler than the street of Maksim Gorki (4).

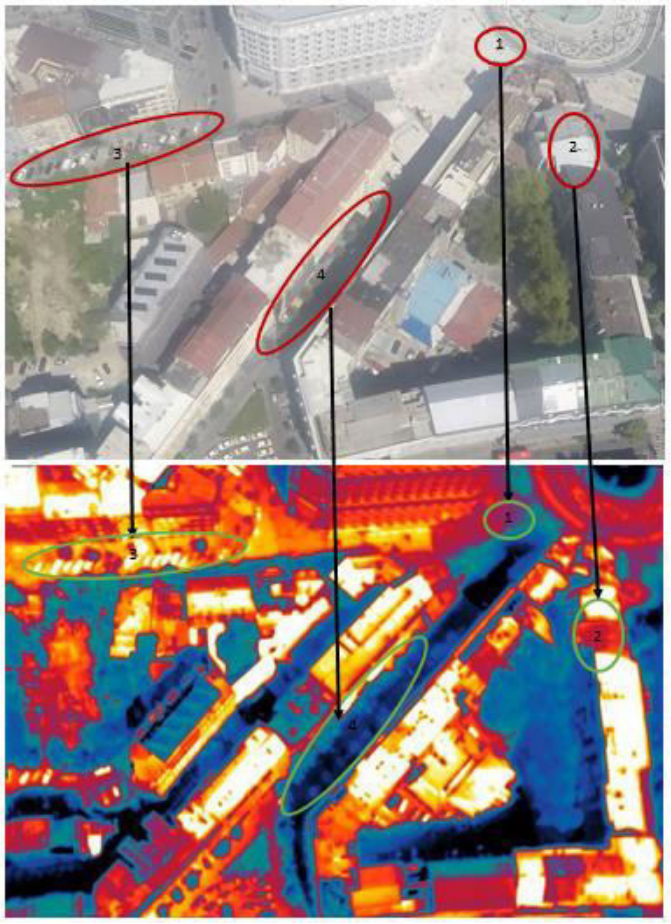


Fig. 5. Comparative analysis of the temperatures in the Skopje's central city area

IV. Conclusion

By using standard and thermal camera, we tried to retrieve the positions of micro areas that are significantly warmer than others in Skopje i.e. to determine the heat islands by recording their infrared radiations. In the analysis of the obtained photos, we concluded that Skopje is one of the cities that contain heat islands which significantly influence to the effect of global warming [12]. Such warmer regions are more dominated in the central area of Skopje where the human activity is more expressed. The range between the maximum surface temperatures in the periphery of the city is approximately 7 degrees, meanwhile the variance regarding Vodno is approximately 12 degrees which proves the concept of global heat island in the center of the city. By execution of the micro-analysis to the photos captured in the center of the city, we concluded that objects made with material with dark color absorb more warmness that the object made with

material with bright color. Also, we should not that the warmth of the surfaces depends on the direct exposure to the solar radiation. Other period of the day may provide different results.

V. FUTURE WORK

As a future work, we plan to continue our research in heat islands determination by using UAV devices like drones which movement direction can be easily controlled remotely. The drone will be equipped with the same cameras but it will increase the resolution of captured areas and can provide detailed analysis of the radiated infrared energy during different periods of the day [14]. Also, we will try to use additional software like Open Drone Map[4] for the purpose of conversion of ordinary photo to three dimensional geographic data. Such representation will provide us better comprehension and detection of the heat islands in micro level. The detection of such micro heat islands will provide information about the spatial variability of surface radiant temperatures caused by the thermal behavior of different land-cover types.

REFERENCES

[1] Weng, Qixhao. "Fractal analysis of satellite-detected urban heat island effect." Photogrammetric engineering & remote sensing 69.5 (2003): 555-566.

[2] Roth, M., T. R. Oke, and W. J. Emery. "Satellite-derived urban heat islands from three coastal cities and the utilization of such data in urban climatology." International Journal of Remote Sensing 10.11 (1989): 1699-1720.

[3] Gallo, K. P., et al. "The use of a vegetation index for assessment of the urban heat island effect." Remote Sensing 14.11 (1993): 2223-2230.

[4] Lo, Chor Pang, Dale A. Quattrochi, and Jeffrey C. Luvall. "Application of high-resolution thermal infrared remote sensing and GIS to assess the urban heat island effect." International Journal of Remote Sensing 18.2 (1997): 287-304.

[5] Matson, Michael, et al. "Satellite detection of urban heat islands." Monthly Weather Review 106.12 (1978): 1725-1734.

[6] Maldague, Xavier PV. "Introduction to NDT by active infrared thermography." Materials Evaluation 60.9 (2002): 1060-1073.

[7] Wong, Nyuk Hien, and Chen Yu. "Study of green areas and urban heat island in a tropical city." Habitat international 29.3 (2005): 547-558.

[8] Reducing Urban Heat Islands: Compendium of Strategies

[9] Buyantuyev, Alexander, and Jianguo Wu. "Urban heat islands and landscape heterogeneity: linking spatiotemporal variations in surface temperatures to land-cover and socioeconomic patterns." Landscape ecology 25.1 (2010): 17-33.

[10] Weng, Qihao, Dengsheng Lu, and Jacquelyn Schubring. "Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies." Remote sensing of Environment 89.4 (2004): 467-483.

[11] Li, Qingxiang, et al. "Detecting and adjusting temporal inhomogeneity in Chinese mean surface air temperature data." Advances in Atmospheric Sciences 21.2 (2004): 260-268.

[12] Wang, Wei-Chyung, Zhaomei Zeng, and Thomas R. Karl. "Urban heat islands in China." Geophysical Research Letters 17.12 (1990): 2377-2380.

[13] Doulos, L., M. Santamouris, and I. Livada. "Passive cooling of outdoor urban spaces. The role of materials." Solar energy 77.2 (2004): 231-249.

[14] Gallo, Kevin P., et al. "Assessment of urban heat islands: a satellite perspective." Atmospheric Research 37.1-3 (1995): 37-43.

[4] http://opendronemap.github.io/odm/

Deep learning based plant segmentation from RGB images

Petre Lameski*, Eftim Zdravevski*, Andrea Kulakov* and Vladimir Trajkovik*
*University of Sts. Cyril and Methodius in Skopje, Faculty of Computer Science and Engineering

***Abstract*—Plant-ground segmentation is one of the most important steps in the process of plant classification. Plant segmentation is used for the process of weed detection in fields, vegetation coverage estimation from satellite images, detection of illness in plants, water level detection in plants, etc. There are quite a few approaches presented in the literature that use different types of color indexes, segmentation and classification techniques for the purpose of weed segmentation. In this paper we apply the SegNet deep learning architecture for plant weed segmentation. The images are taken from approximately 1m height under slightly varying lightning condition from the same field using a smart phone regular RGB camera with auto-focus. The presented results show that it is possible to successfully train a deep learning model based on a single image and obtain similar results to the best plant segmentation techniques available that use color indexes without expensive cameras.**

I. Introduction

The ever increasing need for food production in the world [1] requires innovative approaches to satisfy the current and future food demands. One of the ways to accomplish this is to increase the food production by the process of automation [2]. The process of automation requires adequate sensing technologies. The first step of any such approach that would be able to detect the plant characteristics is the process of plant detection by segmenting the plant pixels from an image from the ground and other object pictures. This step is very important since the accurate segmentation of the plants could improve the accuracy of the plant classification. Also most of the contour descriptors rely on the accurate contour detection of the plant leafs, which is impossible without an accurate segmentation of the leafs. In this paper we use the deep convolutional encoder-decoder architecture (SegNet) proposed in [3] to train a model for vegetation segmentation. The paper is organized as follows: in section II we present the most common and state of the art approaches for vegetation segmentation from images. In section III we describe the dataset and algorithms used for the segmentation model generation. Finally in section IV we discuss the obtained results and conclude the paper.

II. Related Work

There are several available approaches that use color indexes and several machine learning approaches for plant-ground segmentation. The authors in [4] propose the usage of color indexes that are linear combination of the R, G and B bands detected from the normal cameras for plant-ground segmentation under various light conditions. They propose the (Excess Green) ExG index calculated with 1:

$$ExG = 2g - r - b \tag{1}$$

Another vegetation index that is used for plant segmentation is the (Excess Green minus Excess Red) ExGExR calculated with 2:

$$ExGExR = ExG - 1.4r - g \tag{2}$$

Both of these approaches first normalize the channels to accomplish illumination invariance by using 3, 4 and 5:

$$r = \frac{R}{R+G+B} \tag{3}$$

$$g = \frac{G}{R+G+B} \tag{4}$$

$$b = \frac{B}{R+G+B} \tag{5}$$

Finally, they use Otsu thresholding [5] to establish the green from non-green areas in the images. The approaches allow per-pixel plant segmentation with high speed and accuracy.

In [6] authors use a simple Bayesian model from the normalized RGB channels, the Hue Saturation Value (HSV) channels and the value of G-R to generate a naive Bayesian model and classify the pixels to plant and non-plant pixels. Authors in [7] use neural networks to segment the images based on their pixel values. Authors also experiment with Infrared filters that eliminate the visible spectrum and allow near-infrared (NIR) light to cross to the camera sensor. The introduction to the near-infrared spectrum allows the usage of The normalized difference vegetation index (NDVI) [8] which was introduced for vegetation estimation from satellite images. It can also be used for vegetation segmentation by applying equation 6 where VIS is the visible spectrum detected illumination.

$$NDVI = \frac{NIR - VIS}{NIR + VIS} \tag{6}$$

The NIR has values between -1.0 and 1.0 and the higher the number, the higher the probability of detected vegetation from satellite images. The NDVI index is also successfully used in vegetation segmentation from [9] from NIR and Red channel images. Deep learning architectures have already been applied for segmentation of vegetation from images. Authors in [10] use convolutional neural networks and then fully connected layers to classify vegetation and non-vegetation pixel from images. The approach we propose has already been applied

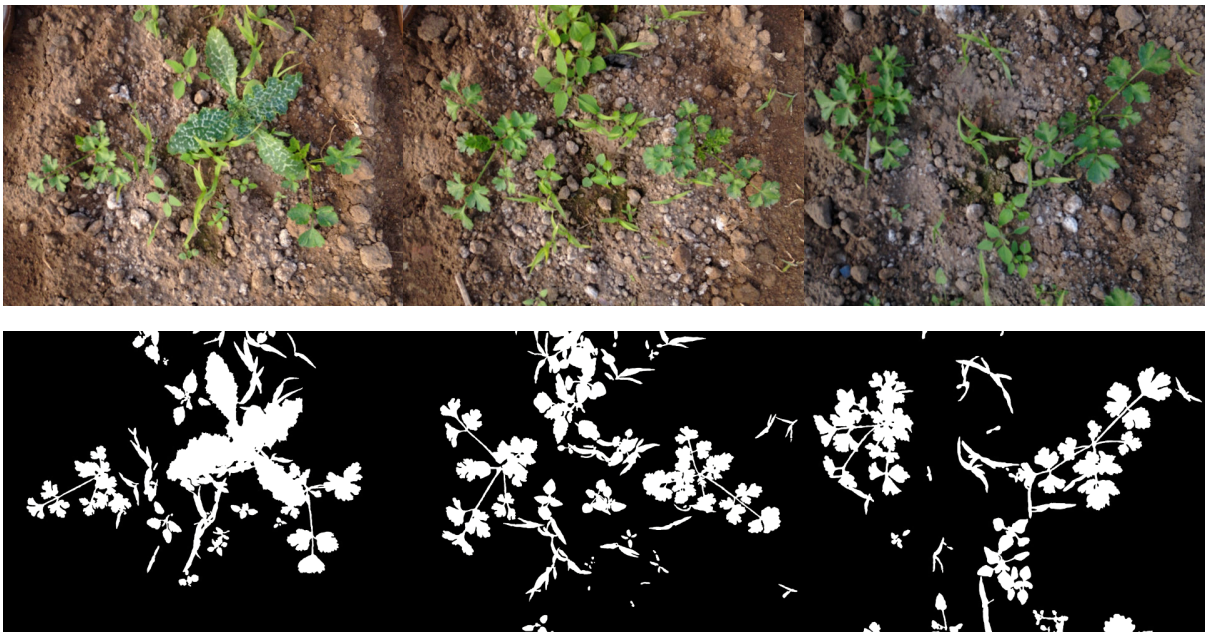


Fig. 1: Examples from dataset images and respective masks

in [11]. The main difference is that we use only one image from the dataset to train the model and we test on a different RGB dataset with single plant under variable light conditions.

III. Materials and Methods

In this paper we use the deep convolutional encoder-decoder architecture to train a model for segmenting plant pixels from other pixels. We use our own dataset consisted of 40 images that contain carrot plants and other weed plants. The images are taken from a 10MP RGB camera using auto focus from a smart phone from distance of approximately 1 meter above the plants. Examples of the images can be found in Fig. 1. The first row in 1 represents the taken images and the second rows represents binary vegetation masks that were manually labeled.

As it can be observed from the images, the images are taken under variable light condition in close time proximity (less than several minutes apart). To generate a training model we divide the images to 256X256 patches and train the SegNet model using these patches. We take the patches with step of 128 and from the images we obtain a total of 46 patches for training. The test set is consisted of randomly selected patches from the rest 38 images. We use the Bayesian SegNet implementation available at: https://github.com/alexgkendall/caffe-segnet. The network is trained with base learning rate=0.01, gamma=0.5, momentum=0.9 and weight decay= 0.005. The learning rate is reduced every 500 iterations. The implementation is done in Python and the GPU used for training is GeForce GTX TITAN X with 12GB of GPU RAM. The machine used for training is Intel Core i7 CPU at 2.67GHz with 12GB of 1333MHz DDR3 RAM memory and 1TB of HDD.

We compare the results obtained from the SegNet model to the results obtained from ExG and ExGExR indices with Otsu threshold.

IV. Discussion

The obtained accuracies from the proposed approach and the ExG and ExgExR approaches are given in Table I

TABLE I: Vegetation segmentation accuracy

ExG and Otsu	ExGExR and Otsu	SegNet
0.974	0.977	0.694

The obtained results show that the used SegNet architecture obtains weaker results when compared to the most common vegetation segmentation techniques on the used dataset. Further investigation is needed to examine if the results can be repeated for other plant types and under different lightning conditions. The overhead of using deep learning for vegetation segmentation seems to be too large since the ExG and ExGExR approaches with Otsu threshold give the results without the need of prior training and the deep learning model needs over 2000 iterations to obtain the given accuracy with the used training parameters. Additional experiments with using combination of RGB channels and some of the vegetation indexes are expected to further improve the per-pixel segmentation accuracy.

Acknowledgments

The work presented in this paper was financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje, Macedonia.

REFERENCES

[1] D. K. Ray, N. D. Mueller, P. C. West, and J. A. Foley, "Yield trends are insufficient to double global crop production by 2050," *PloS one*, vol. 8, no. 6, p. e66428, 2013.

[2] P. Cosmin, "Adoption of artificial intelligence in agriculture," *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Agriculture*, vol. 68, no. 1, 2011.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[4] D. Woebbecke, G. Meyer, K. Von Bargen, D. Mortensen *et al.*, "Color indices for weed identification under various soil, residue, and lighting conditions," *Transactions of the ASAE-American Society of Agricultural Engineers*, vol. 38, no. 1, pp. 259–270, 1995.

[5] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[6] S. Moorthy, B. Boigelot, and B. Mercatoris, "Effective segmentation of green vegetation for resource-constrained real-time applications," in *Precision agriculture'15*. Wageningen Academic Publishers, 2015, pp. 93–98.

[7] F. De Smedt, I. Billiauws, and T. Goedemé, "Neural networks and low-cost optical filters for plant segmentation," *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, vol. 3, p. 4, 2011.

[8] D. W. Deering, *Rangeland reflectance characteristics measured by aircraft and spacecraft sensors*. Deering, 1978.

[9] S. Haug, A. Michaels, P. Biber, and J. Ostermann, "Plant classification system for crop/weed discrimination without segmentation," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 1142–1149.

[10] C. Potena, A. Pretto, and D. Nardi, "Fast and accurate crop and weed identification with summarized train sets for precision agriculture." IAS, 2016.

[11] M. Di Cicco, C. Potena, G. Grisetti, and A. Pretto, "Automatic model based dataset generation for fast and accurate crop and weeds detection," *arXiv preprint arXiv:1612.03019*, 2016.

A Filter for Images Decoded using Cryptcodes Based on Quasigroups

Daniela Mechkaroska Aleksandra Popovska-Mitrovikj Verica Bakeva
Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
Emails: daniela-mec@hotmail.com,{aleksandra.popovska.mitrovikj, verica.bakeva}@finki.ukim.mk

***Abstract*—Cut-Decoding and 4-Sets-Cut-Decoding algorithms are proposed elsewhere. They give improvement in decoding processes of Random Codes Based on Quasigroups (RCBQ). These codes are cryptcodes, so they provide a correction of certain amount of errors in the input data and an information security, all built in one algorithm. Performances of these codes for decoding images transmitted through a binary-symmetric channel are investigated elsewhere.**

In this paper, we consider a transmission of images through Gaussian channel and propose a filter for enhancing the quality of images decoded with Cut-Decoding and 4-Sets-Cut-Decoding algorithms. In the decoding process of these codes, three types of errors appear: more-candidate-error, null-error and undetected-error. With the proposed filter we can visually enhance only pixels damaged by first two kinds of errors.

***Keywords*—filter, cryptcoding, Gaussian channel, image, quasigroup.**

I. Introduction

The need for secure data transmission requires continuous improvement of existing and developing new algorithms that will provide correct and secure transmission of data. Due to necessity of obtaining efficient and secure transmission of data at the same time, the concept of cryptcoding begins to develop. Cryptcoding merges processes of encoding and encryption. A usual way to obtain codes resistant to an intruder attack consists in application of some of the known ciphers on the codewords, before sending them through an insecure channel. Then two algorithms are used, one for correction of errors and another for obtaining information security.

Random Codes Based on Quasigroups (RCBQ) are defined (in [1]) by using a cryptographic algorithm during the encoding/decoding process, i.e., they are cryptcodes. Therefore, they allow not only correction of certain amount of errors in the input data, but they also provide an information security, all built in one algorithm. For improving the performances of these codes, Cut-Decoding and 4-Sets-Cut-Decoding algorithms are proposed in [2], [3].

In this paper, we consider a transmission of images through Gaussian channel using RCBQ with Cut-Decoding and 4-Sets-Cut-Decoding algorithms. In the decoding process of these codes, three types of errors appear: *more-candidate-error*, *null-error* and *undetected-error*. All of them make damages of the images in the form of horizontal lines. Here, we propose a filter that visually enhance only pixels damaged by first two kinds of errors. This filter cannot be applied for *undetected-errors* since we do not know where they appear. The proposed filter is a median filter, i.e., for enhancing of a damaged pixel it uses the median of surrounding pixels.

The paper is organized as follows. In Section 2, we briefly describe Cut-Decoding and 4-Sets-Cut-Decoding algorithms for RCBQ. The explanation how the experiments are made is given in Section 3. In Section 4, we define the filter for enhancing decoded images. The experimental results are given in Section 5. At the end, we give some conclusions for presented results.

II. Description of Cut-Decoding and 4-Sets-Cut-Decoding algorithms

RCBQs are designed using algorithms for encryption and decryption from the implementation of TASC (Totally Asynchronous Stream Ciphers) by quasigroup string transformation ([4]). These cryptographic algorithms use the alphabet Q and a quasigroup operation $*$ on Q together with its parastrophe $\backslash$.

The notions of quasigroups and quasigroup string transformations are given in the previous papers for these codes ([2], [5], [6], [7]). Here, we use the same terminology and notations as there.

A. Description of coding

At first, let describe Standard coding algorithm for RCBQs proposed in [1]. The message $M = m_1 m_2 \dots m_l$ (of $N_{block} = 4l$ bits where $m_i \in Q$ and Q is an alphabet of 4-bit symbols (nibbles)) is extended to message $L = L^{(1)}L^{(2)}...L^{(s)} = L_1 L_2 ... L_m$ by adding redundant zero symbols. The produced message L has $N = 4m$ bits ($m = rs$), where $L_i \in Q$ and $L^{(i)}$ are sub-blocks of r symbols from Q. In this way we obtain (N_{block}, N) code with rate $R = N_{block}/N$. The codeword is produced after applying the encryption algorithm of TASC (given in Fig. 1) on the message L. For this aim, a key $k = k_1 k_2 ... k_n \in Q^n$ should be chosen. The obtained codeword of M is $C = C_1 C_2 ... C_m$, where $C_i \in Q$.

In Cut-Decoding algorithm, instead of using (N_{block}, N) code with rate R, we use together two $(N_{block}, N/2)$ codes with rate $2R$ for coding/decoding the same message of N_{block} bits. Namely, for coding we apply the encryption algorithm (given in Fig. 1) two times, on the same redundant message L using different parameters (different keys or quasigroups). In this way we obtain the codeword of the message as

Encryption	Decryption
Input: Key $k = k_1k_2 \dots k_n$ and $L = L_1L_2 \dots L_m$ **Output**: codeword $C = C_1C_2...C_m$	**Input**: The pair $(a_1a_2 \dots a_r, k_1k_2 \dots k_n)$ **Output**: The pair $(c_1c_2 \dots c_r, K_1K_2 \dots K_n)$
For $j = 1$ to m $X \leftarrow L_j$; $T \leftarrow 0$; For $i = 1$ to n $X \leftarrow k_i * X$; $T \leftarrow T \oplus X$; $k_i \leftarrow X$; $k_n \leftarrow T$ **Output**: $C_j \leftarrow X$	For $i = 1$ to n $K_i \leftarrow k_i$; For $j = 0$ to $r-1$ $X, T \leftarrow a_{j+1}$; $temp \leftarrow K_n$; For $i = n$ to 2 $X \leftarrow temp \setminus X$; $T \leftarrow T \oplus X$; $temp \leftarrow K_{i-1}$; $K_{i-1} \leftarrow X$; $X \leftarrow temp \setminus X$; $K_n \leftarrow T$; $c_{j+1} \leftarrow X$; **Output**: $(c_1c_2 \dots c_r, K_1K_2 \dots K_n)$

Fig. 1. Algorithms for encryption and decryption

concatenation of two codewords of $N/2$ bits. In 4-Sets-Cut-Decoding algorithm we use four $(N_{block}, N/4)$ codes with rate $4R$, on the same way as in coding with Cut-Decoding algorithm and the codeword of the message is a concatenation of four codewords of $N/4$ bits.

B. Description of decoding

The decoding in all three algorithms is actually a list decoding and it is described below.

In Standard decoding algorithm for RCBQs, after transmission through a noise channel (for our experiments we use Gaussian channel), the codeword C will be received as message $D = D^{(1)}\ D^{(2)} \dots D^{(s)} = D_1D_2 \dots D_m$ where $D^{(i)}$ are blocks of r symbols from Q and $D_i \in Q$. The decoding process consists of four steps: (i) procedure for generating the sets with predefined Hamming distance, (ii) inverse coding algorithm, (iii) procedure for generating decoding candidate sets and (iv) decoding rule.

Let B_{max} be a given integer which denotes the assumed maximum number of errors that occur in a block during transmission. We generate the sets $H_i = \{\alpha | \alpha \in Q^r, \quad H(D^{(i)}, \alpha) \leq B_{max}\}$, for $i = 1, 2, \dots, s$, where $H(D^{(i)}, \alpha)$ is Hamming distance between $D^{(i)}$ and α.

The decoding candidate sets S_0, S_1, S_2,...,S_s are defined iteratively. Let $S_0 = (k_1 \dots k_n; \lambda)$, where λ is the empty sequence. Let S_{i-1} be defined for $i \geq 1$. Then S_i is the set of all pairs $(\delta, w_1w_2 \dots w_{4ri})$ obtained by using the sets S_{i-1} and H_i as follows (w_j are bits). For each element $\alpha \in H_i$ and each $(\beta, w_1w_2 \dots\ w_{4r(i-1)}) \in S_{i-1}$, we apply the inverse coding algorithm (i.e., algorithm for decryption given in Fig. 1) with input (α, β). If the output is the pair (γ, δ) and if both sequences γ and $L^{(i)}$ have the redundant zeros in the same positions, then the pair $(\delta, w_1w_2 \dots w_{4r(i-1)}c_1c_2 \dots c_r) \equiv (\delta, w_1w_2 \dots w_{4ri})$ $(c_i \in Q)$ is an element of S_i.

In Cut-Decoding algorithm, after transmitting through a noisy channel, we divide the outgoing message $D = D^{(1)}D^{(2)} \dots D^{(s)}$ in two messages $D_1 = D^{(1)}D^{(2)} \dots D^{(s/2)}$ and $D_2 = D^{(s/2+1)}\ D^{(s/2+2)} \dots D^{(s)}$ with equal lengths and we decode them parallel with the corresponding parameters. In this decoding algorithm we make modification in the procedure for generating decoding candidate sets. Let $S_i^{(1)}$ and $S_i^{(2)}$ be the decoding candidate sets obtained in the i^{th} iteration of both parallel decoding processes, $i = 1, \dots, s/2$. Then, before the next iteration we eliminate from $S_i^{(1)}$ all elements whose second part does not match with the second part of an element in $S_i^{(2)}$, and vice versa. In the $(i+1)^{th}$ iteration the both processes use the corresponding reduced sets $S_i^{(1)}$ and $S_i^{(2)}$.

In [3] authors proposed 4 different versions of decoding with 4-Sets-Cut-Decoding algorithm. The best results are obtained using 4-Sets-Cut-Decoding algorithm#3. In our experiments we use only this version and further on we briefly describe it. After transmitting through a noisy channel, we divide the outgoing message $D = D^{(1)}D^{(2)}...D^{(s)}$ in four messages $D^1 = D^{(1)}D^{(2)}...D^{(s/4)}$, $D^2 = D^{(s/4+1)}D^{(s/4+2)}...D^{(s/2)}$, $D^3 = D^{(s/2+1)}D^{(s/2+2)}...D^{(3s/4)}$ and $D^4 = D^{(3s/4+1)}D^{(3s/4+2)}...D^{(s)}$ with equal lengths and we decode them parallelly with the corresponding parameters. Similarly, as in Cut-Decoding algorithm, in each iteration of the decoding process we reduce the decoding candidate sets obtained in the four decoding processes, as follows. Let $S_i^{(1)}$, $S_i^{(2)}$, $S_i^{(3)}$ and $S_i^{(4)}$ be the decoding candidate sets obtained in the i^{th} iteration of four parallel decoding processes, $i = 1, \dots, s/4$. Let $V_1 = \{w_1w_2 \dots w_{r \cdot a \cdot i} | (\delta, w_1w_2 \dots w_{r \cdot a \cdot i}) \in S_i^{(1)}\}$, ...,
$V_4 = \{w_1w_2 \dots w_{r \cdot a \cdot i}\ | (\delta, w_1w_2 \dots w_{r \cdot a \cdot i}) \in S_i^{(4)}\}$ and $V = V_1 \cap V_2 \cap V_3 \cap V_4$.
If $V = \emptyset$ then
$V = (V_1 \cap V_2 \cap V_3) \cup (V_1 \cap V_2 \cap V_4) \cup (V_1 \cap V_3 \cap V_4) \cup (V_2 \cap V_3 \cap V_4)$.
Then, before the next iteration we eliminate from $S_i^{(j)}$ all elements whose second part is not in V, $j = 1, 2, 3, 4$.

After the last iteration, if all reduced sets $S_{s/2}^{(1)}$, $S_{s/2}^{(2)}$ in Cut-Decoding algorithm (or $S_{s/4}^{(1)}$, $S_{s/4}^{(2)}$, $S_{s/4}^{(3)}$, $S_{s/4}^{(4)}$ in 4-Sets-Cut-Decoding) have only one element with a same second component then this component is the decoded message L. In this case, we say that we have a *successful decoding*. If the decoded message is not the correct one then we have an *undetected-error*. If the reduced sets obtained in the last iteration have more than one element then we have a *more-candidate-error*. If we obtain $S_i^{(1)} = S_i^{(2)} = \emptyset$ in some iteration of Cut-Decoding or $S_i^{(1)} = S_i^{(2)} = S_i^{(3)} = S_i^{(4)} = \emptyset$ in some iteration of 4-Sets-Cut-Decoding algorithm, then the process will finish (a *null-error* appears). But, if we obtain at least one nonempty decoding candidate set in an iteration then the decoding continues with the nonempty sets (the reduced sets are obtained by intersection of the non-empty sets only).

In [6] authors have proposed a method for decreasing the number of *null-errors* by backtracking. Namely, a *null-error* occurs when more than predicted B_{max} bit errors appear during transmission of some blocks. So, in this method if a *null-error* occurs in some iteration (for example i^{th}), then k previous iterations ($(i-1)^{th}$, $(i-2)^{th}$, ..., $(i-k)^{th}$) are canceled. After that the first of canceled iterations ($(i-k)^{th}$)

is reprocessed with $B_{max} = B_{max} + 1$ or $B_{max} = B_{max} + 2$, and the next iterations continue with the old value of B_{max}. But, with this procedure only part of *null-errors* will be eliminated since we cannot know exactly in which iteration the correct sub-block does not enter in the decoding candidate set and exactly how many transmission errors occur in this sub-block. Also, we must note that with this backtracking in some cases instead of *null-error*, *more-candidate-error* or *undetected-error* can be obtained.

Similar method with backtracking in the case of *more-candidate-error* is proposed in [2]. In this method, if the decoding process ends with more elements in the reduced decoding-candidate sets after the last iteration then a few of iterations are canceled and the first of canceled iterations is reprocessed using smaller value of B_{max} (the next iterations use the old value of B_{max}).

III. Description of experiments

All experiments (presented in this paper) are made for code $(72, 576)$ with rate $R = 1/8$, $B_{max} = 5$ and the following parameters:

- In Cut-Decoding algorithm - redundancy pattern: 1100 1100 1000 0000 1100 1000 1000 0000 1100 1100 1000 0000 1100 1000 1000 0000 0000 0000, for rate $1/4$ and two different keys of 10 nibbles.
- In 4-Sets-Cut-Decoding algorithm - redundancy pattern: 1100 1110 1100 1100 1110 1100 1100 1100 0000 for rate $1/2$ and four different keys of 10 nibbles.
- In all experiments we used the same quasigroup on Q given in Table I.

TABLE I
QUASIGROUP OF ORDER 16 USED IN THE EXPERIMENTS

$*$	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	3	c	2	5	f	7	6	1	0	b	d	e	8	4	9	a
1	0	3	9	d	8	1	7	b	6	5	2	a	c	f	e	4
2	1	0	e	c	4	5	f	9	d	3	6	7	a	8	b	2
3	6	b	f	1	9	4	e	a	3	7	8	0	2	c	d	5
4	4	5	0	7	6	b	9	3	f	2	a	8	d	e	c	1
5	f	a	1	0	e	2	4	c	7	d	3	b	5	9	8	6
6	2	f	a	3	c	8	d	0	b	e	9	4	6	1	5	7
7	e	9	c	a	1	d	8	6	5	f	b	2	4	0	7	3
8	c	7	6	2	a	f	b	5	1	0	4	9	e	d	3	8
9	b	e	4	9	d	3	1	f	8	c	5	6	7	a	2	0
a	9	4	d	8	0	6	5	7	e	1	f	3	b	2	a	c
b	7	8	5	e	2	a	3	4	c	6	0	d	f	b	1	9
c	5	2	b	6	7	9	0	e	a	8	c	f	1	3	4	d
d	a	6	8	4	3	e	c	d	2	9	1	5	0	7	f	b
e	d	1	3	f	b	0	2	8	4	a	7	c	9	5	6	e
f	8	d	7	b	5	c	a	2	9	4	e	1	3	6	0	f

In the experiments with the both algorithms we use the following combination of two methods with backtracking explained in the previous section. If the decoding ends with *null-error*, then the last two iterations are canceled and the first of them is reprocessed with $B_{max} + 2$ (the next iterations use the previous value of B_{max}). If the decoding ends with *more-candidate-error*, then the last two iterations of the decoding process are canceled and the penultimate iteration is reprocessed with $B_{max} - 1$. In the decoding of a message only one backtracking is made, except when after the backtracking for *null-error*, *more-candidate-error* appears and more than one decoding candidate set is non-empty. In this case, we make one more backtracking for *more-candidate-error*.

In all decoding algorithms for RCBQ, when a *null-error* appears, the decoding process ends earlier and only a part of the message is decoded. Therefore in the experiments with images we use the following solution. In the cases when a *null-error* appears, i.e., all reduced sets are empty in some iteration, we take the strings without redundant symbols from all elements in the sets from the previous iteration and we find their maximal common prefix substring. If this substring has k symbols then in order to obtain decoded message of l symbols we take these k symbols and we add $l - k$ zero symbols at the end of the message. When the decoding process ends with a *more-candidate-error*, we take a message of l zero symbols as a decoded message. These zero symbols (added in the both types of detected errors) make a horizontal black lines on the image and these zero symbols are used in the definition of the filter proposed in this paper. Since, we do not know the position of *undetected-errors*, with this filter we cannot enhance pixels damaged from this type of errors.

In all presented experiments we consider a transmission through Gaussian channel (with different values of SNR) of the image of "Lenna", given in Fig. 2.

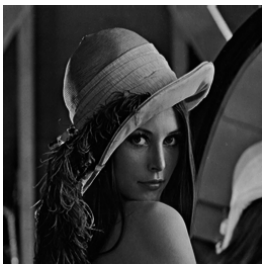


Fig. 2. Lenna

IV. Definition of the filter for images

In order to visually enhance damaged pixels and improve the image, here we define a filter that transforms pixel intensity values of the pixels damaged by both types of detected errors (*null-errors* and *more-candidate-errors*). One pixel is considered as damaged if it belongs in a zero sub-block with at least four consecutive zero nibbles. The basic idea in definition of this filter is to replace damaged pixel intensity value with a new value taken over a neighborhood of fixed size. In this process we use the median of the nonzero gray values of the surrounding pixels, so our filter is a median filter.

For each damaged pixel in the position (i, j), the filter uses the following algorithm:

1. take a 3 x 3 region centered around pixel (i, j);

2. sort the nonzero intensity values of the pixels in the region into ascending order;

3. select the middle value (the median) as the new value of pixel (i, j).

V. Experimental results

In Fig. 3 – 6, we present images obtained with Cut-Decoding algorithm before and after application of the proposed filter for $SNR = -2$, $SNR = -1$, $SNR = 0$ and $SNR = 1$, correspondingly.

Fig. 3. $SNR = -2$

Fig. 4. $SNR = -1$

Fig. 5. $SNR = 0$

In Fig. 7 – 10, we present images obtained with 4-Sets-Cut-Decoding algorithm before and after application of the proposed filter for $SNR = -2$, $SNR = -1$, $SNR = 0$ and $SNR = 1$, correspondingly.

From the presented images we can notice that the proposed filter provides a great improvement of the images for all considered values of SNR. Also, this filter gives better results

Fig. 6. $SNR = 1$

Fig. 7. $SNR = -2$

Fig. 8. $SNR = -1$

Fig. 9. $SNR = 0$

for the images obtained with Cut-Decoding algorithm than with 4-Sets-Cut-Decoding algorithm. The reason for this is the larger number of *undetected-errors* produced with 4-Sets-Cut-Decoding algorithm.

VI. Conclusion

From all presented results we can conclude that the proposed median filter enhances the images decoded with Cut-Decoding algorithm and 4-Sets-Cut-Decoding algorithm after

Fig. 10. $SNR = 1$

transmission through Gaussian channel. We analyze why the results for the images obtained with Cut-Decoding algorithm are better than with 4-Sets-Cut-Decoding algorithm and we notice that the methods with backtracking decrease the number of detected errors, but produce more *undetected-errors* which cannot be filtered. Therefore, for further research we can consider application of the proposed filter on images decoded with both algorithms, but without backtracking.

ACKNOWLEDGMENT

This research was partially supported by Faculty of Computer Science and Engineering at "Ss Cyril and Methodius" University in Skopje.

REFERENCES

[1] D. Gligoroski, S. Markovski, Lj. Kocarev, "Error-correcting codes based on quasigroups", Proc. 16th Intern. Confer. Computer Communications and Networks, 2007, pp. 165-172.

[2] A. Popovska-Mitrovikj, S. Markovski, V. Bakeva, "Increasing the decoding speed of random codes based on quasigroups". S. Markovski, M. Gusev (Eds.), ICT Innovations 2012, Web proceedings, ISSN 1857-7288, 2012, pp. 93–102.

[3] A. Popovska-Mitrovikj, S. Markovski, V. Bakeva, "4-Sets-Cut-Decoding algorithms for random codes based on quasigroups", International Journal of Electronics and Communications (AEU), Elsevier, Vol.69, Issue 10, 2015, pp. 1417–1428.

[4] D. Gligoroski, S. Markovski, Lj. Kocarev, "Totally asynchronous stream ciphers + Redundancy = Cryptcoding". S. Aissi, H.R. Arabnia (Eds.), Proc. Internat. Confer. Security and management, SAM 2007, Las Vegas, CSREA Press, 2007, pp. 446–451.

[5] A. Popovska-Mitrovikj, V. Bakeva, S. Markovski, "On random error correcting codes based on quasigroups", Quasigroups and Related Systems, Vol. 19, 2011, pp. 301–316.

[6] A. Popovska-Mitrovikj, S. Markovski, V. Bakeva, "Performances of error-correcting codes based on quasigroups". D.Davcev, J.M.Gomez (Eds.), ICT-Innovations 2009, Springer, 2009, 2009, pp. 377–389.

[7] A. Popovska-Mitrovikj, S. Markovski, V. Bakeva, "Some New Results for Random Codes Based on Quasigroups", Proc.10th Conference on Informatics and Information Technology with International Participants, Bitola, Macedonia, 2013, pp. 178–181.

[8] J.G. Proakis, M. Salehi, Digital Communications (5th edn), McGrawHill Higher Education, 2008.

Graph theoretical approach for construction of Lyapunov function for a coupled stochastic neural network

Biljana Tojtovska
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
Email: biljana.tojtovska@finki.ukim.mk

***Abstract*—In this paper, we describe a new model of coupled stochastic neural network given by a system of stochastic functional differential equations (SFDE's) and give a way for construction of a Lyapunov function of the system. The considered coupled system is in fact a large system of SFDEs driven by n-dimensional Brownian motion, with impulses and Markovian switching. This complex system consists of large number of interconnected, mutually interacting neural networks with their own dynamics. The considered model is more complex than the ones presented in the literature and thus it is more difficult to analyze its stability properties. We take an approach from the graph theory which will give us an elegant way to construct the Lyapunov function. The result is important since the function can be effectively used to analyze the stability properties of the coupled system.**

Keywords*—coupled systems; coupled neural networks; graph theory Lyapunov function

I. Introduction

Complex networks are large networks of systems with individual dynamics, which interact between each other based on some coupling structure. There are numerous examples of complex networks both in nature and in engineering - electrical power grids, the World Wide Web, biochemical reaction networks, cellular and metabolic networks, large groups of interacting neurons and above all of them, the human brain [1].

The complex systems consist of large number of interconnected, mutually interacting parts and they can be studied effectively using graph theory. This idea came at the end of the last century, from the revolutionary work by Watts and Strogatz [2] and Barabási Albert [3]. In this graph-theoretic approach the system is modeled by a graph, where each node represents a dynamical unit and the edges are build based on the interactions between the nodes. This way, different network models can be studied. For example the network connections may change over time, the links between the nodes can have different signs, weights and directions. This is seen in nervous systems where the synapses can be weak or strong, inhibitory or excitatory. The dynamics of the individual systems can also affect the weights assigned to the connecting edges. A well known example comes from the Hebbian theory in Neuroscience [4] - the neurons that are coupled together, and often fire together, in time strengthen their interconnections. Finally, the whole network can have a complex dynamics, for example, when the nodes themselves are nonlinear or even stochastic dynamical systems. Thus the complex networks are also studied by the theory of nonlinear dynamics where they are defined as coupled dynamical systems. In this approach it is very often assumed that the network has a regular, simple architecture, i.e. the models can be based on grids, chains, lattices or fully-connected graphs. With these assumptions we can focus our study on the collective behavior of the coupled system and the individual dynamics in the nodes, and forget about the possible complex topology, even though it also affects the dynamics of the whole system [1].

One aspect of coupled dynamical systems which received great attention, especially in recent decades, is synchronization. An important issue is the one of complete synchronization, i.e. stability of the synchronous state. Yamada and Fujisaka [5] were among the first who studied stability in synchronous, coupled chaotic systems. Their analysis of the change of the dynamics in the system was based on Lyapunov exponents of the coupled systems. Criteria for the stability of the synchronized motion are very often obtained in terms of Lyapunov exponents [6], [7]. In [8], the author studies how the interaction structure affects the stability and stabilization of the system. All this emphasizes one more time the importance of stability analysis and the vast possibilities for its application. In this paper we will use the graph theory approach to describe a new model of coupled stochastic neural network and explain how to approach the construction of a Lyapunov function, which can then be used for stability analysis of the system.

II. Coupled systems of SNNs with impulses, Markovian switching, and node delays

In this section, we describe a network constructed by coupling of M neural networks which have own internal dynamics. We assume that the kth network, $k = 1, 2, ..., M$ consists of l_k connected neurons. However, to simplify the

notations we define $n = max\{M, l_1, l_2, ..., l_M\}$ and assume that all the neural networks have dimension n and also that the whole system consists of n such networks. This can be achieved by adding "dead" neurons in the system which have zero dynamics modeled by functions which are constantly zero and do not affect the dynamics of the whole system. Such a system can be represented by a directed graph $\mathcal{G}$ with n vertices, where each vertex represents one neural network from the system. The directed edge $(k, j) \in E_{\mathcal{G}}$ exists if the jth network is connected to the kth network. The model is defined on a complete probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq t_0}, \mathbb{P})$ with a natural filtration $\{\mathcal{F}_t\}_{t\geq t_0}$ generated by a standard n-dimensional Brownian motion $W = \{W(t), t \geq t_0\}$.

The dynamics of the kth vertex, $k \in N = \{1, 2, ..., n\}$, is given by the following stochastic differential equation with Markovian switching, delays and impulses at times $t_m \in \mathbb{R}$:

For $t \geq t_0$, $t \neq t_m$, and $i, k \in N = \{1, 2, ..., n\}$

$$
\begin{aligned}
dX_i^{(k)}(t) = &-h_i^{(k)}\big(X_i^{(k)}(t), r(t)\big)\Bigg[c_i^{(k)}\big(t, X_i^{(k)}(t), r(t)\big) \\
&- \sum_{j=1}^{n} a_{ij}^{(k)}(t, r(t)) f_j^{(k)}\big(X_j^{(k)}(t), r(t)\big) \\
&- \sum_{j=1}^{n} b_{ij}^{(k)}(t, r(t)) g_j^{(k)}\big(X_{j,\tau_k}^{(k)}, r(t)\big) \\
&- \sum_{j=1}^{n} d_{ij}^{(k)}(t, r(t)) \int_{-\infty}^{t} l_{ij}^{(k)}(t-s) k_j^{(k)}\big(X_j^{(k)}(s), r(s)\big)\, ds\Bigg] dt \\
&+ \sum_{j=1}^{n} \eta_i^{(kj)}\big(t, X_i^{(k)}(t), X^{(j)}(t), r(t)\big) dt \\
&+ \sum_{j=1}^{n} \sigma_{ij}^{(k)}\big(t, X_j^{(k)}(t), X_{j,\tau_k}^{(k)}, r(t)\big)\, dW_j(t) \\
&+ \sum_{j=1}^{n} \zeta_i^{(kj)}\big(t, X_i^{(k)}(t), X^{(j)}(t), r(t)\big) dW_j(t) \\
&\equiv \Bigg(J_i^{(k)}\big(t, X^{(k)}(t), X_{\tau_k}^{(k)}(t), r(t)\big) \\
&+ \sum_{j=1}^{n} \eta_i^{(kj)}\big(t, X_i^{(k)}(t), X^{(j)}(t), r(t)\big)\Bigg) dt \\
&+ \sum_{j=1}^{n} \Bigg(\sigma_{ij}^{(k)}(t, X_j^{(k)}(t), X_{j,\tau_k}^{(k)}, r(t)) \\
&+ \zeta_i^{(kj)}\big(t, X_i^{(k)}(t), X^{(j)}(t), r(t)\big)\Bigg) dW_j(t), \qquad (1)
\end{aligned}
$$

and for $t = t_m$, $m \in \mathbb{N}$,

$$
\begin{aligned}
X_i^{(k)}(t) = &\mathcal{I}_{im}\big(X_1^{(k)}(t^-), ..., X_n^{(k)}(t^-)\big) \\
&+ \mathcal{J}_{im}\big(X_1^{(k)}(t^- - \tau_k(t^-)), ..., X_m^{(k)}(t^- - \tau_k(t^-))\big), \qquad (2)
\end{aligned}
$$

and an initial condition

$$
X_i^{(k)}(t_0 + s) = \xi_i^{(k)}(s),\ s \in (-\infty, t_0], \qquad (3)
$$

where $X_i^{(k)} \in \mathbb{R}$ is the state stochastic process of the ith neuron in the kth vertex at time t. We denote that $X^{(k)}(t) = (X_1^{(k)}, X_2^{(k)}, ..., X_n^{(k)})^T \in \mathbb{R}^n$ is the process which describes the dynamics in the kth vertex and $X(t) = \big(X^{(1)}(t), ..., X^{(n)}(t)\big) \in \mathbb{R}^{n\times n}$ is the stochastic process which describes the dynamics of the whole coupled system. $\xi_i^{(k)} \in C((-\infty, t_0]; \mathbb{R})$ is an initial condition for the corresponding neuron, and we also denote $\xi^{(k)} = (\xi_1^{(k)}, ..., \xi_n^{(k)})^T$ and $\xi = (\xi_i^{(k)})_{n\times n}$. $X_{\tau_k}^{(k)} = (X_{1,\tau_k}^{(k)}, ..., X_{n,\tau_k}^{(k)})^T \in \mathbb{R}^n$ is the delayed process in the kth vertex, where $X_{i,\tau_k}^{(k)} = X_i^{(k)}(t - \tau_k(t))$ is the delayed process of the corresponding neuron dependent on transmission delay $\tau_k(t)$. For simplification, we assume that the delay τ_k is the same for the whole vertex k and is such that $0 \leq \tau_k(t) \leq \tau$, τ is a constant. For the delayed process in the whole network, we use the notation $X_t = \big(X_{\tau_1}^{(1)}, ..., X_{\tau_n}^{(n)}\big)$.

The switching function $r(t)$ is a right-continuous Markov chain independent of the underlying Brownian motion, taking values in the finite space $\mathcal{P} = \{1, 2, ..., m\}$, with $r(t_0) = \rho_0$ and with a generator matrix $\Pi = (\pi_{ij})_{m\times m}$ of $r(t)$ is given by

$$
\mathbb{P}(r(t+\Delta) = j | r(t) = i) = \begin{cases} \pi_{ij} + o(\Delta), & i \neq j, \\ 1 + \pi_{ij} + o(\Delta), & i = j, \end{cases} \qquad (4)
$$

where $\Delta > 0$, $\pi_{ij} \geq 0$ is the transition rate from i to j if $i \neq j$, while $\pi_{ii} = -\sum_{j\neq i} \pi_{ij}$. To simplify the notation we fix one mode $r(t) = \rho \in \mathcal{P}$ and write $X_i^{(k)}(t) = X_i^{(k)}, X^{(k)}(t) = X^{(k)}, h_i^{(k)}(X_i^{(k)}(t), \rho) = h_{i,\rho}^{(k)}$, $c_i^{(k)}(t, r(t)) = c_{i,\rho}^{(k)}$ and so on. Then, the equation (1) can be shortly written as

$$
dX_i^{(k)} = \left(J_{i,\rho}^{(k)} + \sum_{j=1}^{n} \eta_{i,\rho}^{(kj)}\right) dt + \sum_{j=1}^{n} \left(\sigma_{ij,\rho}^{(k)} + \zeta_{i,\rho}^{(kj)}\right) dW_j \qquad (5)
$$

The meaning of the functions in the model is the following - $h_{i,\rho}^{(k)}$ are amplification functions at time t, $c_{i,\rho}^{(k)}$ are appropriately behaved functions dependent on t and on the state processes $X_i^{(k)}$, while $a_{ij,\rho}^{(k)}, b_{ij,\rho}^{(k)}(t)$ and $d_{ij,\rho}(t)^{(k)}$ describe the strength of the neuron interconnections in the kth vertex network at times t, $f_{i,\rho}^{(k)}, k_{i,\rho}^{(k)}$ and $g_{i,\rho}^{(k)}$ are activation functions of the ith neuron of the kth vertex at time t and $t - \tau_k(t)$, respectively, and $l_{ij}(t)$ are delay kernel functions. The term $\sigma_\rho^{(k)}(t, X^{(k)}, X_{\tau_k}^{(k)}) = \big(\sigma_{ij,\rho}^{(k)}(t, X_j^{(k)}, X_{j,\tau_k}^{(k)})\big)_{n\times n}$ is a diffusion-coefficient matrix. New in this model are the interconnection functions $\eta_\rho^{(kj)} = \big(\eta_{1,\rho}^{(kj)}, \eta_{2,\rho}^{(kj)}, ..., \eta_{n,\rho}^{(kj)}\big)$ and $\zeta_\rho^{(kj)} = \big(\zeta_{1,\rho}^{(kj)}, \zeta_{2,\rho}^{(kj)}, ..., \zeta_\rho^{(kj)}\big)$ which represent the influence from the jth vertex to the kth vertex, $k \neq j$ and we take $\eta_\rho^{(kj)} = \zeta_\rho^{(kj)} \equiv 0$, for $k = j$. Here $\eta_i^{(kj)}, \zeta_i^{(kj)} : [t_0, \infty) \times \mathbb{R} \times \mathbb{R}^n \times \mathcal{P} \to \mathbb{R}$ depend on the time t, the state of the ith neuron in the kth vertex, the state of the vertex j and the switching function $r(t)$.

The impulses in the whole coupled network happen at fixed moments $t_m, m \in \mathbb{N}$ satisfying $t_1 < t_2 < \cdots$ and $\lim_{m\to\infty} t_m = \infty$, $\mathcal{I}_{im}(X_1^{(k)}(t^-), ..., X_n^{(k)}(t^-)) \in \mathbb{R}$ are impulsive perturbations of the ith neuron in the kth vertex at time t_m, where $X_j^{(k)}(t^-)$ is the left limit of $X_j^{(k)}(t)$, $\mathcal{J}_{im}(X_1^{(k)}(t^- - \tau_k(t^-)), ..., X_n^{(k)}(t^- - \tau_k(t^-))) \in \mathbb{R}$ are impulsive perturbations of the ith neuron in the kth vertex at time t_m caused by transmission delays. We denote that $\mathcal{I}_m = (\mathcal{I}_{1m}, ..., \mathcal{I}_{nm})^T$ and $\mathcal{J}_m = (\mathcal{J}_{1m}, ..., \mathcal{J}_{nm})^T$.

If we define $J_\rho^{(k)} := (J_{1,\rho}^{(k)}, \ldots, J_{n,\rho}^{(k)})^T$, the system (1) can be represented in the matrix form,

$$
\begin{aligned}
dX^{(k)}(t) &= \Big(J_\rho^{(k)}\big(t, X^{(k)}(t), X_{\tau_k}^{(k)}\big) \\
&+ \sum_{j=1}^{n} \eta_\rho^{(kj)}\big(t, X^{(k)}(t), X^{(j)}(t)\big)\Big) dt \\
&+ \sum_{j=1}^{n} \Big(\sigma_\rho^{(k)}\big(t, X^{(k)}(t), X_{\tau_k}^{(k)}\big) + \zeta_\rho^{(kj)}\big(t, X^{(k)}(t), X^{(j)}(t)\big)\Big) dW_j \\
&\equiv \big(J^{(k)} + \sum_{j=1}^{n} \eta_\rho^{(kj)}\big) dt + \big(\sigma_\rho^{k} + \sum_{j=1}^{n} \zeta_\rho^{(kj)}\big) dW_j(t), \qquad (6)
\end{aligned}
$$

where $t \geq t_0,\ t \neq t_m, k \in N$. Additionally, there are impulsive perturbations in the points $t = t_m$, $k \in N$, $m \in \mathbb{N}$, given with

$$X^{(k)}(t) = \mathcal{I}_m(X^{(k)}(t^-)) + \mathcal{J}_m(X_{\tau_k}^{(k)}(t^-)), \qquad (7)$$

and initial condition

$$X^{(k)}(t_0 + s) = \xi^{(k)}(s), \quad s \in (-\infty, t_0]. \qquad (8)$$

Since our research is focused on stability problems, we assume with no emphasis on conditions that there exists a unique global solution $X(t; \xi, \rho_0)$ to the system (1) satisfying $\mathbb{E}\sup_{t\in\mathbb{R}} ||X(t; \xi, \rho_0)||^p < \infty$, as well as that all the Lebesgue and Itô integrals employed further are well defined. The considered coupled system is in fact a large system of SFDEs with impulses and Markovian switching and because of that, we can use the known literature which discusses existence and uniqueness of a solution. For more details on SFDEs with Markovian switching, we refer to the work by Mao et. al. [9], [10].

For the stability purpose, we usually assume that $J_\rho^{(k)}(t, 0, 0) = \eta_\rho^{(kh)}(t, 0, 0) = \sigma_\rho^{(k)}(t, 0, 0) = \zeta_\rho^{(kh)}(t, 0, 0) \equiv 0$ so that equation (1) admits a trivial solution $X \equiv 0$. We additionally assume that if there exist "dead" neurons in the system, they are modeled by constant zero functions and have no influence on the dynamics on the other neurons.

III. Construction of global Lyapunov function

In this section we describe a way for construction of Lyapunov function for the system (1), based on known results in the literature for coupled systems of differential equations [11] and coupled systems of stochastic functional differential equations [12].

In [11] the authors consider a complex system described by a directed graph $\mathcal{G}$ with $n \geq 2$ vertices, each with its own dynamics. The dynamics of the kth vertex in coupled system on the graph $\mathcal{G}$ ($k \in \{1, 2, ..., n\}$) is given by

$$
\begin{aligned}
dX^{(k)}(t) &= f^{(k)}(t, X^{(k)}(t))\, dt \\
&+ \sum_{j=1}^{n} g^{(kj)}\big(t, X^{(k)}(t), X^{(j)}(t)\big), k \in N,\ t \in [t_0, \infty), \\
X(t_0) &= x_0, \qquad (9)
\end{aligned}
$$

We assume that the functions $f^{(k)}$ and $g^{(kj)}, k, j \in N$, fulfill the conditions for the existence and uniqueness of a solution to the initial-value problem.

Let $D^{(k)} \subseteq \mathbb{R}^{l_k}$ be an open set and let $V^{(k)} : \mathbb{R} \times D^{(k)} \to \mathbb{R}$ be a Lipschitz function, for which the Lyapunov derivative with respect to the system (9) is given with

$$
\begin{aligned}
\dot{V}^{(k)}(t, x^{(k)}) &:= \frac{\partial V^{(k)}(t, x^{(k)})}{\partial t} \\
&+ \frac{\partial V^{(k)}(t, x^{(k)})}{\partial x^{(k)}} \Big(f^{(k)}(t, x^{(k)}) + \sum_{=1}^{n} g^{(kh)}(t, x^{(k)}, x^{(h)}) \Big) \qquad (10)
\end{aligned}
$$

Let $D = D^{(1)} \times D^{(2)} \times ... \times D^{(n)} \subseteq \mathbb{R}^l$, where $l = l_1 + l_2 + \cdots + l_n$, $x = (x^{(1)}, x^{(2)}, ..., x^{(n)}) \in \mathbb{R}^l$, and $x^{(k)} \in \mathbb{R}^{l_k}$. For a Lipschitz function $V : \mathbb{R} \times D \to \mathbb{R}$, we define

$$
\begin{aligned}
\dot{V}(t, x) &:= \frac{\partial V(t, x)}{\partial t} \\
&+ \sum_{k=1}^{n} \frac{\partial V(t, x)}{\partial x^{(k)}} \Big(f^{(k)}(t, x^{(k)}) + \sum_{h=1}^{n} g^{(k)}(t, x^{(k)}, x^{(h)}) \Big).
\end{aligned}
$$

In [11] the authors have discussed the global stability problem of coupled systems of the form (9). One of the assumptions is that, when isolated, each vertex system is globally stable and has a Lyapunov function $V^{(k)}$. Then, they try to construct a global Lyapunov function as a linear combination of the vertex-Lyapunov functions

$$V(t, x) = \sum_{k=1}^{n} \varsigma^{(k)} V^{(k)}(t, x^{(k)}). \qquad (11)$$

The construction of such global Lyapunov function for the coupled system is possible, based on the assumptions of the following theorem.

Theorem 1: Let the following assumptions be satisfied:

(i) There are functions $V^{(k)}(t,x^{(k)})$, $F^{(kj}(t,x^{(k)},x^{(j)})$ and constants $a^{(kj)} \geq 0$ such that

$$\dot{V}^{(k)}(t,x^{(k)}) \leq \sum_{j=1}^{n} a^{(kj)} F^{(kj)}(t,x^{(k)},x^{(j)}),;$$

for $t>0$, $x^{(k)} \in D^{(k)}, x^{(j)} \in D^{(j)}$, $k,j \in N$

(ii) Along each directed cycle $\mathcal{C}$ of the weighted graph $(\mathcal{G}, A_{\mathcal{G}})$, $A_{\mathcal{G}} = (a^{(kj)})_{n\times n}$ it holds

$$\sum_{(s,r)\in E_{\mathcal{C}}} F^{(rs)}(t,x^{(r)},x^{(s)}) \leq 0;$$

for $t>0$, $x^{(r)} \in D^{(r)}$, $x^{(s)} \in D^{(s)}$

(iii) The constants $\varsigma^{(k)}, k \in N$ in (11), are given by in the Kirchhoff's theorem, as presented in [11].

Then, the function $V(t,x)$ in (11) satisfies

$$\dot{V}(t,x) \leq 0 \quad \text{for } t>0,\ x \in D,$$

i.e. it is a Lyapunov function for the coupled system (9).

In [11] the authors give an example which shows that the existence of Lyapunov functions for each vertex system is not sufficient for the existence of global Lyapunov function for the coupled system. Hence, it is important to give conditions on the network and the coupling structure which will imply existence of a global Lyapunov function, i.e. it will imply stability of the coupled system. In the proof of the results, the authors construct a global Lyapunov function V for the considered coupled system using the Laplacian matrix of the graph $(\mathcal{G}, A_\rho)$ and vertex Lyapunov functions $V^{(k)}$ which are known. In [12] Li et al. extended the ideas from [13] and discussed stability of a system of coupled stochastic functional differential equations (CSFDE) with no Markovian switching, defined on a graph $\mathcal{G}$. As in [13], the authors show that the global Lyapunov function can be constructed as a weighted sum of the vertex Lyapunov functions, which may be known from the studies. The authors also give sufficient conditions for the pth moment and almost sure exponential stability based on the M-matrix method. In addition, they extend their discussion to stochastic coupled systems with time-varying delays, which are present only in the interconnection functions. However, the theory in both [12], [13] does not consider a model with both time-varying delays and Markovian switching, and also impulsive effects have not been assumed in the systems. Even more, both [12], [13] base their models on one-dimensional Brownian motion. This is not very realistic, since the system may be influenced by different independent random sources and thus, as in the previous chapters, the model should include an n-dimensional Brownian motion. Thus, the theory in these papers can not be directly applied for stability of our model.

Motivated by this, we suggest local Lyapunov functions and construct the global Lyapunov function for the system (1). For $V^{(k)}(t,x^{(k)},\rho) \in C^{1,2}(\mathbb{R}^+ \times \mathbb{R}^n \times \mathcal{P}; \mathbb{R}^+)$, we define an operator $LV^{(k)}(t,x^{(k)},\rho)$ associated with the kth vertex of system (1) by

$$\begin{aligned}
&LV^{(k)}(t,X^{(k)},\rho) = \\
&= \sum_{j=1}^{m} \pi_{ij} V^{(k)}(t,X^{(k)},\rho_j) + \frac{\partial V^{(k)}(t,X^{(k)},\rho)}{\partial t} \\
&+ \frac{\partial V^{(k)}}{\partial x^{(k)}} \left[J^{(k)}(t,X^{(k)},X^{(k)}_{\tau_k},\rho) + \sum_{j=1}^{n} \eta^{(kj)}(t,X^{(k)},X^{(j)},\rho) \right] \\
&+ \frac{1}{2} trace \Bigg\{ \left[\sigma^{(k)}(t,X^{(k)},X^{(k)}_{\tau_k},\rho) + \sum_{j=1}^{n} \zeta^{(kj)}(t,X^{(k)},X^{(j)},\rho) \right]^T \\
&\times \frac{\partial^2 V^{(k)}}{\partial (x^{(k)})^2} \left[\sigma^{(k)}(t,X^{(k)},X^{(k)}_{\tau_k},\rho) + \sum_{j=1}^{n} \zeta^{(kj)}(t,X^{(k)},X^{(j)},\rho) \right] \Bigg\},
\end{aligned}$$

where

$$\frac{\partial V^{(k)}}{\partial x^{(k)}} = \Big(\frac{\partial V^{(k)}(t,x^{(k)},\rho)}{\partial x_1^{(k)}}, \cdots, \frac{\partial V^{(k)}(t,x^{(k)},\rho)}{\partial x_{l_k}^{(k)}} \Big), \quad (12)$$

and

$$\frac{\partial^2 V^{(k)}}{\partial (x^{(k)})^2} = \left(\frac{\partial^2 V^{(k)}(t,x^{(k)},\rho)}{\partial x_i^{(k)} \partial x_j^{(k)}} \right)_{l_k \times l_k}. \quad (13)$$

The next step in the stability analysis is to choose local (vertex) Lyapunov functions, give sufficient conditions and prove that under those conditions the constructed function is in fact a global Lyapunov function for the system (1). We suggest the following approach. Let $x^{(k)} = (x_1^{(k)}, ..., x_2^{(k)})^T \in \mathbb{R}^n$ and $x = (x^{(1)}, ..., x^{(n)})) \in \mathbb{R}^{n\times n}$. Let $V_{i,\rho}^{(k)}(x_i^{(k)}) = V_i^{(k)}(x_i^{(k)},\rho) = \theta_\rho^{(k)} |x_i^{(k)}|^p$ is the Lyapunov function corresponding to the ith neuron in the kth vertex, where $\theta_\rho^{(k)} > 0$ are some constants. Also, let $V_\rho^{(k)}(x^{(k)}) = V^{(k)}(x^{(k)},\rho) = \sum_{i=1}^{n} V_{i,\rho}^{(k)}(x_i^{(k)}) = \sum_{i=1}^{n} \theta_\rho^{(k)} |x_i^{(k)}|^p$. Then, for the coupled system (1) we have

$$\begin{aligned}
LV_{i,\rho}^{(k)}(X_i^{(k)}) &= \sum_{j=1}^{m} \pi_{\rho j} V_{i,j}^{(k)}(X_i^{(k)}) + \frac{\partial V_{i,\rho}^{(k)}(X_i^{(k)})}{\partial t} \\
&+ \frac{\partial V_{i,\rho}^{(k)}(X_i^{(k)})}{\partial x_i^{(k)}} \Big(J_{i,\rho}^{(k)} + \sum_{j=1}^{n} \eta_{i,\rho}^{(kj)} \Big) \\
&+ \frac{1}{2} \frac{\partial^2 V_{i,\rho}^{(k)}(X_i^{(k)})}{\partial (x_i^{(k)})^2} \sum_{j=1}^{n} \Big(\sigma_{ij,\rho}^{(k)} + \zeta_{i,\rho}^{(kj)} \Big)^2
\end{aligned}$$

For each $\rho \in \mathcal{P}$, let us denote by $\mathcal{G}_\rho$ the corresponding complete directed graph $\mathcal{G}_\rho$ and assign weight $\alpha_\rho^{(kj)} = \theta_\rho^{(k)}$ to each directed edge (j,k), $j \neq k$. Let $\mathcal{A}_{\mathcal{G},\rho} = (\alpha_\rho^{(kj)})_{n\times n}$ represent the weight matrix of the graph $\mathcal{G}_\rho$. Then, the corresponding Laplacian matrix is given by

$$\mathcal{L}_{\mathcal{G},\rho} = \begin{bmatrix} (n-1)\theta_\rho^{(1)} & -\theta_\rho^{(1)} & \cdots & -\theta_\rho^{(1)} \\ -\theta_\rho^{(2)} & (n-1)\theta_\rho^{(2)} & \cdots & -\theta_\rho^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_\rho^{(n)} & -\theta_\rho^{(n)} & \cdots & (n-1)\theta_\rho^{(n)} \end{bmatrix}.$$

Let $V_\rho^{(k)}(x^{(k)}) = \sum_{i=1}^{n} V_{i,\rho}^{(k)}(x_i^{(k)})$. Then $V_\rho(x) = \sum_{k=1}^{n} \varsigma_\rho^{(k)} V_\rho^{(k)}(x^{(k)})$ for $x \in \mathbb{R}^{n\times n}$ denoted as before, where $\varsigma_\rho^{(k)}$ is the cofactor of the kth diagonal element of $\mathcal{L}_{\mathcal{G},\rho}$.

The discussion presented in this paper should be followed by a detailed analysis of the coupled system (1), which will give sufficient conditions under which the pth moment exponential stability can be proven, with the use of the proposed function.

REFERENCES

[1] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[2] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[3] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[4] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, 2002.

[5] H. Fujisaka and T. Yamada, "Stability theory of synchronized motion in coupled-oscillator systems," *Progress of Theoretical Physics*, vol. 69, no. 1, pp. 32–47, 1983.

[6] L. M. Pecora and T. L. Carroll, "Master stability functions for synchronized coupled systems," *Physical Review Letters*, vol. 80, no. 10, pp. 2109–2112.

[7] T. Pereira, "Stability of synchronized motion in complex networks (lecture notes)," *arXiv preprint arXiv:1112.2297*, 2011.

[8] Z. Lin, "Coupled dynamic systems: from structure towards stability and stabilizability," Ph.D. dissertation, University of Toronto, 2006.

[9] X. Mao, "Stochastic functional differential equations with Markovian switching," *Functional Differential Equations*, vol. 6, no. 3-4, pp. 375–396, 1999.

[10] X. Mao, A. Matasov, and A. B. Piunovskiy, "Stochastic differential delay equations with Markovian switching," *Bernoulli*, vol. 6, no. 1, pp. 73–90, 2000.

[11] M. Y. Li and Z. Shuai, "Global-stability problem for coupled systems of differential equations on networks," *Journal of Differential Equations*, vol. 248, no. 1, pp. 1–20, 2010.

[12] W. Li, X. Qi, M. Pan, and K. Wang, "Razumikhin-type theorems on exponential stability of stochastic functional differential equations on networks," *Neurocomputing*, vol. 131, pp. 278–285, 2014.

[13] W. Li, H. Song, Y. Qu, and K. Wang, "Global exponential stability for stochastic coupled systems on networks with Markovian switching," *Systems & Control Letters*, vol. 62, no. 6, pp. 468–474, 2013.

Binary Invasive Weed Optimization Algorithm Approaches for Binary Optimization

Ismail Koc, Refik Nureddin, Ismail Babaoglu, Sait Ali Uymaz

Department of Computer Engineering,

Faculty of Engineering, Selçuk University,

Konya, Turkey

e-mail: ismailkoc@selcuk.edu.tr, refik.nureddin@gmail.com, ibabaoglu@selcuk.edu.tr, aliuymaz@selcuk.edu.tr

Abstract— **Invasive Weed Optimization (IWO) Algorithm which is one of the population-based optimization algorithms has been recently developed by inspired from weed colonization. In a simple yet powerful optimization algorithm called the IWO algorithm, it is aimed to imitate the stability, adaptability and randomness of weed weeds.**

For the optimization problems with binary structured solution space, the basic IWO algorithm should be modified because its basic version is proposed for solving continuous optimization problems. In this study, three different adapted versions of IWO, IWObin1, IWObin2 and IWObin3 for short, are proposed for binary optimization. In the proposed methods to solve binary optimization problems, despite the fact that artificial weeds in the algorithm works on the continuous solution space, each weed position is converted to binary values, before the objective function is evaluated. In the first approach, search space is binarized by utilizing mod 2 process. In the second approach, sigmoid function is used to transform continuous space into binary search space. As for the last binary approach of IWO algorithm, this approach is carried out by means of tanh function for converting to binary values.

The accuracy and performance of the proposed approaches have been examined on well-known 12 benchmark instances of uncapacitated facility location problem. The results obtained by *IWObin1*, *IWObin2* and *IWObin3* are compared each other by employing well-known small, medium and large sized twelve instances of UFLPs. The performance of the proposed approaches is also analyzed and compared in terms of convergence speed and running time (CPU time). The experimental results and comparisons show that proposed algorithm is an alternative and simple binary optimization method in terms of solution quality and robustness.

Keyword—Invasive weed optimization; Binary optimization; Sigmoid function; Tanh function; Mod process; Uncapacitated facility location problem

I. INTRODUCTION

In the literature, so many heuristic algorithms have been developed to solve the problems of combinatorial optimization. These algorithms can be divided into subgroups depending on the criteria considered such as iterative, deterministic, population-based, trajectory based and stochastic [1]. An algorithm that works with a group of solutions and tries to enhance performance of them is referred to as population-based [2]. Many swarm intelligence-based approaches have been proposed for solving NP-hard optimization problems in recent years [3]. One of the most recently developed population-based meta-heuristic methods is the invasive weed optimization (IWO) algorithm from within a family of algorithms called swarm intelligence based algorithms. IWO was firstly introduced by Mehrabian and Lucas in 2006 for numerical optimization problems [4]. If the solution space of the problem is constructed in binary, the corresponding method must either be operated using binary vectors in the binary solution space or continuous values in solution vectors of the method must be converted to binary values. For a proposed method called *Mod function-based binary IWO* in this paper, the continuous values in the candidate solutions of the population obtained by the IWO were transferred to the binary space by being inspired by a suggested approach for particle swarm optimization by Güner and Şevkli [5]. In addition, the continuous values in solution space were moved to binary values by employing *sigmoid* and *tanh* functions by being inspired by Mirjalili et al. [6]. IWO has been used for solving different problems in recent years. A modified IWO was used for design of non-uniform circular antenna arrays by Roy et al. [7]. Rad and Lucas proposed a recommendation system based on IWO[8]. Basak et al. developed a modified IWO for time-modulated linear antenna array synthesis [9]. In addition, IWO was proposed for linear antenna array synthesis by Pal et al. [10]. Saravanan et al. proposed unit commitment problem solution utilizing IWO algorithm [11]. Ghalenoei et al. developed discrete IWO [12].

To demonstrate the effectiveness of the proposed binary algorithms, experiments are carried out on benchmarked sets of uncapacitated facility location problem (UFLP) which is one of the most extensively studied combinational

optimization problems in the literature. It consists of determining which facilities should be opened from within a particular potential facilities and how the customers will be assigned to these facilities [13, 14]. The purpose is to use the most appropriate facilities to meet the customer's demand by reducing the total fixed and transport costs. A number of optimization algorithms have been developed for UFLP in recent years, including a wide range of techniques.

The paper is organized as follows: the study is introduced in this section. The basic IWO and proposed binary variants of IWO are presented in Section 2 and 3, respectively. In Section 4, a brief mathematical model of the problem dealt with the study is given. The experimental results and discussion are given in Section 5 and finally, the conclusion is given in Section 6.

II. IVASIVE WEED OPTIMIZATION ALGORITHM

IWO is an evolutionary optimization algorithm inspired by the invader and resistance characters of growing and colonizing weeds [15]. As described below, the algorithm consists of four steps:

A. Initialization

A certain number of weeds are dispersed all over the n-dimensional search area as random [16].

B. Reproduction

Weeds which are randomly produced at the initial stage are allowed to generate seeds at this stage. The production of seeds by means of a weed is related to its own fitness and the fitness of its colonies. While the weed with better fitness value generates more seeds, the weed with worse fitness value generates fewer seeds. The seeds generated by weeds grow linearly, starting with the worst fitness and ending with the best fitness. The formula of reproduction is given in Eq. 1.

$$weedi = \frac{fcurrent - fmin}{fmax - fmin} * (Smax - Smin) + Smin \quad (1)$$

where *fcurrent* is the fitness value of the current weed. *Smax* and *Smin* express the maximum and the minimum value of a weed, respectively. *fmax* and *fmin* represent the maximum and minimum fitness value of the population [17].

C. Spatial Dispersal

At this stage, the seeds produced are randomly spread in the search space such that they are located near to the parent plant based on normal distribution with mean equal to zero and varying variance. Here, the standard deviation (σ) of the random function will be decreased over the iterations from a predefined initial value $(\sigma_{initial})$, to a final value (σ_{final}) and it is evaluated in each step by Eq. 2.

$$\sigma_{iter} = \frac{(iter_{\max} - iter)^n}{(iter_{\max})^n}(\sigma_{initial} - \sigma_{final}) + \sigma_{final}, \quad (2)$$

where $iter_{\max}$ is the maximum number of iterations, σ_{iter} is the standard deviation at the current time step and *n* is the non-linear modulation index.

D. Competitive Exclusion

There exists a competition between plants for survival. The first plants in the colony multiply the plants very rapidly and all plants are considered as colonies. The population should not exceed the maximum population (Pmax). Therefore, while plants with more fitness are included in the colony, plants with less fitness are removed from the colony. Finally, in this step, the plants in the colony are regarded as parent plants and steps 2-4 are repeated until the maximum number of iterations is reached.

III. PROPOSED BINARY VERSIONS OF INVASIVE WEED OPTIMIZATION

To solve the binary optimization problem, the continuous values of the solution space in the algorithm must be converted to binary values. Therefore, three different binary methods have been proposed for transforming continuous values into binary space in this paper.

A. Sigmoid function-based binary IWO (IWObin1)

Sigmoid function given in Eq. 3 is used to obtain a probability value for binary conversion.

$$P_{i,j} = \frac{1}{1+e^{-W_{i,j}}} \quad (3)$$

where, $W_{i,j}$ is higher than the 0.5, a temporary decision variables array is defined and its j^{th} dimension is set to 0 before the objective function evaluation. This procedure is performed to adapt IWO (*IWObin1*) to binary optimization.

B. Tanh function-based binary IWO (IWObin2)

The *tanh* function works element-wise on arrays. Domains and ranges of the function contain complex values. For *IWObin2*, *tanh* function is used as in Eq. 4:

$$P_{i,j} = |tanh(W_{i,j})| \quad (4)$$

where, $W_{i,j}$ is higher than the 0.5, a temporary decision variables array is defined as in *IWObin1* and and its j^{th} dimension is set to 0.

C. Mod function-based binary IWO (IWObin3)

Similar to usage of sigmoid function *in IWObin1*, *modulo base2* is used in *IWObin3* to convert the continuous solutions to binary equivalent. This conversion is given in Eq. 5.

$$P_{i,j} = mod(abs(\lfloor W_{i,j} \rfloor), 2) \quad (5)$$

where, $P_{i,j}$ is binary solution obtained for $W_{i,j}$, $\lfloor \ \rfloor$ is rounding operation to down, *abs* is absolute function. The

fitness of $W_{i,j}$ is calculated by evaluation of the objective function related to binary optimization problem by using $P_{i,j}$ binary array of decision variables.

IV. UNCAPACITATED FACILITY LOCATION PROBLEM

The performance and effectiveness of the proposed binary algorithms are examined on UFLP. In basic formulation, the UFLP consists of a series of potential facilities *I* that can open a facility and have no capacity constraints, and a set of customer location *J* that need to be served. The purpose (Eq. (6)) is to determine a subset *F* of *I* facilities that is corresponded demand of customers *J*. The objective function of the problem is to minimize sum of the shipment costs between *F* and *J* and the opening costs of the facilities. The standard model of the UFLP can be stated as follows:

$$f(\text{UFLP}) = \min\left(\sum_{i=1}^{k}\sum_{j=1}^{l} c_{i,j}x_{i,j} + \sum_{j=1}^{l} fc_j y_j\right) \quad (6)$$

subject to :

$$\Sigma_{j=1}^{l} x_{i,j} = 1 \qquad j \in J, \text{ and } x_{i,j} \leq y_i, i \in I \text{ and } j \in J, \quad (7)$$

$$x_{i,j} \in \{0,1\}\ , i \in I \text{ and } j \in J, \text{ and } \ y_i \in \{0,1\},\ i \in I, \quad (8)$$

where $i= 1...k; j= 1,...,l;$ x_{ij} represents the quantity provided from facility *i* to customer *j*; y_j expresses whether facility *j* is located (y_j= 1); otherwise (y_j= 0). The constraint in Eq. 7 ensures that demands of all customers must be satisfied by an open facility. The constraint in Eq. 8 provides the collectivity, as well.

V. EXPERIMENTAL RESULTS

The uncapacitated facility location test suite (12 test problems) obtained from the OR-Library was used in order to examine the performance and accuracy of the proposed binary versions of the IWO algorithm. In the test suite, four problems (Cap71-74) are small-sized, the four problems are medium-sized (Cap101-104) the remaining four problems are large-sized problems (Cap131-134) [18].

In Table 1, IWO parameter values were given for *IWObin1, IWObin2* and *IWObin3*.

TABLE I. IWO PARAMETER VALUES

Quantity	Value
Initial search area (X_{ij})	$-8 < X_{ij} < 8$
Minimum number of seeds	1
Maximum number of seeds	5
Nonlinear modulation index	2
Initial value of standard deviation	0.5
Final value of standard deviation	0.001
Number of initial population	20
Maximum number of plant population	40
MaxFEs	80.000
Run	30

Problem name, size and cost of the optimal solution are defined in Table 2.

TABLE II. DESCRIPTION OF THE TEST SUITE

Problem name	Problem size	Cost of the optimal solution
Cap71	16 x 50	932,615.75
Cap72	16 x 50	977,799.40
Cap73	16 x 50	1,010,641.45
Cap74	16 x 50	1,034,976.98
Cap101	25 x 50	796,648.44
Cap102	25 x 50	854,704.20
Cap103	25 x 50	893,782.11
Cap104	25 x 50	928,941.75
Cap131	50 x 50	793,439.56
Cap132	50 x 50	851,495.33
Cap133	50 x 50	893,076.71
Cap134	50 x 50	928,941.75

The GAP, which is the difference between the cost of the optimal solution and the solution found by the method is given in Eq. 9.

$$GAP(\%) = \frac{f(mean)-f(opt)}{f(opt)} \text{x} 100 \quad (9)$$

TABLE III. THE COMPARATIVE TEST RESULTS OF SMALL-SIZED PROBLEMS

		Best	Worst	Std. Dev.	Mean	GAP (%)
Cap71	*IWObin1*	**932615.750**	937364.400	1129.616	933428.528	0.087
	IWObin2	**932615.750**	957848.375	6837.213	941735.968	0.968
	IWObin3	**932615.750**	**932615.750**	**0.000**	**932615.750**	**0.000**
Cap72	*IWObin1*	**977799.400**	984026.463	2077.684	980085.555	0.233
	IWObin2	**977799.400**	1006716.913	6110.815	985575.124	0.789
	IWObin3	**977799.400**	**977799.400**	**0.000**	**977799.400**	**0.000**
Cap73	*IWObin1*	**1010641.450**	1018454.575	2022.631	1012339.856	0.168
	IWObin2	**1010641.450**	1037301.688	7808.909	1016922.543	0.618
	IWObin3	**1010641.450**	**1010641.450**	**0.000**	**1010641.450**	**0.000**
Cap74	*IWObin1*	**1034976.975**	1053008.938	5256.444	1038995.276	0.387
	IWObin2	**1034976.975**	1095037.738	12707.673	1049584.994	1.392
	IWObin3	**1034976.975**	**1034976.975**	**0.000**	**1034976.975**	**0.000**

In Table 3, comparative test results were given for small-sized problems, named as *Cap71*, *Cap72*, *Cap73* and *Cap74*. The better results are written in bold face font type. In terms of the best value (*best*) for each of the problems, *IWObin1, IWObin2* and *IWObin3* have achieved the same results. However, according to other values *IWObin3* is better than the other two algorithm. As for the standard deviation and GAP, *IWObin3* obtained the best results among the algorithms. Therefore, it can be clearly stated that *IWObin3* is the best one for the small-sized problems in terms of both of standard deviation and GAP.

TABLE IV. THE COMPARATIVE TEST RESULTS OF MEDIUM-SIZED PROBLEMS

		Best	Worst	Std. Dev.	Mean	GAP(%)
Cap101	*IWObin1*	**796648.438**	806026.500	2513.775	800489.729	0.480
	IWObin2	799092.113	840952.375	11088.345	811332.677	1.810
	IWObin3	**796648.438**	**797508.725**	**262.498**	**796734.466**	**0.011**
Cap102	*IWObin1*	**854704.200**	864578.613	2928.641	858174.553	0.404
	IWObin2	856660.013	884088.025	7491.424	864828.935	1.171
	IWObin3	**854704.200**	**854704.200**	**0.000**	**854704.200**	**0.000**
Cap103	*IWObin1*	**893782.113**	904107.200	2947.294	896964.386	0.355
	IWObin2	**893782.113**	939183.125	11336.519	910101.664	1.793
	IWObin3	**893782.113**	**894008.138**	**57.344**	**893797.181**	**0.002**
Cap104	*IWObin1*	**928941.750**	957488.388	8227.766	938504.620	1.019
	IWObin2	**928941.750**	979395.063	14707.327	949074.704	2.121
	IWObin3	**928941.750**	**928941.750**	**0.000**	**928941.750**	**0.000**

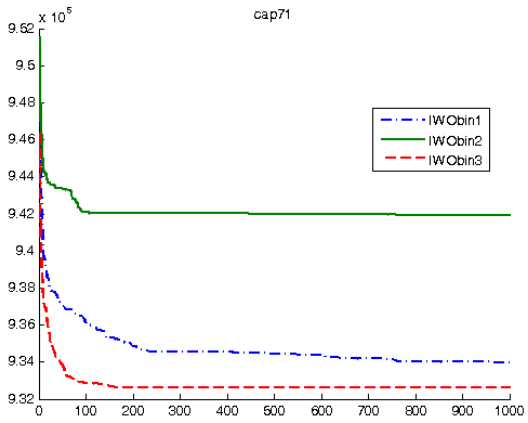
Medium-sized problems were given in Table 4. The worst results have been obtained by *IWObin2* for each of the problems. However in terms of the best value, *IWObin1* and *IWObin3* have achieved same results for each of the problems, *IWObin2* has achieved the same results with *IWObin1* and *IWObin3* in Cap103 and Cap104. *IWObin3* gave results equal to zero in the standard deviation. It is obviously seen that *IWObin3* obtained the best results for all of the medium-sized problems. Hence, *IWObin3* is more successful and robust than the other binary variants.

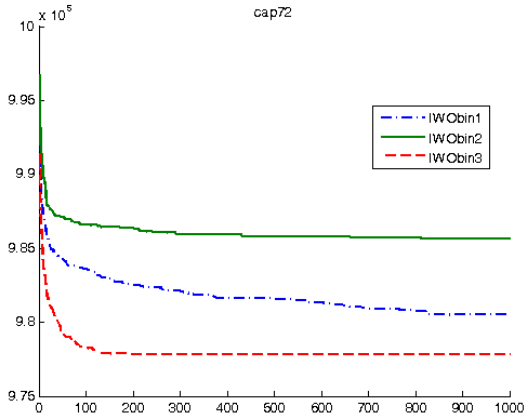
TABLE V. THE COMPARATIVE TEST RESULTS OF LARGE-SIZED PROBLEMS

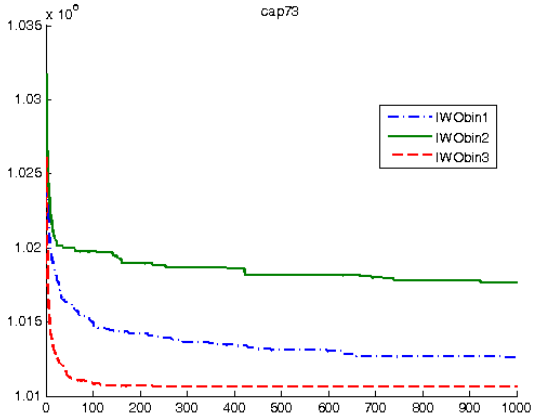
		Best	Worst	Std. Dev.	Mean	GAP(%)
Cap131	*IWObin1*	797570.300	834187.638	7457.799	811066.128	2.173
	IWObin2	800262.113	840728.075	9568.384	818819.122	3.100
	IWObin3	**794956.113**	**805253.438**	**2564.812**	**799113.655**	**0.710**
Cap132	*IWObin1*	857979.813	922919.813	14206.250	878897.360	3.118
	IWObin2	857203.000	915943.713	13995.674	879331.952	3.166
	IWObin3	**854061.125**	**868623.988**	**3202.621**	**858098.754**	**0.770**
Cap133	*IWObin1*	900622.088	973860.338	16993.563	924355.003	3.384
	IWObin2	900090.213	982987.863	19761.139	937772.720	4.766
	IWObin3	**893076.713**	**910383.675**	**4609.327**	**899699.540**	**0.736**
Cap134	*IWObin1*	940980.375	1066364.638	30790.362	998947.348	7.008
	IWObin2	942446.013	1073170.325	32727.773	1003640.578	7.443
	IWObin3	**928941.750**	**955654.525**	**5670.400**	**937415.320**	**0.904**

In table 5, the test results of large-sized problems named as *Cap131*, *Cap132*, *Cap133* and *Cap134* was presented comparatively. According to these results, *IWObin1* and *IWObin2* had the worst results in all the values when comparing with *IWObin3*. When analyzing according to GAP, it is clearly stated that *IWObin3* yielded results approximately to zero.

As seen from the tables, *IWObin3* reached to the best results for each of the problems among the other binary variants of IWO. In small-sized problems and medium-sized problems, *IWObin3* achieved results equal to zero in terms of standard deviation. In addition, *IWObin3* achieved results approximately to zero with regard to the GAP values. However, the worst results were achieved by *IWObin2* in each of the tables. Therefore, it can be seen from the results, that when comparing with binary variants of IWO each other, while *IWObin3* became the most successful binary algorithm, *IWObin2* became the worst algorithms. In addition, as expected, increasing the number of problem dimensions reduces the chance of obtaining the optimal solution by the proposed algorithm. The results are also shown graphically in Fig 1, 2 and 3:







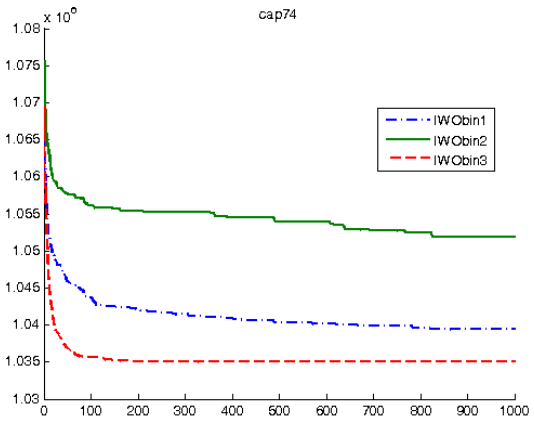
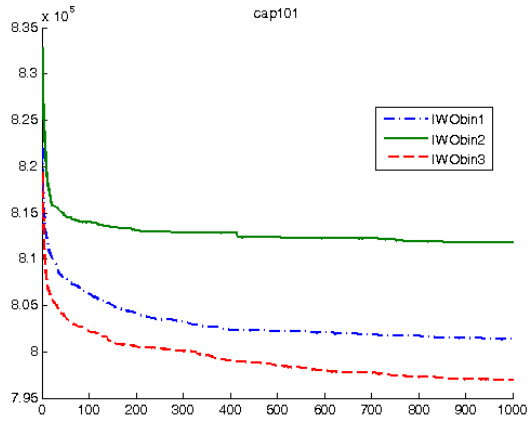
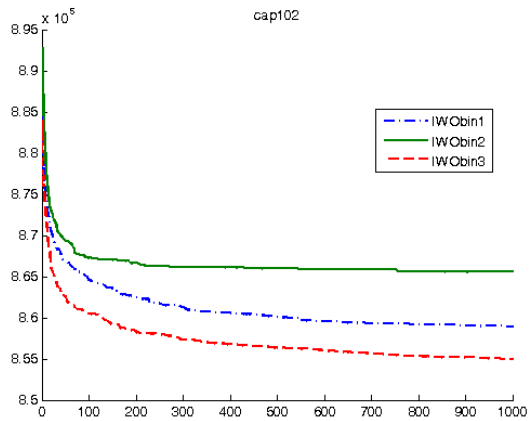


Fig. 1. Convergence curves of proposed binary versions of IWO for small sized problems

It is shown in Fig. 1, while *IWObin3* obtained the optimal solutions for all of the small-sized problems approximately in first hundred iterations, *IWObin1* and *IWObin2* could achieved the optimal solutions after long iterations. In addition, *IWObin1* is more successful than *IWObin2* in terms of convergence speed. Therefore, it can be clearly stated that convergence speed of *IWObin3* is higher than those of the other variants.





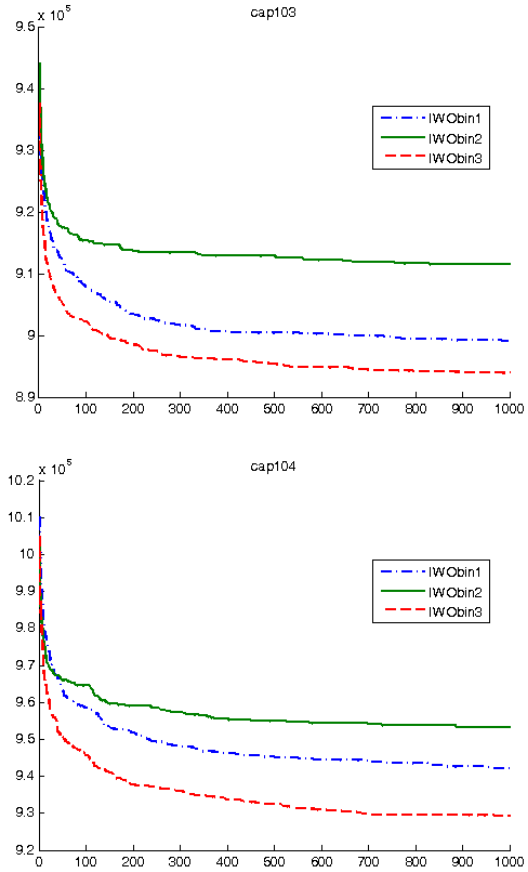
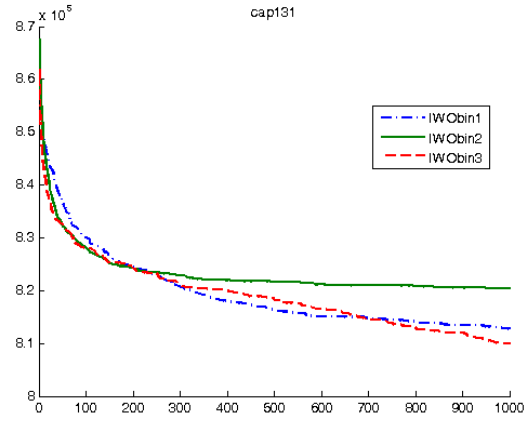


Fig. 2. Convergence curves of proposed binary versions of IWO for medium sized problems

It is shown in Fig. 2, all of the binary variants of IWO continuously converged over all the iterations. It is seen that generally while convergence speed of *IWObin2* is lower, those of the other two variants are higher. However, it is obviously seen that for all of the medium-sized problems, *IWObin3* is the best one among the binary variants of IWO in terms of convergence speed.



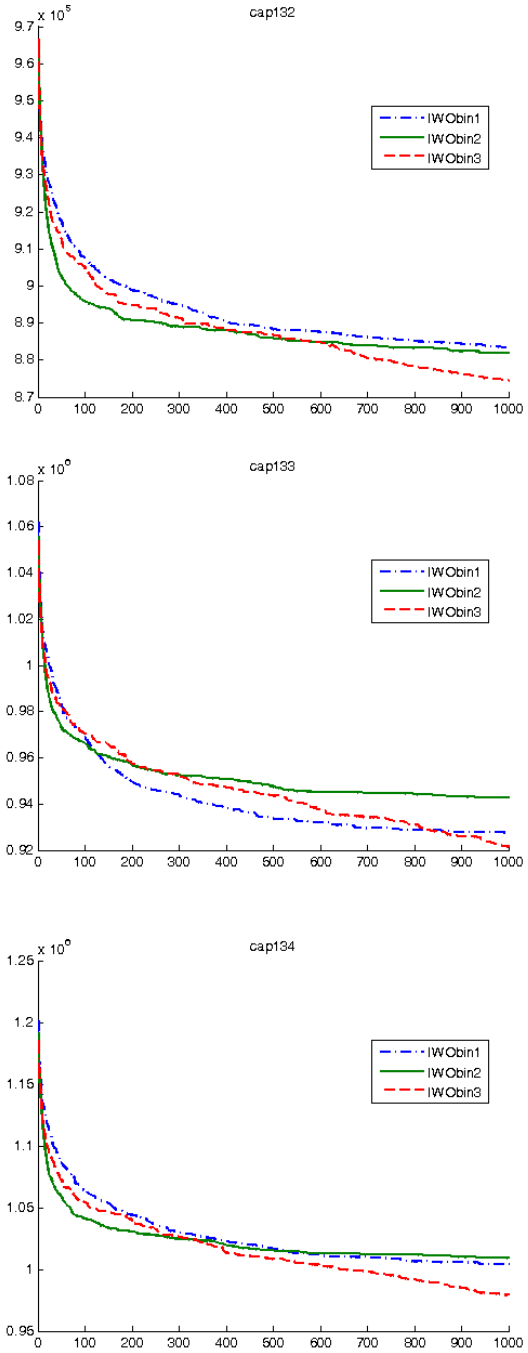


Fig. 3. Convergence curves of proposed binary versions of IWO for large sized problems

It is shown in Fig. 3, as the iterations continued, the superiority of the algorithms to each other changed. Though there is no significant difference between the binary variants, it is seen that *IWObin3* gave better results. In other words, *IWObin3* yielded the best performance for all of the large-sized problems.

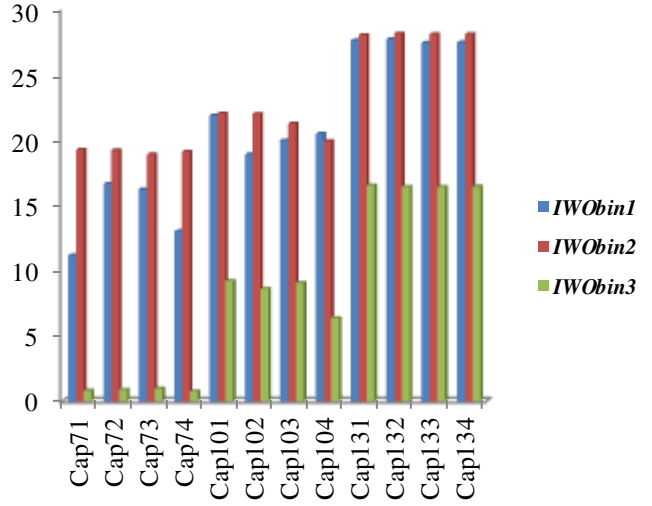


Fig 4. Running time comparison of the methods. (second)

Fig. 4 shows the running time comparison of the methods measured in seconds. While small-sized problems take the least time to process, large-sized problems take the longest amount of time. In addition, it can be clearly stated that *IWObin3* is the best one among the binary variants in terms of CPU time because it obtained optimal results in the shortest times.

VI. CONCLUSION

In this study, three different binary variants of IWO, *IWObin1*, *IWObin2* and *IWObin3* for short have been proposed for binary optimization. In the proposed methods in order to solve problems with binary variables, despite the fact that artificial weeds in the algorithm search for on the continuous solution space, each weed position is transformed to binary values, before evaluation of the objective function.

The performance of the proposed binary variants of IWO have been investigated on UFLP. The results obtained by *IWObin1*, *IWObin2* and *IWObin3* have been compared each other by employing well-known small, medium and large sized twelve instances of UFLPs. The performances of the proposed approaches have been also analyzed and compared in terms of convergence speed and running time. The experimental results demonstrate that proposed binary variant of algorithm for short *IWObin3* which utilizes *mod2* process is an alternative and simple binary method tool in terms of solution quality and robustness.

REFERENCES

[1] M. H. Kashan, N. Nahavandi, and A. H. Kashan, "DisABC: a new artificial bee colony algorithm for binary optimization," *Applied Soft Computing,* vol. 12, pp. 342-352, 2012.

[2] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of global optimization,* vol. 39, pp. 459-471, 2007.

[3] M. S. Kiran, "The continuous artificial bee colony algorithm for binary optimization," *Applied Soft Computing,* vol. 33, pp. 15-23, 2015.

[4] A. R. Mehrabian and C. Lucas, "A novel numerical optimization algorithm inspired from weed colonization," *Ecological informatics,* vol. 1, pp. 355-366, 2006.

[5] A. R. Guner and M. Sevkli, "A discrete particle swarm optimization algorithm for uncapacitated facility location problem," *Journal of Artificial Evolution and Applications,* vol. 2008, 2008.

[6] S. Mirjalili and A. Lewis, "S-shaped versus V-shaped transfer functions for binary particle swarm optimization," *Swarm and Evolutionary Computation,* vol. 9, pp. 1-14, 2013.

[7] G. G. Roy, S. Das, P. Chakraborty, and P. N. Suganthan, "Design of non-uniform circular antenna arrays using a modified invasive weed optimization algorithm," *IEEE Transactions on antennas and propagation,* vol. 59, pp. 110-118, 2011.

[8] H. S. Rad and C. Lucas, "A recommender system based on invasive weed optimization algorithm," in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, 2007, pp. 4297-4304.

[9] A. Basak, S. Pal, S. Das, A. Abraham, and V. Snasel, "A modified invasive weed optimization algorithm for time-modulated linear antenna array synthesis," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*, 2010, pp. 1-8.

[10] S. Pal, A. Basak, S. Das, and A. Abraham, "Linear antenna array synthesis with invasive weed optimization algorithm," in *Soft Computing and Pattern Recognition, 2009. SOCPAR'09. International Conference of*, 2009, pp. 161-166.

[11] B. Saravanan, E. Vasudevan, and D. Kothari, "Unit commitment problem solution using invasive weed optimization algorithm," *International Journal of Electrical Power & Energy Systems,* vol. 55, pp. 21-28, 2014.

[12] M. R. Ghalenoei, H. Hajimirsadeghi, and C. Lucas, "Discrete invasive weed optimization algorithm: application to cooperative multiple task assignment of UAVs," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, 2009, pp. 1665-1670.

[13] D. Ghosh, "Neighborhood search heuristics for the uncapacitated facility location problem," *European Journal of Operational Research,* vol. 150, pp. 150-162, 2003.

[14] A. Rahmani and S. MirHassani, "A hybrid firefly-genetic algorithm for the capacitated facility location problem," *Information Sciences,* vol. 283, pp. 70-78, 2014.

[15] E. Pourjafari and H. Mojallali, "Solving nonlinear equations systems with a new approach based on invasive weed optimization algorithm and clustering," *Swarm and Evolutionary Computation,* vol. 4, pp. 33-43, 2012.

[16] S. Roy, S. M. Islam, S. Das, and S. Ghosh, "Multimodal optimization by artificial weed colonies enhanced with localized group search optimizers," *Applied Soft Computing,* vol. 13, pp. 27-46, 2013.

[17] Y. Zhou, H. Chen, and G. Zhou, "Invasive weed optimization algorithm for optimization no-idle flow shop scheduling problem," *Neurocomputing,* vol. 137, pp. 285-292, 2014.

[18] J. E. Beasley, "OR-Library: distributing test problems by electronic mail," *Journal of the operational research society,* vol. 41, pp. 1069-1072, 1990.

Hash functions and their application in digital signatures and digital forensics

Trajche Roshkoski, Snezana Savoska, Blagoj Ristevski, Tome Dimovski
Faculty of information and communication technologies,
University "St. Kliment Ohridski" – Bitola, 7000, Partizanska bb,
Republic of Macedonia
roskoskitrajce@gmail.com, snezana.savoska@fikt.edu.mk,
blagoj.ristevski@fikt.edu.mk, tome.dimovski@fikt.edu.mk

***Abstract* – Hash functions are used as building blocks in certain cryptographic systems. Because of their complexity, program languages with high accuracy performance have to be used. A Python based software application, created by the authors is presented. The main purpose of the application is to ensure the authenticity and integrity of digital objects as well as to offer an alternative to high cost forensic software tools. The accent is placed on the usage of hash functions in forensic applications and their importance in the area of digital signatures and digital forensics. Hash-based methods are attractive for this application because of their performance and their memory efficiency.**

Keywords - cryptographic hash functions; digital signatures; digital forensics; MD5; SHA1; message authentication

I. INTRODUCTION

The security of today's communication is based on cryptographic protocols that use hash functions as building blocks. Hash functions play an important role in certain cryptographic systems and are widely used in digital signatures, storing passwords, message authentication and digital forensics [20]. The main feature of any hash function is that it is particularly difficult for an attacker to get the original data value from the hash value. However, we are witnesses that the computing power doubles every few years and computers become powerful enough to be able to practically realize some of the attacks that are deemed to be practically unfeasible. It is therefore necessary to make a detailed analysis of hash functions and analyze the level of safety that they provide. By analyzing cryptographic hash functions, we can find some weaknesses in the design that can further be exploited by malicious people to realize a successful attack on them. Attacks on these functions can be prevented by constantly upgrading or designing new features that will offer increased security [14].

Digital signatures are some kind of the digital equivalent of a handwritten signature or seal, but with the only exception that they offer far greater security. The purpose of a digital signature is to solve the problem of falsification or misrepresentation in digital communication. The digital signature can provide proof of origin, identity and status for the electronic document, transaction or message [18], [21]. Additionally it can offer proof of the signatory's consent regarding the document, transaction or message. To create a digital signature a person can use different hash functions. Thus the security of the digital signature depends on the security of the hash function.

Digital forensics is a separate branch of forensic science and represents the process of discovery, analysis and presentation of digital data that can be later used in court [11], [13]. Generally, this investigation should answer questions like what and when it happened, how and who did the crime that is the subject of the research. With the increasing trend of digitalization of all information, the digital forensic analysis is especially important in exploring the full range of illegal activities, from minor offenses to major criminal cases. In criminal cases which include digital data, it is necessary to make rapid and effective investigation by forensic scientists [13], [17]. Fast, efficient scanning of digital data and digital evidence is based on hashing techniques. Specially designed algorithms are based on hash functions and are used in digital forensics for various applications from proving the authenticity and integrity of digital evidence to detection of malware and detection of violations of privacy.

As the generation of new digital content increases, the volume of digital data that ends in forensic laboratories increases as well. Often forensic scientists face enormous amounts of data with only a small percentage of that data being relevant. Therefore there is a need for rapid methods that would allow the elimination of irrelevant data and that will highlight the data that is in the best interest of the digital investigation. Hash-based methods are attractive for this application because of their performance and their memory efficiency [15].

Lead by these findings, we propose developing a Python based application that will be simple enough for non-technical users and can be used to ensure the authenticity and integrity of digital objects. The software application proposed in this paper will be based on methods that rely on hash functions. The paper is organized as follows. In the second section, related works are reviewed and explained. The third section reviews the problem that will be solved by the proposed software tools. The fourth section describes the organizational structure of the proposed Python based application, all prerequisites for the

application and the software platform. The fourth section also describes the functionality of the proposed application and the results are explained in the next section. Finally, some conclusions are drawn and future researches are proposed.

II. Related Works

Bart Preneel [1] demonstrates the importance of hash functions in protecting the authenticity of the data. The subject of research in his paper is the practical application of hash functions in protecting and checking the integrity of conventional messaging and digital signatures. The main contribution of this paper is the study of practical constructions of hash functions. In addition, the paper gives an overview of existing attacks, depicting possible new attacks on hash functions and provides possible solutions and schemes.

Vassil Roussev [2] considers hash-based tools that can quickly, accurately and reliably connect the growing number of digital data and provide information on their similarity. Currently NIST has a suggestion for two different solutions that can be used for this purpose. SSDEEP produces hash values with a fixed size based on random polynomials while SDHASH produces hash values of variable size based on statistical identified properties and packaged in a Bloom filter. This paper examines the significance of these two tools and gives an assessment of their qualifications based on data from a controlled environment and on data derived from the real world. The results show that similarity hash functions are significantly better in terms of accuracy in all test scenarios and show stable behavior.

Wenjun Lu, Avinash L. Varna and Min Wu [3] give a proposal for a tool that will enable detection of changes in digital content. Digital images and videos are widely available online and have a significant impact on society, which can be evidenced by the growing number of social networking sites. Because of the ease way for changing and manipulating images and video, a process has to be defined as process that will ensure the security of multimedia information. To that end, their work is a proposal of a tool for digital forensics on multimedia content.

Christoph Zauner [4] points out the application and importance of perceptual hash functions. These functions generate hash values depending on the visual appearance of the image. The functions will produce similar hash values for similar images and vice versa. Those hash values are then compared using a similarity function or distance which can calculate if the images are similar or not. The author tests these hash functions in cases where rotations are made or the dimensions of the images are changed. The test result shows that all these functions are stable and provide accurate results in the case where one of the images is inverted horizontally.

Ilya Mironov [5] explores the theory and application of cryptographic hash functions, paying special attention to hash functions such as MD5 and SHA1. Special attention is also given to their resistance to attacks by detecting collisions. Additionally, the author presents definitions, design principles, generic attacks and recent attacks on specific functions.

Praveen Gauravaram [6] considers the various types of attacks on hash functions and suggests possible schemes that can be used to prevent those attacks. Furthermore, he explores alternative approaches in the construction of hash functions and explores the use of hash functions in messages authentication codes.

Besides the previously mentioned works, there are many other papers [7-10] that testify of the importance of secure hash functions and their practical applications. All these studies indicate that as one of the most important building blocks, hash functions require constant analysis and upgrading in order to increase the safety of transmission and storage of data.

III. The problem defining

For better conducting the digital researches and investigations, developers have created many computer forensics tools. Police departments and investigation agencies select the tools based on various factors including budget and available experts on the team. During our research we came across the following software tools:

- X-way forensics [25] is an advanced platform for digital forensic examiners, intended for the Windows Platform. The advantage of this software is that it allows the user to perform a variety of advanced forensic tasks. The disadvantages are that this software is only limited to the Windows Platform and requires a certain level of expertise from the user.
- Cofee [25] represents a set of tools designed by Microsoft specifically for digital forensic examiners. Microsoft works together with INTERPOL and NW3C to provide this tool to law enforcement agencies around the world at no cost. The disadvantage of this tool is that it is not available to the general public and can be obtained only by contacting INTERPOL.
- EnCase [25] is a multipurpose forensic platform which comes with all the necessary tools needed for a digital investigation. The disadvantage of this software is a high level cost and unavailability for people or agencies with a low level budget.
- SIFT [25] is a multipurpose forensic system based on Ubuntu which comes with a variety of tools that might come in handy with the digital investigation. The advantage of this software is that there is a free version available to the public but the software requires a user that has previous training in order to properly conduct and analyze the data.

Based on the previously described software tools, we came to a conclusion that these software tools offer a variety of methods for any type of digital investigation. However they have disadvantages like high level costs and the need for trained professionals. Also some of the software tools are limited by operating platforms and can't be used on all computer systems. Lead by these findings we propose developing a simple software tool that can be used without training and can run on different platforms. This software can

be used by individuals or companies with low budget as an alternative to high cost forensic tools.

For fast and efficient data scanning, forensic scientists rely on hash-based techniques. The hashing process represents a primary but many times underestimated tool in the process of digital investigation. Proper and effective implementation of some of the hash techniques can simplify the digital forensic investigation. Basically cryptographic hash functions in digital forensics are used for two purposes. The first usage of hash functions in the digital investigation is to ensure authenticity and integrity of digital traces. Another usage is to identify well-known objects (such as illegal documents). Because of these reasons, the software application described in this paper will rely on methods based on hash functions. The hash-based methods are attractive for this application because of their performance and their memory efficiency.

IV. PRACTICAL APPLICATION OF HASH FUNCTIONS IN PYTHON

Python is a high level interpreted programming language. The syntax of this language allows developers to express concepts in fewer lines of code, compared to programming languages such as C++ or Java [16]. It supports multiple programming paradigms such as object-oriented, procedural or functional programming. It features automatic memory management and has a large and comprehensive standard library. Python is available on many operating systems which allow the code to run on different systems. The code written in this programming language can additionally be packed in standalone executable for some of the most operating systems.

Due to the large selection of modules, Python is the perfect choice for making the software application. Some of the modules included in the development of the software tool for digital forensics include:

- Hashlib.py is a module that comes with the standard Python library and allows the use of different hash algorithms which include MD5, SHA, RIPEMD, DSA and WHIRLPOOL.
- PyCrypto.py is a collection of hash algorithms and various encryption algorithms. This module allows calculation of hash values (SHA and RIPEMD), creation of digital signatures (SHA and RSA) and encryption techniques (AES, DES, RSA, etc.). The purpose of this module is to provide tools for creating safe and effective software. This module does not come with the standard Python library and needs to be additionally installed.
- Ssdeep.py is a module which allows the use of algorithms or fuzzy hashing context triggered piecewise hashes (CTPH). Such algorithms can detect similar files by comparing their hash values. This module needs to be further installed to be available for use because it does not come with the standard Python library.
- Sdhash.py is a module that provides two sets of data to be compared for similarity based on common arrays of binary data. This module is designed to provide quick results in the initial stages of a digital investigation

A. *Organizational structure of the Python-based software application*

Due to the large selection of modules, Python is the perfect choice for making the application. The application provides the following functionalities:

- Creating a hash sums of files in a local directory
- Creating digital signatures
- Detection of similarity with SSDEEP

In order to provide the previously discussed functionality, we will be using modules such as PyCrypto, Hashlib and ssDeep. For the development of the graphical interface will use Tkinter module which allows the use of the most popular elements such as buttons, labels, input fields and custom dialog windows. With the help of this module we can create the windows of the software application. Fig.1 shows the organizational structure of the created application.

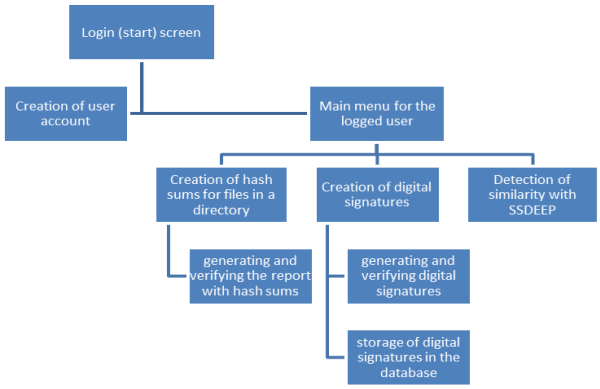


Fig.1. Organizational structure for the Python-based software tool

User data and digital signatures will be stored in a database. To create the database we are going to be using Microsoft Access in combination with the pyodbc module which allows the user to connect and manipulate the database directly from Python. The data will be stored in two tables. Korisnik table will be used to store user data and table Potpisi will be used for storage of digital signatures.

The password field from the Korisnik table will be filled with hash values for each password. Keeping the hash value instead of the password is useful and prevents unauthorized people to have access to the raw form of the passwords. Any unauthorized person to access the database will not be able to get the passwords based on the hash values.

The table Korisnik has the field ID as a primary key, that field is unique for each user. The table Potpisi has the field ID as the primary key and the field Avtor as an external key. Both tables are connected in a relationship one to many. This means that each user can create and store countless signatures. On Fig. 2 we can see the organizational structure of the database.

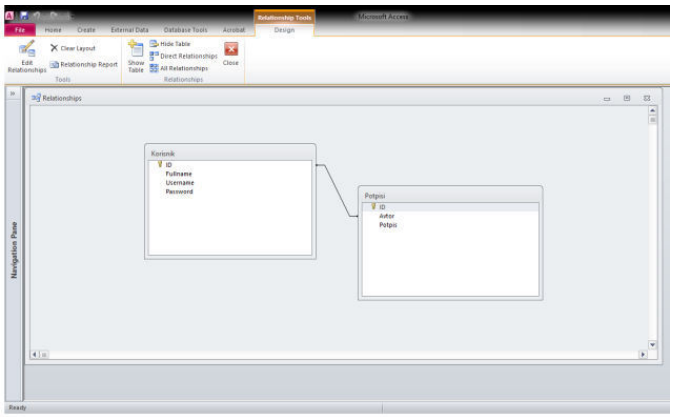


Fig. 2. Organization structure of the database. Table Korisnik has the fields ID, Fullname, Username and Password. Table Potpisi has the fields ID, Avtor and Potpis. The field Avtor is an external key and connects the table with the ID field from the Korisnik table. The relationship is one to many.

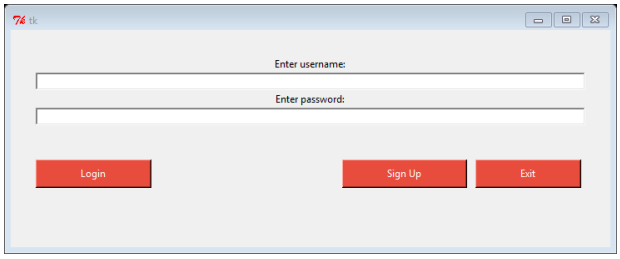


Fig. 3. Login screen for the application designed using the Tkinter module. It allows the user to login and access the main menu using his user credentials. Otherwise the user will be able to navigate to the sign up screen by pressing the Sign Up button where he can create a new user account.

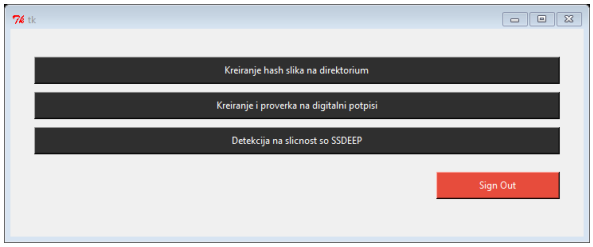


Fig. 4. Main menu screen that appears for the logged in user. Once the user uses his login credentials, the main menu screen will open. The user will be able to use this screen to access the features of the application.

B. Creating hash sums from files in a local directory

One of the tasks of digital forensics is to ensure the authenticity and integrity of digital evidence. In forensic investigations we need a mechanism that will ensure that digital evidence has not changed during the investigation. Often in investigations, items such as computers, laptops, mobile phones and various digital media are confiscated. These devices can be the source of various data that may or may not have to be in the interest of the forensic investigation. If a person manages to get access to a seized device then it is very easy to manipulate the data that are stored on it. For this reason at the beginning of the digital forensic investigation, the investigator does a "snapshot" of all confiscated devices which makes it known what data and what files were located on the device at the moment of the confiscation. The easiest way to ensure data integrity of an electronic device is by using a hash function.

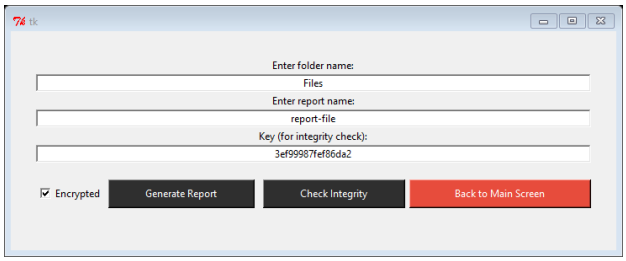


Fig. 5. Creation of a digital signature for the folder Files. The generated report will be saved under the name report-file.txt and will be encrypted using AES algorithm.

The application will allow the user to choose whether to encrypt the report or not. The encrypted report offers increased security because only the people who have the key can decrypt and view the content of the report. The key is generated using a random number generator and then calculating the hash value for that number. That way the key will always have a fixed size. It is important for the user to memorize the key because without it the report can't be decrypted. Figure 6 shows both encrypted and decrypted report. The report contains information about each file in the directory, the size in bytes and the MD5 and SHA-256 hash value.

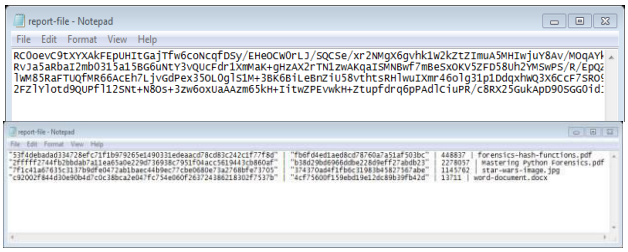


Figure 6. Encrypted report (top) and decrypted report (bottom) for the folder Files which contains 4 files. The report is generated in the same file as the directory in which the application is started.

The report actually represents a snapshot or an image of the present state of the directory. The report maps each file into a hash value. Later, the report can be used as a proof that nobody has manipulated the data located in the directory. The integrity of the data can be checked with a simple comparison. The application allows the user to check the report and confirm the integrity of the directory. If the report is encrypted, the user will also have to provide the decryption key. Once the integrity check has finished, the user will get the integrity check report which is shown on Figure 7. As shown on Fig. 7 the integrity of the file in the report is confirmed.

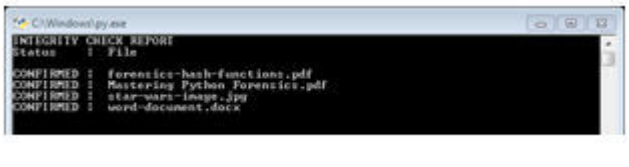


Fig. 7. Integrity check of the previously generated report. For each file in the folder Files, a new hash value is calculated. That value is later compared with the one found in the report. If the two values are the same then the integrity of the file is confirmed.

C. *Creating and verifying digital signatures*

Digital signatures represent the electronic equivalent of a signature which is put on a paper [18]. To create a digital signature, we need to have two elements: a way of creating a hash value and method of signing the hash value [21]. From Hashlib and PyCrypto libraries, we can use the SHA and RSA algorithms. With the first algorithm, we can generate a hash value of the message, while the RSA algorithm will be used to sign the generated hash value. The resulting value will be a digital signature.

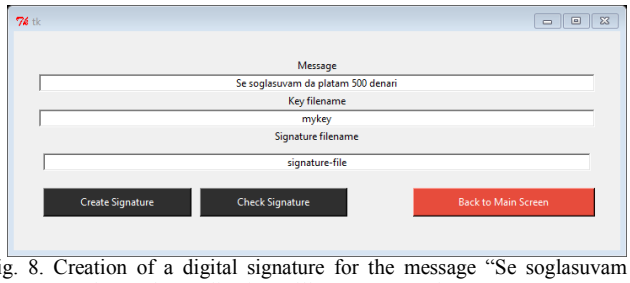


Fig. 8. Creation of a digital signature for the message "Se soglasuvam da platam 500 den". The application will generate two keys. One will be used to sign the hash value calculated from the message while the other will be saved into mykey.pem. The digital signature will be saved in the file signature-file.txt.

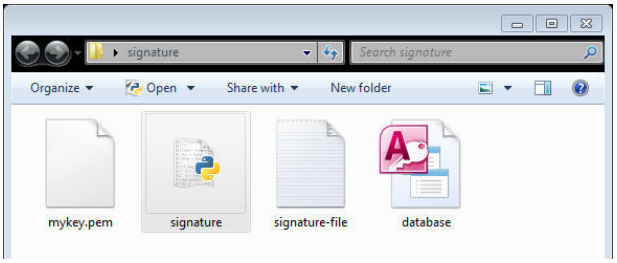


Fig. 9. Generated mykey.pem and signature-file.txt. The user can now distribute the key and the signature to another user. Using these files, the recipient can check the authenticity of the data.

First of all, the application calculates the hash value for the message using the SHA algorithm and later signs the hash value using the RSA algorithm. The result represents a digital signature. The signature can also be stored in the database using the pyodbc module. For easier manipulation the signature is transformed into base64 encoding and then it is saved into the database and the local txt file. If the user has to check the signature, the application first decodes the signature from base64 into binary form.

The user can use the application to verify any digital signature. In order to verify the digital signature, the user has to provide the key and the message. After that the application will check if the digital signature is authentic. Additionally the application will check if the signature is located in the database. If the signature is found in the database, the application will display the name of the person that created the signature. Otherwise the user will get a prompt asking if he wants to add the signature into the database.

D. *Detection of similarity with SSDEEP*

The ability to detect a similarity between any two files is especially important in the detection of plagiarism and the detection of malware [19], [22]. Cryptographic hash functions enable identification of identical digital objects but because their design, they are bad at detecting similar objects. The ssdeep algorithm allows detection of similarity between two files by calculating their hash values and comparing them [23], [24]. This algorithm comes with the ssdeep module which should be additionally installed in order to be used in the Python environment. This module allows calculation of fuzzy hash values or context triggered piecewise hashing (CTPH). It is particularly useful in comparing text files and can't be used to compare images or videos. The sdhash module is recommended for detection of similarity between images and videos [23], [24].

The application allows the user to detect similarity between two text files. First, it loads the content from the files and splits the content into segments. A hash value is calculated for each segment. The final hash sum is calculated and based on all the hash values from all segments. The same process is used on both files. The final two hash values are then compared and analyzed. The ssdeep algorithm then produces a value between 0 and 100. Bigger value indicates a higher chance of similarity between the two documents. Figure 10 shows the result of the similarity check between two documents. One of the disadvantages of this algorithm is that the text file has to be large enough in order to be split into segments. Otherwise the algorithm can show false positive results.

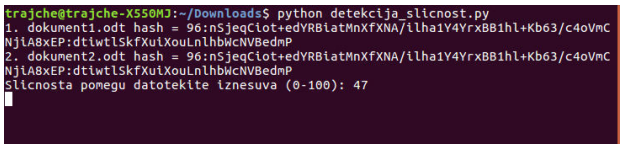


Fig. 10. Detection of similarity between document1.odt and document2.odt. The result of the comparison is 47 which shows that there is a high similarity between the two documents.

V. Results and Discussion

The software application presented in this paper consists of 3 different options which offer functionality that can be used by a forensic examiner in different situations. At the beginning of the forensic investigation, examiners should secure any digital evidence against future tempering. The examiner needs a way to make a list of all the files which are located on the confiscated digital device and use that list in the future to detect any tempering with the data by a third person. For that purpose, the software application in this paper relies on hash functions to create a list of files ("disk image" or "disk snapshot") and detect any tempering with the files.

Additionally it is very important for the examiner to be able to detect similarities between the data located on a confiscated device and a potentially dangerous file. Specially designed hash functions can be used for this purpose. The software application described in this paper relies on the ssdeep algorithm to check and detect similarity between two files. The usage of this algorithm can be detection of plagiarism, detection of incriminating images and videos with explicit content, illegal documents.

The presented software application also demonstrates the usage and creation of digital signatures. They represent a code that can be attached to electronically transmitted document to verify its contents and the sender's identity. For that reason they can be useful in the process of exchanging messages or documents because the authenticity of the data can be easily confirmed.

The paper shows that Python is a suitable, modern, mature and complete scripting language that is fully prepared for research on hash functions and their usage in digital signatures and digital forensics. The software tool presented in this paper offers functionality that can help scientists or law enforcement agencies in forensic investigations and can provide assistance in gathering and securing digital evidence. The presented software tool offers an alternative solution for law enforcement agencies and companies who have low budget.

VI. CONCLUSION

With the increased use of computers, large amounts of data are generated and exchanged each day. The explosive growth of computer use provided a fertile ground for the emergence of electronic or cyber-crime. Consequently, the amount of data to be found in a digital forensic laboratory is constantly growing. The practical application of hash functions in digital forensics is to quickly and efficiently search and scan data, to ensure the authenticity and integrity of digital evidence, to detect violations of privacy and to identify known objects such as illegal documents.

The paper makes an analysis of some of the most commonly used hash functions and their application in digital signatures and digital forensics. The application presented in this paper offers an alternative tool compared with many high cost software solutions. The application relies on hash based methods to offer functionality that can help individuals or law enforcement agencies in the process of forensic examination of digital data. Future work will focus on further developing the functionality of this application, mainly focusing in the direction of hash functions used for similarity detection between images and videos. With the increased usage of computers in everyday live, cyber-crime is also experiencing growth. In the future forensic scientists will need more reliable software tools which will provide even further assistance with the processing of digital data and securing digital evidence.

REFERENCES

[1] B. Preneel and M. Lowry, "Analysis and Design of Cryptographic Hash Functions", Doctor Dissertation, University of Leuven, 2003, [Online] http://homes.esat.kuleuven.be/~preneel/phd_preneel_feb1993.pdf, accessed: 12.2.2017.

[2] V. Roussev, "An evaluation of forensic similarity hashes In Digital Investigation". Vol. 8, 2011.

[3] W. Lu, A.L. Varna and M. Wu, "Forensic hash for multimedia information" in Proceedings of the SPIE, Volume 7541, 2010 © SPIE – The International Society for Optical Engineering, doi: 10.1117/12.838745

[4] C. Zauner, "Implementation and benchmarking of perceptual image hash functions", M.S. thesis, Upper Austria University of Applied Sciences, 2010.

[5] I. Mironov, "Hash functions: Theory, attacks, and applications", Microsoft Research, Silicon Valley Campus, p.1-22, 2010.

[6] P. Gauravaram, "Cryptographic Hash Functions: Cryptanalysis, Design and Applications", Ph.D dissertation, Queensland University of Technology, 2007.

[7] M. Naor and M. Yung, "Universal one-way hash functions and their cryptographic applications", Proceedings of the twenty-first annual ACM symposium on Theory of computing - STOC '89, 1989, p.33-43. [Online] http://portal.acm.org/citation.cfm?doid=73007.73011.

[8] A.F.A.L. Mohamed, et al., "Testing the forensic soundness of forensic examination environments on bootable media", In Digital Investigation. Elsevier Ltd., 2014.

[9] M.D. Kohn, M.M. Eloff and J.H.P. Eloff, "Integrated digital forensic process model", Computers & Security, 38, 2013, p.103-115. [Online] http://www.sciencedirect.com/science/article/pii/S0167404813000849.

[10] J. Fridrich and M. Goljan, "Robust hash functions for digital watermarking", In Proceedings - International Conference on Information Technology: Coding and Computing, ITCC 2000, Institute of Electrical and Electronics Engineers Inc.

[11] A.T.S Ho, "Handbook of Digital Forensics of Multimedia Data and Devices", published by Wiley, September 2015, ISBN: 978-1-118-64050-0.

[12] K. Matusiewicz, M.Sc., "Analysis of Modern Dedicated Cryptographic Hash Functions", Macquarie University, August 2007.

[13] A. Philipp, D. Cowen and C. Davis, "Hacking Exposed Computer Forensics - Secrets & Solutions", (Second Edition), 2010, ISBN: 978-0-07-162678-1.

[14] M. Stevens, "Attacks on Hash functions and Applications", Ph.D dissertation thesis, 2012, ISBN: 978-94-6191-317-3.

[15] A.J. Menezes, P.C. van Oorschot and S.A. Vanstone, "Handbook of Applied Cryptography", August 2001, ISBN: 0-8493-8523-7.

[16] M. Spreitzenbarth, J. Uhrmann, "Mastering Python Forensics", October 2015, ISBN: 978-1-78398-804-4.

[17] J. Pieprzyk, T. Hardjono, J. Seberry, "Fundamentals of Computer Security", 2013, ISBN: 9783662073247.

[18] P. Leskov, "Introduction to computer security - Hash functions and digital signature", (lectures), Universitat Tubingen.

[19] V. Roussev, "Hashing and Data Fingerprinting in Digital Forensics", [Online] http://roussev.net/pubs/2009-IEEE-SP--hashing.pdf, accessed: 10/07/2016.

[20] http://csrc.nist.gov/publications/nistbul/March2009_cryptographic-hash-algorithm-family.pdf - [Online], accessed: 16/06/2016.

[21] https://www.sans.org/reading-room/whitepapers/vpns/overview-cryptographic-hash-functions-879 - [Online], accessed: 28/07/2016.

[22] https://digital-forensics.sans.org/summit-archives/2012/practical-use-of-cryptographic-hashes-in-forensic-investigations.pdf - [Online], accessed: 15/08/2016.

[23] http://research.ijcaonline.org/volume68/number23/pxc3887433.pdf - [Online], accessed: 16/07/2016.

[24]. https://www.fbi.h-da.de/fileadmin/personal/h.baier/Lectures-winter-11/WS-11-Forensics/vorlesung_forensik_ws11-12_kap08_hash-handout.pdf - [Online], accessed: 16/08/2016.

[25]. http://resources.infosecinstitute.com/computer-forensics-tools - [Online], accessed: 28/02/2017.

A Comparison of the Results Obtained by Two Types of Low-Discrepancy Sequences in Quasi-Monte Carlo Method

Vesna Dimitrievska Ristovska
Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
Email: vesna.dimitrievska.ristovska@finki.ukim.mk

***Abstract*—Quasi-Monte Carlo method is a well-known method for numerical integration and solving many numerical problems using low-discrepancy sequences. The main idea in Quasi Monte Carlo method is to approximate the integral of a function f as the average of the function evaluated at a set of points $x_1, \ldots, x_n$.**

In this paper we consider the absolute errors between the exact value of an definite integral and the results obtained with a numerical integration with Quasi-Monte Carlo method using low-discrepancy sequences: Halton and Sobol sequences. In the experimental computations we choose some different functions on the interval [0,1] in one dimensional case, with different number of points in the sequences.

Numerical results and graphical figures verify theoretical results: Quasi-Monte Carlo has a bigger rate of convergence than the rate for Monte Carlo method. Almost all our computations show that Sobol sequence produces better results, with smaller absolute errors than Halton sequence in numerical integration for the all chosen type of functions.

Keywords:Quasi-Monte Carlo method; low-discrepancy sequence; Halton sequence; Sobol sequence; numerical integration

I. Introduction

Quasi-Monte Carlo method is a well-known method for numerical integration and solving many numerical problems using low-discrepancy sequences (also called quasi-random sequences or sub-random sequences). The difference between Quasi-Monte Carlo and regular Monte Carlo method is that the second method is based on sequences of pseudo-random numbers. The main idea in the both these methods is the same: to approximate the integral of a function f as the average of the function evaluated at a set of points $x_1, \ldots, x_n$:

$$\int_0^1 f(x)dx \approx \frac{1}{n} \cdot (f_1 + f_2 + \ldots + f_n), \qquad (1)$$

where $f_k = f(x_k)$, for $k = 1, \ldots n$.

The benefit of using low-discrepancy sequences is a faster rate of convergence to the exact value of the definite integral. Quasi-Monte Carlo has a rate of convergence close to $\mathcal{O}(\frac{1}{n})$, while the rate for the Monte Carlo method is $\mathcal{O}(\frac{1}{\sqrt{n}})$.

A. Sequence of Halton

We will introduce this sequence following Halton [1].
Definition 1 *Let $b \geq 2$ is an integer number. If an arbitrary integer number $i \geq 0$ has its representation in a number system with a base b given by $i = \sum_{j=0}^{k} a_j(i)b^j$ then the sequence of Halton is defined as: $S_b(i) = \sum_{j=0}^{k} a_j(i)b^{-j-1}$.*

For example, if $b = 2$, the few first elements are: $\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \ldots$ In other words, the i-th element of this sequence is the number i written in binary representation, inversed, and written after the decimal point. This is true for arbitrary base.

As an example, to find the fourth element of the above sequence, we write $4 = 1 * 2^2 + 0 * 2^1 + 0 * 2^0 = 100_2$, which can be inverted and placed after the decimal point to give $0.001_2 = 0 * 2^{-1} + 0 * 2^{-2} + 1 * 2^{-3} = \frac{1}{8}$. So, the sequence above is the same as 0.1_2, 0.01_2, 0.11_2, 0.001_2, 0.101_2, 0.011_2, 0.111_2, 0.0001_2, 0.1001_2, ... Here on the Fig. 1 is an example of Halton sequence with base $b = 3$ and number of points $n = 27$.

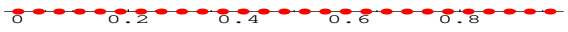


Fig. 1. Halton sequence $b = 3, n = 27$

If we consider Halton sequence of fixed base b, with length equal to power of b, it will fill the interval [0,1] uniformly. In this way the formula (1) gives an approximation of the definite integral with integral sum.

B. Sobol sequence

Sobol sequence (also called $\Lambda\Pi_\tau$ sequence or $(t, s)-$sequence in base 2) was first introduced by the Russian mathematician Ilya M. Sobol [5] in 1967. For this sequence a base two is used to form successively finer uniform partitions of the unit interval. The main property of this sequence is good uniform distribution in the $s-$dimensional unit hypercube.

For our numerical experiments, we will use generalized Sobol sequence, called $(t, s)-$sequence with arbitrary prime base b. Following Niederreiter [4] we will give the concept of this 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. It is necessary b to be a prime number. In the following we give the definition of a $(t, m, s)-$net and a $(t, s)-$sequences.

Definition 2 *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 3 *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.*

We notice that a $(t, m, s)-$net is extremely well uniformly distributed if the quality parameter t is small.

Here on the Fig. 2 is an example of generalized Sobol sequence with base $b = 23$ and number of points $n = 529$.

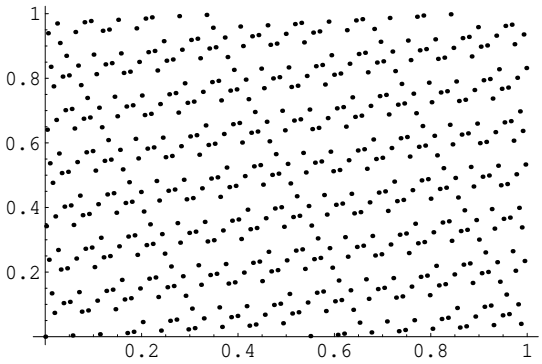


Fig. 2. Sobol (0,2)-net, s=2, b=23, n=529, a={6,16}

II. Main results

In this paper we present results for absolute errors of numerical integration with Quasi-Monte Carlo method in one dimensional case using two types of low-discrepancy sequences: Halton sequence and Sobol sequence.

In our paper we use software program written in [7] for generalization the constructive approach of Sobol to generate classes of $(t, s)-$ sequences over the field $\mathbb{F}_b$ by using monocyclic difference operators over $\mathbb{F}_b$, where b is a prime number.

In the experimental computations some different classes of functions are considered: $\sin x, \log(1+x), \exp(x), \sqrt{x}, x^2$, but the interval of integration for all of them is [0,1]. As a base for generation of the quasi-random sequences the prime numbers 11, 7 and 2 are chosen. The number of points in the sequences is $11^4, 7^5, 2^{13}$, respectively.

On the Fig. 3-15, the results obtained for absolute errors in computing $\int_0^1 f(x)dx$ by the regular Monte Carlo method are shown with red color, the results obtained by the Quasi Monte Carlo method using the Halton sequence are shown with green color and the results obtained by the Quasi Monte Carlo method using generalized Sobol sequence are shown with blue color.

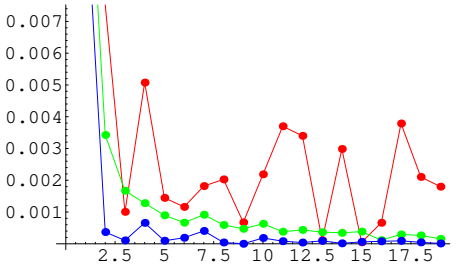


Fig. 3. Absolute errors for $\sin x$, $b = 11$, $n \leq 11^4$

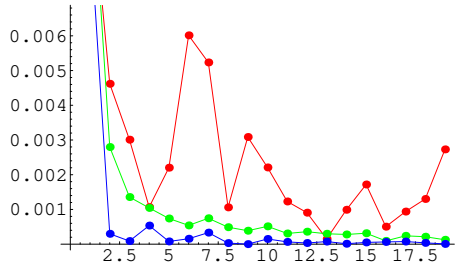


Fig. 4. Absolute errors for $\log(1+x), b = 11, n \leq 11^4$

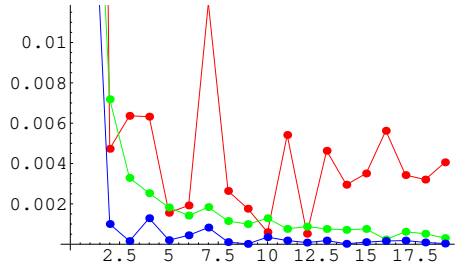


Fig. 5. Absolute errors for $e^x, b = 11, n \leq 11^4$

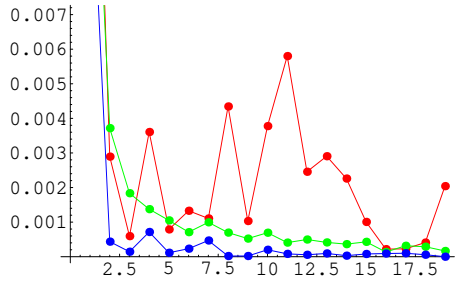


Fig. 6. Absolute errors for $\sqrt{x}, b = 11, n \leq 11^4$

On the Fig. 3, 4, 5 and 6 the absolute errors obtained in computing $\int_0^1 f(x)dx$ for functions $\sin x, \log(1+x), e^x, \sqrt{x}$, for base $b = 11$, and number of points in the sequence $n \leq 11^4$ are presented.

On the Fig. 7, 8, 9, 10 and 11 the absolute errors obtained

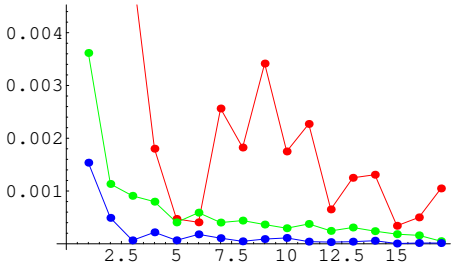


Fig. 7. Absolute errors for $\sqrt{x}$, $b = 7$, $n \leq 7^5$

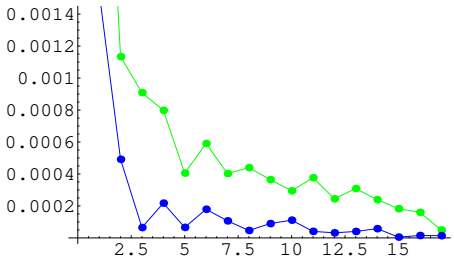


Fig. 8. Absolute errors for $\sqrt{x}$, $b = 7$, $n \leq 7^5$

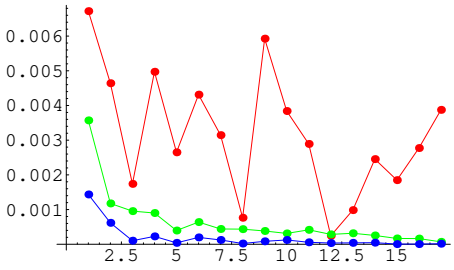


Fig. 9. Absolute errors for x^2, $b = 7$, $n \leq 7^5$

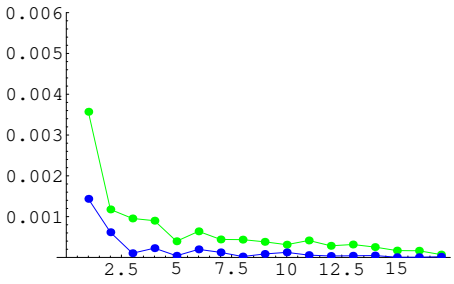


Fig. 10. Absolute errors for x^2, $b = 7$, $n \leq 7^5$

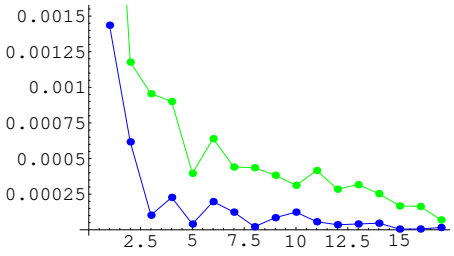


Fig. 11. Absolute errors for x^2, $b = 7$, $n \leq 7^5$

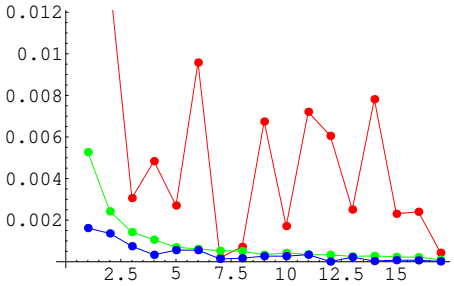


Fig. 12. Absolute errors for $\sin x$, $b = 2$, $n \leq 2^{13}$

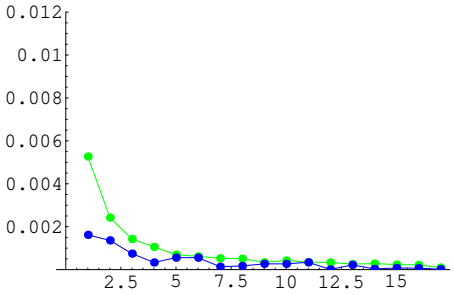


Fig. 13. Absolute errors for $\sin x$, $b = 2$, $n \leq 2^{13}$

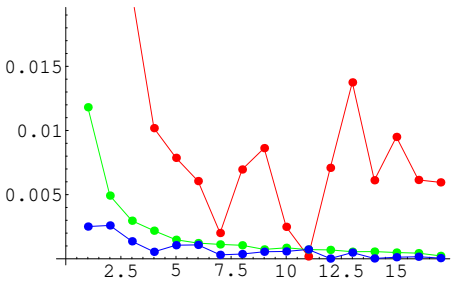


Fig. 14. Absolute errors for e^x, $b = 2$, $n \leq 2^{13}$

in computing $\int_0^1 f(x)dx$ for functions $\sqrt{x}$ and x^2, for base $b = 7$, and number of points in the sequence $n \leq 7^5$ are presented.

On the Fig. 12, 13, 14 and 15 the absolute errors obtained in computing $\int_0^1 f(x)dx$ for functions $\sin x$ and e^x, for base $b = 2$, and number of points in the sequence $n \leq 2^{13}$ are presented.

On the table on the Fig. 16, as a contribution to the comparison, results obtained for the absolute errors in computing $\int_0^1 e^x dx$ by the regular Monte Carlo method are presented in the second column; results obtained by the Quasi Monte Carlo method using the Halton sequence and Sobol sequence are shown in the third and the fourth column.

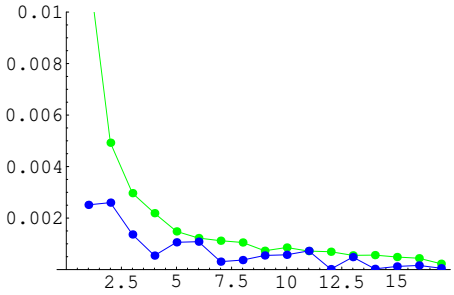


Fig. 15. Absolute errors for e^x, $b = 2$, $n \leq 2^{13}$

Numb. of points	Monte Carlo	QMCarlo Halton seq.	QMCarlo Sobol seq.
91	0.0379853	0.0118126	0.00251292
591	0.0221876	0.00492606	0.00260029
1091	0.0206925	0.00296867	0.0013639
1591	0.0101833	0.00219064	0.0005436
2091	0.00786837	0.00148285	0.00105895
2591	0.00605873	0.00122469	0.0010841
3091	0.00201137	0.00112107	0.000308212
3591	0.00696161	0.00105259	0.000370098
4091	0.00862636	0.000730971	0.000548657
4591	0.00249027	0.000857023	0.00057431
5091	0.000183324	0.000717911	0.000726263
5591	0.00708973	0.000692858	0.000018648
6091	0.0137498	0.000552531	0.000485038
6591	0.00612145	0.000562301	0.0000281534
7091	0.00949842	0.000488113	0.000122458
7591	0.00614588	0.000440701	0.000161182
8091	0.00595625	0.000226931	0.0000527024

Fig. 16. Absolute errors for e^x, $b = 2$, $n \leq 2^{13}$

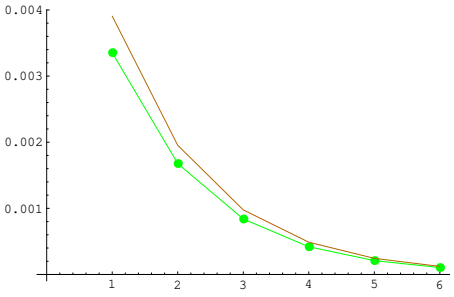


Fig. 17. QMC and $\mathcal{O}(\frac{1}{n})$: abs. err. for e^x, $b = 2$, $n \leq 2^{13}$

Two curves on the Fig. 17 verify asymptotic behavior of the Quasi-Monte Carlo method according to $\mathcal{O}(\frac{1}{n})$. The green points show the values obtained by the Quasi-Monte Carlo using Halton sequence and the brown curve is the graphics of the function defined by $\mathcal{O}(\frac{1}{n})$, i.e. the graphics of the function $n \mapsto \frac{1}{n}$.

III. Conclusion

Our numerical evaluations and visualizations verify theoretical results: that Quasi-Monte Carlo has a bigger rate of convergence than the rate for the Monte Carlo method. Almost all our computations show that the Sobol sequence produces better results in numerical integration for all chosen type of functions.

Consequently the next question is: what is the reason for higher quality results obtained using Sobol sequence in relation with the obtained results using the Halton sequence?

The answer is in the way of construction the sequences. The definition and the algorithm for constructing Sobol sequence is more complex and with more requirements than the algorithm for constructing Halton sequence. Since of these precise steps in construction, Sobol sequence is more uniformly distributed than Halton sequence, and respectively the absolute errors in numerical integration using Quasi-Monte Carlo are smaller.

Acknowledgment

This research was partially supported by Faculty of Computer Science and Engineering at "Ss Cyril and Methodius" University in Skopje.

References

[1] J. H. Halton, "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals", Numer Math., Vol. 2, 1960, pp. 84–90
[2] L. Kuipers, H. Niederreiter: Uniform distribution of sequences, John Wiley & Sons, New York, 1974
[3] H. Faure, "Discrépance de suites associées à un systeme de numeration (en dimension s)", Acta Arithmetica, XLI, 1982, pp. 337-351.
[4] H. Niederreiter, "Random Number Generator and Quasi-Monte Carlo Methods", CBMS - NSF Series in Applied Mathematics, 63, SIAM, Philadelphia, 1992.
[5] I. M. Sobol', Mnogomernye kvadraturnye formuly i funktsii Haara, Izdat. Nauka, Moscow, 1969.
[6] I. M. Sobol', On the calculations of multi-dimensional integrals, DAN SSSR, Vol. 139, Iss. 4, Moscow
[7] V. Dimitrievska Ristovska, V. Grozdanov, A. Atanasov, "An effective algorithm for constructing of (t, s)-sequences over $\mathbb{F}_b$", Proc. V Congress of mathematicians of Macedonia, 2014
[8] $https://en.wikipedia.org/wiki/Quasi-MonteCarlo-method$
[9] $https://en.wikipedia.org/wiki/Halton-sequence$
[10] $https://en.wikipedia.org/wiki/Sobol-sequence$
[11] J. G. Van der Corput, "Verteilungsfunktionen", Proc. Kon. Ned. Akad. Wetensch., Vol.38, 1935, pp. 813–821

Lightweight OAI-PMH repository server implementation

Nikola Popovski
Faculty of Information
and Communication Technologies,
St. Kliment Ohridski University - Bitola
Email: nikola.popovski86@gmail.com

Ilija Jolevski
Faculty of Information
and Communication Technologies,
St. Kliment Ohridski University - Bitola
Email: ilija.jolevski@fikt.edu.mk

***Abstract*—In this paper we will describe what comprises the minimal standard compliant OAI-PMH server implementation and demonstrate a standalone Java based implementation of a OAI-PMH digital repository server, primarily aimed at small online publishers. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol established for harvesting (or collecting) metadata descriptions of records in an archive so that services can be built using metadata from a number of repositories. It provides facilities for metadata exchange between applications, ensuring metadata standardization and validation. The built proof-of-concept server is very fast and lightweight since it does not rely on external client-server implementations for database access; but rather it uses embedded SQLite database. It provides the repository functions with only the fraction of the load and dependencies of other similar implementations.**

keywords: OAI-PMH, digital repository, library, online publishing

I. Introduction

Digital repository, also known as institutional repository (IR) is a network accessible archive of digital items. The purpose of a digital repository is to collect, preserve and disseminate digital copies of intellectual output of an institution. Some of the main objectives of developing an institutional repository is to provide open access to institutional research output by self-archiving it, to provide global visibility to the research and store and preserve materials that are unpublished or otherwise lost, such as theses, technical reports etc. The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of digital content. Although OAI has its roots in the Open Access and IR movements, over time it has expanded to promote broad access to digital resources. Such effort is the OAI Protocol for Metadata Harvesting (OAI-PMH). The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [1] is a protocol designed to provide application-independent interoperability for metadata exchange. The concept of metadata exchange is divided in two entities in the context of OAI-PMH:

- Data provides
- Service providers

II. OAI-PMH

This document addresses data providers, particularly repositories, in a manner of defining, describing and providing implementation analysis. In order to familiarize with data providers as a concept in context of OAI-PMH, we should begin by listing the relative entities. Let us begin by outlining the protocol itself. As mentioned before, the OAI-PMH serves the purpose of providing means for metadata exchange. But what is metadata?

A. Metadata

Metadata is most simply defined as data about data. In more technical terms, metadata is a collection of information which provides additional description about a resource. A resource can be anything from a physical object to a digital document.

B. Metadata standards

Because each of these processes can be done in a countless ways, metadata itself can be a subject of arbitrary definition. However, many specialized and well-defined models arouse from the need for standardization. Such standard is implied by the Dublin Core Metadata Initiative [7] (DCMI) as a set of predefined vocabulary terms that can be used to describe resources. The set is comprised of 15 predefined elements. Additional information about the set can be found at: http://purl.org/dc/elements/1.1/. OAI-PMH does not define a metadata standard of its own, but rather relies on the Dublin Core Schema; it is required by the protocol, that metadata could be at least disseminated in DC format.

C. Usage

OAI-PMH is designed to work on top of HTTP. It defines a set of six REST based [2] actions (verbs) and a number of parameters that can be used in valid communication. Using any allowed combination of verb and parameters inside a standard HTTP request, by this protocol, results in metadata exchange. It should be noted that OAI-PMH does not rely on the HTTP methods of error handling, but rather defines its own set of errors. That being said, the validity of the OAI-PMH request/response does not mirror the validity of the HTTP request, and vice versa. The six actions (verbs) are allowed in OAI-PMH are:

- **Identify** – Identifies the data provider (repository). It provides metadata about the repositorys configuration.

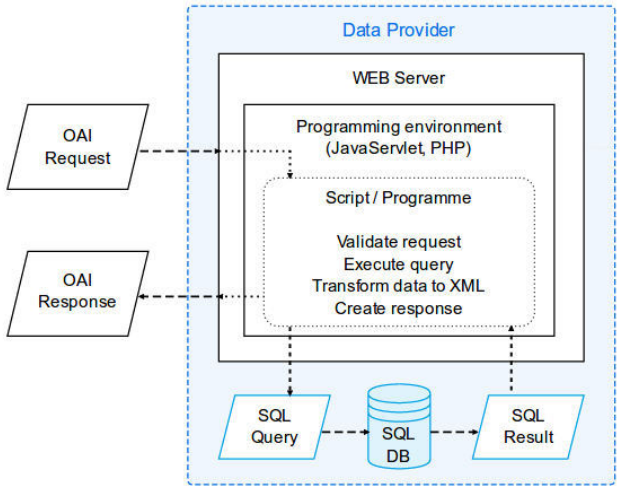


Fig. 1. Typical structure of a OAI-PMH compliant data provider

- **ListMetadataFormats** – Lists the metadata formats supported by the repository or the metadata formats that are available for a resource (if any).
- **ListIdentifiers** – Retrieves a list of headers (descriptions) of records.
- **ListSets** – Used to retrieve the set structure of the repository.
- **GetRecord** – Retrieve an individual metadata record.
- **ListRecords** – Retrieve a list of metadata records, which conforms certain criteria.

D. Service providers

Service providers use the metadata harvested via the OAI-PMH as a basis for building value added services. Service providers rely on harvesters for metadata harvesting. A harvester is a client application that issues OAI-PMH requests which combine the allowed verbs and parameters to further refine the metadata harvesting. In this document, however, we are mostly concerned about repositories. A repository is a network accessible application that can process the six OAI-PMH requests. Data providers rely on repositories to expose metadata to harvesters. Three distinct entities relate to the metadata exposed via OAI-PMH:

- **Resource** – The object in question. The object which the exposed metadata is about.
- **Item** – a part, from which metadata about a resource can be disseminated.
- **Record** – metadata in a specific metadata format.

III. Digital repositories

Repositories conforming to the definition of the OAI-PMH can be implemented in a number of ways. The typical structure of a OAI-PMH compliant data provider is shown on Fig. 1

Many features mentioned before are optional; a repository can have minimal implementation (as we created it), dealing only with the absolute necessities for metadata exposure. Minimal implementation implies that the repository is able to identify itself, expose small amount of unstructured metadata and provide basic flow control and exceptions. Or, in more technical outline:

1) Our minimal implementation: The minimal repository implementation requires that metadata can be exposed at least in DC format. Repositories are free, however, to implement any other metadata standard, or provide community driven metadata formats. Sets provide mechanism for grouping, and thus partitioning the repositorys contents. While this is a powerful mechanism for providing selective harvesting, it is often rather complicated to implement and not all harvesters will make use of the feature. Like non-DC metadata, sets are most likely to be useful within specific communities. OAI-PMH states that set hierarchy is optional for any repository, thus it is recommended that implementation of sets is omitted unless a real need is presented.

IV. Features of our server implementation

A. Response compression

Response compression is the ability to compress the data sent to harvesters, aiming for performance optimization. This is rather aim, than actual optimization because two possible scenarios apply to this context:

- Not all harvesters elect to implement acceptance of compressed responses. As selective harvesting can significantly diminish the results being acquired from each request, it is often unnecessary to compress the response as it is sufficiently small.
- Lengthy responses, that are candidates for compression, do also require additional time and resources to compress and decompress. Unless explicitly required, in such an instance, as a responses that have large collections (e.g. thousands of records), compression and decompression impose execution complexity in both, harvesters and repositories and should therefore omitted.

B. Flow control

Flow control deals with validation of requests, execution exceptions and the idempotence (definition of idempotence [6]) of list retrieval. It is intuitive that validation of requests and exception handling are a necessity. Incomplete lists, however, are neither a necessity nor a performance booster. As with compression, implementation of resumptionToken (a mechanism to retrieve partial lists and ensure idempotence) mainly depends on the amount of data the repository id designed to expose. Suffice to say, that if a repository contains a large collection of records, it becomes somewhat necessary to implement partial lists. However, what a large collection is, and how many records comprise such a large collection is an arbitrary definition.

C. Date-stamp and granularity

OAI-PMH permits two date-stamp granularities: days and seconds, formatted **yyyy-mm-dd** and **yyyy-mm-ddThh:MM:ssZ** respectively. Each new or modified items

date-stamp, must reflect the granularity used for the metadata. The date-stamp values used in OAI-PMH are for the purpose of selective harvesting. They should be changed, in order to reflect any change in the item which comprises any metadata format. For example, even if the underlying master-record is not changed, a change in the way that some metadata formats are generated on the fly should result in date-stamp change (update) for all items that comprise the generated metadata.

D. Incomplete lists

Let us, however, assume that we need full repository implementation. As the definition for repository in OAI-PMH context states, such application should be able to process the six actions (verbs) mentioned before. Although, some actions are fairly straightforward to handle, listing actions in particular should be implemented holding in regard scalability. That is, they should be implemented under the assumption that a repository can, and in fact will, grow continuously. This implies that listing requests, although partitioned via selective harvesting, can still be relatively long, and should therefor be further divided into a set of incomplete-list responses.

E. Resumption token

OAI-PMH conceptualizes this strategy via the use of resumptionToken elements. Suppose a harvester requests a list of all records between lowest and highest time-stamp in the repository. Let us further assume, that this request matches a 1000 records. In order to serve the entire record set, the repository needs to query a database, allocate memory for the entire 1000 element set, and subsequently disseminate metadata for each record in a particular format. It is apparent that this approach is memory and time consuming. Instead, by partitioning the response, the repository is free to serve additional requests, while the harvester handles the response it has been served. The OAI-PMH, as a protocol merely provides the concept of incremental harvesting. It does not in any way suggest an implementation strategy or syntax for the resumptionToken element. There are mainly two strategies that are naturally suitable for implementing incomplete list responses. Depending on the scenario, they can deliver stateful or stateless communication between the parties. Because HTTP is stateless by definition, it would be a good idea that the implementation strategy should allow the repository maintain stateless communication. Stateful implementation means that each response result is cached, and delivered to a number of subsequent requests from the same client. This approach is a good choice for maintaining idempotence, as records that are cached cannot be accessed until the list is completed. This means, that any changes in the repository will not be reflected until the original request is satisfied. This however, means that repositories should implement logic for caching and cleaning which further implies that the resumptionToken element should expire at some point, to avoid cache overflow. Stateless implementation trades machine resource usage and complexity for caching results, for possible increased time in recomputing the state for the next response. In the previous example, that included a set of 1000 records, the following course of subsequent actions provides implementation where state is encoded in each response:

- A client (harvester) issues a OAI-PMH request: List all records between your earliest date-stamp (provided in the Identify response) and now that are in DC metadata format.
- The repository matches a thousand records, but is configured to serve only a 100 at a time. I responds with a 100 element set, providing along the resumption token.

This encodes the state of the original request by including the from, until and metadataPrefix arguments, while storing information in cursor about the last id that has been delivered (assuming that ids are in incremental order), thus providing information for the next request where the list should be resumed. This however imposes a problem on maintaining idempotence. Suppose that by the time the incomplete list is processed by the harvester, a record is updated to move out of the interval. Although in this scenario strict idempotence is not required according to the OAI-PMH, it is good practice that this is avoided. One strategy to accomplish that is to maintain scheduled update intervals in the repositories, invalidating all resumptionToken elements that have been issued. This can be done by including the expirationDate attribute in the resumptionToken element, which matches the time when the next update interval is scheduled. In this period, the repository may choose not to process requests, and instead respond with HTTP status code 503 Service Unavailable, also providing the Retry-After header. When all subsequent requests are served, and the list is completed, an empty resumptionToken element is included in the final response, to signify the completion of the list.

F. Sets

Sets provide a mechanism for selective harvesting. Set hierarchy is optional in the context of implementation and even if present, harvesters may choose to ignore the structure if exposed. Sets can conform to a certain hierarchy which is denoted by the colon (:) character. Repositories that implement sets, may choose not to implement set hierarchy, and should not include this character in the setSpec elements. Set description may be included in the setDescription element of the ListSets response. If the whole repository represents a single collection, then it might be more appropriate to include the description container in the Identify response to describe the collection.

G. Error handling

Errors and exceptions are inevitable part of every implementation. The OAI-PMH defines error handling by using one or more error elements. Although a single error is enough to stop processing the request, it is strongly recommended that all errors are reported. Furthermore, each error element should include a detailed and helpful error message describing the nature and cause of the error, in addition to the mandatory error code. According to the OAI-PMH repositories are permitted

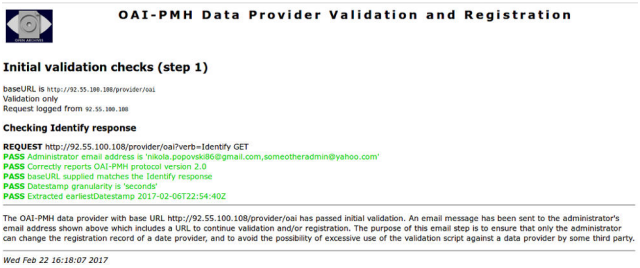


Fig. 2. Initial validations check Step 1

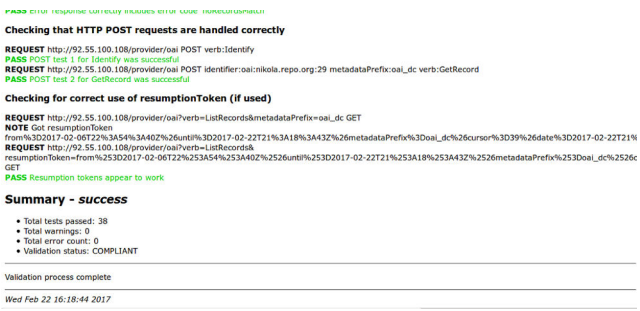


Fig. 3. Automated tests for compliance to the OAI-PMH protocol

to include just a single error element even when multiple error occur. However, to aid debugging, it is advised that all errors are reported in a list of error elements, even if they share the same element code, such as badArgument.

H. Our implementation

The authors have built very fast and lightweight implementation of the minimal OAI-PMH standard compliant repository server. The server was built in Java and since it does not rely on external client-server implementations for database access; but rather it uses embedded SQLite database, it is very lightweight and fast. It provides the repository functions with only the fraction of the load and dependencies of other similar implementations.

I. Verification and registration

After a repository is implemented to satisfy the institution's needs, it needs to be verified that it is OAI-PMH compliant, and subsequently registered if all tests are passed. There is a two step verification process that can be initiated at [5]. After providing the address, where the repository can be found on the network, the first step is to Identify the repository and determine if it is properly configured. A verification email is sent to all administrators, requesting confirmation to initiate the next step in the process.

Step 2, performs a number of automated tests to determine if the repository is in fact OAI-PMH compliant. After all tests are executed, a report is generated and sent to the administrators via email. Our Java based implementation met all requirements proposed by the standard, so the repository hosted via our solution can be registered with the Open Archives Initiatives directory.

V. Further development

As the protocol allows optional implementation of sets, it is highly recommended that if no immediate need exists, implementation of the sets feature should, in fact, be omitted. Because a basic repository seldom addresses the notion of grouping the items, implementation should be provided only as a means for scalability; meaning that it should be predicted in the project planning, such as database design, even if the repository elects not to provide this feature.

VI. Related Work

There are a number of already available and much bigger implementations of OAI-PMH repositories. Although all of them serve the defined purpose of institutional repository software, they vary in complexity, licensing and other services they provide. Some of the most popular IR software solutions are:

- Digital Commons [4] A hosted IR platform. It supports a variety of publication and editorial workflows, as well as peer review. Content is uploaded through batch uploads, by linking to external sites, or via a customizable submit form.
- DSpace An open source repository software. It serves a specific need as a digital archives system, focused on the long-term storage, access and preservation of digital content.
- EPrints [3] - A free and open-source software package for building open access repositories that are compliant with the OAI-PMH.

VII. Conclusion

This document outlines the general guidelines for implementing a OAI-PMH compliant digital repository and our proof of concept implementation in Java. Addressing the needs of the institution, correlating with the OAI-PMH, it is clear that many forms of implementation are possible. From small portable applications with embedded database, to large distributed systems. The main goal is to provide interoperability framework for metadata dissemination and thus resource discovering and sharing.

References

[1] OAI-PMH Initiative, https://www.openarchives.org/pmh/
[2] Tilkov, Stefan, A brief introduction to REST, InfoQ, Dec 10 (2007).
[3] McCutcheon, V and Eadie, M, Managing open access with EPrints software: a case study, Insights, 2016, 29(1), 4552; DOI: http://dx.doi.org/10.1629/uksg.277
[4] White, Deborah L. and Ohira, Yumi, Creating Open Access to Institutional Scholarship using Digital Commons (2016). Criss Library Faculty Proceedings & Presentations. 73. http://digitalcommons.unomaha.edu/crisslibfacproc/73
[5] Repository Validation, https://www.openarchives.org/Register/ValidateSite
[6] Idempotence, https://en.wikipedia.org/wiki/Idempotence
[7] Dublin Core Metadata Initiative, http://dublincore.org/

Analysis of server and network performance for HTTP-based streaming

Sasho Gramatikov

Faculty of Computer Science and Engineering, Skopje,
University "Ss. Cyril and Methodius", Skopje, Macedonia
Email: sasho.gramatikov@finki.ukim.mk

***Abstract*—The Video on Demand (VoD) is a service that requires significant amount of data to be processed by the streaming servers and delivered thought the Internet or the privately managed network. There are different mechanisms for processing and delivery of the video streams, which imply different processing and network resource demands. In this work, we created an environment for real-case measurement of the server performance for VoD streaming using HTTP progressive and adaptive streaming. We measured the CPU utilization, the generated in-bound and out-bound traffic, the number of established connections and the memory requirements. From the analysis of the obtained data, we concluded that although the adaptive streaming is more CPU demanding streaming mechanism, it provides smoother traffic pattern, less simultaneous connections and less memory. Moreover, the server performance for the adaptive streaming can be further improved by increasing the duration of the video segments.**

***Keywords*—Performance, Efficiency, Video, adaptive streaming**

I. Introduction

The rapid advances of the network technologies created suitable conditions for dominant presence of the Video on Demand (VoD) streaming services on the Internet and the privately managed networks. The main characteristic of this service is that it generates large amount of traffic since each request for video requires a dedicated stream. There are various mechanisms for delivering the video streams to the clients. The traditional streaming uses the RTP (Real-time Transport Protocol) protocol with combination of the RTSP (Real Time Streaming Protocol) and delivers contents as long as the client watches the video. The drawbacks of this delivery method for VoD services are the reliability, the use of UDP (User Datagram Protocol) as transport protocol and the usage of proprietary streaming servers. Although aimed for transfer of text files on the Internet, HTTP (Hyper Text Transfer Protocol) became a handy protocol for delivery of videos. Since it uses the well known port number 80, the video contents could easily pass through the firewalls of even the most restrictive administrators. Another advantage of the protocol is that it does not require dedicated streaming server and uses the same servers for hosting web contents. The protocol is also very reliable because it uses the TCP (Transport Control Protocol) as transport protocol, guaranteeing that the data between the server an the client will arrive in order and with no errors. This kind of streaming is called progressive streaming because it tends to download as much of the video as possible. From the nature of the TCP protocol, which tends to maximally utilize the available throughput, the clients download significantly higher portion of the video at the very beginning of the transfer, and therefore, it is very common that the video is completely downloaded at a moment when the client watched only a small portion of it. If the client decides to stop watching, the remaining downloaded data is wasted.Another disadvantage comes from the fact that the downloaded video can be further distributed by the client, which is a copy-right related issue.

The most widely accepted solution for overcoming the disadvantages of the previous delivery methods is the HTTP Adaptive Streaming (HAS). It divides the videos into segments with duration of several seconds and delivers entire segments as long as the client is watching the video. Apart from the reliable delivery of small video segments, the main advantage of the HAS is that the streaming rate can be adapted to the quality of the link between the client and the server. This is achieved by coding the video with different qualities, dividing each version into segments and generation manifest file that contains information about the available video qualities and locations of all the segments. Thus, for every segment of the video there are several qualities which can be requested from the server independently, a feature that is impossibly to implement in the traditional and progressive streaming.

There are many commercial HAS implementations such as Apple's HTTP Live Streaming (HLS) [1], Microsoft's Http Smooth Streaming (HSS) [2] and Adobe's HTTP Dynamic Streaming (HDS) [3] used in the popular video streaming platforms. In order to overcome the different segment and manifest file formats, the HAS was standardized by MPEG into the open standard MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH) [4]. The standard defines the format of the manifest file called Media Presentation Description (MPD). The DASH standard is currently widely accepted by a large community of leading streaming companies which formed the DASH Industry Forum (DASH-IF) [5].

No matter which mechanism is used for delivery of the videos, each request requires dedicated network and server resources. To our knowledge, there are no articles in the literature that treat the measurement of the server performance during the streaming process. Most of the work that considers the performance issue is focused on measurements on client's side. They treat the Quality of Experience (QoE) received by the clients for various streaming quality selection algorithms based on the network and the server conditions [6] [7] [8]. Therefore, the goal of this paper is to measure and compare the resources that are required by the server for VoD streaming us-

ing progressive and HAS streaming service. For that purpose, we created test environment consisting of a streaming server and clients that simultaneously generate requests for videos from the server. Then, we measure the resources required by the streaming server for serving these requests, such as CPU utilization, memory required by the server process, out-bound, in-bound bandwidth and number of established TCP connections. The results of these measurements give an overall picture of the efficiency of the different streaming methods.

II. Measurement Environment

In order to create real-case scenarios for measurements of progressive and adaptive streaming, we designed a network of server and clients which required additional software for playing the received streams, measuring and recording the performance data. We also provided contents which are adapted for HAS streaming.

A. Network and hardware requirements

Our testbed network consists of one streaming server and four client machines within the campus perimeters. The clients machines simulate multiple clients that simultaneously request videos. The streaming server is virtual machine on the faculty cloud with four cores and 16 GB RAM. Its is running Ubuntu 16.04 operating system with Apache web server for hosting the videos. Two of the clients are also virtual machines on the same cloud with the same specifications as the server machine. They are connected to the server with a virtual network. The other two clients are desktop and laptop computers connected to the cloud via the campus network. The connections of the sever and the clients are shown in the diagram in Fig.1. Two of the clients run Ubuntu 16.04 and the other two run Windows operating systems.

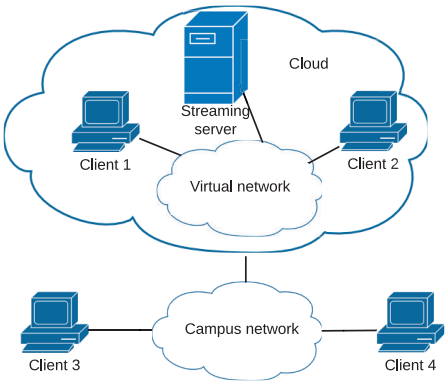


Fig. 1. Network diagram of streaming server and clients

B. Software requirements

Due to its simplicity, the progressive streaming can be viewed on any media player. For the purpose of our work, we use the Chrome web browser and its native video player. The source of the HTML5 video tag is set to the URL of the MP4 video. The HAS streaming, however, is not supported by all the media players and the native players of the web browsers. Therefore, a specific web player or libraries have to be used. There are many proprietary and open-source players. In our work, we use the Dash.js v.2.4.0 player [9] for DASH streaming, developed and maintained by the DASH Industry Forum. The web browser player uses client-side JavaScript libraries. Dash.js is still not supported by all web browser, but, since it is supported by the majority of them, including Chrome, we use it as a tool for playing DASH contents. To initiate a DASH session, the source of the HTML video tag is set to the URL of the MPD file. Then, the playing of the DASH streaming is handled by the dash.js library.

The disadvantage of using a web browser for measuring purposes is that it limits the number of simultaneous connections to single host. The current maximum allowed number of connections in Chrome is 6, which is the number of simultaneous streaming sessions per host. With 4 different clients, we are limited to 24 simultaneous sessions. It is important to emphasize that there can be more the 6 concurrent streaming sessions, however, if they are all initiated in near time proximity, the most recent sessions will suffer significant delay until the previous sessions fetch enough data and temporary release the connections. This limitation especially affects the progressive streaming.

In order to run multiple video sessions, we created operating system specific scripts that enable running an arbitrary number of sessions for specific video with arbitrary inter-session interval. The disadvantage of the current implementation of the scripts is that they have to be executed manually on every client.

In order to measure and record the system performance during the video streaming, we created a dedicated program that runs on the server. The program uses the SIGAR library [10] which provides a cross-platform, cross-language programming interface to low-level information on computer hardware and operating system activity. The SIGAR library offers functions for accessing process specific data, like CPU time and memory utilization, as well as system specific data such as overall memory utilization, received bytes, transmitted bytes on the network interface and established TCP connections. The program gathers this data periodically (every second) and writes them in CSV file that is used for latter analysis.

C. Content requirements

The server hosts one sample 1280x720 video with duration of 60 seconds. The video is encoded with the H264 codec in MP4 format with playback rate of 1380 kpbs and total size of 10 MB. In order to provide DASH streaming, we further divided the sample video into segments and created corresponding MPD files for every value of the segment duration. In our work we used the open-source MP4Box tool [11] to divide the video in segments with durations of 1, 2, 5 and 10 seconds. Each version of the segmented video is accessed by downloading the MPD file that contains the meta data and the url of every segment for that quality.

III. Measurements and Results

Every measurement scenario is conducted by simultaneous execution of the scripts for generating requests from the client. After the first request, the clients generate new request every 2 seconds. The program for measuring and recording the server performance runs until the last video session is over. For every

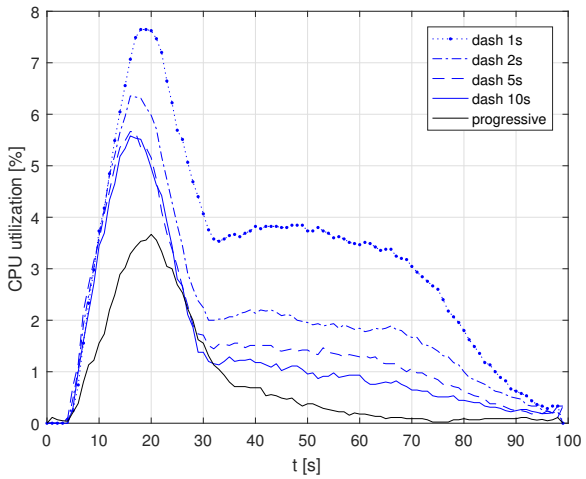


Fig. 2. CPU utilization for progressive streaming and DASH streaming with segment duration of 1, 2, 5 and 10 s.

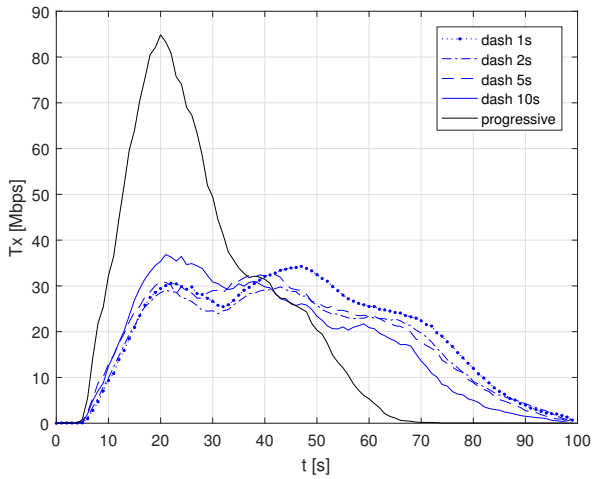


Fig. 3. Out-bound traffic for progressive streaming and DASH streaming with segment duration of 1, 2, 5 and 10 s.

specific scenario we run 5 independent measurements and use their mean value for the analysis. In order to reduce the periodic spikes that appear in most of the analyzed parameters, especially the traffic data, we applied a smoothing function that takes the average of the proximate points of the measurement data.

The results of the server performance for the scenarios of progressive streaming and adaptive streaming for different durations of the segments are shown in Figure.2-Figure.6.

In Figure. 2 the CPU utilization of the Apache process during the streaming is shown. The figure shows that the CPU utilization pattern is the same for all the scenarios, with a significant CPU load in the initial phase and a slow decrease afterwards. The peak of the CPU utilization is reached after the last request of the clients. If we compare the progressive and the adaptive streaming, we can conclude that the adaptive streaming is more processing power demanding. The shorter the segment duration, the more CPU power is required. The reason for this behaviour is that the streaming of shorter segments imposes more requests by the clients that have to be processed by the server. Unlike the adaptive streaming where clients make request in regular fashion, in the case of progressive streaming, most of the video content is requested, downloaded and buffered at the very beginning, alleviating the servers in the remaining time while clients are watching the video.

Figure 3 shows the time dependence of the server out-bound traffic. As it is expected from download of large files using the TCP protocol, the progressive streaming reaches a peak immediately after the last request since all the clients tend to download as much of the content as the link permits. Since the measurements are conducted in the campus network where the links between the server and the clients have large capacities, the out-bound traffic peak reaches values above 80 Mbps. We can observe that the traffic rapidly reduces even before the end of the videos. The DASH streaming, on the contrary, follows different out-bound traffic pattern which has even changes during the streaming process. This behavior is consequence of the fact that the players request only one

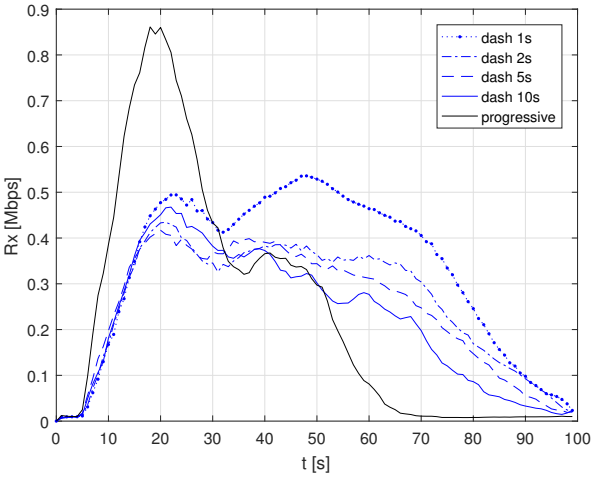


Fig. 4. In-bound traffic for progressive streaming and DASH streaming with segment duration of 1, 2, 5 and 10 s.

segment at a time.

The amounts of in-bound traffic originating from the clients during the streaming process is shown in Figure. 4. One can notice that the figure has very similar shape to Figure 3, but has two orders of magnitude smaller values. The curves for the DASH traffic show that when the segments have shorter duration, there is more traffic in the network due to the larger number of segments that have to be requested for the entire video.

Another parameter that is considered in our measurements is the number of simultaneous TCP connections for distribution of the data segments to the clients shown in Figure 5. As expected from the limitation of the Chrome browser to allow maximum number of 6 simultaneous connections, serving the 4 clients with progressive streaming reaches the maximum number of 24 connections in the initial phase. Afterwards, this number reduces with the same intensity as it increased at the beginning of the streaming process. If there were

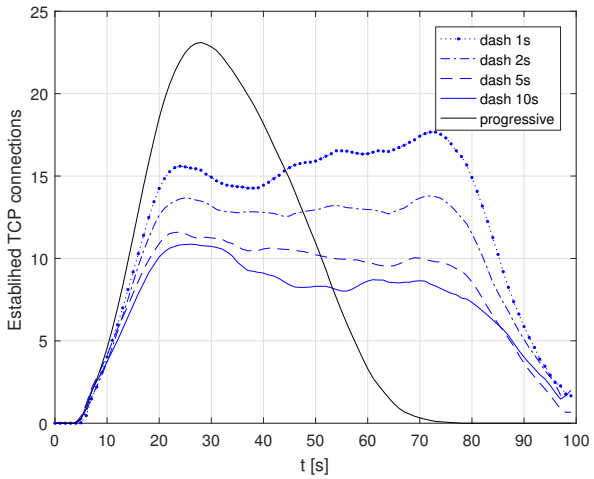


Fig. 5. Number of concurrent TCP connections for progressive streaming and DASH streaming with segment duration of 1, 2, 5 and 10 s.

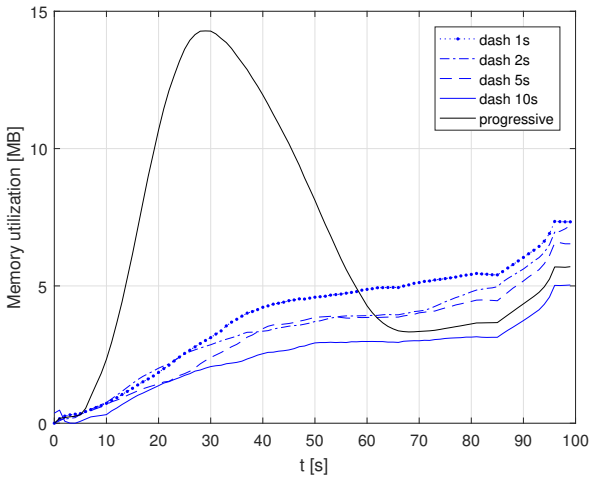


Fig. 6. Memory utilization for progressive streaming and DASH streaming with segment duration of 1, 2, 5 and 10 s.

more clients requesting videos from the server in the initial phase, they would stall until some of the players would have buffered enough content and released the connections. The adaptive streaming, on the other side, served the same number of clients with significantly smaller number of connections which is maintained during most of the streaming process. We can justify this behaviour by the fact that only a portion of the clients are simultaneously requesting segments, while the rest of the clients are watching the buffered segment from the previous request. Once the buffer reaches critical level, the player takes a connection from the players that were finished downloading a segment. The number of simultaneous connections depends on the segment duration. The shorter segments impose more frequent request, hence they maintain higher number of simultaneous connections.

The different streaming scenarios also require different memory resources by the server, which can be shown in Figure 6. The progressive streaming demands substantially more RAM memory space in the initial phase compared to the adaptive streaming in the initial phase due to the higher number of simultaneous connections and higher amount of data that has to be simultaneously read from the secondary memory and served. After the initial phase, the memory demands reduce to values equal with the demands for the adaptive streaming. Comparing the memory requirements for the different segment durations, we can conclude that the longer segments are less memory demanding.

IV. Conclusions

From the analysis, we can conclude that the progressive streaming requires less CPU power on the server, however it generates bursts of out-bound traffic, number of connections and memory utilization. On the contrary, the adaptive streaming requires more CPU power, less connections and less memory. The values of these parameters and the generated traffic are more even and hence more predictable. The analysis show that the negative effects of the server performances can be reduced by increasing the segment duration. Although the adaptive streaming requires more CPU power, it compensates with the capability to serve more simultaneous clients. Another key advantage is that clients can switch quality of the video during the streaming depending on their link quality.

Acknowledgment

The author thanks the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje, under the EEAVS (“Energy Efficiency of Adaptive Video Streaming”) project for financial support.

References

[1] (2016) Apple HTTP Live Streaming. [Online]. Available: https://developer.apple.com/streaming

[2] (2016) Microsoft Smooth Streaming. [Online]. Available: http://www.iis.net/downloads/microsoft/smooth-streaming

[3] (2016) Adobe HTTP Dynamic Streaming. [Online]. Available: http://www.adobe.com/products/hds-dynamic-streaming.html

[4] A. Vetro and I. Sodagar, “Industry and Standards The MPEG-DASH Standard for Multimedia Streaming Over the Internet,” *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, 2011.

[5] (2016) DASH Industry Forum. [Online]. Available: http://www.dashif.org

[6] L. Yitong, S. Yun, M. Yinian, L. Jing, L. Qi, and Y. Dacheng, “A study on Quality of Experience for adaptive streaming service,” *Communications Workshops (ICC), 2013 IEEE International Conference on*, pp. 682–686, 2013.

[7] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, “Probe and adapt: Rate adaptation for HTTP video streaming at scale,” *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, 2014.

[8] C. Zhou, C. W. Lin, and Z. Guo, “MDASH: A Markov Decision-Based Rate Adaptation Approach for Dynamic HTTP Streaming,” *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 738–751, 2016.

[9] (2017) Dash.js MPEG DASH player. [Online]. Available: https://github.com/Dash-Industry-Forum/dash.js

[10] (2017) SIGAR. [Online]. Available: http://sigar.hyperic.com/

[11] (2017) MP4Box. [Online]. Available: https://gpac.wp.imt.fr/mp4box/

Cyber Attacks On Power Grids

Will There Be Light In The Future?

Goce Kiseloski, Dobre Blazhevski, Veno Pachovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Republic of Macedonia
gkiseloski@gmail.com, {dobre.blazevski, pachovski}@uacs.edu.mk

Abstract—Electricity plays a very important role in today's world. Functioning of the military, medical and home appliances, computers and lap-tops depend on electricity. Electric power runs banks, industries, factories, hospitals, governmental facilities, water supply systems. Therefore, it is safe to conclude that power plants and power grids play important role in the modern society. Lack of electrical power will affect more than one segment in the society, since these segments are all connected in some way. Therefore, long term power failure will have cascade effect and will have very big impact on the society. Hacking attacks on Ukraine power grid proved that cyber threats can cause black-out for longer periods of time. Usually, attackers plan attacks from outside via Internet, but they can plan attack from the inside as well by using insiders. Typical question is how safe power plants (and power grids) are from the hacker attacks and its answer is very important for every country. In order to answer this question it is very important to consider how we control and command our power plants and our power distribution network, communication channels and most importantly, how secure communication channels are? Therefore, this paper discusses SCADA and VPN based vulnerabilities. In addition, publically available information regarding control, command and communications exposes data attackers can use. We found that such information is available on Internet for 40% of Balkan countries. At the end recommendations are given to mitigate some of the vulnerabilities.

Keywords—power grid; ICS; SCADA vulnerability; electrical distribution

I. INTRODUCTION

Electricity is vital to modern life and it is almost considered a necessity. It powers the lights and appliances in our households. It is used to power many industry related processes and it powers huge transport links and infrastructure such as railway transport systems. The need for electricity drives a never-ending growing demand for generation and transmission of it, with thousands of new power plants that are built around the globe to satisfy the need of this valuable resource.

On the way from production then transmission to the end users, electricity has to travel many kilometers, and has to change its form, ex. increasing and decreasing of the voltage, and making it suitable for use to the end users. In this vast network it comes to various electrical transmission and generation objects such as: power plants, high voltage substations and low voltage substations with transformers (Fig. 1).

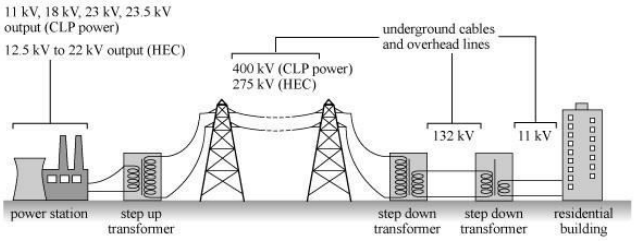


Fig. 1. Example of electrical transmission and generation objects [1]

Very important question to be asked is what will happen if a major power outage occurs that will leave a large urban area without electricity for a couple of hours or even days. Then, it is almost certain that the life there will come to a complete stop and massive social disruption will result, because social structures in most developed countries rely on high-reliability electricity [2]. Blackout or a power outage can be described as the total loss of power to an area. This can be result of overloading the network, faults in the devices that are used for transport and protection of the electrical grid or even human and computer-controlling error. The blackouts are very difficult to recover from, because the grid design does not allow easily switching on after a blackout. Cascade or domino effect as a result of blackout at even a small part of the power grid is already proven in practice [3] and happens very often. Namely, when a certain part of the network has a fault, then it will trip more and more failures in the network, if some of its components cannot handle the additional current [2] which leads to cascade of shutting down major parts of the grid. Therefore, this will reflect as cascade effect on different society segments as well and will have big negative impact.

Cyberspace as a domain where a war can be fought in the later years has proven to be a very interesting and developing concept. Until now "old" war techniques were in place where it was most certain that a fire weapon needs to be used to insert damage to the opponent. In the later years computer and internet based technology have come to a huge advancement and the use of them in war related purposes is ever increasing [4]. Electrical power system has always been high priority target for militaries. Cyber-attack is new addition to possible

attacks that offers attacks with a low cost and a long range. In addition to this, there is evidence that some entities have probed computer network of certain power grids and have done reconnaissance necessary for cyber-attacks [3].

After several cyber-attacks have been conducted in the past years, such attacks on electrical power grids have gotten attention. In this paper, brief description of SCADA (Supervisory Control and Data Acquisition) systems and their usage in electrical transmission and distribution systems will be presented. The presentation will first contain explanation of known attacks on SCADA systems - the attack on the Iranian enrichment plant in Natanz and Ukrainian power grid attack. Each attack will be explained in order to point out the capabilities of the attackers and to notify the shortages that allowed these attacks to happen. Then vulnerability of SCADA, field devices and VPNs as components for control and communication will be presented. We will also present the findings from conducted Internet research regarding publically available information for power grids in Balkan. And on the final section there will be a conclusion and recommendations for protection of the power grids from further attacks based on critical analysis of the covered literature and on the findings.

II. Cyber Attacks on SCADA Systems

A. *SCADA Systems*

According to [5], "Industrial control system (ICS) is a general term that encompasses several types of control systems, including supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and other control system configurations such as Programmable Logic Controllers (PLC) often found in the industrial sectors and critical infrastructures". SCADA systems are category of computer programs that analyze and display conditions of the process. SCADA systems interfere with different actuator devices (pumps, motors, valves, etc.) by using industrial controllers [6]. Typical hardware includes a control server which is usually placed at a control center, different communications equipment (e.g., radio, telephone line, cable, satellite, etc.), and one or more distributed field sites consisting of Remote Terminal Units (RTUs) and PLCs, which controls actuators and monitors sensors [5]. Namely, control server processes the information from RTU, whereas the RTU/PLC controls the local process. SCADA systems are designed to display collected information to operator in a graphical or textual form in real time. Such systems are used in water distribution systems, wastewater collection systems, oil and natural gas pipelines, electrical transmission and distribution systems, and in public transportation system [5].

B. *Natanz Case and The Stuxnet*

Not so long ago there was an example of cyber-attack that was targeting SCADA systems by using sophisticated worm known as Stuxnet. Earliest Stuxnet samples were seen in 2009 [7]. There are only assumptions about who conducted the attack and who created the worm [8]. The goal of the worm was to damage centrifuges for enriching uranium. Therefore, Stuxnet was designed to attack specific Siemens SCADA systems by performing several operations [8]. Fig. 2 presents summary of Stuxnet operation.

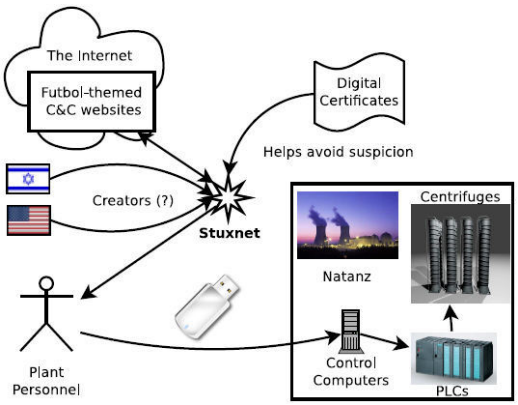


Fig. 2. Summary of Stuxnet operation [8]

According to [8] the worm exploits zero-day vulnerabilities, it modifies system libraries, attacks Siemens SCADA control software known as Step7 and it can automatically update old version of itself to newer version that is available on the local network. In order to update itself, the worm runs RPC (remote Procedure Call) server and communicates with so called command and control servers (www.mypremierfutbol.com and www.todaysfutbol.com) in encrypted form (Fig. 2). Moreover, this way the worm provides information about its spread and computers it has infected. In order to perform its final task – reach PLCs, the worm installs stolen signed drivers on computers running Windows OS. Stuxnet used sophisticated techniques known as rootkits in order to hide itself from users and from anti-malware software.

As it is specified in [9, 7, 8] the Stuxnet uses six methods in order to spread itself. First method of spreading is by using USB flash drives. This is shown in Fig. 2. In [3] it is assumed that the infected flash drive was inserted in the Natanz's control computers via outside contractors and that is how infection started. The second method of spreading is by attacking WinCC database using SQL command. This helps the worm to upload itself to a remote machine and start its own copy. WinCC is Siemens interface to Siemens SCADA systems. The third method is by using network shares, i.e. Windows shared folders as a way to propagate over a local network. The worm can copy itself on a remote computer by using MS10-061 print spooler zero-day vulnerability as well and then it executes the copy of itself in order to infect the remote machine. In order to propagate and execute on a remote machine, it can also use MS08-067 vulnerability of the SMB protocol. This protocol is used for sharing files and other resources between computers. Lastly, the Stuxnet can propagate by infecting Step7 projects that are open on the infected computer. By performing all of these sophisticated operations, the malware forces WinCC monitor program to show normal operation results on the operators' screen whereas in reality it modifies PLCs operation (Fig. 2) in order to cause damage on centrifuges [8].

With its sophisticate operation, the Stuxnet has increased awareness of the ICSs vulnerability but it also increased the

possibility that similar attacks will be performed by the attackers in the future [8].

C. Cyber Attack on The Power Grid in Ukraina

One of the first publicly acknowledged attacks in the field of cyberwarfare on the electrical grids was the attack on the Ukrainian Kyivoblenergo electrical distribution company [10]. This attack occurred on December 23, 2015 in the Ivano-Frankivsk area in western Ukraine and left around 30 substations without power which supplied electricity to around 230.000 people, for a period up to six hours. The attacks were conducted on a three different voltage distribution levels, which means that the attackers had a vast knowledge of the system, the way it works and all of the devices that are used in the system for transmission of electricity. Russian security services were accused of implementing this attack, but that was never proven to be true.

Attackers have demonstrated a lot of techniques and capabilities (Fig. 3) in order to steal vital information of the company's system and conduct the attack [10]. As can be seen in Fig. 3 the first step in gathering information was the usage of spear phishing e-mails that contained malicious software (Word Documents and Excel spreadsheets) that when opened dropped variants of Black Energy3 malware on the computers [11, 10]. In [10] Black Energy is described as attack tool that is used to conduct DoS attack, create botnets and steal bank credentials.

Fig. 3. Ukraine attack techniques and capabilities [7]

In Ukraine case, Black Energy3 was used to steal legitimate user credentials [11, 10]. It is assumed that attackers remained in the network environment for about six months prior the attack. At this point of the attack the only task that attackers were doing was harvesting and downloading information about the infested systems. They managed to steal administrator credentials and create their own users to whom gave full access privileges. As it is shown in Fig. 3, with the stolen credentials the attackers were able to infiltrate trough virtual private networks (VPNs) as fully authorized users. As it is stated in [10], VPN connections between Ukraine power companies' ICS and enterprise networks did not use two-factor authentication and there was no monitoring of the network traffic. Therefore, attackers were able to issue commands directly from a remote workstation to all of the substations of the affected company, to manipulate the SCADA systems and to operate the circuit breakers, disconnectors, and other components in the substations.

The attackers demonstrated capability and power to target even hardware operational devices at the substations. Namely, they developed and managed to upload malicious firmware in to serial-to-ethernet gateways. With this they were able to secretly take over control on the diverse devices, make them inoperable and unrecoverable such that operators of the electrical company were not able to send remote commands to the substations [10]. This way the recovery from that attack was delayed. Important software that attackers have used to hide their tracks was a modified version of the KillDisk program. This software tool was used to delete master boot record of the attacked system and particular logs [11, 10]. The attackers managed to damage all of the backup systems such as UPSs, and generators by disabling them in the time frame of the attack. Telephone distributed denial-of-service attack (DDoS) was implemented (Fig. 3) on the call center of the electrical company during the time of the attack and therefore users were not able to report any of the power outages [11, 10].

The attack on the power grid in Ukraine showed vulnerabilities of the SCADA systems used in electrical systems and it is considered as escalation from some past destructive attacks that impacted general purpose computers and servers. On the other hand, the attack methodology, tactics, techniques and procedures are employable in infrastructures around the world [10] and therefore such attack could happen to anyone [12]. In order to lower costs as economic reason, many industries are using public networks (Internet) and commercial software for remote access to control systems. This approach is increasing systems' vulnerability and therefore they become easy to attack [3]. Moreover, it is believed that a lot of information about the devices used in substations and control centers (circuit breakers, type of transformers, protection relays, SCADA system version and RTUs) was publicly released and could have been accessed by any individual with internet access [10].

The experts in [10] suggests several defense measures as response to attack in Ukraine: application whitelisting, using YARA forensic tool for searching BlackEnergy3 infections, controlling access to information about hosts, devices, distribution substations, etc., two-factor authentication for VPNs, disabling remote access at the hosts and at the perimeter firewall, required authentication from the operator, disabling remote control for unnecessary devices, etc. However, they also predict next possible attacks. In addition to this, [13] states that attacks only become better and improve with time.

D. Vulnerability of SCADA and Field Devices

Industrial operations use SACDA systems in order to control remote equipment and to collect data about equipment performances [14]. As it is explained in [15] there are many examples that show vulnerability of SCADA systems, such as Ohio power plant and breaking down part of Austrian and German power grid. SCADA attacks target power plants, factories and refineries and therefore they tend to be political in

nature [14]. Attacks on SCADA systems are rising [15, 14] and it is possible that many of them have gone undetected. Moreover, in many cases attacker only monitor, observe activity and wait for months and years before taking the action [15]. As SCADA attacks increased from 91.676 in January 2012 to 163.228 in January 2013, and 675.186 in January 2014, more SCADA attacks should be expected in the future 14]. Moreover, in [14], 16 different known key attack methods on SCADA systems are pointed out.

Despite of knowing this, a lot engineers and SCADA technicians are ignoring this fact, until it becomes a threat which can deploy a great amount of damage to a certain system. Main and most common threats and vulnerabilities in the SCADA systems arise from [15]: lack of monitoring in the system, slow updates for the systems, lack of knowledge about devices (problem of pairing a five with twenty year old technology to work together has become a problem, and knowledge about devices and infrastructure is often incomplete which leaves with a lot of devices not to be set-up and work properly) and authentication holes.

Field devices (Programmable Logic Controllers - PLC, Programmable Automation Controllers - PAC, Remote Terminal Units - RTU and Intelligent Electronic Devices - IED) are components of control systems that communicate with sensors and actuators in order to monitor and control a process [16]. The vulnerability of ICS devices has its roots decade ago. U.S. Department of Homeland Security identified vulnerability in ICSs back in 2007 dubbed Boreas [17]. This vulnerability allows permanent disabling controllers by simply loading manipulated firmware [18]. Field devices are designed for closed, trusted networks with little thought for security issues. In [16] it is explained that if attacker is capable to ping such device, then he/she would be able to send any command for reading, writing, diagnostic or configuration, due to the lack of source/data authentication. It also explains attacks and vulnerabilities found on filed device Ethernet cards including loading malicious firmware on it. Once exploited, such cards can be used to attack other cards on the field device and other devices on the control system LAN and WAN.

E. VPN Vulnerability

VPNs allow security and privacy to networks and they enable data transition across shared private/public networks. However, there are known vulnerabilities among VPNs which persist even today. VPNs can use PPTP (Point-to-Point Tunneling Protocol). PPTP achieve authenticity and confidentiality by applying MS-CHAPv2 (Microsoft Challenge Handshake Authentication Protocol) and MPPE (Microsoft Point-To-Point Encryption) accordingly. Vulnerabilities of PPTP are analyzed and presented in [19] where some critical flaws of MS-CHAPv2 are given. It is also stated that the authentication and encryption of PPTP is only as secure as user password and that authentication can be captured and broken by dictionary attack. Moreover, ChapCrack tool for parsing and decrypting MS-CHAPv2 network handshakes can be used do drop PPTP encryption in an easy way [20].

On the other hand, IPSec (Internet Protocol Security) as protocol for insuring confidentiality, integrity and authenticity can also be used in VPNs. IKE (Internet Key Exchange) is a sub-protocol used in IPSec that is responsible for peer authentication. This sub-protocol supports two modes: Main mode and so called Aggressive mode. The later allows shortened authentication process and therefore introduces critical security flaw. Some vendors are shipping VPNs with Aggressive mode set as default [21].

Among latest attacks that threatened infiltration in organization's intranet servers, network devices, client machines and VPN servers in order to steal valuable data was Heartbleed. Heartbleed is implementation flaw in OpenSSL version 1.0.1 and 1.0.2. beta. It leaks contents of the memory from the server to client and from client to server, which can expose sensitive data such as passwords and private key of SSL server [22]. According to [23], Heartbleed allows attackers to bypass two-factor authentication system on organization's VPNs.

III. Open Source Information

We conducted a research on Internet in order to look for available information regarding power grids of Balkan countries. We found that there was available information for 40% of the countries.

Among other, we were able to find schematics and diagrams for electrical transmission and distribution systems structures from various sources. One such example is shown on Fig. 4.

Fig. 4. Example of electrical transmission and distribution systems structures available on Internet [adopted from 24]

Moreover, we were able to find different information for computer networks that connects SCADA systems. One example is shown on Fig. 5.

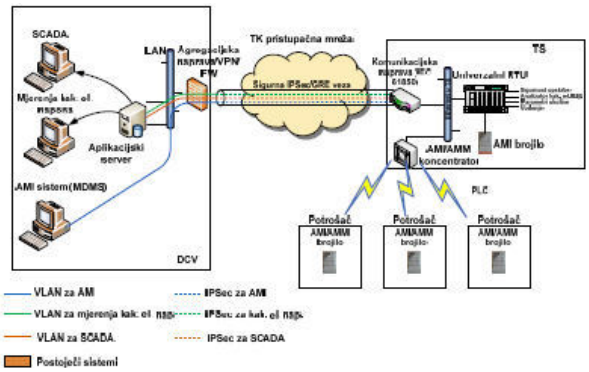


Fig. 5. Example of computer network that connects SCADA system available on Internet [25]

One other important finding was that it was very easy to gather even detail specifications regarding equipment used for controlling and monitoring of power grids by using published documents for biddings.

IV. Discussion

This paper presents vulnerability of power grids trough pointing out vulnerability of different SCADA components that are used for controlling and monitoring. Production and distribution of electricity is very important. Stable and reliable electricity runs industries, different services and our home appliances. Major power outage can cause cascade effect on the power grids and can leave a large urban area without electricity. On the other hand this can cause massive social disruption and negative cascade effect in the society. Power grids were always been important military targets. Nowadays, cyberspace has proven to be a very interesting and developing concept for conducting low cost and low range attacks on power grids by giving a chance to anyone to cause damage and live no provable evidence about itself, as it was case both with power grid attack in Ukraine and the Natanz case.

One proven fact that affects vulnerabilities on control systems is usage of public networks (Internet) and commercial software to provide remote access. Therefore, they become easy to attack. Namely, attack on Ukrainian power grid started by sending phishing e-mail over Internet in order to deploy BlackEnergy3 malware. Moreover, the attackers were able to remain in the environment undetected for at least six months and harvest credentials, which were used to access VPN. Among other things VPN had no two factor authentication and there was no traffic monitoring. In this attack it is proven that whole security network failed – was overtaken by the attackers and they were able to control the outage for a couple of hours.

There are known issues regarding VPNs that are using PPTP and IPSec. Moreover, latest Heartbleed attack allows infiltration in VPNs and even bypassing of two-factor authentication which was recommended from experts in order to mitigate attacks such the one that happened in Ukraine power grid. Although we are using VPNs to secure our networks, above mentioned security issues indicate that communication channels used in electrical transmission and distribution systems might not be secure.

Ukraine attack showed vulnerabilities of the SCADA systems used in electrical systems and it is considered as escalation from some past attacks. Namely, among others, attacker used vulnerability of field devices and managed to upload malicious malware. Attacks on such devices like Boreas and similar newer that relies on it are known for longer time now. Moreover, newer attacks on field devices allow attackers even more capabilities. However, it appears that users of such devices are not upgrading firmware on regular basics and are not following security recommendations. There are many examples that show vulnerability of SCADA systems, whereas researches are showing that attack methods on SCADA systems are fast growing, which becomes big consideration. Important finding is that attacks on SCADA systems have political nature and on the other hand there is assumption that many of these attacks have been undetected. This indicates that there are attack methods on SCADA systems that are still unknown for security experts. Since attacks evolve from each other and cannot be predicted, it can be concluded that in near future there will be new and more powerful attacks that will exploit yet undetected VPN and SCADA vulnerabilities.

In order to attack SCADA systems at Natanz, the worm was delivered to the system by using removable media and it used six methods in order to spread and copy itself to other machines. Moreover, hiding its existence from users and from anti-malware software was accomplished by using sophisticated rootkits techniques. In order to attack specific Siemens SCADA systems the worm exploits zero-day vulnerabilities, modifies system libraries, attacks Siemens SCADA control software known as Step7 and even installs stolen signed drivers on computers running Windows OS. This proves that hackers are capable of conducting very sophisticated attacks by using only one malware tool that is capable to perform several operations.

Regarding attack on Ukrainian power grid it was mentioned that a lot of information about the devices used in substations and control centers was publicly released. It was interesting to see how such information was available on Internet for Balkan countries. It was found that schematics for electrical power grid structures and different information about computer networks that connects SCADA systems were available for 40% of the countries. On the other hand, detail specifications about controlling and monitoring equipment were available by following published documents for biddings. Such publically available information can be used as starting point for conducting an attack.

V. Conclusion

It appears that attacks will grow and attackers will be inspired to exploit new vulnerabilities. Having in consideration all mentioned issues from above, we recommend building own infrastructure for communication networks and avoiding public infrastructures. This will avoid attacking from Internet and accessing electrical transmission and distribution systems from outsiders. Updating firmware on used components and equipment on regular basis, as well as last available patches would be necessary in order to avoid known vulnerabilities. On the other hand, manufacturers must implement security patches as soon as possible and consider field devices usage in open and vulnerable networks. Although above was shown that malware can hide from anti-malware software, we recommend network traffic monitoring and implementation of auditing tools. This will help in gathering knowledge about unusual network traffic and unusual behavior. All commands towards control devices must be authenticated in order to avoid unauthorized issuing of commands, whereas protocol implementation should follow all latest findings regarding security issues. Using removable storage media must be minimized and only at dedicated points. It has to be under strong supervision of IT and Information Security personnel. As addition, personnel should be trained to respect security policies in order to avoid social engineering and prevent insider attacks. Since public information about electrical production, transmission and distribution systems and their components can be exploited by attackers, such information must be treated

as confidential. Lastly, all communication paths (cables, fiber optic, and wireless) between control center, remote facilities and units must be encrypted by using very strong hardware encryption. As it was mentioned previously, electrical power systems present valuable military target. But, there is no conventional military defense in the cyber space. Therefore, in order to protect our electrical systems from cyber-attacks, communications that control these systems must be protected with military grade encryption. This means that electricity companies should abandon "cheapest way" of doing things, due to the economic reasons. The above mentioned security measures are just few among many, but we believe they will help in providing stronger security of our power grids.

REFERENCES

[1] Physics World, "Power production: Electricity generation and transmission in Hong Kong", Physics World, 2010, available at: http://www.hk-phy.org/energy/power/print/elect_phy_print_e.html, [last accessed on March 02, 2017].

[2] P. Hines, K. Balasubramaniam and E.C. Sanchez, "Cascading failures in power grids", IEEE, 2009, available at: https://www.uvm.edu/~phines/publications/2009/hines_2009_potentials.pdf, [last accessed on October 27, 2016].

[3] J.A. Lewis, "The electrical grid as a target for cyber attack", Center for Strategic and International Studies, 2010, available at: https://pdfs.semanticscholar.org/36f2/1b2deb3eb35bf6043cb9b743fc031695f9a3.pdf, [last accessed on October 15, 2016].

[4] J. Carr, Inside Cyber Warfare, 2nd Edition. NY, O'Reilly Media Inc, 2011.

[5] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams and A. Hahn, "Guide to industrial control systems (ICS) security", NIST Special Publication 800-82 Revision 2, 2015, available at: nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r2.pdf, [last accessed on December 8, 2016].

[6] R. Langner, "To kill a centrifuge: A technical analysis of what Stuxnet's creators tried to achieve", The Langner Group, Arlington, 2013, available at: www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf, [last accessed on December 18, 2016].

[7] N. Felliere, L.O. Murchu and E. Chien, "W32.Stuxnet dossier", Symentec Security Response, Version 1.4, February 2011, available at: https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/security-response-w32-stuxnet-dossier-11-en.pdf, [last accessed on December 23, 2016].

[8] P. Mueller and B. Yadegari, "The Stuxnet worm", CSc 466-566 Computer Security 2012, Resources, 2012, available at: http://www2.cs.arizona.edu/~collberg/Teaching/466-566/2012/Resources/presentations/2012/topic9-final/report.pdf, [last accessed on December 19, 2016].

[9] A. Matrosov, E. Rodinov, D.Harley and J. Malcho, "Stuxnet under the microscope", Revision 1.1 ESET, 2011, available at: http://download.esetnod32.ru/company/virlab/analytics/Stuxnet_Under_the_Microscope.pdf, [last accesed on December 20, 2016].

[10] E-ISAC, "Analysis of the cyber attack on the Ukrainian power grid", Electricity Information Sharing and Analysis Center, E-ISAC, 2016, available at https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf, [last accessed on November 2, 2016].

[11] Antiy Labs, "Comprehensive analysis report on ukraine power system attacks", ANTIY, 2016, available at http://www.antiy.net/p/comprehensive-analysis-report-on-ukraine-power-system-attacks/, [last accessed on November 3, 2016].

[12] K.J. Higgins, "Lessons from the ukraine electric grid hack", DarkReading, 2016, available at: http://www.darkreading.com/vulnerabilities---threats/lessons-from-the-ukraine-electric-grid-hack/d/d-id/1324743, [last accessed on November 2, 2016].

[13] K. Paterson, "TLS security – where do we stand" University Of Cambridge, Security Seminars, 2013, available at: http://www.cl.cam.ac.uk/research/security/seminars/archive/slides/2013-10-15.pdf, [last accessed on December 13, 2016].

[14] Dell, "Dell security annual threat report 2015", Dell Inc., 2015, available at: https://software.dell.com/docs/2015-dell-security-annual-threat-report-white-paper-15657.pdf, [last accessed on November 15, 2016].

[15] D. Adams, "Common SCADA system threats and vulnerabilities", Patriot Tech, 2016, available at: http://patriot-tech.com/blog/2015/10/27/common-scada-system-threats-and-vulnerabilities/, [last accessed on November 3, 2016].

[16] D. Peck, "Leveraging ethernet card vulnerabilities in field devices", 4th Annual SCADA Security Scientific Symposium - S4, 2009, available at: https://www.digitalbond.com/wp-content/uploads/2011/05/1_PLC_final.pdf, [last accessed on December 8, 2016].

[17] R. Langner, Robust control system networks. How to achieve reliable control after Stuxnet. New York, Momentum Press, 2012.

[18] R. Langner and P. Pederson, "Bound to fail: Why cyber security risk cannot simply be managed away", Center For 21st Century Security and Intelligence, Cyber Security Series, 2013, available at: www.langner.com/en/wp-content/uploads/2013/06/Bound-to-fail.pdf, [last accessed on January 9, 2017].

[19] B. Schneier, Mudge, "Cryptanalysis of Microsoft's PPTP authentication extensions (MS-CHAPv2)", CQRE '99, Springer-Verlag, 1999, pp. 192-203.

[20] M. Marlinspike, "Chapcrack", 2012, available at: https://github.com/moxie0/chapcrack, [last accessed on September 6, 2016].

[21] N. Schiess, "Vulnerabilities & attack vectors of VPNs (Pt 1)", INSINUATOR, 2013, available at: https://insinuator.net/2013/08/vulnerabilities-attack-vectors-of-vpns-pt-1/, [last accessed on September 3, 2016].

[22] K. Jackson, "Heartbleed's intranet & VPN connection", DARKReading, 2014, available at: http://www.darkreading.com/analytics/heartbleeds-intranet-and-vpn-connection/d/d-id/1204457, [last accessed on October 2, 2016].

[23] M.J. Schwartz, "Heartbleed's attack targeted enterprise VPN", DARKReading, 2014, available at: http://www.darkreading.com/attacks-breaches/heartbleed-attack-targeted-enterprise-vpn-/d/d-id/1204592, [last accessed on October 2, 2016].

[24] R. Goić, D. Jakus, I. Penović, "Distribution of electrical energy" (in Croatian), Fakultet elektrotehnike, strojarstva i brodogradnje Split, 2008, available at http://marjan.fesb.hr/~rgoic/dm/skriptaDM.pdf, [last accessed on January 10, 2017].

[25] V. Lovrenčić, S. Rapoša, P. Ceferin, S. Ceferinn and M. Dečman, "Conceptual design of a pilot project of transformation station in Maribor as support for advanced network concept" (in Croatian), Hrvatski ogranak međunarodne elektrodistribucijske konferencije, 3. (9.) savetovanje, SO6-11, 2012, available at: http://www.ho-cired.hr/3savjetovanje/SO6-11.pdf, [last accessed on January 10, 2017].

VPN server versus Proxy server privacy

Slavcho Andreevski, Adrijan Bozhinovski, Biljana Stojchevska
School of Computer Science and Information Technology
University American College Skopje, Macedonia
sandreevski @outlook.com, {bozinovski, stojcevska}@uacs.edu.mk

***Abstract*—Virtual Private Networks and Proxy Servers are being widely used by Internet users to secure their own privacy. Today we have millions of proxy and VPN servers which can be which can be accessed free of charge or may require access fees. Both connect the user to a remote computer, although there are many differences between them that are going to be explained in the following paper. Also, this study is going to include a comparison between two VPN servers and two proxy servers, along with their advantages and disadvantages. The VPN servers are: Private Internet Access and PrivateVPN, which are purchased from their websites and the proxies' website is sockslist.net, which is free. Mask-my-ip is another proxy, purchased from their website.**

Keywords—Virtual Private Network; Proxy; Server; Privacy; Speed; Ping;

I. Introduction

Everyone who uses the Internet wants their privacy to be at a top level. To achieve that, some programs/applications that provide this kind of privacy must be used. To test if this principle really secures us 100%, this is what will be tested and explained in details in this paper. Two ways of implementing security are the uses of a proxy and a VPN server.

A proxy server is a server that acts as an intermediate, so the Internet activities would appear as if coming from somewhere else. It helps the user to bypass local censorship and restrictions made by a local network proxy. These proxies must be configured to help hide the users' IP address [1].

There are two types of proxies based on communication direction: Open Proxy and Reverse proxy.

Open proxy – this is a forwarding proxy which is an interconnection between a client and a server. When the client sends a request, the request is sent to the proxy. Then the proxy requests the content from the server and returns it to the client. This type of proxy also needs configuration.

Reverse proxy – this type of proxy does not need configuration because it acts like an ordinary server. The client makes a request in the name of the proxy and then the reverse proxy decides where to send that request and return the content as if being sent by the original server.

Proxies can also be differentiated, based on the protocol they use, to HTTP and SOCKS proxies.

HTTP proxies – this type of proxy interprets the traffic at an HTTP level, for example http:// or htpps:// traffic. If the web browser is configured with a proxy, the whole traffic will be routed through the remote proxy.

SOCKS proxies – This type of proxy is different than the http one. It does not only route the web traffic through the remote proxy, but it also handles all the traffic that passes through, like web server, torrent clients etc.[2]

Another way of providing privacy is a virtual private network (VPN) which extends a private network and uses public network, e.g. the Internet, to connect to sites or users. It is a technology which can be applied to LAN and WLAN. A VPN keeps the user privacy though security procedures and tunnels. The data which is sent is encrypted and then forwarded though a VPN tunnel. The level of security can also be upgraded by encrypting the network addresses [3].

There are two types of VPNs:

Remote Access VPN – This type allows the users to establish a secure connection and connect to a private network to gain access to their data and resources. Private users of VPN primarily use these VPN services to bypass some restrictions on the Internet or particular websites.

Site-to-Site VPN – Also called router-to-router VPN, which is widely used by companies and corporations that are placed at different geographical locations. This network connection is called extranet. But, if more offices from one company are connected via Site-to-Site VPN, this connection is called intranet [4].

II. Privacy Level

In this research, a battery of tests was carried out in order to test the privacy level of the proxies and the VPN servers. These tests are: rDNS, WIMIA, Tor, Loc, Header and DNSBL test.

rDNS or Reverse DNS lookup is a reverse test of the forward DNS lookup which requires the DNS to check if the domain name is associated with an IP address. This test uses PTR records (pointer to a canonical name). The database of the reverse DNS is rooted in the ARPA top level domain [5].

WIMIA is a propitiatory Application program interface that whatismyipaddress.com uses to see if the IP is used by more than one person, indicating that the IP is indeed a possible proxy sharing device. Because of lack of information, it is supposed that WIMIA does not unmask the real IP [6].

The **Tor** test checks if a specific IP address is a Tor exit node. In other words, it lets the user know if some connection is coming from the Tor network [7].

Loc is a test which checks the geolocation providers. IP geolocation is a mapping of an IP to the real geographic location of an Internet connected to some device. This works

with databases which are commercially available and the accuracy may vary depending on which database is being used [8].

The **Header** test may work in two modes: basic and advanced. The basic mode analyses the info which the browser sends when it makes a web request. The "User-Agent" and the "Referrer" headers of the requesting IP address are saved in the web server's log. The advanced mode carries out a more active analysis, but the results are purely informational and are up for further interpretation [9].

DNSBL or Domain System Blacklists checks if the proxy is an open proxy which means that the proxy is accessible by any Internet User. This test includes blocking lists and its first purpose was to block spam mails and educate people about spam [10].

III. Test Results

These six tests were used on around 80 proxies and 80 VPN servers belonging to four different VPN and proxy providers: Private Internet Access, PrivateVPN, Socketlist.net and Mask My IP. Two random VPN servers were chosen (one detected by one or more of the tests and the other not detected) and 255 IP addresses were tested with the previous mentioned tests.

The results are shown in the following tables and graphs:

Private Internet Access (application bought from their web site - privateinternetaccess.com) – contains 39 servers, but 2 of them were not working.

Table 1: Tests performed in order to detect a proxy or vpn

Tests	Detected	Not Detected
rDNS	0	37
WIMIA	37	0
Tor	0	37
Loc	0	37
Header	0	37
DNSBL	0	37

Table 2: VPN server speed test (pinging www.google.com on every one of 37 VPN servers used in Table 1)

Average RTT	310 ms
Minimum RTT	100 ms
Maximum RTT	520 ms

Proxy/VPN or network sharing device was detected on every 37 servers just by the WIMIA test. All the other tests got negative results.

PrivateVPN – contains 40 servers, 2 of them were not working

Table 3: Tests performed in order to detect a proxy or vpn

Tests	Detected	Not Detected
rDNS	0	38
WIMIA	12	26
Tor	0	38
Loc	0	38
Header	0	38
DNSBL	0	38

Table 4: VPN server speed test (pinging www.google.com on every one of the 38 VPN servers used in Table 3)

Average RTT	284 ms
Minimum RTT	50 ms
Maximum RTT	518 ms

Proxy/VPN or network sharing device was detected on just 12 servers with the WIMIA test. All the other tests got negative results.

First Random VPN server (1 server from privateVPN with 255 ip addresses) – X.X.X.X (X.X.X.1-X.X.X.255) – one VPN server's IP was not detected by the tests. Thorough analyses were made by examining all other 244 IP addresses:

Table 5: Tests performed in order to detect a proxy or vpn

Tests	Detected	Not Detected
rDNS	0	255
WIMIA	0	255
Tor	0	255
Loc	0	255
Header	0	255
DNSBL	0	255

Table 6: VPN server speed test (pinging www.google.com on every one of the 255 IPs used in Table 5)

Average RTT	253 ms
Minimum RTT	69 ms
Maximum RTT	437 ms

The random server was tested on one IP address and was not detected by the tests. Then all 244 other IPs were tested and it showed that none of them were detected by the 6 tests.

Second Random VPN server (1 server from Private Internet Access with 255 ip addresses)

– X.X.X.X (X.X.X.1-X.X.X.255) – one VPN server IP was detected by one or more of the tests. Thorough analyses were made by examining all other 244 IP addresses:

Table 7: Tests performed in order to detect a proxy or vpn

Tests	Detected	Not Detected
rDNS	0	255
WIMIA	255	0
Tor	0	255
Loc	0	255
Header	0	255
DNSBL	0	255

Table 8: VPN server speed test (pinging www.google.com on every one of the 255 IPs used in Table 7)

Average RTT	253,5 ms
Minimum RTT	77 ms
Maximum RTT	430 ms

Initially, the other random server was tested on one IP address and was detected just by the WIMIA test. Then all 244 other IPs were tested and showed that all of them were detected by the WIMIA test.

Socketlist.net – contains more than 100 proxies – 72 were tested – 40 working, 32 not working

Table 9: Tests performed in order to detect a proxy or vpn

Tests	Detected	Not Detected
rDNS	0	40
WIMIA	40	0
Tor	0	40
Loc	0	40
Header	40	0
DNSBL	0	40

Table 10: Proxy server speed test (pinging www.google.com on every working proxy used in table 9)

Average RTT	367 ms
Minimum RTT	97 ms
Maximum RTT	637 ms

Two of six tests detected a proxy/VPN on every working proxy server. The tests are WIMIA and Header tests, which show that that server was surely a proxy/VPN.

Mask my IP – contains around 80 proxies – 40 working and 40 not working

Table 11: Tests performed in order to detect a proxy or vpn

Tests	Detected	Not Detected
rDNS	0	40
WIMIA	39	1
Tor	0	40
Loc	0	40
Header	38	2
DNSBL	0	40

Table 12: Proxy server speed test (pinging www.google.com on every working proxy used in Table 11)

Average RTT	364,5 ms
Minimum RTT	86 ms
Maximum RTT	643 ms

WIMIA test detected a proxy/VPN on 39 out of 40 working proxy servers. Also header test detected a proxy/VPN on 38 out of 40. Just 1 proxy server was not detected by the tests.

IV. Comparisons Among the 2 VPNs and the 2 Proxies

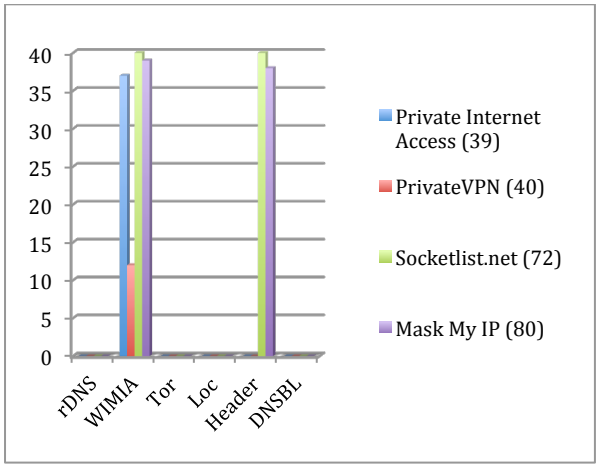


Fig. 1. Test detection comparison among the 2 VPNs and the 2 proxies used in this paper

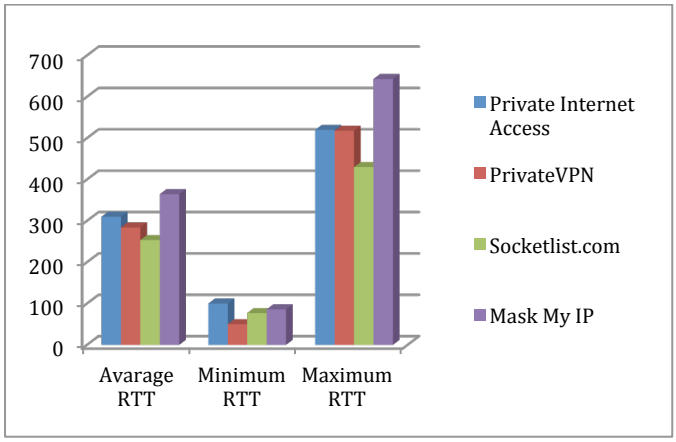


Fig. 2. Ping comparison among the 2 VPNs and the 2 proxies used in this paper

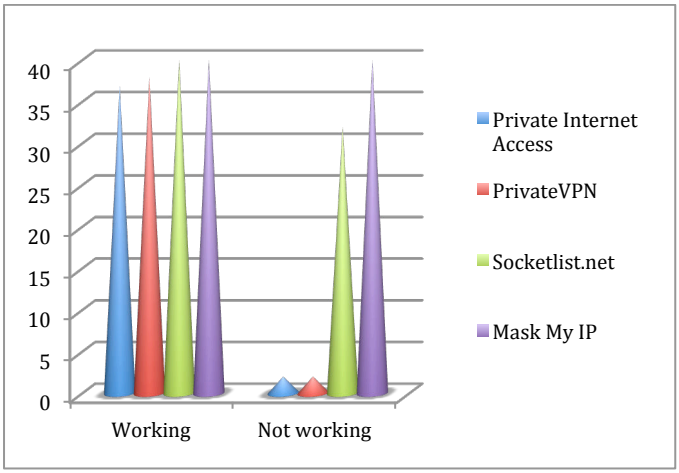


Fig. 3. Comparison of working and non-working servers among the 2 VPNs and the 2 proxies used in this paper

V. Conclusion

To provide security on the Internet, the users must use some software like VPN or Proxy server. But, just using this kind of security measures does not guarantee that their privacy is on the top level. To support this, six tests were made which make some analysis and show if the user is behind a proxy/VPN or not. Those tests are rDNS, WIMIA, Tor, Loc, Header and DNSBL test. These tests were made on every working VPN and proxy server and, as shown, the WIMIA test is the most accurate test which showed that most of the servers were VPNs and proxies. While the VPNs were just detected by the WIMIA tests, the proxies were also detected by the Header test.

Because of a lack of information about the WIMIA test, it is presumed that this test does not reveal the original public address. It just shows if the user is behind a proxy/VPN. Also it is assumed that if one VPN or proxy that is not recognized by the tests is being actively used by many users, after a limited time, it will fail the WIMIA test. To test this assumption, further analysis should be made.

The speed (packet transmission) was another way to see if a VPN or proxy is better. In the results, it could be seen that the two VPNs have better results in the packet transmission. The average VPN ping RTT is smaller than the proxies ping RTT.

Overall, these two parameters do not prove if one is better than the other. If a user wants to be sure whether s/he is 100% private, at least these six tests should be performed. If the VPN or the proxy fail one or more of the tests, the privacy has been compromised and the user cannot be sure if her/his privacy is maintained or not. Also, if one VPN server is not detected to be a VPN by the tests on one IP address, it will not be detected on the 254 other IP addresses either.

As a final remark, we can say that if a user wants her/his privacy at top level, just using a proxy or a VPN is not the solution. The VPN or proxy which is going to be used first needs to be analyzed with the six tests presented in the paper, and then the user can proceed with other work. Also from four tested VPNs and proxies, the PrivateVPN is the best for using because just this one had VPN servers that were not detected by the aforementioned tests.

References

[1] I. Cooper, I. Melve, and G. Tomlinson, “Internet Web Replication and Caching Taxonomy,” Internet Engineering Task Force RFC 3040, January 2001, www.ietf.org/rfc/rfc3040.txt, last accessed on March 3, 2017

[2] J. Fitzpatrick, “What’s the Difference Between a VPN and a Proxy?”, 2016, acessible at: https://www.howtogeek.com/247190/whats-the-difference-between-a-vpn-and-a-proxy/, last accessed on March 3, 2017

[3] Microsoft, “Virtual Private Networking: An Overview”, 2001, accessible at: https://technet.microsoft.com/en-us/library/bb742566.aspx, last accessed on March 3, 2017

[4] admin@kryptotel.net, “Types of VPN and Types of VPN Protocols”, 2016, acessible at: https://www.vpnoneclick.com/types-of-vpn-and-types-of-vpn-protocols/, last accessed on March 3, 2017

[5] AOL, “Web Tools”, accessible at: https://www.dmoz.org/Computers/Internet/Protocols/DNS/Web_Tools, last accessed on March 3, 2017

[6] WhatIsMyAddress.com, Advanced Proxy Check, 2011, accessible at: http://whatismyipaddress.com/proxy-check, last accessed on March 3, 2017

[7] NPM, “tor-test”, 2016, accessible at: https://www.npmjs.com/package/tor-test, last accessed on March 3, 2017

[8] BrandMedia Inc.,Where is Geolocation of an IP Address?, 2017, accessible at: https://www.iplocation.net/, last accessed on March 3, 2017

[9] E. Fulkerson, “Display your proxy server information”, 2017, accessible at: http://www.whatismyproxy.com/, last accessed on March 3, 2017

[10] CGP Holdings Inc., “What is a DNSBL?”, 2017, accessible at: http://www.dnsbl.info/, last accessed on March 3, 2017

Optimal Parallel Wavelet ECG Signal Processing

Ervin Domazet
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
1000, Skopje, Macedonia
Email: ervin_domazet@hotmail.com

Marjan Gusev
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
1000, Skopje, Macedonia
Email: marjan.gushev@finki.ukim.mk

***Abstract*—Real time detection of heart abnormalities can prevent serious health problems. This requires real time processing of ECG data by a corresponding web service. Considering the case of wearable devices to collect ECG data, the signal is actually contaminated by noise. Noise can seriously change the ECG signal and occur in the form of a baseline drift representing various physical movements and breathing. Unless it is removed, correct analysis on ECG data is impossible. Being characterized by very low frequencies, its elimination can not be efficiently realized by simple DSP filters, such as Finite Response Filters (FIR) or Infinite Response Filters (IIR).**

Wavelet Transformation is a promising technique to eliminate the noise with very low frequencies, and its digital version (DWT) is capable of efficient removing the ECG baseline drift. In this paper, we set a research question to investigate the dependence between the nodes in the DWT implementation (and therefore to their corresponding threads) and the available number of cores that can execute the code. This analysis leads to valuable conclusions that will allow construction of even better optimizations. Results indicate that proper allocation of cores can yield faster code.

***Index Terms*—Wavelet Transformation, ECG, Heart Signal, Parallelization, OpenMP**

I. Introduction

Information and Communication Technologies (ICT) is an emerging field, which stimulates innovative solutions in the domain of healthcare. In this paper, we analyze solutions based on wearable Electrocardiogram (ECG) sensors that continuously stream data to the server and huge data quantities are being processed by a corresponding web service [1].

It is scientifically proven that the detection of heart abnormalities can prevent serious health problems [2], [3]. This requires real time processing of ECG data by a corresponding web service. Considering huge data quantities coming in a certain velocity, optimization is inevitable.

The pre-processing phase in processing of ECG signals is mainly responsible to eliminate the noise stemming from different sources and DSP filters are primary used tools.

Noise can seriously change the ECG signal and occur in the form of a baseline drift representing various physical movements and breathing. Unless the noise is removed, correct analysis on ECG data is impossible. This noise is characterized by very low frequencies, and its elimination can not be efficiently realized by simple DSP filters, such as Finite Response Filters (FIR) or Infinite Response Filters (IIR). Wavelet Transformation is a promising technique to eliminate the noise with very low frequencies, and its digital version (DWT) is capable of efficient catching and removing the ECG baseline drift. Additionally, DWT is also used in Feature Space Reduction phase in order to locate the QRS characteristics of the ECG signal.

Milcheski and Gusev [4] propose a new version of DWT implementation using a circular buffer and obtain a significant speedup. In our previous study [5], we have optimized this implementation of the DWT algorithm by optimizing the Initialization part for additional 20% faster code using OpenMP. However, the problem of synchronization between different iterations prevents even higher speedup.

In this paper, we set a research question to investigate the dependence between the nodes in the DWT implementation (and therefore to their corresponding threads) and the available number of cores that can execute the code. This analysis leads to valuable conclusions that will allow construction of even better optimizations. We give a detailed analysis and also realize experimental testing to analyze the practical implementations. Evaluation of the results are compared with the results of previously parallel code.

The paper is organized as follows, Section II presents background information and shortly the previous study. In Section III, the optimization approaches are listed. Section IV gives the details about conducted tests. Evaluation and discussion regarding the results are presented in Section V. Related work is given in Section VI and, lastly, in Section VII, the paper is concluded with future considerations.

II. Background

ECG holds vital information related to the cardiovascular condition of a living person. This section gives a general overview of ECG signal processing and briefly explains Wavelet Transformation and its importance in ECG feature reduction and extraction.

A. ECG signal processing

Methodologies for processing and analyzing ECG signal consist of three stages: data pre-processing, feature space reduction and feature extraction [3].

DSP filters are generally used in the data pre-processing phase. Low pass filters are usually used to eliminate the noise with high frequencies, such as the electrical switching and radio waves. High pass filters eliminate the noise initiated

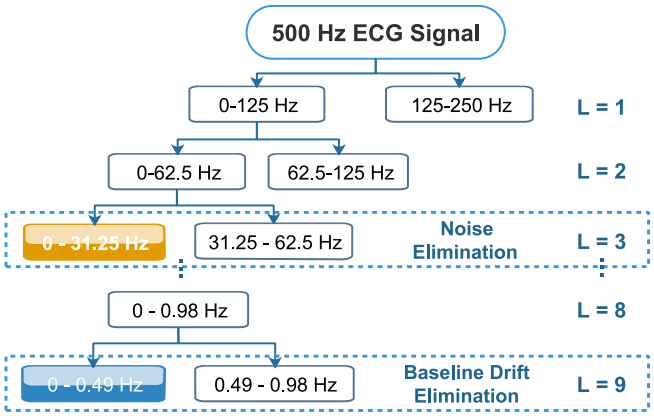


Fig. 1. Wavelet decomposition tree. High pass filter is used to eliminate baseline drift and low pass filter for noise elimination.

by physical movement and breathing, mainly interpreted as baseline drift elimination. Bandpass filters, as a combination of high pass and low pass filters are considered as effective DSP tools for noise elimination. Although, DSP filters eliminate the noise to a certain extent, they provide a relatively clear signal, which can be further processed for feature extraction.

In the feature space reduction phase, the signal is analyzed by detecting the peaks of QRS complexes and locating the peaks of individual P and T waves. A QRS complex is used as the starting point for further analysis, and, therefore, it's exact detection is of a high importance [6]. For example, a Wavelet transformation can be used for baseline drift elimination in this stage. In the final phase, QRS features are extracted, and the ECG signal precisely characterized.

The quality of extracted features, is directly dependent on the correct rate of eliminated baseline drift. Thus, focusing on this step is vital.

Digital filtering is essential for both the first two steps of the ECG signal processing. Wavelet Transformation is an efficient method used in both the elimination of baseline drift and QRS complex extraction.

B. Baseline drift and noise removal of an ECG signal

An important fact about Wavelet Transform is that the number of iterations needed to decompose the signal into smaller frequency bands is higher for smaller bands. In the case of ECG, a smaller number of iterations (and therefore time) are needed to analyze higher frequencies, and the opposite for lower frequencies. To eliminate the baseline drift one needs to deal with very low frequency bands and filter the lower frequency components.

An example of an implementation of a DWT algorithm that eliminates the baseline drift of an ECG signal is given next.

In case of a ECG sensor functioning at a 500Hz sampling frequency, it is known that spectrum of a real valued signal is symmetric [7]. Since the symmetric part is the mirror image of the first half, it does not provide additional information.

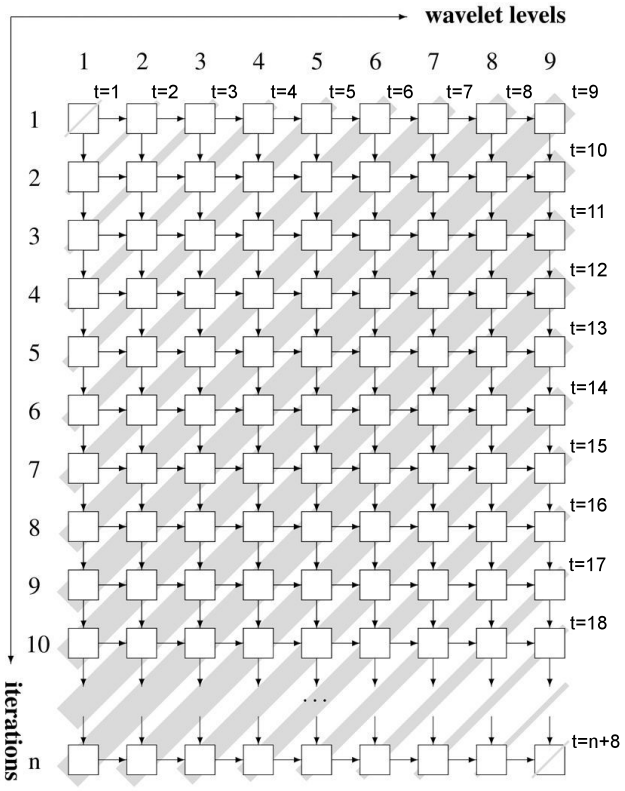


Fig. 2. Simultaneous Execution of nodes on the Processing phase of DWT code.

Thus, working on a 250 Hz frequency is enough to extract data from the signal.

DWT is based on decomposition, and reconstruction of the signal [4]. In each step, a signal is decomposed into high pass and low pass coefficients, from which approximation and detail coefficients are calculated. In order to eliminate the baseline drift of 0.5 Hz, it requires at least $L = 9$ wavelet levels. Noise can also be canceled with an additional $L = 3$ wavelet levels and a delay of 6. Details are presented on our previous work [5].

C. Discrete Wavelet Transform Analysis

Our analysis on the previous study [5], showed that DWT algorithm contains two bottlenecks, exposed in the *Initialization* and the *Processing* phase.

Observation is that the former phase does not include data dependencies between iterations. This was a vital information for pure parallelization. Though, the latter phase is highly dependent, preventing direct parallelization. This is presented in Fig. 2 where data dependence is vizualised as A $\rightarrow$ B, with the meaning *B depends on A*.

D. Previous Parallel Algorithm

Our previous parallel algorithm [5] was based on optimizing both of the bottlenecks. The *Initialization* phase was parallelized by a straightforward approach. Nevertheless, high data

dependency on the *Processing* phase required re-arrangement on the nodes for a concurrent computation.

Computation waves can flow with 45 degrees to axes where each wave contains independent computations and can be executed simultaneously at a given time stamp. This ensures that previous nodes (found on the left) are already calculated. Due to this pipelined structure, the first output will be ready after L iterations, where L is the number of wavelet levels.

The proposed implementation requires that each block of independent nodes to be synchronized between iterations. However, this is a costly operation and prevents theoretical speedup of L, when executed on L cores.

Next section gives further optimization strategies, in order to achieve the best efficiency through the parallel algorithm.

III. Optimization Approaches

The methodology for testing the parallel algorithm on the previous study [5] was based on executing both the bottlenecks on the same number of cores.

The algorithmic and storage complexity of the DWT is $O(L*2^L)$, making it nearly hard to increase the Wavelet levels.

One interesting approach is to keep the core numbers for *Initialization* phase high. This would increase the efficiency, simply because the data independent iterations.

In the *Processing* phase the maximum available nodes that can concurrently be processed is restricted to the number of wavelet levels. Thus, increasing the core numbers, will only increase the number of idle cores. However, executing this region with less number of cores can decrease the burden of barrier synchronization.

Our previous work did not address the effect of filter length. Theoretically, increasing the filter length will directly increase the percentage of processing compared to the percentage required to synchronize iterations.

Moreover, OpenMP provides built in optimization strategies [8]. Previous study did not considered using them. It would be interesting to test their effect on the barrier synchronization.

IV. Testing Methodology

Let the response time required to process the parallel algorithm be denoted by T_p, and the response time required to process the optimized parallel algorithm with p cores, be denoted by T_{op}. Then, the speedup is defined as the ratio of the execution times by (1).

$$S_{OP} = \frac{T_p}{T_{op}} \tag{1}$$

The proposed optimization approaches are tested on an Amazon C3 c3.8xlarge instance. It consists of a high-frequency Intel Xeon E5-2680 v2 (Ivy Bridge) Processor with 32 cores, 60GB of memory. OpenMP library is used for testing the proposed optimizations.

The following optimisation approaches will be tested:

OA1: Using more core numbers for Initialization phase.
OA2: Using less core numbers for Processing phase.
OA3: Increasing the filter length.
OA4: Using compiler optimizations.
OA5: Combined Effect.

On the previous study, we observed the effect of input size is negligible as the wavelet levels increase. Due to this, the input size will be fixated to 10.000 sample length ECG signal. Throughout the tests, wavelet levels vary from 3 up to 24, with incremental steps of 1 level.

On the test environment, the maximum number of available cores is 32. Considering this, the test configuration is presented in Table I.

TABLE I. Test Environment Setup

Optimization Approach	Description of The Testing Methodology
OA1	Core numbers from 2 to 30, incremental steps of 2
OA2	Core numbers from 2 to 10, incremental steps of 2
OA3	Daubechies filters of length 4, 8, 16 , 32 and 64
OA4	OpenMP's built-in O1, O2 and O3 optimizations
OA5	Combination of the most efficient approaches

Each test case was tested ten times and an average value of measured times was calculated and used for further processing. Moreover, functional verification was conducted to verify the functional characteristics of the executions obtain identical results.

V. Evaluatuation and Discussion

Figure 3 presents the speedup values when running the parallelized *Initialization* phase with fixed number of cores. These values are calculated by comparing with the case when running on core numbers equal to Wavelet levels. It is observed that, from wavelet levels ranging from 3 to 10, the fixed 2 core execution is faster by an average speedup of 12%. Similar speedup is obtained between wavelet levels 11 and 20, when running on 4 cores. On wavelet levels higher than 20, the average speedup is calculated to be 14%, though on executions with fix 10 cores.

These results indicate that, running the initialization phase on fixed number of cores yield faster code. It is observed that, as the wavelet increase, using higher number of cores becomes efficient. The speedup which is obtained is nearly 12%.

Next, on the Figure 4, speedup values which correspond to running parallelized *Processing* phase with fixed number of cores. Again, values are calculated by comparing with the case when running on core numbers equal to Wavelet levels. It is observed that this optimization approach is effective especially when the Wavelet level is greater than 13. Fix core 2 case yields the best results. This was expected, since it decreases the negative effect of barrier synchronization. On wavelet levels higher than 13, the average speedup is calculated to be 10%, though on executions with fix 2 cores.

Figure 6 presents the effect of filter length. Filter length has an effect only on the *Processing* phase, thus test is conducted on the complete parallelization case. As expected, increasing the filter length has a direct effect on the speedup of the parallelization. This is primarily due to the fact that, as filter length increase, the effect of synchronization becomes less

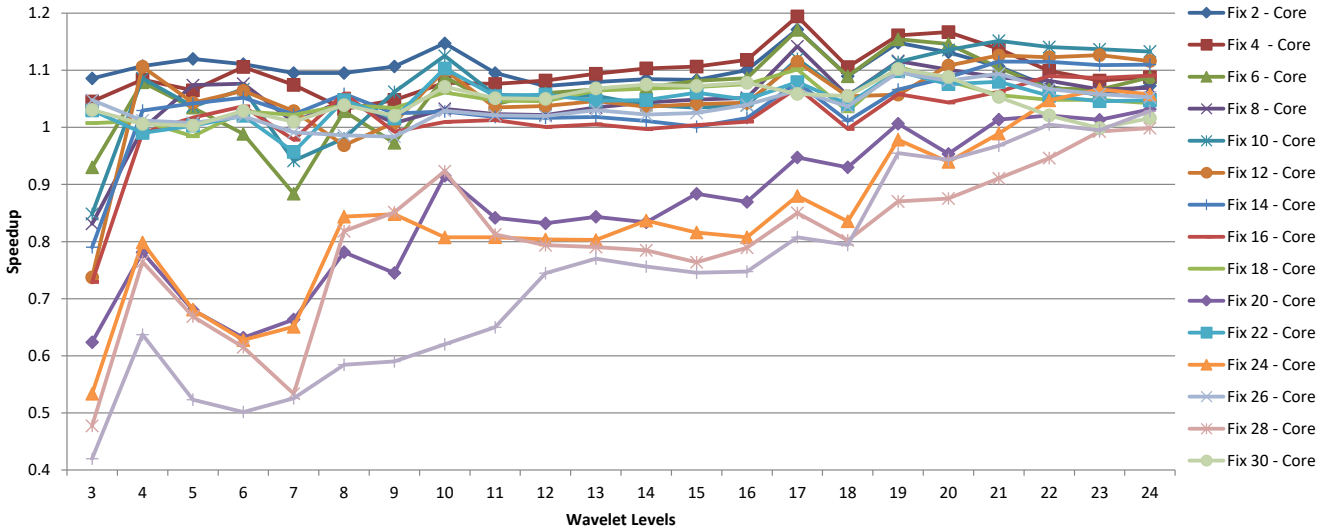


Fig. 3. Speedup of the paralleled Initialization phase with fixed number of cores compared to an implementation with cores equal to the Wavelet levels.

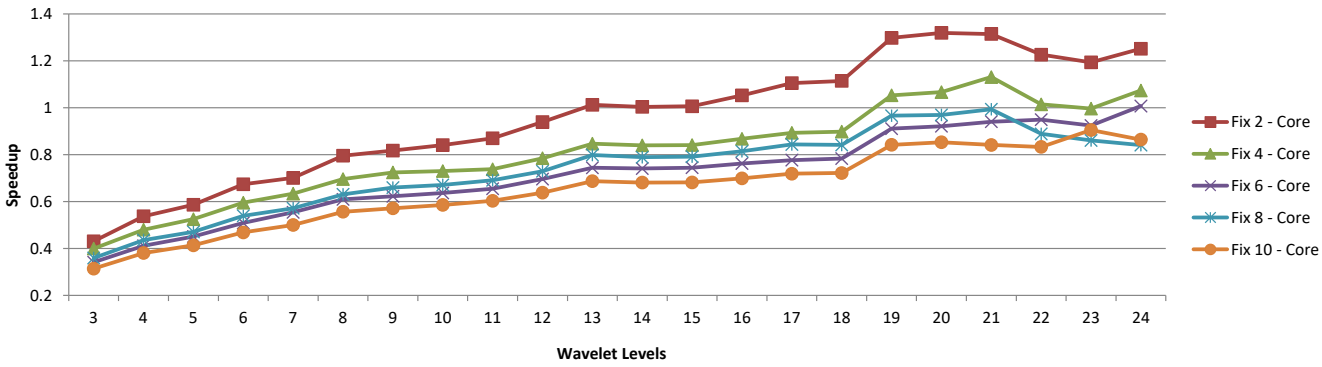


Fig. 4. Speedup of the paralleled Processing phase with fixed number of cores compared to an implementation with cores equal to the Wavelet levels.

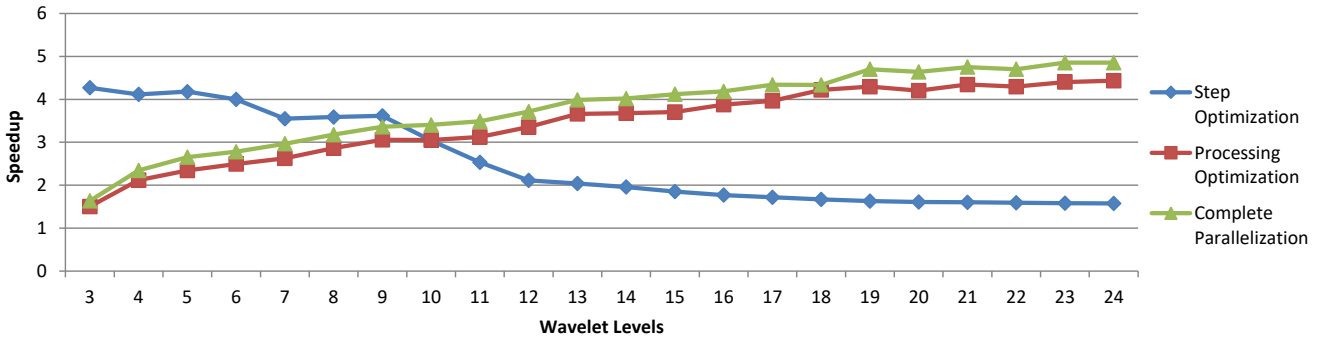


Fig. 5. Speedup of combining the best optimisation approaches compared to an implementation with cores equal to Wavelet levels without optimisations.

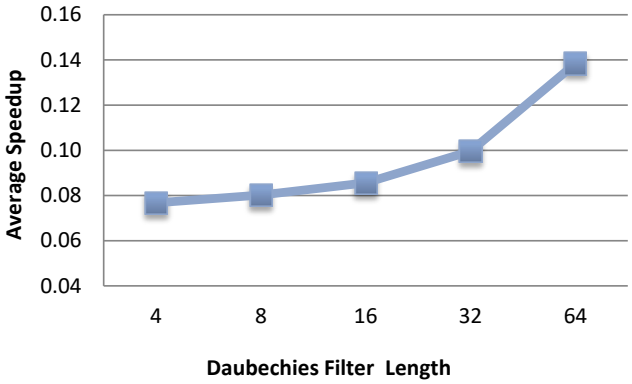


Fig. 6. Average speedup of the paralleled Processing phase with different filter lengths.

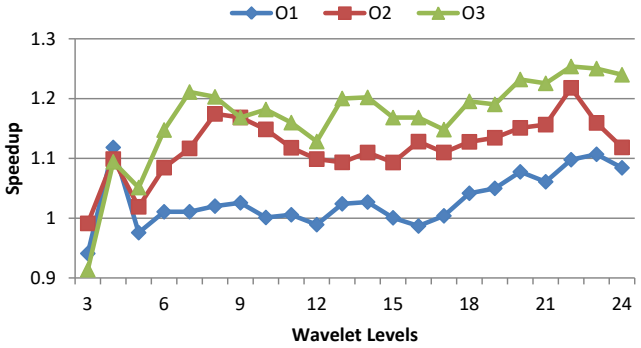


Fig. 7. Speedup of the parallel algorithm with using built-in OpenMP optimization flags.

important. On filters of length 64, the completely parallel algorithm performs 2 times faster.

The effect of compiler optimizations is shown in Figure 7. Results indicate that the build-in **O3** optimization, fastens the completely parallel algorithm by at least 15%.

Figure 5 shows the combined effect of the optimization approaches, where the code is tested on the best configuration, i.e. fix 2 cores, with 64-length Daubechies filter and built-in compiler optimization flag **O3**. Observation is that on Wavelet levels less than 10, the *Step Optimization* algorithm has an average of 4 speedup. What is more interesting, when the Wavelet levels are greater than 10, the optimizations yield a speedup of 4 *Complete Parallelization*, in a scalable nature.

Observation is that proposed optimisation approaches can yield faster codes when run on proper configurations.

VI. Related Work

Previously, we have focused on parallelizing DSP filters on Maxeler dataflow cores [9] and NVIDIA CUDA platform [10], [11]. In both cases, we achieved faster codes with a scalability depending on the number of used cores.

Pan and Tompkins, [12], have presented a real time algorithm for ECG QRS detection. Their algorithm considers the slope, amplitude and with information, and adaptively adjusts to the thresholds and parameters. It uses integer arithmetic in order to operate without requiring much computation power. There are no execution times presented, though their analysis is concentrated in the quality, where their correctness rate 99.3 percent.

An efficient implementation of DWT's in Field Programmable Gate Array(FBGA) devices [13]. They have optimised the power consumption and throughput. Additionally, a three level DWT algorithm with 4 Daubechies length filter is presented.

Milcheski and Gusev [4] have proposed an efficient DWT implementation using a circular buffers. They obtained significant speedups of at least 15. This is further improved in our previous study [5] by 20%. These papers serve as a basis for the current paper.

Several studies in literature addressed the delineation concept of ECG signal. Alfaouri and Daqrouq [14] present algorithms which produces better quality output, however no consideration is made for the performance.

Kayhan and Ercelebi [15], proposed lifting scheme based DWT algorithm for ECG denoising. Tests were conducted with an 360Hz ECG signal with 216.000 samples. Their algorithm provided fast executions where on Daubechies filter of 8, 0.141s execution times were provided.

An interesting approach was reported by Rajmic and Vlach [16], where a segmented wavelet transform analysis was presented. This approach is very attractive in the real-time case, however the authors have only published the concept.

VII. Conclusions

This work contributes OpenMP optimization for the baseline drift elimination of ECG heart signals. Totally five optimization approaches were proposed. Results indicate that, each of them can yield faster codes on proper configuration.

Approaches *OA1* and *OA2* yields speedup values of at least 10%. On the other hand, results showed that as filter length increases, the proposed parallel algorithm's efficiency increases. To be more specific, the *OA3* approach speeds up the parallel code by a factor of 2, when the filter length is increased from 4 to 64.

The effect of compiler's built-in optimization strategies were tested. The outcome of this *OA4* approach was that the *O3* flag performs best, with at least a 15% performance gain.

Lastly the combined effect was tested as the proposed *OA5* approach. Observation was that the combined effect yields a speedup of 4.

As a future work, we plan to further optimize the DWT by considering real-time segmented wavelet transform analysis concept. Additionally, it would be interesting to port the code to dataflow engine and test the parallel DWT algorithm.

References

[1] M. Gusev, A. Stojmenski, and I. Chorbev, "Challenges for development of an ecg m-health solution," *Journal of Emerging Research and Solutions in ICT*, vol. 1, no. 2, pp. 25–38, 2016.

[2] P. Laguna, N. V. Thakor, P. Caminal, R. Jane, H.-R. Yoon, A. Bayés de Luna, V. Marti, and J. Guindo, "New algorithm for qt interval analysis in 24-hour holter ecg: performance and applications," *Medical and Biological Engineering and Computing*, vol. 28, no. 1, pp. 67–73, 1990.
[3] T. S. Lugovaya, "Biometric human identification based on ECG," 2005.
[4] A. Milchevski and M. Gusev, "Improved pipelined wavelet implementation for filtering ECG signals," University Sts Cyril and Methodius, Faculty of Computer Sciences and Engineering, Tech. Rep. 27/2016, 2016.
[5] E. Domazet and M. Gusev, "Parallelization of digital wavelet transformation of ecg signals," in *MIPRO, 2017 Proceedings of the 40th Jubilee International Convention*. Opatija, Croatia, in press: IEEE, 2017.
[6] P. Mehta and M. Kumari, "Qrs complex detection of ECG signal using wavelet transform," *International Journal of Applied Engineering Research*, vol. 7, no. 11, pp. 1889–1893, 2012.
[7] R. Polikar, "The wavelet tutorial," 1996.
[8] Intel, "Quick-reference guide to optimization with intel compilers version 12," 2010, https://software.intel.com/sites/default/files/compiler_qrg12.pdf.
[9] E. Domazet, M. Gusev, and S. Ristov, "Dataflow DSP filter for ECG signals," in *13th International Conference on Informatics and Information Technologies*, in press, Bitola, Macedonia, 2016.
[10] E. Domazet, M. Gusev, and S. Ristov, "CUDA DSP filter for ECG signals," in *6th International Conference on Applied Internet and Information Technologies*, in press, Bitola, Macedonia, 2016.
[11] E. Domazet, M. Gusev, and S. Ristov, "Optimizing high-performance CUDA DSP filter for ECG signals," in *27th DAAAM International Symposium*. in press, Mostar, Bosnia and Herzegovina: DAAAM International Vienna, 2016.
[12] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.
[13] D. Shah and C. Vithlani, "Efficient implementations of discrete wavelet transforms using fpgas," *International Journal of Advances in Engineering & Technology*, vol. 1, no. 4, pp. 100–111, 2011.
[14] M. Alfaouri and K. Daqrouq, "ECG signal denoising by wavelet transform thresholding," *American Journal of applied sciences*, vol. 5, no. 3, pp. 276–281, 2008.
[15] S. Kayhan and E. Ercelebi, "Ecg denoising on bivariate shrinkage function exploiting interscale dependency of wavelet coefficients," *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, vol. 19, no. 3, pp. 495–511, 2011.
[16] P. Rajmic and J. Vlach, "Real-time audio processing via segmented wavelet transform," in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*. Citeseer, 2007.

Virtual machine migration in Cloud – techniques, challenges and CloudSim migration simulation

Dejan Stamenov, Magdalena Kostoska
Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, Macedonia
Email: stamenov.dejan@outlook.com, magdalena.kostoska@finki.ukim.mk

***Abstract*—Recent revolution in virtualization has transformed the way data centers operate by providing in depth control of the data center resources, improving performance and enhancing flexibility. Virtualization provides valuable solution to different IT challenges by minimizing complexity and lowering operational costs.**

In this paper, we will discuss about all the techniques for achieving successful migration, not only between servers in the same data center, but also between servers in distinct data centers, located on different geographical locations. Along with these techniques, the migration challenges and metrics will be introduced, on which the success of the migration depends. We will also present our extension of the CloudSim framework in order to be able to simulate part of the techniques presented in this paper.

***Index Terms*—Cloud computing, green computing, portability.**

I. Introduction

Cloud computing is the new way for providing on-demand resources and services through Internet. The recent revolution in virtualization technology is one of the key factors in cloud computing. With virtualization, we are able to achieve better cloud performance and utilize cloud computational resources, such as: CPU, memory and I/O demands [1].

Virtualization allows multiple instances of different virtual machines to work on a single server (host) at the same time. Each virtual machine has its own operating system, called guest operating system. A virtual machine behaves like a physical machine, having its own allocated (virtual) CPU, memory and network interface card (NIC) [2]. The virtualization software on the server separates the guest operating system from the underlying physical hardware using hypervisor. Virtualization provides in depth control of the data center resources, improves performance and enhances flexibility of the cloud.

The benefits of using virtual machines are [2]:

- Multiple operating system environments can exist simultaneously on the same server;
- Virtual machines that exist on the same server are isolated from each other; if one of them fails and is in error state, the other machines will not be affected by this state and will continue to run;
- The provided architecture for the virtual machine can be different than the architecture of the virtual machine itself;
- Each virtual machine is an encapsulated system with its own resources, which makes the virtual machine portable and easy to manage;
- Virtual machines are completely independent from their underlying physical hardware.

CloudSim [3], [4] is a framework for modeling and simulating cloud computing infrastructure and services. The framework is developed in Java, which supports:

- Modeling and simulation of large scale Cloud computing data centers;
- Virtualization of servers with policies for host provisioning;
- Energy-aware computational resources;
- Data center network topologies and message-passing applications.

II. Related Work

CloudSim as a framework, and the results produced by this tool, has been used for several different researches. Beloglazov and Buyya use CloudSim to validate efficiency of their adaptive heuristics for dynamic consolidation of VMs based on an analysis of historical data from the resource usage by VMs [5]. Calheiros et al. use CloudSim in order to provide experimental results for their proposed Virtual Machine provisioning model using analytical performance (queueing network system model) and workload information to supply intelligent input about system requirements [6]. Wu et al. use CloudSim to simulate and observe performance the cloud computing environment that utilizes resource allocation algorithms they propose, intended for SaaS providers who want to minimize infrastructure cost and SLA violations [7].

A few extensions of CloudSim have been created and published. CloudSimEx [1] represents an extension of CloudSim in order to enable MapReduce simulation, Web session modeling etc. Jayalakshmi and Srinivasan propose additional extension to both CloudSim and CloudSimEx in order implement geo-distributed MapReduce [8]. WorkflowSim [2] represents another

[1] https://github.com/Cloudslab/CloudSimEx
[2] https://github.com/WorkflowSim/

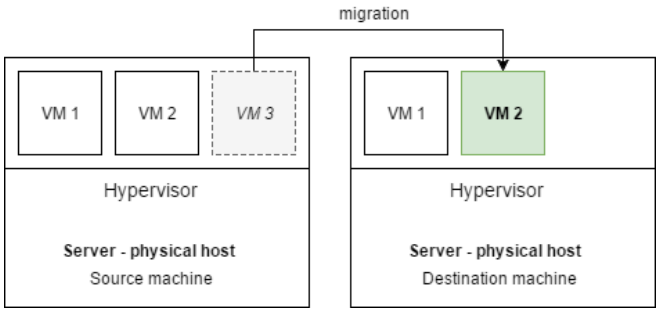


Fig. 1. The process of virtual machine migration

extension of CloudSim. It introduces support for workflow preparation and execution with an implementation of a stack of workflow parser, workflow engine and job scheduler. Cloud2Sim [3] is another extension dedicated to distributed concurrent architecture by enabling multiple instances to execute the Cloudlet and VM workloads from multiple nodes, and submit them to the DatacenterBroker, while executing the core simulation segments.

III. Virtual Machine Migration

A. The need of virtual machine migration

Despite the large number of benefits of using virtual machines, there are some disadvantages too:

- Degradation of performance on the server;
- Virtual machine is not that efficient compared to a real, physical machine when accessing the physical hardware of the server.

The first one is a major drawback of the virtual machines. When multiple virtual machines run simultaneously on the same server, each virtual machine may demand different amount of physical resources, based on the virtual machine peak moments and user requirements. These additional requirements increase the workload of the server itself, which leads to degradation of performance to the other virtual machines hosted on the same server.

Virtual machine migration is a powerful technique that allows data center administrators to optimize cloud performance and utilize cloud resources by reallocating virtual machines from one server to another. The migration, as an important process in the cloud, is useful when a server is overloaded or the performance of the system falls below the minimum accepted level.

B. Virtual machine migration patterns

There are two patterns in the process of virtual machine migration: **non-live migration** and **live migration**. Both patterns depend on the following migration metrics: down time and total migration time.

The down time metrics represents the period during which the service is unavailable due to the migration process. This period is visible to the clients as service interruption.

[3] http://sourceforge.net/projects/cloud2sim/

The total migration time is represented by the time taken to complete full migration from the source to the destination server; down time is included in the total migration time [1], [2].

In the process of non-live migration, the virtual machine on the source server is paused, and then, the state of the virtual machine is transferred to the destination server. After the data transfer of the memory pages is done, the virtual machine is resumed on the destination server. The drawback in this pattern is the down time that is noticeable from the user's point of view. In this pattern, the down time and the total migration time metrics are equal. In the process of live migration, the virtual machine is migrated from the source server to the destination server with minimal possible disruption of service. In this pattern, the down time and the total migration time metrics are not equal.

IV. Live Virtual Machine Migration Pattern in Local Area Network (LAN)

The live migration pattern brings the revolution in managing of clouds, while achieving impressive performance with minimal service down time. In the live migration pattern of virtual machines, each machine that is migrated must complete three general phases before the migration is successful [2], [9], [10]:

1) **Push phase**: Memory pages of the virtual machine that is migrated are being copied from the source server to the destination server. In this phase, the virtual machine is still running on the source server. Because the virtual machine is still operational, to keep up the consistency of the memory, the pages that have been modified must be resent to the destination server before the migration is done.
2) **Stop-and-copy phase**: The virtual machine is stopped on the source server. Then, the memory pages that were not sent, or were changed meanwhile the machine was operational, are resent to the destination.
3) **Pull phase**: The virtual machine is resumed on the destination server. If the virtual machine accesses a memory page that has not been copied from the source server, that particular memory page is faulted, and then the data for the page is pulled from the source server.

A. Live migration methods

To achieve live migration of a virtual machine from a source to a destination server, there are different methods, each defining strict steps for successful migration. In this paper, we will take a deeper look at the three most popular methods: **pre-copy**, **post-copy** and **hybrid** migration [1], [10].

1) Pre-copy method: In the pre-copy method, the memory pages from the source server are iteratively being copied to the destination server. While the pages are being copied, the virtual machine is still running and is in operational state on the source server. To keep up the consistency of the memory pages while the virtual machine is still running, page level protection is used to ensure that snapshots created from the memory are consistent for transfer over the network to the

destination server. At the final phase of this method, the virtual machine is stopped on the source server and resumed at the destination server. At this moment, the memory pages transferred to the destination server are consistent, and the virtual machine can immediately start working. Some of the memory pages may not be transferred in the first phase of the method, or some of them may have changed while the virtual machine was still operational on the source server. When the machine on the destination server tries to access such a memory page, the memory page is faulted and a request is sent to the source server from which the page is pulled again.

The pre-copy method is split on stages [2], [10]:

- **Pre-migration**: Target destination server is preselected where the memory pages will be copied when the process of live machine migration begins. This stage guarantees available resources on the destination server when the virtual machine will be resumed;
- **Reservation**: In this stage, a request is issued to the destination server where the virtual machine will be migrated. With this request, the source server confirms that the required resources will be available on the destination server when the machine is migrated, in a form of a virtual machine container. If the resources on the destination server cannot be confirmed, the virtual machine will keep running on the source server, and live migration process will not start;
- **Iterative pre-copy**: The first iteration stage, where all the memory pages are transferred from the source to the destination server. At this stage, a consistent snapshot is created. After this stage, only the modified memory pages are being copied to the destination server;
- **Stop-and-copy**: The virtual machine is paused on the source server. The modified memory pages are being copied again to the destination server, along with the CPU state. The network traffic of the virtual machine is redirected from the source to the destination server. Both servers keep the same resources for the virtual machine, in case of failure of the virtual machine on the destination server. If the machine fails to start, the state of the machine is still kept alive on the source server, so that the migration can be reverted and the virtual machine to keep running on the source. If the virtual machine is successfully resumed on the destination server, the state of the machine is still kept alive on the source server, in case of inconsistency of resources or faulty memory pages. In such cases, the source server is requested to resend the faulty resources and memory pages. At this stage, the source server is still considered as a primary server for the virtual machine, in case of failures of the destination server;
- **Commitment**: When the destination server is ensured that the virtual machine has consistent resources and copy of memory pages, indicates the source server to discard the resources of the virtual machine. With this, the destination server becomes the primary server of the virtual machine;
- **Activation**: The migrated virtual machine is now resumed on the destination server, which is now the primary server for the machine. The new IP address on the destination server is advertised over the network as the primary IP address of the virtual machine.

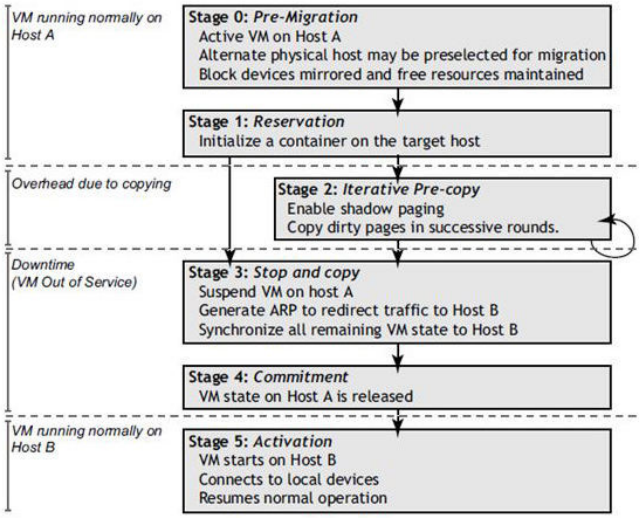


Fig. 2. Pre-copy method overview [2], [10]

This method ensures failure management by keeping consistent virtual machine data on at least one server until the live migration is successfully finished.

2) Post-copy method: The post-copy method is a vice versa of the pre-copy method. In this method, the live migration begins with pausing the virtual machine on the source server. Then, small amount of the state of the virtual machine is transferred to the destination server, along with all the CPU registers. After the transfer is done, the virtual machine is immediately resumed at the destination server. Since most of the memory pages have not been transferred, for each page the virtual machine will create page faults and will ask the source server for the memory pages to be transferred. This process is in progress until all the memory pages from the source are successfully transferred at the destination server.

This method degrades the performance of the services that are running inside the virtual machine, because of the amount of memory page faults [1], [9].

3) Hybrid method: Hybrid method is a combination of pre-copy and post-copy methods. This method is based on five stages [9]:

- **Preparation**: Reservation of virtual machine resources is done on the destination server;
- **Bounded pre-copy rounds**: Defining working set of resources of the virtual machine that will be pre-copied to the destination server before the migration begins;
- **State transfer**: Minimal state of the virtual machine is transferred to the destination server, including CPU

registers;
- **Resume**: At the destination server, the virtual machine is resumed;
- **On demand memory paging**: Since the memory pages are not completely transferred, the virtual machine at the destination server generates page faults and asks the source server for the memory pages to be transferred.

The first two stages in this method represent the pre-copy method. The other three stages represent the post-copy method.

V. Live Virtual Machine Migration in Wide Area Network (WAN)

A. Motivation for machine migration over Wide Area Networks (WANs)

The machine migration over the wide area networks is a relatively new concept in the cloud computing. Provides the data centers increased availability and reliability, along with the power to transport compute environments from one data center to another. With this, we can create dynamic environment, and we can move our compute environments across different data centers, based on the user's needs. Such a cloud system can easily respond to load balancing needs and specific computing demands, by focusing the computation power on the place where it is needed the most [11].

B. Challenges in Wide Area Network (WAN) migration

Live virtual machine migration between distinct networks brings lots of challenges. The biggest challenge is maintaining network connectivity, while preserving open connections in the process of migration. When migrating virtual machine from one data center to another, the IP address of the virtual machine must be changed, thus breaking the initial connection of the users with the source server of the virtual machine. The virtual hard drive that each virtual machine has, brings even more challenges. When migrating from one data center to another, the destination data center does not have access to the storage of the source data center [11].

C. Solutions to the challenges in the WAN migration

The usage of Dynamic DNS [4] solves the problem with the change of the IP address of the virtual machine after the migration is done. Dynamic DNS allows automatic update of a name server in the Domain Name System (DNS) in real time. In addition to the usage of Dynamic DNS, IP tunneling is an important method in this scenario. With tunneling, we can create a redirection scheme from the source server to the destination server, so that all the IP packets that are received at the source server are automatically re-routed to the destination server, where the virtual machine is migrated. Mobile IPv4/IPv6 [5] [6] are standards that enable a machine to maintain the same IPv4/IPv6 address when moving between different networks. In these standards, the machine that is moving in different networks is called a mobile node, while the source from where the node initially moved is called a home. Only the home agent knows the current, real address of the node that moves. All the users still send requests to the home agent, which then forwards all the requests to the real location of the node. The main advantage here is that the clients will never lose the connection because of the movement of the node; they will always send the requests to the home agent [11].

There are different methods for migrating the virtual hard drive of the virtual machine.

Three-phase migration scheme is a method very similar to the pre-copy method of live virtual machine migration in LAN. The virtual hard drive of the machine is iteratively copied to the destination server. The blocks that were modified meanwhile the virtual machine was still operational, are tracked and written as block bitmap. When the virtual machine is resumed on the destination server, the block bitmap is used to resend the modified blocks from the source server [12], [13]. Another method represents transfer of the virtual hard drive on efficient manner: identifying drive sectors that are more likely to change from the sectors that are less likely to change. The sectors that are less likely to change are transferred first to the destination server. Later, the sectors that were identified as more likely to change are finally transferred to the destination server [12], [14].

VI. CloudSim Simulation

We have extended the core of CloudSim to achieve virtual machine migration between hosts in one data center, and also, migration between distinct data centers. The virtual machine migration is based on calculating the MIPS (Millions Instructions Per Second) rating of each host in each data center.

A. Simulation core extension

To achieve successful migration between hosts in the same data center, we have extended the Host class of CloudSim to support functionality for host utilization. When a host is created in the simulation, we define utilization threshold. If the utilization threshold is met, virtual machines must migrate from the host in order to not degrade the performance of the machine. Along with the utilization threshold, the Host class has been extended with two methods: *isHostOverUtilized()* and *isHostOverUtilizedWithNewVm(Vm newVm)*, presented in Listing 1. Both methods are used when the migration is in progress, to check which hosts from the data center are already over utilized, so that the migration will not happen on these hosts. In addition to this, when underutilized host is found, we check if the addition of new virtual machine will make the host over utilized.

The DatacenterBroker class is used to start the process of virtual machine migration in the cloud simulation. The *processVmCreate(SimEvent ev)* method is extended to support virtual machine migration. When all the virtual machines in the cloud are created, a method *checkIfHostIsOverUtilized()*

[4] https://www.ietf.org/rfc/rfc3007.txt
[5] https://tools.ietf.org/html/rfc3344
[6] https://tools.ietf.org/html/rfc6275

Listing 1. Extensions of *Host* class

```
protected boolean isHostOverUtilized() {
  double totalRequestedMips = 0;
  for (Vm vm : this.vmList) {
     totalRequestedMips += vm.getMips();
  }
  double utilization = totalRequestedMips /
      this.getTotalMips();
  return utilization >
      this.getHostUtilizationThreshold();
}

protected boolean
        isHostOverUtilizedWithNewVm(Vm newVm) {
  double totalRequestedMips = 0;
  for (Vm vm : this.vmList) {
     totalRequestedMips += vm.getMips();
  }
  totalRequestedMips += newVm.getMips();
  double utilization = totalRequestedMips /
      this.getTotalMips();
  return utilization >
      this.getHostUtilizationThreshold();
}
```

is invoked which starts the process of checking host's over utilization. In the method we iterate the virtual machines that are created in the cloud, and retrieve their hosts. For each host, the MIPS utilization is checked. If the host is over utilized, for the current virtual machine a new host is found and the migration process starts. The new host where the machine will be migrated is found by the following criteria: (1) the new host must not be already over utilized, (2) must not be over utilized when the migration process is done, and (3) must not be the same source host from which the virtual machine will be migrated. If such host is found in the data center, the virtual machine is updated to hold data about the new host. Listing 2 presents these extensions.

Then, the *MigrateVM(Vm vmToMigrate)* method, shown in Listing 3, from the Host class is invoked, which prepares the migration data (the virtual machine that is migrated and the host to which the machine will be migrated).

To achieve virtual machine migration between different data centers, the *findHostToMigrateVm(Vm vmToMigrate, Host sourceHost)* method is extended as shown in Listing 4.

B. Cloud setup and simulation scenarios

For the purpose of simulation migration within one data center, as well as between two data centers, we have created two different use cases.

1) Use case 1 - VM migration within one data center: The configuration for this CloudSim simulation is shown in Table I.

The output of the simulation shows that at the beggining, in order to simulate over utilization, all the virtual machines will be allocated to the hosts in the data center, based on the amount of free processing elements (PEs). Due to the utilization threshold set to the hosts in the data center and the MIPS values of each virtual machine, there will always be a host that will be over utilized after the allocation of

Listing 2. Extensions of *DatacenterBroker* class

```
// all the requested VMs have been created
  if(getVmsCreatedList().size()==
      getVmList().size() - getVmsDestroyed()) {
     this.checkIfHostIsOverUtilized();
     submitCloudlets();
}

protected void checkIfHostIsOverUtilized() {
  for(Vm tmpVm : getVmsCreatedList()) {
    Host tmpHost = tmpVm.getHost();

    if(tmpHost != null &&
        tmpHost.isHostOverUtilized()) {
      Log.printConcatLine("The Host # ",
          tmpHost.getId(), " is over utilized! ",
          "The VM # ", tmpVm.getId(), " will be
          migrated to new Host in Datacenter # ",
          tmpHost.getDatacenter().getId());

      Vm vmToMigrate =
          VmList.getById(getVmsCreatedList(),
          tmpVm.getId());
      Host hostToMigrateVm =
          this.findHostToMigrateVm(vmToMigrate,
          vmToMigrate.getHost());

      if(hostToMigrateVm != null) {
        Log.printConcatLine("The VM # ",
            tmpVm.getId(), " will be migrated ",
            "to new Host #",
            hostToMigrateVm.getId(), "in Datacenter
            # ",
            hostToMigrateVm.getDatacenter().getId());

        vmToMigrate.setHost(hostToMigrateVm);
        tmpHost.MigrateVM(vmToMigrate);
      }
      else {
        Log.printConcatLine("No Host can be found
            in the Datacenter # ",
            tmpHost.getDatacenter().getId(), " to
            migrate the VM # ", tmpVm.getId());
      }
    }
    else {
      Log.printConcatLine("The Host # ",
          tmpHost.getId(), " is not over utilized!
          ", "VM migration is not needed..");
    }
  }
}
```

TABLE I
USE CASE 1 CONFIGURATION

Property	Value
Number of data centers	1
Number of hosts per data center	2
MIPS rating of Host 1	3000
MIPS rating of Host 2	3500
Utilization threshold of Host 1	0.6
Utilization threshold of Host 2	0.9
Number of VM	8
MIPS rating of VM	390-470
Number of Cloudlets per VM	1

the virtual machines is completed. In this use case, the over

Listing 3. Migrate VM method

```
protected void MigrateVM(Vm vmToMigrate) {
  Log.printConcatLine("[Host.MigrateVM] The Host #
      ", getId(),
" will try to migrate the VM # ",
    vmToMigrate.getId(), " to new Host ", "in the
    Datacenter # ", this.datacenter.getName());

  HashMap<String, Object> migrationData = new
      HashMap<>();
  // The VM that will be migrated.
  migrationData.put("vm", vmToMigrate);
  // The host where VM will be migrated.
  migrationData.put("host", vmToMigrate.getHost());

CloudSim.send(
   getId(),
   ((Datacenter)((Host)
   vmToMigrate.getHost()).getDatacenter()).getId(),
   0, CloudSimTags.VM_MIGRATE, migrationData);
}
```

Listing 4. New method for selecting Host

```
protected Host findHostToMigrateVm(Vm vmToMigrate,
    Host sourceHost) {
  for(Vm tmpVm : getVmsCreatedList()) {
    Host tmpHost = tmpVm.getHost();

  if(!tmpHost.isHostOverUtilizedWithNewVm(vmToMigrate)
      && tmpHost.getId() != sourceHost.getId()) {
         // Cross-data center virtual machine
             migration.
         Datacenter sourceVMDatacenter =
             ((Host)vmToMigrate.getHost()).
             getDatacenter();
         if(sourceVMDatacenter.getId() !=
             ((Datacenter)tmpHost.getDatacenter())
             .getId()) {
                 return tmpHost;
         }
  }

  return null;
}
```

utilized host will be the *Host 1*. Because of this scenario, the virtual machine migration process will start, and at least 2 (two) of the machines from the *Host 1* will be migrated to the *Host 2*, which has higher utilization threshold.

2) Use case 2 - VM migration between two data centers: The configuration for this CloudSim simulation is shown in Table II.

The output of the simulation shows that at the beggining, in order to simulate over utilization, all the virtual machines will be allocated to the hosts in the first data center, based on the amount of free processing elements (PEs). Due to the utilization threshold set to the hosts in the first data center and the MIPS values of each virtual machine, there will always be a host that will be over utilized after the allocation of the virtual machines is completed. Because of this scenario, the virtual machine migration process will start, and at least 3 (three) of the machines from the first data center will be migrated to new host in the second data center, which has higher utilization threshold.

TABLE II
USE CASE 2 CONFIGURATION

Property	Value
Number of data centers	2
Number of hosts per data center	3
MIPS rating of Host 1 (in each data center)	3000
MIPS rating of Host 2 (in each data center)	3500
MIPS rating of Host 3 (in each data center)	3500
Utilization threshold of Host 1 (in each data center)	0.6
Utilization threshold of Host 2 (in each data center)	0.9
Utilization threshold of Host 3 (in each data center)	0.8
Number of VM	13
MIPS rating of VM	650-1300
Number of Cloudlets per VM	1

VII. CONCLUSION

In this paper, we have presented techniques for achieving successful virtual machine migration between distinct data centers. With all the techniques and patterns presented, we can optimize cloud performance and utilize cloud resources by reallocating virtual machines from one server to another. The techniques presented in this paper are not the only techniques available for achieving virtual machine migration; different techniques with custom algorithms are available, each of them supporting the cloud needs. Each of these techniques have the same main objectives: achieving less down time, less total migration time while providing seamless service to the users of the cloud.

We have successfully extended the CloudSim toolkit to include two different mechanisms for virtual machine migration - 1) migration of virtual machines from one physical host to other within one data center and 2) migration between two (or more) distinct data centers.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University, Skopje, Macedonia through the project CCIoT(Cloud computing solution for streaming IoT).

REFERENCES

[1] L. Mallu and R. Ezhilarasie, "Live migration of virtual machines in cloud environment: A survey," *Indian Journal of Science and Technology*, vol. 8, no. S9, 2015. [Online]. Available: http://www.indjst.org/index.php/indjst/article/view/65579

[2] A. Agarwal and S. Raina, "Live migration of virtual machines in cloud," *International Journal of Scientific and Research Publications*, vol. 2, no. 6, 2012.

[3] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, no. 1, pp. 23–50, 2011. [Online]. Available: http://dx.doi.org/10.1002/spe.995

[4] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities," in *2009 International Conference on High Performance Computing Simulation*, June 2009, pp. 1–11.

[5] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurr. Comput. : Pract. Exper.*, vol. 24, no. 13, pp. 1397–1420, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1002/cpe.1867

[6] R. N. Calheiros, R. Ranjan, and R. Buyya, "Virtual machine provisioning based on analytical performance and qos in cloud computing environments," in *2011 International Conference on Parallel Processing*, Sept 2011, pp. 295–304.

[7] L. Wu, S. K. Garg, and R. Buyya, "Sla-based resource allocation for software as a service provider (saas) in cloud computing environments," in *2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, May 2011, pp. 195–204.

[8] D. S. Jayalakshmi and R. Srinivasan, *Simulation of MapReduce Across Geographically Distributed Datacentres Using CloudSim*. Cham: Springer International Publishing, 2017, pp. 70–81. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-50472-8_6

[9] R. Singh, K. S. Kahlon, and S. Singh, "Comparative study of virtual machine migration techniques and challenges in post copy live virtual machine migration," *International Journal of Science and Research (IJSR)*, vol. 5, no. 3, 2016.

[10] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in *Proceedings of the 2Nd Conference on Symposium on Networked Systems Design & Implementation - Volume 2*, ser. NSDI'05. Berkeley, CA, USA: USENIX Association, 2005, pp. 273–286. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251203.1251223

[11] E. Harney, S. Goasguen, J. Martin, M. Murphy, and M. Westall, "The efficacy of live virtual machine migrations over the internet," in *Proceedings of the 2Nd International Workshop on Virtualization Technology in Distributed Computing*, ser. VTDC '07. New York, NY, USA: ACM, 2007, pp. 8:1–8:7. [Online]. Available: http://doi.acm.org/10.1145/1408654.1408662

[12] T. Mohammad and C. S. Eati, "A performance study of vm live migration over the wan," Master's thesis, , Department of Communication Systems, 2015, 0763472814.

[13] Y. Luo, B. Zhang, X. Wang, Z. Wang, Y. Sun, and H. Chen, "Live and incremental whole-system migration of virtual machines using block-bitmap," in *2008 IEEE International Conference on Cluster Computing*, Sept 2008, pp. 99–106.

[14] S. Akoush, R. Sohan, B. Roman, A. Rice, and A. Hopper, "Activity based sector synchronisation: Efficient transfer of disk-state for wan live migration," in *2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems*, July 2011, pp. 22–31.

Overview of Workflow Management Systems

Tina Ranic and Marjan Gusev
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
1000 Skopje, Macedonia
e-mail: tinaranic@gmail.com, marjan.gushev@finki.ukim.mk

***Abstract*—Performing in a data-driven world, companies need continuous execution of jobs to improve their business. This arises the need of having a robust scheduler that will be able to execute tasks reliably. Big companies, such as Spotify and Airbnb, that need to process complex and business critical data, have already encountered this problem. Their solutions are available for the general community. This paper compares the provided open-source solutions, examines the pros and cons, and aims to provide any company facing this issue to easier make a decision.**

***Index Terms*—workflows; scheduling; workflow management systems; styx; airflow; luigi**

I. Introduction

Everything that needs to be done repeatedly should be automated. People who work with data need to write jobs that should run on a given schedule. Most often a job will require other datasets as dependency. This introduces the need of: *a reliable scheduler*; and a *wiring library* for jobs and inter-job dependencies definition.

This is important especially in the following use cases:

1) Finding all customers that haven't been active in the last month and sending them an email reminder (a monthly scheduled procedure).
2) Database dump to a file system on a daily basis.
3) Sending invoices every hour.
4) Fetch latest source of information needed for further processing or just refreshing every 10 minutes.

All of these jobs need to be scheduled at a given frequency, i.e. hourly, daily, weekly etc. For those familiar with the Unix systems, most probably will first think of the well-known Cron scheduler.

Cron is a Unix program that allows executing commands or bash scripts at repetitive intervals. It is used for removing logs, creating backups, sending emails etc.

The next example presents how to send an email everyday at 07:30 AM using Cron schedulеding:

```
30 7 * * * echo ''Don't forget to eat heathy'' | mail -v -s ''chocolates'' martin@somewhere.com
```

It is comprised of two parts: `30 7 * * *` is the first part that Cron understands when to schedule the job. There is a place for five stars, and they represent different date parts in the following order:

- minute (from 0 to 59)
- hour (from 0 to 23)
- day of month (from 1 to 31)
- month (from 1 to 12)
- day of week (from 0 to 6) (0=Sunday)

Thus, `30 7 * * *` would mean that a job will get scheduled at hour 7, 30 minutes, every day, every month, every day of the week. The second part is the command itself, or most commonly a shell script is provided.

It is a powerful program, and has been used in production across companies for running data pipelines. However, as everything else it comes with some limitations. Some of the problems that come with cron are:

- How does cron cope with a multi-node cluster?
- How to tell on which node should execute a job?
- How to access logs?
- How to retry jobs?
- How to track job configuration for each job?
- How to handle dependencies between jobs? Especially common in ETL/Big Data pipelines where there are several data inputs (upstreams) and intermediate crunching until producing result. [1]
- Where to see statistics?

All these questions are something that cron hasn't responded well to, but required the developer to handle them himself. Therefore, existing big data oriented companies have put their efforts to build reliable workflow management systems that satisfy these or some of these needs.

There are several commercial and open source solutions on the market. In this paper, we analyze the open source solutions and develop a methodology to compare them and evaluate which is the best solution in a specific use case scenario.

The paper is organized as follows. Section II describes Styx, a plain Docker scheduler open sourced by Spotify. Section III gives an overview of Luigi, another open source product by Spotify for defining jobs dependencies, that is a wiring library. Section IV examines the competitor system Airflow, open sourced by Airbnb which is a combination of scheduler and a wiring library. A relevant discussion is given in Section V. Section VI discusses conclusions and directions for future work.

II. Styx

Styx is a scheduler that schedules Docker containers in Kubernetes. It is designed and built by Spotify with the aim to support smooth migration for their platform from Hadoop to Google Cloud Platform (GCP) [2], [3].

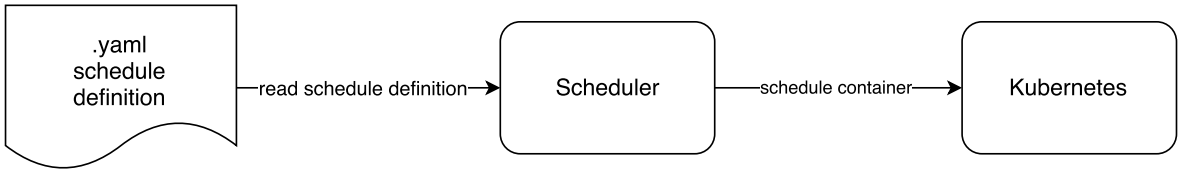


Fig. 1. Styx high level architecture

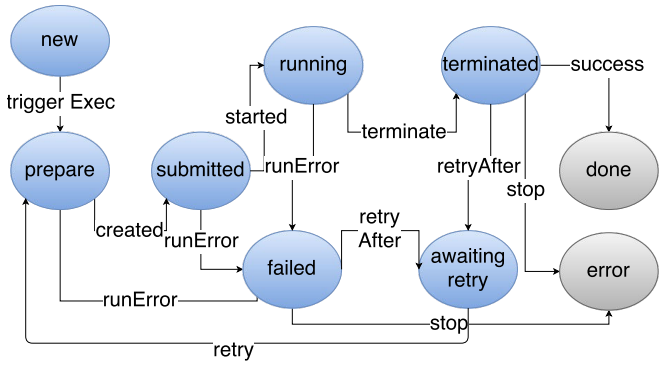


Fig. 2. State lifecycle of a workflow instance

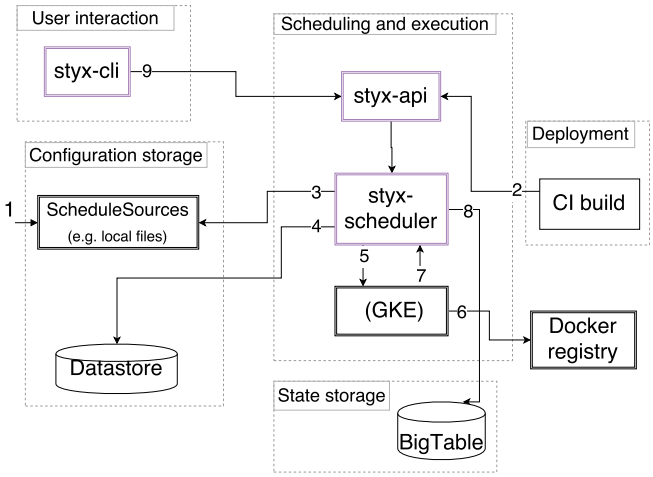


Fig. 3. Styx architecture

A. *Styx High Level Architecture and Concepts*

Fig. 1 presents the Styx high level architecture. The input of the scheduler is defined in a Yaml file that presents the workflow definition. The output of the scheduler is the scheduled container presented by Kubernetes.

Fig. 1 represents a high level representation of Styx. To schedule a workflow (job), a schedule definition needs to be defined. A schedule definition defines all the needed information for Styx to schedule a job, i.e. schedule a Docker container. The schedule definition is defined in a Yaml file and it should contain the following:

- `data_endpoint id` is the name of the workflow,
- *partitioning* represents the frequency at which the workflow should be scheduled,
- `docker_image` is the image that will be used to start a container,
- `docker_args` is a specification of the arguments that will be used to start the container.

Once Styx has this information available, it stores it in a persistent state. A workflow definition has a state, and it can be either enabled or disabled. If it is enabled, Styx will schedule the workflow with the provided frequency. Whenever it is a time to schedule a workflow, Styx will create all the needed metadata to spin off a container.

A workflow is the definition itself (like a class definition), triggering and executing a container is a specific instance of the workflow, called workflow instance. A workflow instance is associated with a parameter, the date for which it runs, e.g. 2017-01-01 or 2017-01-01T15. A workflow instance is associated with a running state as well, which represents the lifecycle of an execution.

Styx is an Apollo service built to smoothly integrate with the Google Cloud Platform.

The building blocks in Styx are Apollo, Docker, Kubernetes and GCP.

B. *Apollo*

Apollo is a set of libraries that easy writing micro - services. It has been used at Spotify for internal use and nowadays is open-sourced. It includes an HTTP server and a URI routing system, making it easy when writing microservices.

C. *Docker*

Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries and anything one can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in [4].

Fig. 4 compares virtual machines and Docker. Both virtual machines and containers aim to provide resource isolation and resources allocation benefits. Virtual machines are comprised of a whole different operating system (OS) independent from the host OS, all application dependencies (binaries and libraries), and the application itself [4]. On another hand, a Docker container includes the application and all of its dependencies - but resides in the host operating system itself. Under the hood, it utilizes the power of the Linux nature - using cgroups and namespaces - which helps creating the wall between Docker containers. Namespaces ”wrap a set of system resources and present them to a process to make

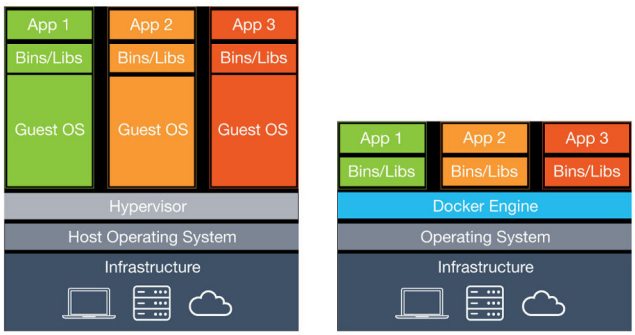


Fig. 4. VM vs Docker

it look like they are dedicated to that process." [2], while Cgroups (developed by Google govern isolation and usage of system resources. Namespaces deal with resource isolation for a single process, while cgroups manage resources for a group of processes.

Companies know that they must build, deploy, and scale applications faster to be competitive. The monolithic architecture of many existing applications hampers innovation and increases time to market, so many companies are migrating to a microservice architecture. DevOps methodologies and container virtualization technologies like Docker don't only make it easier to implement microservices, they also reduce risk and speed up continuous deployment and delivery.

Docker has become a synonym for microservices development. In order to be successful, companies need to develop and deploy their software applications continuously, but the monolithic architecture increases the time to the market. Therefore, many companies are migrating to a microservice architecture, and the nature of Docker satisfies its needs. More specifically, building containerized applications using Docker eases the process of:

- Deployment - The development, testing and deployment cycle gets shorter, which speeds up the release of new versions.
- Reliability and Availability - If one of the services is malfunctioning, it will not affect the whole service. Additionally, when deploying a new version the downtime is significantly lower compared to deployment of a monolith.
- Scalability - Every microservice can scale independently
- Autonomy - In an environment with many teams, maintaining a monolith system slows down the development process. Using microservices, these teams can employ technology and tools they want.

D. Kubernetes

Kubernetes is a platform that manages deployment and scaling of Docker containers. A docker container is any service, job, or whatever that has been dockerized. Its been developed by Google and open-sourced [5].

A *pod* is a group of containers that are scheduled onto the same host. Pods serve as units of scheduling, deployment, and horizontal scaling/replication. Pods share fate, and share some resources, such as storage volumes and IP addresses. In Styx, there exists one-to-one mapping between a pod and a container, so we use the names interchangeably. [6]

The importance of understanding of this concept is to understand how Styx integrates with Kubernetes, more specifically understanding the lifecycle of a Styx workflow instance. A *pod phase* can be one of the following pod's lifecycle phases: [6]

- Pending: The pod has been assigned for scheduling, but at least one Docker image hasn't been created yet.
- Running: All the containers associated with the pod have been created.
- Succeeded: All containers in the pod have terminated in success, and will not be restarted.
- Failed: All containers in the pod have terminated, at least one container has terminated in failure (exited with non-zero exit status or was terminated by the system).
- Unknown: For some reason the state of the pod could not be obtained, typically due to an error in communicating with the host of the pod.

E. Google Cloud Platform

Google Cloud Platform (GCP) is a cloud computing service by Google that offers hosting on the same supporting infrastructure that Google uses internally for end-user products like Google Search and YouTube [7]. GCP provides developer products to build a range of programs from simple websites to complex applications [8]. It offers many products, the one that Styx utilizes are given below.

Google Container Engine (GKE) is a powerful cluster manager and orchestration system for running Docker containers. GKE schedules containers into the cluster and manages them automatically based on defined requirements (such as CPU and memory). It's built on the open source Kubernetes system, giving the flexibility to take advantage of on-premises, hybrid, or public cloud infrastructure [9].

Google Cloud Datastore is a NoSQL document database built for automatic scaling, high performance, and ease of application development. Cloud Datastore features include:

- atomic transactions
- high availability of read and writes
- massive scalability with high performance - i.e. the queries scale with the size of the result set, not the size of the data set. Cloud Datastore uses a mix of indexes and query constraints so your queries scale with the size of your result set, not the size of your data set.
- flexible storage and querying of data - maps naturally to object-oriented and scripting languages.

Google Cloud Bigtable is a compressed, high performance, and proprietary data storage system built on Google File

System, Chubby Lock Service, SSTable (log-structured storage like LevelDB) and a few other Google technologies. Today there is a commercial version offered on the GCP. It is characterized with low latency and high throughput. It underlies many core Google services like Search, Analytics, Maps and Gmail [10].

F. Execution process explained

After explaining the different systems and technologies that Styx uses, this section will explain the execution process for a workflow instance.

In Fig. 3 the start of the process is referring the the Schedule sources, step 1. Schedule sources is the place wherefrom Styx fetches the schedule definitions, including the workflow name, partitioning, Docker image and Docker arguments, as it was stated before. Additionally, Styx provides an API to supply the Docker image for a specific workflow (2). This is very useful, since data pipelines can make use of CI system that will leverage running a script that will deploy the latest image of the pipeline. Internally, Styx stores the workflow configuration in Datastore. (step 2.)

To schedule a workflow, Styx submits its configuration to GKE - step 5. Then a Docker container (a pod, the names are used interchangeably since the mapping is one to one) is spin off and it enters 'pending phase', which includes the time to fetch the image over the network from the Docker registry.

To track the status of the execution, there is a Kubernetes watcher that sends events to Styx about the pod phase. There is additionally a pod poller in Styx because of reliability issues with the Pod Watcher.

During state transitioning, all events that are emitted are stored in Bigtable in a transition log (step 8). This is very powerful since it allows for restoring the workflow instance state by replaying the state machine in a dry mode. Meaning if Styx goes don't it can recover by replaying the states in dry mode.

The Styx scheduler service is deployed on one instance, while the Styx API is high availability consisting of 3 instances. The command line interface (CLI) is a powerful tool where one can manually trigger execution of workflows (for testing purposes), track execution status, i.e. manipulating workflows.

III. Luigi

Luigi is an execution framework that supports tasks definition, tasks dependency definition, tasks execution and visualisation [11].

A *Task* is a basic unit of work. It defines:

1) Its dependences - in the requires() method. In building data pipelines, it is very common that there is a job that takes initial input, and then its output is chained as another input to another job. Luigi helps to easily model these dependences.
2) What gets done - in the run() method
3) The output location where the job will write the result - output()

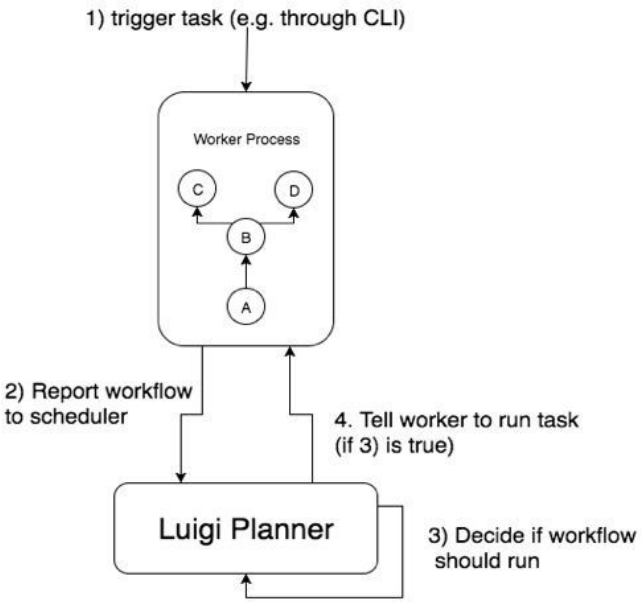


Fig. 5. Luigi - high level architecture

```
43 class ExampleTask(luigi.Task):
44     def requires(): ...
45     def run(): ...
46     def output(): ...
```

Fig. 6. An example of a Luigi Task

The *target* class corresponds to a file on a disk, a file on another distributed file system or maybe a record in a database. The only method it should override is the exists() method, that returns true if the file exists.

Fig. 5 represents the execution process of a task. The triggering of a job is done by using a CLI command, e.g. `luigi --module myModule MyTask --local-scheduler` . Executing this command triggers creation of a worker process. The worker builds the dependency graph and uploads it to the scheduler (Fig. 5, step 2), and then in an endless loop asks the scheduler for work. If there are no other workers running the same task (Fig. 5 step 3), the worker will be assigned this dependency graph for execution (Fig. 5 step 4).

IV. Airflow

Airflow is a workflow management system open-sourced by Airbnb. It is able to build, run and monitor data pipelines. [12] The platform is easily extensible, one could write jobs that interact with Hive, Presto, MySQL, HDFS, Postgres and S3. It offers a rich UI supporting pipeline dependences visualisation, monitoring progress, triggering tasks and so on. It also provides a command-line interface. Airflow provides what Styx and Airflow offer together, of course with a totally different philosophy.

There are several essential components in the airflow architecture:

- Python code is used for defining workflows, which represents a Directed Acyclic Graph (DAG) of tasks.
- Airflow scheduler that fires up tasks from the DAGs.
- Metadata repository which is a relational database that keeps track of job statuses and other persistent information. It uses SQLAlchemy that abstract away the choice of database.
- Web application where you can see the DAGs definitions and the dependencies between tasks, you can track progress, see metadata and logs. It is built on top of the Flask Python Web framework. It is an independent gunicorn process which connects to the metadata database.

Another component is a CLI to test, run, backfill and describe of defined DAGs [13].

Basic unit of execution is called *Task*. The instantiation defines specific values when calling the abstract operator, and the parameterized task becomes a node in a DAG.

A *task instance* represents a specific run of a task and is characterized as the combination of a DAG, a task, and a point in time (execution_date). Task instances also have an indicative state, which could be ”running”, ”success”, ”failed”, ”skipped”, ”up for retry”, etc.

A *Direct Acyclic Graph* (DAG) is a graph of all the tasks to be run in a logical structure where the tasks interdependences are clear. Basically a DAG defines how a workflow is going to run, not what. A DAG contains:

- `start_date` the first trigger for the DAG its tasks will be scheduled
- `schedule_interval` The frequency with which the DAGs tasks will run. The scheduler doesn't wait for an already running task to finish before it runs the next scheduled. If a task is scheduled every hour, but it takes longer than one hour to finish, it will start the next one before waiting for the first one to finish. An instance of a DAG is called a DAGRun. A DAGRun is identified by the id of the DAG suffixed with with execution date

An *operator* describes the unit of task that will get done. Airflow is very versatile with the different provided operators, like Docker Operator that executes a Docker image, Bash Operator which executes a bash script or bash command, furthermore Email Operator, HTTP operator, Sql operator, Sensor. An operator is assigned to a DAG. Once it is assigned, the DAG will make sure that operators run in the correct certain order. All operators extend from the main, BaseOperator, which in turn extends from the SQLAlchemy base class. (objects can be pushed to the database - investigate in source code). Although the operators are derived from an SQLAlchemy base class, they don't contain class to the database. This part is implemented in hooks.

There are three main types of operators:

1) *Sensor* Waits for events to happen, via polling. It could be any appearance of a file/directory in a file system, or some other existence check. There are 2 things that need to be defined, frequency and timeout.
2) *Remote Execution* This triggers an operation on a remote system, for example a Dataflow job in Google Cloud Platform, or maybe BigQuery query in GCP as well.
3) *Data Transfers* Imports and exports of data between different system. For example it can be dumping data from a database to HDFS.

V. Discusion

All three systems have different advantages and disadvantages and it is important to choose the right one depending on the challenge. This section examines the biggest strengths of the different systems and the right choice of a workflow management system.

A. Styx

Styx integrates smoothly with Google Cloud Platform (GCP), and is the right choice when one is already using GCP or planning migration to GCP. Styx makes workflow deployment easy, packaging it in a Docker image brings you one step away from scheduling it. This means there wont be any dependency hell, and there wont be a need for distributing the code dependencies to executors. It has a powerful CLI, where one can easily trigger, halt, and examine status of currently running workflows and backfills.

B. Luigi

Luigi has a straightforward approach solving the job dependencies problem. The concepts are simple, and the implementation as well. There is no special technology behind one needs to depend on, everything that is needed is a Linux server. It provides a nice UI for tracking job dependencies and execution.

C. Airflow

Airflow is the right choice when there are high demanding requirements on job dependencies. It has a bit more complex concepts then Luigi, but more powerful features at the same time. It is straightforward to use it in its simplest setup, using a local executor. Using a distributed message queue introduces complexity because there is an additional work where all the code dependencies need to be distributed on the workers. This is disputable if one feels comfortable with Celery. Airflow wins with its rich UI, manages job execution, shows the code for every DAG, and manages variables that can be used in the DAGs. It also supports Gantt charts, where one can see statistics about the jobs duration.

VI. Conclusion

This paper overviews open-source solutions for workflow scheduling and orchestration. A lot of architecture and implementation details are given for Styx, Luigi and Airflow. Their concepts and intended use is compared and advantages and disadvantages are discussed.

As a future work, we plan to analyze other commercial solutions, including Amazon Web Services' Data pipelines and Microsoft Azure Data Factory, and realize a more comprehensive overview.

REFERENCES

[1] Danidelvalle, "I'm sorry Cron, i've met AirBnB's Airflow," 12 Sep 2016. [Online]. Available: https://danidelvalle.me/2016/09/12/im-sorry-cron-ive-met-airbnbs-airflow/

[2] A. Konrad, "Why Spotify really decided to move its core infrastructure to Google Cloud," *Forbes, Tech, #InTheCloud*, 29 Feb 2016. [Online]. Available: https://www.forbes.com/sites/alexkonrad/2016/02/29/why-spotify-really-chose-google-cloud/#5b350a0f3ee4

[3] The Spotify Team, "Announcing Spotify infrastructure's Googley future," *Spotify News*, 23 Feb 2016. [Online]. Available: https://news.spotify.com/us/2016/02/23/announcing-spotify-infrastructures-googley-future/

[4] Docker, "What is Docker?" as seen on 27 Feb 2017. [Online]. Available: https://www.docker.com/what-docker

[5] Kubernetes, "What is Kubernetes?" as seen on 27 Feb 2017. [Online]. Available: https://kubernetes.io/docs/whatisk8s/

[6] ——, "Pods," as seen on 27 Feb 2017. [Online]. Available: https://kubernetes.io/docs/user-guide/pods/

[7] Google Cloud Platform, "Why Google Cloud platform?" as seen on 27 Feb 2017. [Online]. Available: https://cloud.google.com

[8] ——, "Products and services," as seen on 27 Feb 2017. [Online]. Available: https://cloud.google.com/products/

[9] ——, "Container Engine," as seen on 27 Feb 2017. [Online]. Available: https://cloud.google.com/container-engine/

[10] ——, "Cloud Bigtable," as seen on 27 Feb 2017. [Online]. Available: https://cloud.google.com/bigtable/

[11] E. Bernhardsson, "More Luigi alternatives," 02 Jul 2015. [Online]. Available: http://nerds.airbnb.com/airflow/

[12] cogNiTioN, "Newbie: Intro to cron," 30 Dec 1999. [Online]. Available: http://www.unixgeeks.org/security/newbie/unix/cron-1.html

[13] M. Beauchemin, "Airflow: a workflow management platform," 02 Jun 2015. [Online]. Available: http://nerds.airbnb.com/airflow/

Cloud services for faculty workflow automatization

Kostadin Mishev[1], Aleksandar Stojmenski[1], Vesna Dimitrova[1], Ivica Dimitrovski[1], Ivan Chorbev[1]
[1]Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
{kostadin.mishev,aleksandar.stojmenski,vesna.dimitrova, ivica.dimitrovski, ivan.chorbev}@finki.ukim.mk

***Abstract*—This paper presents a brief overview of the concepts for collaboration between various systems developed for the Faculty of Computer Science and Engineering in Skopje. Web technology such as the HTTP, originally designed for human-to-machine communication, is utilized for machine-to-machine communication, more specifically for transferring machine-readable data in web service formats such as JSON. A Central Authentication Service is being used for identity management and single sign-on for all integrated applications. Various guidelines and the whole process of integration of existing systems along with their interconnection and interoperability are covered. The interface that we are using in the integration of the multi-platform system pose no hard dependencies between the various applications, thus allowing easily integration and intercommunication protocols.**

Keywords—collaboration; systems integration; web services; cross-platform;

I. INTRODUCTION

System integration is a complex process where a cohesive platform is created from components that were not specifically designed to work together. Components of an integrated platform are often stand alone systems that operate on different computer environments. This paper describes the platform integration of various applications and their interconnection and interdependability.

In order to collaborate between each other, these applications need to specify suitable protocols for exchanging data, as well as protocols for flow control. This paper gives an overview of the transformations of the data structures, the data itself and the impact of the different applications over the data. For this purpose, different design patterns are used to facilitate the communication between systems and to harmonize endpoint data formats that this paper examines. Different problems and their possible solutions are presented, regarding systems lifecycle, architecture, process, interface, synchronization and security. Each application endpoint exports OAuth2 security protocol functionalities for system authentication and authorization. Following the principles of the OAuth2 protocol, each server authenticates the users using bearer tokens. Furthermore, the communication protocol adopts the JSON data format as a primary exchanging throughput over a HTTP communication channel.

II. BACKGROUND WORK

Although a lot of work and progress has already been done in the area of web services in the past years, efforts have been mostly focused on service description models and languages, and on automated service discovery and composition [1]. The term Web services is used frequently nowadays, although sometimes it is very ambiguous. Existing definitions of the terms vary from generic to specific and restrictive. One definition is that a Web service is seen as an application accessible to other applications over the Web [2]. This is a very open definition meaning that anything with a URL address is a Web service. It can include a CGI script or refer to a program accessible over the Web with a stable API, published with additional descriptive information on some service directory. A more precise definition is provided by the UDDI consortium, which characterizes Web services as “self-contained, modular business applications that have open, Internet-oriented, standards-based interfaces” [3]. This definition is more detailed, placing the emphasis on the need for being compliant with Internet standards. A step further in refining the definition of Web services is the one provided by the World Wide Web consortium (W3C), and specifically the group involved in the Web Service Activity: “a software application identified by a URI, whose interfaces and bindings are capable of being defined, described, and discovered as XML artifacts. A Web service supports direct interactions with other software agents using XML-based messages exchanged via Internet-based protocols”. The W3C definition is quite accurate and also hints at how Web services should work. The definition stresses that Web services should be capable of being “defined, described, and discovered,” thereby clarifying the meaning of “accessible” and making more concrete the notion of “Internet-oriented, standards-based interfaces.” It also states that Web services should be “services” similar to those in conventional middleware. Not only they should be “up and running,” but they should be described and advertised so that it is possible to write clients that bind and interact with them. In other words, Web services are components that can be integrated into more complex distributed applications.

The W3C also states that XML is part of the solution. Indeed, XML is so popular and widely used today that, just like HTTP and Web servers, it can be considered as being part of Web technology. There is little doubt that XML will be the data format used for many Web-based interactions. Note that

even more specific definitions exist. For example, in the online technical dictionary Webopedia, a Web service is defined as "a standardized way of integrating Web-based applications using the XML, SOAP, WSDL, and UDDI open standards over an Internet protocol backbone. XML is used to tag the data, SOAP is used to transfer the data, WSDL is used for describing the services available, and UDDI is used for listing what services are available" [4]. Specific standards that could be used for performing binding and for interacting with a Web service are mentioned here. These are the leading standards today in Web services. As a matter of fact, many applications that are "made accessible to other applications" do so through SOAP, WSDL, UDDI, and other Web standards. However, these standards do not constitute the essence of Web services technology: the problems underlying Web services are the same regardless of the standards used. This is why, keeping the above observations in mind, we can adopt the W3C definition and proceed toward detailing what Web services really are and what they imply.

Web services were developed as a solution to (or at least as a simplification of) the system integration problem [5]. The main benefit they bring is that of standardization, in terms of data format (JSON), interface definition language (WSDL), transport mechanism (SOAP) and many other interoperability aspects. Standardization reduces heterogeneity and makes it therefore easier to develop business logic that integrates different (Web service-based) applications. Web services also represent the most promising technologies for the realization of service-oriented architectures (SOAs), not only within, but also outside companies' boundaries, as they are designed to enable loosely-coupled, distributed interaction [6].

While standardization makes interoperability easier, it does not remove the need for design patterns that include adapters and mediators. Different Web services may still support different interfaces and protocols. For example, although two map or driving direction services may support JSON or XML and use SOAP over HTTP as transport mechanism, they may still provide operations that have different names, different parameters, and different business logic or protocols. In addition, other opportunities enabled by Web services have an implication in terms of adaptation needs. In fact, having loosely-coupled and B2B interactions imply that services are not designed having interoperability with a particular client in mind (as it was often the case with CORBA-style integration) [7]. They are designed to be open and possibly without knowledge, at development time, about the type and number of clients that will access them, which can be very large. The possible interactions that a Web service can support are specified at design time, using what is called a business protocol or conversation protocol. [8] A business protocol specifies message exchange sequences that are supported by the service, for example expressed in terms of constraints on the order in which service operations should be invoked. Another studied solution is to make system integration with ActiveXML which utilities peer-to-peer interaction between nodes and specifies special data design and ActiveXML web services [8].

III. SYSTEM ARCHITECTURE

The Faculty of Computer Science and Engineering continues the development of e-platform for student and staff services by providing new e-services and their adaptation with machine interfaces to the central data repository. Such services provide simplification and acceleration of the Faculty administrative workflows by providing easy-to-use interfaces avoiding congestion and bottle-neck scenarios. The system architecture that is discussed in this paper consists of several different subsystems which work as a part of the architecture provided in [8]. That means that the core of the subsystems is a common part which ties the entities as soft links providing scalable and reusable patterns for the purpose of interoperable services

A. Architecture Core

The core of the service architecture is implemented in Microsoft .NET MVC technology. The authentication process is handled by the Central Authentication Service (CAS) which is implemented in Java. The CAS service involves a back-end service, that does not have its own HTTP interface, but communicates with a web application. The Service manager implements a protocol which is platform independent (JSON based). All applications and services in the system are communicating and synchronizing using this protocol. The service manager is implemented in C# Web Api[1] and is used as a mediator for control messages exchange, storing permission access rules, identifying the status of the services (running, failed, blocked…) and enabling intra service communication. As a result of successful authentication, the user obtains JWT token which is passed as authentication header, providing stateless communication between the client browser and the server. The JWT token is used as a key reference for the user credentials. Users identity management is handled by Active Directory. Active Directory is interconnected with the CAS service and serves information about the user credentials. CAS service queries the AD to enable user single sign on authentication for multiple third-party services.

[1] https://www.asp.net/web-api

User authorization i.e. the role of each user for specific service is handled by each service individually. That means that each services implements its own many-to-many relationship which stores the information about the grant tickets associated to each user in the application. If the user does not contain any grant to the application, he will not be able to access it. The authorization is handled immediately after successful authentication to the CAS service. It is provided authentication by the user group. That means that is the user is a part of the group students, he will obtain different grant that the user from the group professors. This properties can be overridden by specifying the grant for each user individually. The grant with higher weight i.e. with stronger permissions wins the authorization process.

The intercommunication among this services is realized with REST JSON-based web services by using the ASP.NET Web API Framework.

Such core enables development of software applications that will facilitate the workflow among students, administrative and teaching staff in the faculty with implementation of several use-case scenarios.

B. Consultation management service

Such service improves the collaboration between students-teaching staff providing on-time delivery of information about consultation terms providing two-way communication. The students are informed about the ordinary or additional terms of consultations for appropriate course by appropriate teacher, and the teacher is informed about the total number of students that will be present on the appropriate consultation term. Also, the teacher has an opportunity to cancel appropriate term by specifying the reason, so the students will be on-time informed. The teachers have appropriate views for terms editing meanwhile the students use user-friendly consultations terms representation for each teacher. Also, for the needs of this application, it was developed an Android application which contains the same views for different roles.

C. Student Requests Service

The one of the most extensive tasks for the students affair office was the processing of the student requests in the start of each semester. There are multiple types of student requests such:

- Student programme change
- Course professor change
- Approval of additional student credits
- Enrollment of course without complied preconditions
- Group change
- Payment in installments for semester enroll.

Each semester, there are approximately 2000 students requests, so the students affair office should process each of them in a paper form. Afterwards, they should transmit the sheet of requests to the vice-dean for teaching staff or vice-dean for finances for approval. Finally, they should inform each of the student for the each of the request for the status of approval. Such task is tremendous and ineffective respecting the technology development and usage. Leading by the following problems, it was developed a student service available on public address where the students can submit a request online for each of the types described above. The vice-deans review the requests and give appropriate approval considering the circumstances provided by the students and immediately responds to the request. The students is also immediately informed for the status of the request by e-mail. Now, the workflow is reduced by excluding the students affair office in processing of each of the request. Also, the paper form of the request is excluded by replacing it with online application form authorized by the Central Authorization System (CAS).

D. Absence report workflow automatization

Absence service is intended for faculty staff providing set of workflows that semi automate the absence registration of the employee at the faculty. The system provides workflow for absence approval, travel expense payment, daily wage etc. The travel report was the final step which was not automated and generated by the application. The employee should fulfil by hand a specific report specifying the days of absence, visited places, travel costs and a short description of the realized activities. We try to automate such process enabling automatic print of the report and its archiving. The employee fulfil the required information on a specific application form of the travel report online, when he finishes the travel absence. We developed a Java service that generates a pdf version of the final travel report with fulfilled content. The look and feel is the same as the original report due the fact that the report should look just the same as the purposed template. We obtained the required functionality by appending the appropriate content as a stamp above the scanned copy of the travel report. Afterwards, the final result of the report can be printed and it is ready for archiving.

IV. USE-CASE SCENARIO OF DATA-FLOW PROCESS SERVICE INTEROPERABILITY

Each semester, FCSE creates a new courses schedule by using a specialized timetable software. The software allows export to a consolidated view in CSV format which as 3 star data type can be easily integrated in other software data-management products. Each record in the CSV export presents a time slot when the course is helded, containing information about the teacher and the student group. First at all (Fig. 1), the CSV format is preprocessed by using specialized data-processor whose role is to clear the errors in merging the data with central database. It tries to fix the typographic errors in each record to extract the containing information and link to material entities saved in the databases. The required matching entities are the student group, teacher and course.

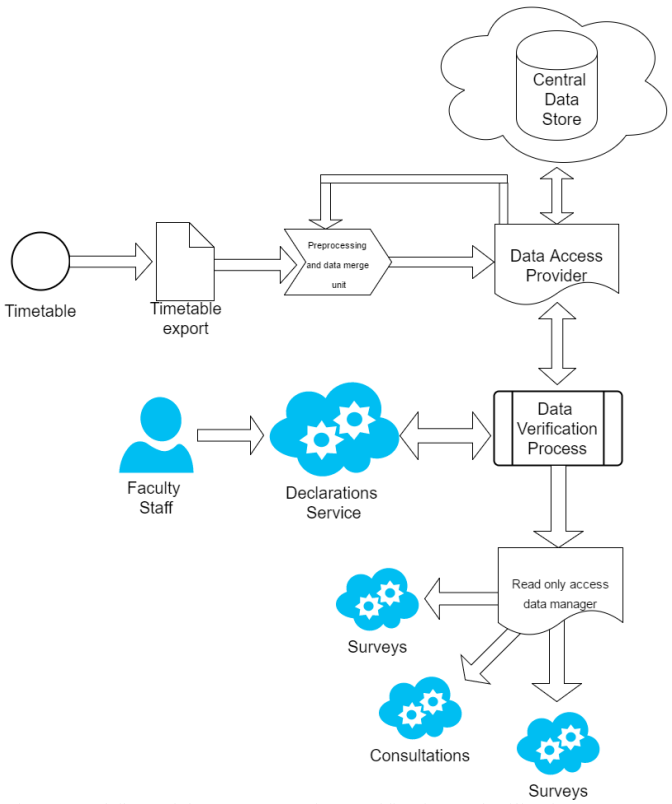


Fig. 1. Workflow of data preprocessing, verification and utilization

After successful matching of all entities, the schedule timeslot are ready for import. The imported data will help us to link schedule and consultation application to pull data from the central database. Also, the declaration software uses the same data to provide pre inserted records intended for the faculty staff. In this process, all records are previews and additional changes if needed are done. So, by finishing this step, all information is consistent and it can be used for survey generation intended for teacher evaluation by students in the end of the semester. This is complete scenario of interservice communication among multiple data service providers including process of data import, verification and reusability of the data which is enabled by using the proposed faculty service architecture.

V. Conclusion

Faculty of Computer Science and Engineering continues the development of e-platform for student and staff services by providing new e-services and their adaptation with machine interfaces to the central data repository. Such services provide simplification and acceleration of the Faculty administrative workflows by providing easy-to-use interfaces avoiding congestion and bottle-neck scenarios. The student requests is one of the services provided from FCSE to its students for implementation of the workflow that manages the adaptation of the needs of the students related to the faculty. The absence software provided for FCSE's staff has its updates that automates the generation of the travel warrant and increases the semantic of the reports by providing additional fields in the travel report. The teaching declaration improves and facilities the process of assertion and determination of the state of academic affairs by enabling concise information for salary calculation.

In this paper it is presented the usage of such software applications as a part of the global e-platform and their interconnections and interoperability by providing the complete architectural design and implementation. Also, it is described complete scenario of interservice communication among multiple data service providers including process of data import, its verification and reusability of the data which is enabled by using the proposed faculty service architecture.

References

[1] Benatallah, Boualem, Fabio Casati, Daniela Grigori, Hamid R. Motahari Nezhad, and Farouk Toumani. "Developing adapters for web services integration." In Advanced Information Systems Engineering, pp. 415-429. Springer Berlin Heidelberg, 2005.

[2] Papazoglou, M.P., 2003, December. Service-oriented computing: Concepts, characteristics and directions. In Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on (pp. 3-12). IEEE.

[3] UDDI Consortium. "UDDI Executive White Paper, Nov. 2001." (2004).

[4] Al-Masri, E. and Mahmoud, Q.H., 2008, April. Investigating web services on the world wide web. In Proceedings of the 17th international conference on World Wide Web (pp. 795-804). ACM.

[5] G. Alonso, F. Casati, H. Kuno, V. Machiraju. Web Services: Concepts, Architectures, and Applications. Springer Verlag. 2004.

[6] B. Benatallah, F. Casati, and F. Toumani. Web services conversation modeling: A Cornerstone for EBusiness Automation. IEEE Internet Computing, 8(1), 2004.

[7] L. Bordeaux et al. When are two Web Services Compatible?. VLDB TES'04. Toronto, Canada. 2004. [CFPT03] C. Canal, L. Fuentes, E. Pimentel, J. Troya, A. Vallecillo. Adding Roles to CORBA Objects. IEEE TSE, 29 (3), 2003, IEEE Press

[8] Stojmenski, A. et al. "Cross platform system integration using web services", CIIT 2016

The Impact of Flood and Earthquake Catastrophes on the Macedonian Insurance

Sanja Tanchevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
sanja.tanchevska@triglav.mk

Marija Mihova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
marija.mihova@finki.ukim.mk

***Abstract*— In the last two summers, (2015 and 2016), two great flood catastrophes in Macedonia led to a substantial material losses of the inhabitants of the affected regions. The most of the Macedonian citizens felt series of earthquakes with high magnitudes, and there was a huge fear among the population for quite some time. It is expected that such disasters will increase the premiums and the number of policies for floods and earthquakes. The objective of this paper is to analyze the changes in the premiums and policies before and after these disasters, and if there exist some changes, how long did that trend last. We are also interested whether the interest of such policies is higher in the affected regions.**

***Keywords* – flood catastrophes; flood insurance; Earthquake catastrophes; Earthquake insurance policies.**

I. INTRODUCTION

The natural catastrophes are unpredictable and can cause high damage, so there is no guarantee that the private market players are willing to take on such risks. Therefore the flood and earthquake insurance products are not included in the basic insurance packages and they are also significantly expensive. Moreover, there are examples where such risks are covered with additional government programs as National Flood Insurance Program (NFIP) in USA.

The companies in Macedonia offer additional insurance for floods and earthquakes, although this is not catastrophe-prone region and there is not a great interest in such types of products. Moreover the majority of Macedonian citizens do not have a habit to buy any insurance if it is not necessary. Usually they insure their property only when the insurance is a part of a credit or a leasing arrangement.

But when a disaster occurs, most citizens are curious whether the insurance is worthwhile. There are some extreme examples in the world when after huge disasters the premiums grew enormously [3]. Therefore multiple studies in the literature have analyzed the correlations between disasters and insurance [1, 2]. In this paper we analyzed the impact of the greatest floods and earthquakes that have happened in Macedonia in the last years on the insurance business.

The paper is organized as follows. In the next chapter we give an overview of the products available in Macedonia and we are analyzing the popularity of these products. The disasters that have effected most of the citizens in Macedonia are regarded in the third chapter. In the fourth chapter we test several hypotheses intended to demonstrate the impact of the disasters on the insurance business in Macedonia. At the end we conclude this paper with summary.

II. PRODUCTS AND COSTUMER'S INTEREST FOR THEM

Flood and earthquake damage is excluded in standard homeowners and renters insurance policies. Flood coverage, however, is available in the form of a separate policy and in all insurance companies there are two products for risk of flood and windstorm, the first one as a part of the homeowners' insurance and the second one as a part of crops and fruits insurance [4, 5, 6].

- The basic package of the homeowners' insurance include: Fire and Lightning, Escape of water from plumbing system in buildings, Explosion except nuclear, Storm and Hail, Falling aircrafts and Civil commotion, while perils which must be insured additionally are flood, torrent and high waters, landslide, subsidence, snow avalanche, earthquake, burglary and robbery, glass breakage, liability to third party, alternative accommodation expenses. The only insurance that includes all additional perils in it is the Casco insurance for cars.
- The basic package of crops and fruits insurance includes indemnification in case of hail, fire and thunderbolt. For risk of flood and windstorm as well as spring and fall frost, loss of seed quality and risk package after harvesting one need special insurances.

We would like to emphasize that although the product for additional insurance of crops and fruits from flood and earthquake is available on the Macedonian market, there is no interest in this product, i.e. there is no sold policy of this type.

TABLE I. PERCENTAGE OF FLOOD AND EARTHQUAKE INSURANCE BETWEEN 2013 AND 2016

year	Flood	Earthquake	Both
2013	5%	0%	0.00%
2014	6.8%	1.1%	0.30%
2015	7.5%	1.6%	0.37%
2016	8%	2.6%	0.57%

Table I shows the percentage of the policies with additional flood insurance and earthquake insurance. We will further omit the data from 2013, since we have only 21 policies from that year. It is evident that we have very small percentage of policies with additional insurance. Even smaller is the percentage of the policies with both flood and earthquake insurance. In order to see whether there is connection between these two additional insurances, we test the following hypothesis:

Hfe: The decision of a costumer to take additional flood insurance depends on his decision to take additional earthquake insurance.

- The Chi-square and Mantel- Haenszel tests of independency both rejected it, with $p<0.001$. Therefore having in mind the crosstab showed in Table II, we can conclude that there is a tendency to take both insurances.

TABLE II. CROSSTABS BETWEEN FLOOD AND EARTHQUAKE POLICIES

			earthquake	
			Yes	No
flood	Yes	% Count	0,4%	7,05%
		% Expected Count	0,13%	7,35%
	No	% Count	1.38%	91.13%
		% Expected Count	1.67%	90.84%

Depending on the costumers there are 3 types of property insurances in Macedonia: home, civil and industry insurance. Next we analyze whether the group have an influence on the decision to take additional insurance.

Hf-t/He-t: The type of insurance does not depend on the decision to take additional flood/earthquake insurance.

- The Chi-square tests for both hypothesis rejected the hypotheses with $p<0.001$ for flood and $p=0.001$ for earthquake. Therefore, from crosstabs in Table III, we may conclude that companies and industry buy additional insurance more often than households.

TABLE III. CROSSTABS BETWEEN TYPE AND ADDITIONAL INSURANCE

			Type		
			home	civil	industry
earthquake	No	% within earth.	43,7%	50,9%	5,4%
		% within type	98,1%	98,4%	97,2%
	Yes	% within earth.	46,1%	45,5%	8,3%
		% within type	1,9%	1,6%	2,8%
flood	No	% within flood	45,4%	50,1%	4,4%
		% within type	96,0%	91,3%	75,9%
	Yes	% within flood	23,1%	59,4%	17,4%
		% within type	4,0%	8,7%	24,1%

III. FLOOD AND EARTHQUAKE DISASTERS IN MACEDONIA AND ITS SURROUNDINGS

We have collected data from the last 3 years, from 2014-2016. Although the additional insurances for flood and earthquake are available from earlier, there is no evidence for additional insurance before the end of 2013. Therefore we will consider only the disasters from this period.

A. *Floods*

Low – intensity floods are common in Macedonia, but such disasters do not cause great damages. In the period of our interest we could distinguished few bigger floods with casualties that affected on most of the population.

- Floods in Bosnia and Herzegovina and Serbia caused of cyclone "Tamara". Although this cyclone did not reach Macedonia, the numerous victims and the huge damages it caused in surrounding countries have occupied the media and the people in Macedonia.
- The flood on the 3th of August, 2015, in Tetovo area. This flood reached several villages near Tetovo and apart the huge material damage, it took 4 victims [8].
- The bigger flood in 2016 in Skopje happened on August 6. This catastrophe took more than 20 lifes and caused huge damage on the highway around Skopje [8].

B. *Earthquake*

There were no sensitive earthquakes in Macedonia during 2014 and 2015, but in 2016 we have two significant earthquakes, without any victims.

- 21st of May, an earthquake with magnitude 4.7 was felt in south west Macedonia and caused some small damage.
- On the 11th of September 2016 the people felt one moderate earthquake with magnitude 5.4 [7] and a series of light and minor earthquakes. A number of little damages caused from this earthquake were also registered.

IV. EFFECT OF CATASTOPHES ON ADDITIONAL FLOOD AND EARTHQUAKE INSURANCE

In this chapter we analyze whether and how the catastrophes mentioned in the previous chapter had an effect on premiums.

Let us observe the time series of the percentage of the policies with additional insurance. Fig. 1 shows this percentage for flood, while Fig. 3 shows this percentage for earthquake in each month during the years from 2014 to 2016. It is evident that in the period of the most remarkable catastrophes, this percentage significantly grows.

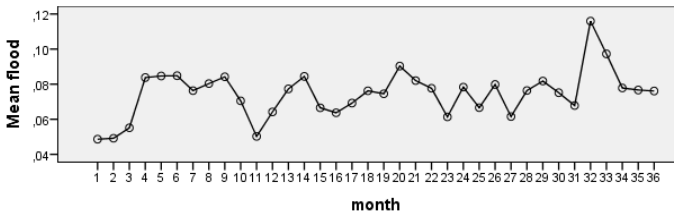


Fig. 1. Mean number of insurance with additional flood insurance during the period 2014-2016 .

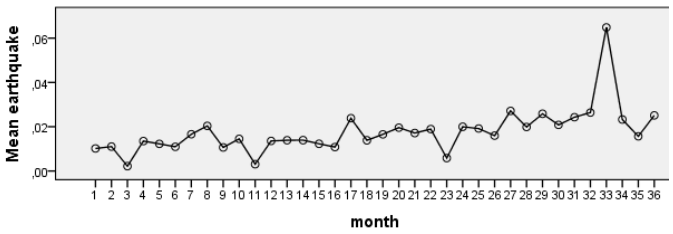


Fig. 2. Mean number of insurance with additional earthquake insurance during the period 2014-2016 .

In order to choose more appropriate test for the effect of the catastrophes on the insurance, we need to consider the changes during the time (years or periods of the year). Therefore we test the following hypotheses:

Hf-y/ He-y: The percentage of additional flood/earthquake insurance is equal during the years.

- Both Pearson Chi-Square test of independency between policies with additional flood and earthquake insurance and years (2014-2016) and ANOVA test for the percentage of additional flood and earthquake insurance do not accepted the hypothesis. The appropriate statistics for flood are $\chi^2(2)=0.005$ and p=0.005, and for earthquake are $\chi^2(2)=0.000$ and p=0.000. Therefore we may conclude that this percentage grows up during the years, and this growth is statistically significant. But if we take only flood insurance in 2015 and 2016, we will obtain that there is no significant difference in percentage of a flood insurance between these two years. The Pearson Chi-Square test accepts that with $\chi^2(2)=0.225$, while the t-test for equality of percentage also accepts that with p=0.225.

In addition we tested the hypotheses of equal distribution of all insurances as well as additional flood and earthquake insurances between months and quarters. All hypothesis were rejected, so we can conclude that the periods of the year have influence on the decision to buy an insurance policy.

We also want to see whether there is any connection in insurance frequency between different years and periods of the year (months or quarters). For that purpose we used Chi-Square test of independency between these two variables. The test was performed on all insurance policies and the policies with additional flood and earthquake insurance.

Hiy/Hqy: The decision of a costumer to buy property insurance in specific month/ quarter, do not depends on the year when it was bought.

- The Chi-Square test of independency between years and months performed on all policies accepted the hypothesis with $\chi^2(22)=0.204$, while the Chi-Square test of independency between years and quarters performed on all policies accepted the hypothesis with $\chi^2(6)=0.816$.

Hiyf/Hqyf: The decision of a costumer to buy additional flood insurance in specific month/ quarter, do not depends on the year when it was bought.

- We extract the instances with additional flood insurance and made Chi-Square test of independency between years and months. The hypothesis that for each flood insurance policy, the month when it is bought is independent of the year when it is bought is not accepted with significance level 0.05, (but we can accept it with significance level 0.01, since $\chi^2(22)=0.014$.
- The hypothesis that for each flood insurance policy, the quartile when it is bought is independent of the year when it is bought is not accepted ($\chi^2(6)=0.003$). But if we take only 2015 and 2016 in to consideration, then the same hypothesis is accepted ($\chi^2(3)=0.191$).

Hiye/Hqye: The decision of a costumer to buy additional earthquake insurance in specific month/ quarter, depends on the year when it was bought.

- We extract the instances with additional earthquake insurance and performed Chi-Square test of independency between years and months. The hypothesis that for each earthquake insurance policy, the month when it is bought is independent of the year when it is bought is rejected ($\chi^2(22)=0.006$).
- The hypothesis that for each earthquake insurance policy, the quartile when it is bought is independent of the year when it is bought is accepted ($\chi^2(6)=0.374$).

The above results indicated that there are certain factors has effect on decision of costumers to buy some insurance product. So we want to analyze could the natural catastrophes have been one of that factors.

A. *Floods*

Let us regard the time series of frequencies of the policies during all three years of observation. It is obvious, from Fig. 1, that the greatest percentages in 2014 are in the months April, May and June, which coincides with the great floods caused by the cyclone "Tamara", and the announcements that the Balkans expect mayor cyclone activity. On the other hand, in 2015 and 2016 this maximum is in September, and then it drastically decreases. As it is expected, the maximum is reached in the months when flood catastrophes with human victims have occurred. But an interesting observation is that just after the thread passing, the number of policies drastically decreases.

The pervious analysis have shown that periods have effect on additional flood insurance, but not on general property insurance. In order to see whether the occurrence of a catastrophe increases the percentage of additional policies, we analyzed the data before and after catastrophic floods. Since the biggest floods in 2016 and 2015 are both in August, the data from June 1 to December 31 are grouped into 3 groups:

- Group 1 - before the disaster, June and July
- Group 2 - during and just after the disaster, August and September
- Group 3 - after the disaster, October and November.

The descriptive statistics for the data from 2016 are given in Table IV. We are testing the following three hypotheses:

Hf1: The mean number of policies with additional flood insurance in Group 2 in 2016 is bigger than the mean number of policies with additional flood insurance in Group 1 in 2016.

- The p value of the t-test of equality is $p<0.001$. Therefore we can conclude that the percentage of policies with additional flood insurance in the period during the disaster and just after the disaster is higher from the percentage of policies with additional flood insurance in the period before the disaster.

TABLE IV. DESCRIPTIVE STATISTICS FLOOD 2016 IN GROUPS BEFORE, DURING AND AFTER THE CATASTROPHIC FLOOD

	period	Mean	Std. Dev.	Std. Error
flood	Group 1	,07	,257	,006
	Group 2	,11	,308	,007
	Group 3	,08	,267	,006

Hf2: The mean number of policies with additional flood insurance in Group 2 in 2016 is larger than the mean number of policies with additional flood insurance in Group 3 in 2016.

- The p value of the t-test of equality is $p=0.001$. Therefore we can conclude that the percentage of policies with additional flood insurance in the period during the disaster and just after the disaster is higher from the percentage of policies with additional flood insurance in the period two months after the disaster.

Hf3: The mean number of policies with additional flood insurance in Group 1 and Group 3 in 2016 are equal.

- The p value of the t-test of equality is $p=0.464$. Therefore we can conclude that the percentage of policies before and after the disaster are equal.

TABLE V. DESCRIPTIVE STATISTICS FLOOD 2015 IN GROUPS BEFORE, DURING AND AFTER THE CATASTROPHIC FLOOD

	period	Mean	Std. Dev.	Std. Error
flood	Group 1	,08	,264	,006
	Group 2	,09	,280	,006
	Group 3	,07	,255	,006

The descriptive statistics for the data from 2015 are given in Table V. We are testing the following three hypotheses:

Hf4: The mean number of policies with additional flood insurance in Group 2 in 2015 is larger than the mean number of policies with additional flood insurance in Group 1 in 2015.

- The p value of the t-test of equality is $p=0.233$. Therefore we can conclude that the percentage of policies with additional flood insurance in the period during the disaster is not statistically higher from the percentage of policies with additional flood insurance in the period before.

Hf5: The mean number of policies with additional flood insurance in Group 2 in 2015 is larger than the mean number of policies with additional flood insurance in Group 3 in 2015.

- The p value of the t-test of equality is $p=0.053$, so the conclusion is that the percentage of policies with additional flood insurance in the period during the disaster is not statistically higher from the percentage of policies with additional flood insurance two months after.

Hf3: The mean number of policies with additional flood insurance in Group 1 and Group 3 in 2015 are equal.

- The p value of the t-test of equality is $p=0.491$, so we can conclude that the percentage of policies before and after the disaster in 2015 are equal.

These three results also leads to the conclusion that the floods in more rural areas do not increase the number of policies with additional flood insurance.

One of the most important things for the insurance companies is the effect in additional premiums after the disasters. It is obtained that this additional number of policies does not increase the total premium significantly. In fact if we take the total flood premiums in the Group 2, T_2, and the total flood premiums in in Group 1, T_1, than the difference $T_2 - T_1$ do not cover the liquidated damage caused during the flood from August 6. But if we take the total premiums for the policies with additional flood insurance (flood premium + gross premium) in Group 2, S_2, and the same premiums in in Group 1, S_1, than the difference $S_2 - S_1$ is 2.8% higher than the liquidated damage for flood from August 2016. The situation is even worse in 2015. In fact, the total liquidated damage for the flood on 03.08.2015 is more than 10 times larger than the difference in additional premiums for flood between Group 2 and Group 1 in 2015. Even more, the total premium in Group 1 in 2015 is larger than the total premium in Group 2.

All these results indicates that the companies do not have a profit after catastrophic floods. Moreover, they may have significant losses after such disasters.

B. Earthquakes

The growth of the number of policies with additional earthquakes insurance just after a disaster is much more evident, Table VI. From Fig. 2 we can see that in all other months the mean percentage of polices with additional insurance for earthquake ranges between 0.3 and 2.4, but in September 2016, this percentage reached 6.5%. The next month the percentage falls back into 2%. Fig. 3 shows the mean percentage of policies with additional earthquake insurance during September 2016. It is remarkable that the highest percentage is in September 14, 3 days after the earthquake with magnitude 5.4, and that tendency had lasted for three days, and then quickly decreases.

TABLE VI. CROSSTABLE FOR EARTHQUAKE 2016 AND GROUPS BEFORE, DURING AND AFTER THE EARTHQUAKE

Earthquake		groupEarthquake1		
		Group 1	Group 2	Group 3
0	% within earthquake	30,0%	32,9%	37,2%
1	% within earthquake	25,2%	49,5%	25,2%

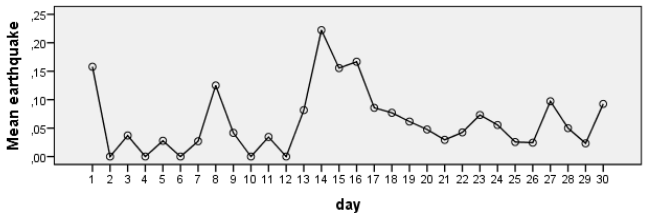


Fig. 3. Tendency of earthquake insurance in September 2016.

To illustrate that the occurrence of a catastrophe increase the percentage of policies with additional earthquake insurance, we analyzed the data before and after the earthquake on 11.09.2016. Therefore again we group the data from July 1 to December 31 into 3 groups:

- Group 1 - before the earthquake, July and August
- Group 2 - during and just after the earthquake, September and October
- Group 3 - after the earthquake, November and December.

The descriptive statistics for the data from 2016 are given in Table VII.

TABLE VII. DESCRIPTIVE STATISTICS EARTHQUAKE 2016 IN GROUPS BEFORE, DURING AND AFTER THE EARTHQUAKE

	period	Mean	Std. Dev.	Std. Error
flood	Group 1	,03	,157	,003
	Group 2	,04	,206	,004
	Group 3	,02	,142	,003

We test following three hypotheses:

He1: The mean number of policies with additional earthquake insurance in Group 2 is larger than the mean number of policies with additional flood insurance in Group 1.

- The p value of the t-test of equality is p=0.001, so we may conclude that the percentage of policies with additional earthquake insurance in the period during and just after the earthquake is higher from the percentage of policies with additional earthquake insurance in the period before it.

He2: The mean number of policies with additional earthquake insurance in Group 2 is larger than the mean number of policies with additional earthquake insurance in Group 3.

- The p value of the t-test of equality is p=0.000, and our conclusion is that the percentage of policies with additional earthquake insurance in the period during and just after the earthquake is higher from the percentage of policies with additional earthquake insurance in the period two months after it.

He3: The mean number of policies with additional earthquake insurance in Group 1 and Group 3 are equal.

- The p value of the t-test of equality is p=0.279. Therefore we may conclude that the percentage of policies before and after the earthquake are equal.

Once again we want to see the effect of the growth of policies on the companies' profit. The situation here is opposite than the flood situation. In fact if we take the total earthquake premiums in the Group 2, T'_2, and the total earthquake premiums in in Group 1, T'_1, than the difference $T'_2 - T'_1$ is more than 27 times larger than the liquidated damage caused from the earthquake from September 11. This indicates that the people are much more afraid of earthquakes, probably because they are less predictable or because their effect is felt by more people. But the companies have a lot more profits after such a disaster.

V. CONCLUSION

Our analysis shows that the Macedonians are insured more when they feel threatened by some natural disaster, and moreover, that they have short-lived memory about the disaster. The fact that a statistically significant incensement of insurance policies follows after a disasters closer to a big city, we can conclude that the people start to think about insuring their property when they feel the disaster. On the other hand, the tendency to insure themselves after an earthquake is much greater than after a flood. Moreover, our data showed not only more significant growth of number of insurance policies than after the catastrophic floods, but also noticeable profit for the insurance companies.

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] M. Gurtler, M. Hibbel, C. Winkelvos, "The Impact of the Financial Crisis and Natural Catastrophes on CAT Bonds", The Journal of Risk and Insurance, pp. 579–612, 2014.

[2] F. Barthel, E. Neumayer, "A trend analysis of normalized insured damage from natural disasters", Climatic Change, pp. 215–237, 2012

[3] L. de la Cruz Tendero, How the insurance sector found success following disaster in the Philippines, World finance, 2014

[4] http://www.aso.mk/index.php?option=com_fjrelated&view=fjrelated&id=0&Itemid=91&lang=en

[5] http://www.triglav.mk/mk/osiguruvanje/

[6] http://www.eurolink.com.mk/page_cat_en.asp?mID=&cID=1

[7] http://earthquaketrack.com/quakes/2016-09-11-13-10-07-utc-5-3-10

[8] http://cuk.gov.mk/mk/odnosi-so-javnosta/mesecni-bilteni.html

Sharing economy

Vase Pandev, Smilka Janeska Sarkanjac
Faculty of Computer Science and Engineering - FINKI
Skopje, Republic of Macedonia
pandev.vase93@gmail.com
smilka.janeska.sarkanjac@finki.ukim.mk

***Abstract* — By definition, sharing economy is a socio-economic ecosystem built around the sharing of human, physical and intellectual resources. It includes the shared creation, production, distribution, trade and consumption of goods and services by different people and organizations. The internet technologies provided platforms for sharing businesses models. Also the social, consumption, living and economic lifestyles of the Millennials strongly support this new economy.**

It is argued that sharing economy is one of the concepts that will change the world. This paper offers a brief review of the sharing economy definitions, key features, drivers, critics, and e-business models, its pioneer businesses on the global market and the first several sharing business models in the Republic of Macedonia.

Keywords — sharing economy; P2P economy; business model; developing countries; Republic of Macedonia

I. INTRODUCTION

The sharing economy concept is a way of direct exchange of goods and services using online market places over the Internet. This concept, also known as "collaborative consumption", "on-demand economy," "gig economy," "access economy", "access-to-excess economy" or "peer-to- peer (P2P) economy"—refers to a hybrid market model of peer-to-peer exchange and has noticed extremely large growth in the last few years.

Scholars and practitioners agree on little about this economic model, as it has provoked fierce controversy. On one side of the spectrum are the sharing economy ambassadors, claiming that this is an economic model of the future, innovative, transformative and choice-enhancing, also altruistic, communal, and environment-friendly. On the other side, its opponents argue that it is an enjoyable way for greedy capitalists to monetize the desperation of people in the post-crisis economy while sounding generous; others file lawsuits concerning labor law violations or zoning regulations, confront P2P companies with huge demonstrations across the globe, and pressure governments for direct regulation. Government responses vary from non-intervention, to creating new regulatory regimes, to sporadically cracking down on some of these services, to complete bans [1].

In 2015, PwC conducted large-scale survey [2] on the sharing economy and proclaimed it's here to stay. According to them, the online sharing economy will be worth $335 billion by 2025, if you consider the offline or physical sharing economy it comes to more than $1 trillion. Several P2P companies are getting close to the largest incumbent competitors on traditional parts of their markets. AirBnB, for example, is worth about $10 billion, getting close to Hilton ($25 billion) and Marriott ($20 billion) or surpassing Intercontinental Hotels Group, an owner of Holyday Inn chain ($9 billion) [3].

The governments around the globe are balancing between the two kinds of potentials of the sharing economy – trying to protect the traditional, incumbent companies on the markets that are threatened by the new P2P companies, while at the same time trying to produce the regulatory environment that will capture potential economic gain from the sharing economy [4].

In that manner, the European Commission, seeking to examine the sector's aggregate economic contribution and the current social and legal state of play regarding the sharing economy in the European Union, in January 2016 has launched a formal assessment of the sharing economy, with a self-explanatory name "The Cost of Non-Europe in the Sharing Economy". Their findings are the following: "The assessment of existing EU and national legislation confirms that there are still significant implementation gaps and areas of poor economic performance. The subsequent examination of areas where it was believed that an economic potential exists highlighted that substantial barriers remain, hindering the achievement of the goals set out in the existing legislation. Moreover, some issues are not or are insufficiently addressed (e.g. status of workers employed by sharing economy service providers). Consequently, more European action would be necessary to achieve the full economic potential of the sharing economy. In doing so, policy-makers should seek to ensure an adequate balance between creative freedom for business and the necessary regulatory protection" [4].

The European Commission prefers to use the expression 'collaborative economy', unlike the U.S. Department of Commerce that uses the term “digital matching firms” [5].

There are scholars that argue that sharing economy will need new metrics. For example GDP, as being classical metric of monetary value of all the finished goods and services produced within a country's borders in a specific time period, takes into consideration only newly produced goods. The sharing economy creates value from existing assets, so it cannot be captured by conventional economics. The questions we need to answer are how can we measure the sharing economy and its benefits? How can we measure both the positive and negative spin offs for the society [6]?

In this paper, we will explain the history of the sharing economy, its features, its advantages and disadvantages, the business models of sharing economy; we will present a few pioneers whose international success is due to this concept, and list several newcomers in the P2P business in Macedonia.

II. THE SHARING ECONOMY

It is not a coincidence that sharing economy started to grow exponentially after the world’s biggest financial and economic crisis in 2008. In 2004, the US President Bush won re-election in part by proclaiming an "ownership society": "The more ownership there is in America, the more vitality there is in America." But, the ownership society, pushed by the major banks and their subprime mortgages and the credit-default swaps, collapsed in 2008. “Ownership hadn't made the U.S. vital; it had just about ruined the country.” [7]

Historically, sharing economy is not a completely new phenomenon. It has its predecessors in bartering from ancient times, and in more recent forms of organizations and activities such as cooperatives, mutual societies, associations and foundations, tontines. These practices have remained from very early times thanks to the working class, poor people and minority communities. The technological predecessors of the sharing economy concept are companies like eBay (1995), Wikipedia (2001), PayPal (1998), Facebook (2004), YouTube (2005) and Couchsurfing and Freecycle (both 2003) etc.

The sharing economy of the 21st century does not innovate the types of services and goods that are exchanged, but rather the way and the scope of doing the exchange. The P2P economy’s innovation lies in its process of connecting consumers and providers—and in the social benefits that the transactions confer. It started as a concept of sharing unused resources between individuals, to later evolve into “consumer-to-consumer” and “supplier-to-consumer” collaboration. Common threads of the sharing economy are disintermediation, the sharing of excess capacity, and increased productivity [8].

The motives for being a part of the sharing economy are different, which is not surprising given the diversity of platforms and activities that the concept offers. Web sites for doing e-business and applying the concept of sharing offer generally lower prices than other market alternatives. The goods and services that are offered can be distributed through the supply chain to producers or consumers very easily and away from the so-called “mediators” and therefore the costs of this type of trade are lower. The main technological enablers of the sharing economy are mainly Internet-related and some of them are: cloud databases, online data analysis, the usage of social media and mobile devices [8].

III. FEATURES OF THE SHARING ECONOMY

The sharing economy as a concept consists of a number of features. We will consider six of them: people, production, values and systems of exchange, distribution, communications and culture [9].

A. People

People are at the heart of a sharing economy; it is a peer-to-peer, person-to-person (P2P) economy. Cooperation is central to this concept; people connect as creators, collaborators, producers, co-producers, distributors and re-distributors. Within business, people – both co- owners, employees and customers – are highly valued, with their opinions and ideas respected and integrated into the business at all levels of the supply chain, organizations and development. Sharing economy is strongly supported by the Millennials, with their social, consumption, living and economic lifestyle. Millennials trust people over brands and value more having the experience over ownership.

B. Production

In a sharing economy, people, organizations and communities as active participants produce or co- produce goods and services collaboratively or collectively or co-operatively. Internet technologies and networks support collective development of products and services, both locally and globally. It is valued that production has positive or minimal environmental impacts, with the available natural resources, not at the expense of the planet.

C. Value and systems of exchange

Value is seen not purely as financial value, but wider economic, environmental and social value. The sharing economy is based on both material and non-material or social rewards and encourages the most efficient use of resources. In a sharing economy, waste has value; it is viewed as resource in the wrong place. It enables ‘waste’ to be reallocated where it is needed and valued. Social responsibility is strongly supported.

D. Distribution

In a sharing economy, resources are distributed and redistributed via a system that is both efficient and equitable on a local, regional, national and global scale. Shared ownership models such as cooperatives, collective purchasing and collaborative consumption are highly valued. Idle resources are re-allocated or traded with those who want or need them to create an efficient, equitable, closed loop or circular system. Recycling, up cycling and sharing the lifecycle of the product are features common to a Sharing Economy.

E. Communications

In a sharing economy information and knowledge is shared, open and accessible. Good, open communications are central to the flow, efficiency and sustainability of this economic system.

F. Culture

The Sharing Economy promotes a collectivist culture where the wider community and the greater good are considered. Business culture is based around the most efficient use of resources. Conscious business, social business, sustainable business, ethical business, social enterprise, business as a force for good are also features of a sharing economy.

IV. ADVANTAGES AND DISADVANTAGES

A. Advantages

Supporters of the sharing economy believe that the concept sooner or later will bring impressive results. They believe that the sharing economy has the potential to become a new socio-economic system that is based on sharing and collaboration and will lead to a fair distribution of values, democratic organized businesses and raising awareness of people by linking them through many different ways and forms of sharing. Also, the companies that apply sharing economy improved in offering lower prices and have the potential to liberate society from so-called "hyper-consumption". The concept offers a new way in the process of sustainability through more efficient utilization of resources, the benefits it brings to the environment by reducing the economic and increase entrepreneurial activity, increasing equity and a fair distribution of goods and services [10].

B. Disadvantages

The main arguments against the sharing economy are an unclear regulatory picture and the lack of a legal framework to regulate the concept, and the disruption of the traditional economy. Also, people who argue against the sharing economy concept only see an opportunity to make personal economic gain and are not interested in using the advantages and the opportunities offered by it. Sharing economy opponents argue that even if there are some altruistic or communal motives among those in the P2P economy, the heart of the industry is financial gain and not altruistic exchanges. All sides to the transaction are motivated, to some extent, by individualism and pursuit of self-interest. Facilitators gain from the increasing number of dealings, as they take a slice of each transaction [10].

V. BUSINESS MODELS OF THE SHARING ECONOMY

A business model consists of many different components that describe the way the company creates, delivers and receives certain values. Key features of the sharing economy business models are: usage of inactive resources, economy of excess capacity, meeting the needs and desires of consumers, fulfilling expectations through dynamic pricing, its regulation and the revolution that can bring [11].

Some of the business models that comprise sharing economy are not profit-based businesses, they are part of the social economy; others are for-profit companies that have higher ethical goals, and they are part of social entrepreneurship. Others are classical for-profit business, as in the example of exchange platforms, that do not have distinct organization, but they share the objects of their activity. The typical transaction in the P2P economy includes three parties: the provider (supplier), the user (consumer), and the facilitator (the website platform). The predominant business models of a sharing economy are: access based models, services, subscription, rental, collaborative and peer- to-peer models. Disruptive innovation, sharepreneurship, creative entrepreneurship, intrapreneurship and micro-entrepreneurship are common features of a sharing economy business models.

VI. PIONEERS OF THE SHARING ECONOMY

Today, the list of businesses that are founded on the concept of sharing economy grows with rapid pace. The website ThePeopleWhoShare.com lists more then 4.000 "amazing services that will transform your everyday life through sharing…" [12] grouped in six categories: Places to stay (603), Transport (807), Communities and Networks (1223), Finance (888), Food and Drink (597), Pets (32).

Fig.1. Sharing economy platforms

In this section we will briefly analyze the several well-known companies that are called "pioneers" of the sharing economy.

1) Uber: It is a platform that serves to connect passengers and drivers. The Uber company does not own any of the vehicles it offers. It works through a mobile app that facilitates the coordination process among the independent drivers and passengers who need transport. The company also have platform called UberRUSH, which is a service that is a form of "supplier of the products". There is also a service called UberBOAT that allows users to request so-called "water taxis" [13].

2) Airbnb: It is a platform for networking and coordinating the people who want to rent property in the short period of time, and those who have property for rent. The so-called "hosts" of the platform offer apartments, rooms, castles, houses and even igloo. On the other hand, people who travel can register on the site and search accordingly. The platform does not own any of the assets offered [13].

3) Open Shed: It is a platform that serves to facilitate the process of renting used tools among individuals. The platform includes various tools for rent such as projectors, machinery and tools for agriculture, trailers, electronic tools and so on. Both parties included in the process of sharing are owners and renters of the tools [13].

4) Zopa: It is an app for the exchange of funds between two parties which operates by "user-to-user" model. Both sides are lenders as well as the ones who lend. This app allows users to skip banks that act as intermediaries in the process of taking a loan and make direct transactions personally i.e. to generate cash loan from individuals, avoiding the traditional and costly banking systems [13].

5) Kickstarter: It is a website that serves to connect investors with potential projects owners. People who offer projects for financing are publishing them on the website, with the aim to come up with funding for their implementation. If the project funded by the investors achieves its goals, all the investors receive a certain reward [13].

6) Airtasker: This platform facilitates the exchange of daily tasks between individuals and businesses. The website of the company allows many individuals or businesses to set tasks that need to be completed, their deadline and price. People who think they can complete the tasks on time, and are satisfied with the price, apply to get the job. There are no prerequisites to apply for a specific task [13].

7) Getaround: A platform which serves to connect people who need a rental car. The app allows users to rent cars from private owners who rent their cars for a certain ammount. Car owners set the rental price, and receive 60% of the amount paid for renting [14].

8) DogVacay: The company through its website allows people to find other people who will take care of their dogs while they are on vacation or busy with work. The people who apply share past experiences and information about themselves, and they are validated through personal interviews [14].

9) Poshmark: It is a mobile and web app that serves as a market of fashion for women. The company goal is selling clothes and other women accessories. The people only need to have profile on the website to start with clothes selling, and those who need it can also buy [14].

VII. SHARING ECONOMY IN MACEDONIA

The concept of the sharing economy is still ascending in Macedonia. There are few platforms, with even fewer users. Some of them are following:

1) Avtostop: A platform that serves as a carpooling aid. The application is good for the people who are traveling with their own vehicle as they will reduce their costs, and for the people who need to reach the same destination in a way cheaper than traditional forms of transport [15].

2) Brainster: A platform for offline classes where the

people can teach and learn skills from digital marketing and design, to entrepreneurship and technology. The platform enables professionals to share their knowledge with the community and each individual to be able to enrich their practical knowledge, using it in their search for work [16].

3) VoziMe: Another platform in Macedonia that is created in order to facilitate and simplify the process of sharing transport. The platform serves to help people save travel costs and to raise awareness about air pollution. It connects people who need cheaper transport to certain destinations with the drivers who offer their vehicles [17].

These platforms have not yet achieved a major success in Macedonia and the main reason for that is lack of budget for running a successful advertising campaign which is a key factor for penetration of existing or create new markets. Brainster can be singled out as the most successful since many young people manage to find their dream job position, and the knowledge they gained was through this platform.

VIII. CONCLUSION

As the sharing economy ambassadors claim, sharing economy is very likely to be an economic model of the future. Some of them claim that it is a part of the Fourth Industrial Revolution that will enable the least developed regions to leapfrog the developed world. Of course, there are even more obstacles for the sharing economy in the developing countries: lack of knowledge and distrust of these new collaborative business models, limited access to finance for these start-ups, few platforms that allow secure payment. In the former socialist countries, there is a culture of "entitlement" - individuals expect the government to solve their or their community's problems, lack of initiative for self-organization and community organization as well as lack of entrepreneurship initiatives.

The developing countries, including Macedonia, should pose the same question as the EU does – what would be the cost to Macedonia if it does not partake in the sharing economy?

REFERENCES

[1] E. Aloni. "Pluralizing the 'Sharing' Economy" Washington Law Review, vol 91, no. 1397, 2016, p. 63.

[2] The Sharing Economy. [Online]. Available: https://www.pwc.com/us/en/technology/publications/assets/pwc-consumer-intelligence-series-the-sharing-economy.pdf. [Accessed: 20 February 2017].

[3] M. A. Cusumano. "How Traditional Firms Must Compete in the Sharing Economy". Communications of the ACM, vol. 58, no. 1, 2015, pp. 32-34

[4] P. Goudin. "The Cost of Non- Europe in the Sharing Economy Economic, Social and Legal Challenges and Opportunities" European Added Value, 2016, p. 206.

[5] R. Telles Jr. "Digital Matching Firms: A New Definition in the Sharing Economy", U.S. Department of Commerce Economics and Statistics Administration, 2016.

[6] Sharing Economy – A global Paradigm Shift. [Online]. Available: https://medium.com/@Cybali/sharing-economy-a-global-paradigm-shift-e88f53b60711. [Accessed: 22 February 2017].

[7] B. Walsh. "Today's smart choice: Don't own share", Time International, vol. 177, no. 12, 2011, p. 45.

[8] J. Schor. "Debating the Sharing Economy", Journal of Self- Governance and Management Economics, vol. 4, no. 3, 2016, pp. 7-22

[9] What Is The Sharing Economy? [Online]. Available: http://www.thepeoplewhoshare.com/blog/what-is-the-sharing-economy/. [Accessed: 24 February 2017].

[10] A. Kosintceva. "Business models of sharing economy companies". Master dissertation. Norwegian School of Economics, 2016.

[11] Fourteen Vital Elements of Sharing Economy Business Model. [Online]. Available: http://www.innovationtactics.com/sharing-economy-business-model/?gclid=CjwKEAjw2qzHBRChloWxgoXDpyASJAB01Io0GvAQllM2DR58k5sBfUU5FN4c9tWdzlro36PbGR1GMRoCIcPw_wcB. [Accessed: 25 February 2017].

[12] The People Who Share. [Online]. Available: http://www.thepeoplewhoshare.com/sharing-economy-guide/. [Accessed: 26 February 2017].

[13] D. Allen and C. Berg. "The sharing economy". Institute of Public Affairs, 2014.

[14] J. Lee. "Six Amazing Pioneers of the Sharing Economy", Tharawat Magazine, Aug. 07, 2015.

[15] Macedonia 2025. [Online]. Available: http://www.macedonia2025.com/new_version/media/single/798/news. [Accessed: 27 February 2017].

[16] Meet the Founder of Brainster – The Platform That Strives to Start a Revolution in Education. [Online]. Available: http://www.huffingtonpost.com/kosta-petrov/meet-the-founder-of-brain_b_8898408.html. [Accessed: 27 February 2017].

[17] VoziMe. [Online]. Available: http://www.vozime.mk/how-it-works. [Accessed: 28 February 2017].

The Level of Maturity of the ISMS in the Private Sector in Albania

Rovena Bahiti
Salijona Dyrmishi
ALCIRT & University of Tirana
Tirana, Albania

Enkeleda Ibrahimi
Corporate Security
Vodafone Albania
Tirana, Albania

***Abstract* — Protection of confidentiality, integrity and availability (CIA) of information and data has become a crucial part of every organization. The numbers of threats and successful attacks have risen dramatically in the latest years and every company has a reason to consider with high priority the protection of CIA of the information. In Albania this is a new challenge and due to the globalization technology has brought, the threats and risks may come from different parts of world. For this reason, Albanian companies have to adapt their policies, procedures, controls and monitoring as per global standards and be prepared for the worst. The implementation of a mature Information Security Management System (ISMS) is the standard approach a company can adapt in order to cover all different aspects of Information Security and ensure all appropriate measures have been taken. Considering the above, we are doing a study in the private sector in Albania trying to measure the level of maturity of the ISMS the companies have implemented. We will provide in this paper the preliminary results of the study, gaps and areas for improvement in order for companies to be compliant to international security standards and local legal and regulatory requirements.**

Keywords* — *Information Security Management System (ISMS), maturity, Albanian companies, compliance.

I. INTRODUCTION

Technology has become a crucial part of every company. Now all the information of any type is saved and processed in information systems. This has raised the concerns abour systems security and governance. In the overall management processes of these companies, the controls over IT actvities and not only has been the new challenge nowadays.

Corporate governance is a set of responsibilities and practices exercised by the board and executive management with the goal of providing direction, ensuring that objectives are accomplished, assuring that risks are managed appropriately and verifying that the enterprise's resources are used with responsibility [1]. Information Security Management System is one of the disciplines of corporate governance focused on the control and management of the technolgical environment.

In this paper we will try to present the importance of the Information Security Management System (ISMS) and level of implementation in Albanian Companies. This paper is organized as follows: Starting with a theoretical chapter regarding Information Security Management System (ISMS) and followed with methodology for evaluating Albanian companies, presenting some preliminary results of the implementation of ISMS adaption.

II. METHODOLOGY

The research types that will be used in this paper is qualitative research and quantitative research. Qualitative researchers aim to gather an in-depth understanding of number of companies in the private sector in Albania that have implemented a Information Security Management System (ISMS). Besides this, the researchers will also examine the level of maturity of the ISMS. Data will be gathered through surveys, inspections and interviews. The results will be provided through statistical analysis processed with SPSS program.

III. INFORMATION SECURITY MANAGEMENT SYSTEMS

The management team encountering security issues inside their organizations need to implement documented procedures and standards to ensure the existence of a safe environment. By ISO definition an ISMS is part of the overall management system, based on a business risk approach, to establish, implement, operate, monitor, review, maintain and improve information security. [2][3]

Implementing a mature Information Security Management System includes six steps:

1. Define the policy
2. Define the scope of ISMS
3. Undertake a risk assessment
4. Manage the risk
5. Select control objectives and control to be implemented
6. Prepare a statement of applicability

The implementation process of ISMS should be inclusive and in order to be effective the management team should create a committee with members from all departments and experts of Information Security. We nominate main five well adopted standards by companies regarding the implementation of ISMS: ISO27001, BS 7799, PCIDSS, ITIL and COBIT. [3]

Companies certified to ISMS standard benefit a high level of image trustworthiness not only on customers perception but also on other stakeholders. It ensures they have a custom model adapted to the company structure and objectives, which applies to all stages of securing information. It also minimizes the impact of any internal and external attacks by prompting a fast reaction. Being that ISMS is not responsibility to only high level managers, it engages everyone in the hierarchy, elevating awareness on the role of information security.

IV. ISMS IN ALBANIAN COMPANIES

Albanian Companies have a very different and unique IT Environment. The new trends and technologies take some time to be implemented in Albania but especially policies, procedures, and controls of the ISMS are difficult to be fully followed by all employees.

It is very important for every company, not only to implement the latest solutions for their systems and networks but also implement a managing system for information security adapted for their size, type of business and the IT Environment they are operating. [4] In this research we have studied data and information for 36 companies operating in Albania (refer to Table I. Companies that will be studied in this research are part of different industries in Albania and use information systems for their daily business activities and operations.

TABLE I. COMPANIES AS PER INDUSTRY

Industry	Nr. of Companies
Telecommunication	4
Financial Services	4
Insurance	2
Retail & Wholesale	18
Manufacturing	1
Other	7

V. PREMINILARY RESULTS

We have made surveys and inspections about their IT Department for the information systems and ISMS in place. We have grouped them and provided the results for each type of industry. After gathering and processing the data we have the results shown in Table II for the implementation of ISMS in these companies as per industry. As per these preliminary results in percentages, 77% of the companies in our study do not have in place an ISMS framework and only 11% of them have implemented it in a high level.

TABLE II. ISMS IMPLEMENTATION

Industry	Companies with ISMS	Level of Implementation
Telecommunication	2	High
Financial Services	2	High
Insurance	0	None
Retail & Wholesale	2	Medium
Manufacturing	0	None
Other	2	Low

VI. CONCLUSION

In conclusion of this short paper it is emphasized the importance of implementing the appropriate ISMS for companies who use the technology in their business as a crucial component of their processes. 77% of the companies in our study result to not have an ISMS implemented for their IT Environment. Based on the security standards and data protection laws and regulations in Albania companies should implement frameworks considering international standards and local factors. The results presented in this paper are only preliminary, the study is in progress and in future publications further data and analysis will be presented.

REFERENCES

[1] IT Governance Online, "ISO 27017 Security in Cloud Computing", ISO/IEC, 2012.

[2] B. Nguyen, "A comparison of the business and technical drivers for ISO 27001, ISO 27002, CobiT and ITIL", 2010.

[3] ISO/IEC, "ISO 27001, Information technology – Security techniques - Information security management systems – Requirements", 2013.

[4] E. Ibrahimi, E. Martiri, "The importance of IT General Controls for Information Security", 6th International Conference "Information Systems and Technology Innovations: inducting modern business solution, 2015.

Agent-based solution of caregiver scheduling problem in home-care context

Aleksandra Stojanova, Natasha Stojkovikj, Mirjana Kocaleva, Saso Koceski
Faculty of Computer Science
"Goce Delcev" University
Shtip, Macedonia
{aleksandra.stojanova, natasa.maksimova, mirjana.kocaleva, saso.koceski}@ugd.edu.mk

***Abstract*—The increased number of elderly, the prolongation of life and the urge of costs reduction have changed the way of how the healthcare services are provided. Nowadays, home care service model has rising popularity. Main challenges of this model are the problems of caregivers' scheduling, their geo-routing, people management etc.**

Despite various technological advances, especially in the fields of Internet services, sensors and Internet Of Things, the problem of caregivers' scheduling is still open and demands for practical solutions. The process of the scheduling of caregivers to elderly people can be considered as Job Shop Scheduling problem, which is NP-hard for solving.

In this paper, we are making analogy of the problem to job shop scheduling problem and we propose agent-bases approach, where corresponding entities are caregivers and elderly people, and they will be represented as agents. The paper proposes a simulation for the problem based on agents using the Anylogic software

Keywords*— *agent-bases model; job shop scheduling; caregiver; elderly.

I. INTRODUCTION

The number of elderly people today is rapidly increasing. This is due to increasing average life age. Demographic aging at first, after the Second World War, was noticed at more developed countries, but nowadays this situation is typical for developing countries too.

According to estimations of the United Nations, every tenth person in the world is an elderly person over the age of 60 years. According to the same data, by 2050 is expected, every fifth person to be an old person over the age of 60 years, and in 2150, every third person to be over the age of 60 years.

The social systems in the countries make an effort helping these people and trying to answer to a larger demand, meeting the needs of elderlies.

Republic of Macedonia is not excluded in this trend of demographic aging. According to population census of 1994, the elderly population over 60 years in Macedonia is 13%, and according to the population census of 2002 elderly population is 15%, and according to population census for 2008, 16.6%. These people have their rights for normal life, but in Macedonia as a developing country, elderly people are a strong risk of poverty and socially isolated group. Therefore, Macedonia is working on National strategy to help these persons with mobile care services, social support and institutional support. However, these services are hardly available because of limited number of caregivers and nursing homes from one side and increasing number of elderly people, from the other.

In order to satisfy the needs of elderly and to provide service of appropriate care, especially healthcare, is necessary to make optimisation of service, taking into account available limited resources and number of elderly.

Therefore, we are trying to offer an optimization method to handle the increasing demand of supplying healthcare services to elderly. That means, by using the limited number of caregivers and making proper scheduling, as more as possible, elder patients to be serviced.

The main problem here is scheduling patients (elderly people) to the available caregivers. Scheduling is the allocation of shared resources over time to competing activities. Our problem is similar to a well-known job shop scheduling problem (JSSP) [1]. Therefore, we are making analogy of this problem to job shop scheduling problem. In addition, we propose new agent-bases approach, where caregivers and elderly people are corresponding entities.

II. JOB SHOP SCHEDULING PROBLEM (JSSP)

The n×m minimum-makespan general job-shop scheduling problem, referred to as the JSSP, can be described by a set of n jobs J_j, where $1 \leq j \leq n$, and each job has to be processed on a set of m machines M_r, where $1 \leq r \leq m$. Each job has a sequence of machines that must be processed. The processing of job J_j on machine M_r is called the operation O_{jr}. Operation O_{jr} requires the exclusive use of M_r for an uninterrupted duration p_{jr}, where p_{jr} is its processing time. A schedule is a set of completion times for each operation $MS_{jr,}$ where $1 \leq j \leq n$ and $1 \leq r \leq m$ that satisfies those constraints. The time required to complete all the jobs is called the makespan *MS*. The scheduling objective is makespan minimization, which means to minimize the completion time of the last operation of any job.

This problem is not only NP-hard, but it, also, is considered as being one of the most computationally stubborn combinatorial problems.

There are different approaches and methods for solving JSSP. Brucker and Schile [1] were the first authors to describe this problem in 1990. They developed a polynomial graphical algorithm for a two-job problem. Several heuristic procedures have been developed in recent years for the JSSP. The methods in this category include dynamic programming and the branch-and-bound method, simulated annealing (SA) and genetic algorithm (GA) [2] [3]. Tabu search [4], and Particle swarm optimization problem [5] [6] are another group of metaheuristic methods used for solving the flexible job shop scheduling problem. Meta-heuristics usually take less time than algorithmic methods to find a good solution for larger problems. However, they do not guarantee optimality.

For better representation of the solution of the problem simulation-based scheduling (SBS) approaches is proposed. Discrete-event simulation is a highly effective tool for modeling complex systems, understanding their behavior over time, and discovering the impact of changes to their configuration. In [7] classical deterministic job shop operations are modelled as a discrete-event simulation model. In this paper, iterative simulation with an optimization approach for solving problem of job shop scheduling is developed. The authors use Linear Programming method for solving scheduling problem. In the simulation model, each machine in the shop is modelled as a unit resource of capacity with an infinite capacity queue in front of it and each job is modelled as an entity, associated with the sequence of machines, which has to visit, and the process time of the job on the machines. If a job arrives to find machine busy, it is placed in a queue. This is a deterministic model with exactly one entity created per job. The entity which successfully seizes the machine begins processing and the rest are put in a queue before the machine, ordered in first-come order. Once the job finishes processing at a machine it proceeds to the next machine as per its sequence. If the job finds the required machine busy, it joins the queue for that machine. Every job proceeds in this manner through the job shop until all the machines in it's sequence are visited. The simulation ends when the last job or entity completes processing at the last machine in it's sequence and the end time indicates the makespan.

In [8] the CHESS algorithm is presented. Here, a series of brief simulation "look-aheads" are used to predict the future impact of scheduling in a particular operation. This approach can be viewed as an advanced application of the dispatch heuristic, where the dispatching rule is to schedule the operation that might be a cause for the some future resource conflicts.

The experimental results of this study offer makespan performance improvements of over 20 percent as opposed to fixed heuristic algorithms. Another simulation approach similar to this is given in [9], where monitoring scheduling performance are improved and also, dynamically adjusting local dispatch parameters are added.

III. Problem Scheduling Caregivers' Services to Patients (Elderly)

Our problem can be mapped to a job shop scheduling problem, where corresponding entities are as follows: Jobs are Patients (elderly people) and Machines are Caregivers (nurses, doctors). There is some similar mapping in [10]. There, an approach for determining a tour for caregiver in a given working day is proposed. In order to optimize multiple criteria, as optimizing caregivers' tours and limiting patients' waited time between two different visits, this new approach is proposed, similar to the problem of finding routes for vehicles, to satisfy all the customers with a minimal travel time, without violating customers' time windows. Where corresponding mappings are customers - patients, vehicles -caregivers, and a warehouse -HCS.

A. Mathematical representation of the problem

Considering job shop scheduling problem, we can formulate scheduling problem of caregivers to patients. Here, scheduling problem consists of m caregivers, which need to serve n elderly people. Let $C = \{C_1, C_2, \ldots, C_m\}$ is a finite set of caregivers and $E = \{E_1, E_2, \ldots\ldots E_n\}$ is a finite set of elderlies. Let x denote the set of all sequential assignments of elderly to caregivers, such that every elderly people is served by each caregiver exactly once. The element $x \in X$, may be written as m x n matrices, in which row i lists the elderly people that caregiver C_i will serve, in order. For example, the matrix

$$x = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix},$$

Means that caregiver C_1 will serve the three elderly people E_1, E_2, E_3, in the order E_1, E_2, E_3 while caregiver C_2 will serve the elderly people in the order E_3, E_1, E_2. Suppose also that there is some cost function MS: $X \rightarrow [0, \infty]$. The cost function, might be interpreted as a total processing time or makespan, and may have some expression in terms of time. MS_{ij}: E x $C \rightarrow [0, \infty]$ is the cost /time for caregiver C_i to serve elderly people E_j. The job-shop problem is to find an assignment of elderly people $x \in X$ such that $MS(x)$ is a minimum, that is, there is no $y \in X$ such that $MS(x) > MS(y)$.

The MIP formulation is often used to model the classical deterministic JSSP [11], i.e. to minimize total processing time MS. MIP model for our problem is as follows:

1) Parameters:

r_{ilk} has a value one, if elderly i requires task l from caregiver k , and zero otherwise.

p_{ik} is servicing time in which an elderly i has to be serviced from caregiver k.

2) Decision variables:

s_{ik} is start time of servicing an elderly i by caregiver k.

y_{ijk} has a value 1 when an elderly j precedes elderly i for caregiver k.

3) MIP model:

The goal is to minimize makespan (to obtain min MS)

$$\sum_{k-1}^{m} r_{imk}(s_{ik} + p_{ik}) \le MS \qquad i = 1,2,\ldots,n \tag{1}$$

$$\sum_{k=1}^{m} r_{ilk}(s_{ik} + p_{ik}) - \sum_{k=1}^{m} r_{i,l+1,k} s_{ik} \le 0, \quad i = 1,2,\ldots,m; \quad l = 1,2,\ldots,m-1; \tag{2}$$

$$K(1 - y_{ijk}) + s_{jk} - s_{ik} \ge p_{ik}, \quad k = 1,2,\ldots m; 1 \le i < j \le n \tag{3}$$

$$Ky_{ijk} + s_{ik} - s_{jk} \ge p_{jk}, \quad k = 1,2,\ldots m; 1 \le i < j \le n \tag{4}$$

Constraint 1 gives the lower bound for the function *MS*. Constraint 2 ensures that the starting time of servicing an elderly *i* with task *l* + 1 is not earlier than its finish time in its predecessor, task *l*. Constraints 3 and 4 ensure that only one elderly is served from caregiver at any given time. The parameter *K* is a large number, sometimes taken as the sum of all processing times.

The MIP model yields optimum solutions for small problem instances, but it's not good model for large problem size.

IV. AGENT BASED APPROACH OF THE PROBLEM

Agent-based models (ABMs) consist of a set of elements (agents) characterized by some attributes, which interact each other through the definition of appropriate rules in a given environment. ABMs can be useful to reproduce many systems related to economics and social sciences, where the structure can be designed through a network. Through ABMs, it is possible implementing an environment with its features, forecasting and exploring its future scenarios, experimenting possible alternative decisions, setting different values for the decision variables and analyzing the effects of these changes [12].

Agents have to interact and communicate among each other. Communication capabilities include the abilities to receive and send messages. This is necessary to ensure a coordination mechanism among agents themselves, in order to prevent and avoid conflicts among agents. In the most general context, agents are both adaptive and autonomous entities who are able to assess their situation, make decisions, compete or cooperate with one another on the basis of a set of rules, and adapt future behaviors on the basis of past interactions [13].

In [14] and [15] is presented another hybrid form of solving job shop scheduling problem combining agent based modeling with heuristic methods like genetic algorithms. Here they parallelizing the genetic algorithms, using agent based modeling, to enhance the performance of these algorithms. Also in [16] Lin and Solberg, proposed an autonomous multi-agent architecture for shop floor dynamic scheduling. Their model represents jobs, resources, and parts by agents. Job agents negotiate with resource agents via a contract net.

Our agent based approach of the problem is consist of two types of agents. One group of agents are elderly people (patients) and another group of agents is caregivers (doctors, nurses). They communicate among each other in a manner that caregivers can provide different services to elderlies. This approach is applied in combination with discrete event simulations mentioned earlier.

There is a similar approach in [17] where integration of discrete event and agent-based simulation is used in order to enhance outpatient service quality in an orthopedic department.

Because we are focusing on job shop scheduling problem and making analogies to our problem, for our needs there is simulation in AnyLogic, solving JSSP problem [18]. This simulation is combinations of discrete event and agent based simulation, which exactly meets our needs.

In this simulation, because it deals with job shop scheduling problem, two types of agents are machines and jobs. Jobs are distributing on the machines and that makes communication between agents. In our case, corresponding agents are caregivers instead of machines and elderly people instead of jobs.

The user interface of the simulation is given in Fig. 1. At first, user can choose the number of caregivers, the number of elderlies and can change service time. The user also can choose which scheduling algorithm to be used in simulation.

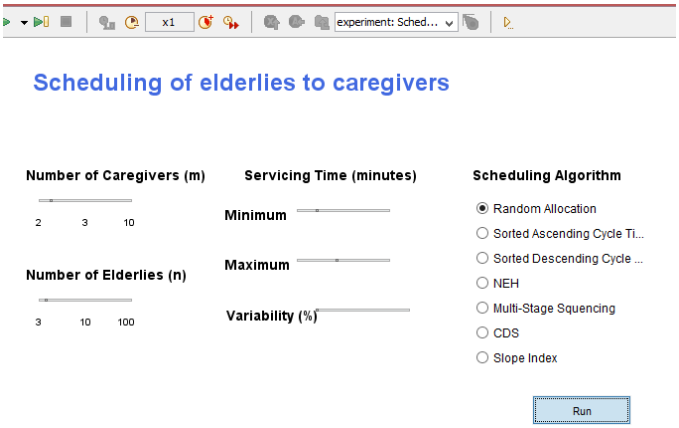


Fig.1. Part of user interface of simulation in AnyLogic.

Here is presented simulator of scheduling n elderlies and m caregivers testing different scheduling strategies. The Caregivers and Elderly people are modeled as agents and can react to changes on the planed schedule.

In Fig 2 is given a screenshot from simulation. Each caregiver is marked by red color if there is elderly in service in the moment, and is marked by green if he/she is free, which means there isn't elderly serviced by him/her in the moment. Each serviced elderly is presented in front of caregiver, in different color. Each caregiver can give service to one elderly in the moment. Simulation ends when all elderly people finished with all services. At the end of simulation, all

caregivers are colored green. How much time each caregiver is busy or free can be noticed in time plot colored chart.

There is also indicators presented in the table of simulation, which changes every moment according to a current situation. There, easily can be seen how much elderlies are in service in the moment, haw much of them finished, and how much are waiting to be served. Total makespan during the whole simulation is 0 and is changed at the end of simulation, representing total time spent for all elderlies to be serviced by caregivers.

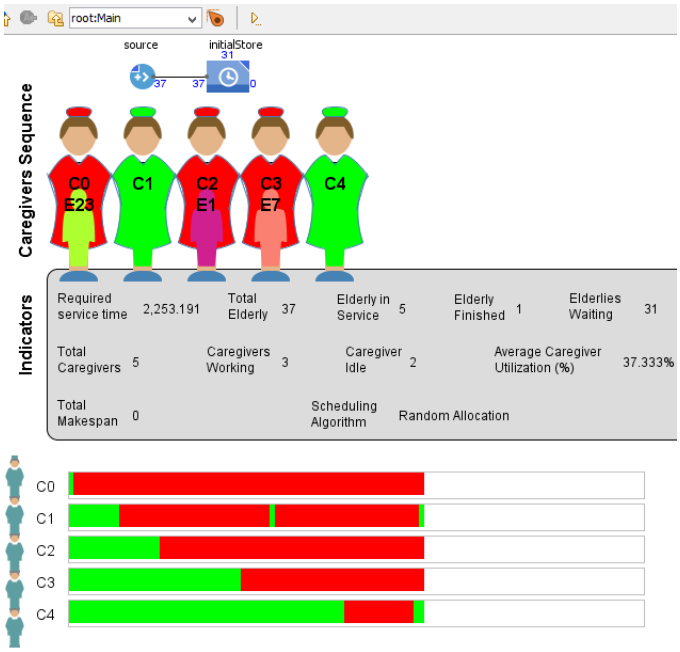


Fig. 2. Screenshot from simulation of scheduling elderlies to caregivers in AnyLogic using agent-based approach and discrete events.

V. Conclusions and Future Work

In this paper, we presented a problem of health care supply to the elderly people. Because of the limited number of caregivers and rapidly increasing number of elderly population, this problem is much more evident. Therefore, finding an optimal solution of the problem is real and necessary need. This problem has many similarities with job shop scheduling problem, therefore we focused on reviewing of different existing solutions of the problem. In addition, we are giving a new approach to the problem using agent based modeling simulation and we demonstrate simulation in AnyLogic.

This presented simulation is adaptation of simulation for job scheduling problem, according to needs of our problem. Our future work will be improving visualizations and functionalities of simulation, and making some improvements in mathematical models in order to obtain better optimization of scheduling. In addition, we are going to add priorities of services and constraints according to the real-life problem needs. We also foresee parallelization of the algorithm in order to be able to cope with large-scale environments.

References

[1] P. Brucker R. Schlie “Job-shop scheduling with multi-purpose machines” Computing. 1990

[2] F. Pezzellaa, G. Morgantia, G. Ciaschettib, “A genetic algorithm for the Flexible Job-shop Scheduling Problem” Computers & Operations Research 35, Elsevier, 2008.

[3] D. Lei, “A genetic algorithm for flexible job shop scheduling with fuzzy processing time”, International Journal of Production Research Vol. 48, No. 10, 15 May 2010, 2995-3013, Taylor & Francis

[4] M. Saidi-Mehrabad , P. Fattahi “Flexible job shop scheduling with tabu search algorithms”, Springer-Verlag London Limited, 2006.

[5] B. S. Girish, N. Jawahar. “A particle swarm optimization algorithm for flexible job shop scheduling problem”, Automation Science and Engineering, IEEE International Conference 2009.

[6] G. Zhang, X. Shao, P. Li, L. Gao, “An effective hybrid particle swarm optimization algorithm for multi-objective flexible job-shop scheduling problem”, Computers & Industrial Engineering 56, Elsevier, 2009.

[7] K. Kulkarni, J. Venkateswaran, “Iterative simulation and optimization approach for job shop scheduling”, Proceedings of the 2014 Winter Simulation Conference, IEEE 2014.

[8] D. J. Toncich, “CHESS: A Methodology for Resolving Scheduling and Dispatch Problems in FMSs”, The International Journal of Flexa'ble Manufacturing Systems, 1996.

[9] R. Lei Sun, H. Xiong Li, Y. Xiong, “Performance-Oriented Integrated Control of Production Scheduling”, IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 36, no. 4, 2006

[10] R. Redjem, S. Kharraja, X. Xie, E. Marcon, “Coordinated Multi-criteria Scheduling of Caregivers in Home Health Care Services”, 2011 IEEE International Conference on Automation Science and Engineering Trieste, Italy, 2011

[11] K. Kulkarni, J. Venkateswaran, “Hybrid approach using simulation-based optimization for job shop scheduling problems”, Journal of Simulation, 2015

[12] C. M. Macal, M. J. North, “tutorial on agent-based modeling and simulation part 2: how to model with agents”, Proceedings of the 2006 Winter Simulation Conference, 2006.

[13] M. R. Friesen, R. D. Mcleod, “A Survey of Agent-Based Modeling of Hospital Environments”, IEEE Access, Volume 2, Pages 227-233, 2014

[14] L. Asadzadeh, K. Zamanifar, “Anagent-based parallel approach for the job shop scheduling problem with genetic algorithms”, Jurnal of Mathematical and Computer Modelling 52,2010.

[15] L. Asadzadeh, “Solving the job shop scheduling problem with a parallel and agent-based local search genetic algorithm”, Journal of Theoretical and Applied Information Technology, 2014

[16] G.Y. Lin, J. J. Solberg. “An agent-based flexible routing manufacturing control simulation system”, Proceedings of the 1994 Winter Simulation Conference, 1994.

[17] C. Kittipittayakorn, K.Ching Ying. “Using the Integration of Discrete Event and Agent-Based Simulation to Enhance Outpatient Service Quality in an Orthopedic Department”, Journal of Healthcare Engineering Volume 2016, Article ID 4189206, 2016.

[18] https://www.runthemodel.com/models/3063/

Automatic POS tagging of Macedonian Language

Martin Bonchanoski

Bitola, Macedonia
martinboncanoski@gmail.com

Katerina Zdravkova
Faculty of Computer Science and Engineering
University Sts Cyril and Methodius
Skopje, Macedonia
katerina.zdravkova@finki.ukim.mk

***Abstract* —This paper presents research work that has led to creating system for automatic disambiguation of the part-of-speech tags for Macedonian language. First, the need for this kind of system is explained. Next, there is given information about the characteristics of Macedonian language. It introduces the pre-processing of the lexical corpus, continues with explanation of the systems for manual tagging and disambiguation of crowd-sourced results. This work has resulted with 96.90% accuracy which is comparable to the state-of-the-art taggers for other languages. The paper contains information about the techniques of machine learning that were applied in order to get these results. The list of models that were built includes TnT tagger, averaged perceptron, model based on neural network implemented by Syntaxnet and model built upon guided learning for bidirectional sequence classification. The final results of this work led to a system that automatically assigns part-of-speech tags to unlabeled text.**

Keywords—POS tagging; disambiguation of crowd-sourced manual tagging; comparison of various POS taggers; Web-based presentation

I. INTRODUCTION

Part-of-speech (abbreviated as POS) tagging is the process of assigning word classes or syntactic categories of lexical items [1]. In many systems, POS tagging is associated with additional morphological information, which carry the particulars about word formation [2]. They lead towards the morpho-syntactic annotation of texts. POS tagging is not trivial since words can play different syntactic roles in different contexts. Information about the word itself and about the context are the keys for building a system that can successfully disambiguate the POS tags.

POS tagging is usually the first step towards further syntactic parsing and processing of texts and speech. It enables sentence parsing, elimination of lexical and functional ambiguities. Practically, it's a process that is required in almost any other task for Natural Language Processing or Understanding (NLP) or (NLU). POS tagging is a required step for preprocessing of the text for information extraction, name-entity recognition and machine translation [1], [3].

This paper presents the full process of creating a very efficient learning system for POS tagging of Macedonian language, with an average accuracy of 96.90%, which is better than that of other languages. Next section presents Macedonian language resources that were used as a basis for current system. It is followed by brief presentation of the algorithms used during the learning stage. Fourth section presents the corpus used during training, and its preprocessing for further manual tagging. It illustrates the manual tagging, which was performed in two independent stages. The comparative analysis of the obtained accuracy of the POS tagger models is presented in the fifth section. The Web-based system that enables tagging of unknown words is presented in the sixth section. The paper concludes with the advantages of the tool for corpora annotation with POS information, and the intentions for its further development.

II. MACEDONIAN LANGUAGE RESOURCES

Macedonian language is a South Slavic language that is spoken by nearly 2.5 million people. This is one of the reasons for the very few research work that has been done in the area of NLP for Macedonian languages.

Macedonian language is part of MULTEXT-East project, which creates multilingual datasets for language engineering research and development. In the earlier stages, morpho-syntactic specifications (MSDs) of the languages which express word-class syntactic information were determined (http://nl.ijs.si/ME/V4/msd/html/index.html). Although the nouns, adjectives and verbs were manually extracted, and the rules for their morphological analysis and synthesis were established, full POS tagging was never done [4]. Vojnovski [5] has also contributed towards the creation of POS tagger for Macedonian language. He built upon the work done in MULTEXT-East project and created a digital annotated corpus that contained Orwell's novel "1984". However, manual disambiguation of the corpus hasn't been done, instead the first available usage of the word was taken.

There has been another attempt to create a POS tagger for Macedonian language. Noemi Aepli's [6] approach is based on multilingual parallel corpora, automatic word alignment, and a set of rules (majority vote). In this work English, Bulgarian, Czech, Slovene and Serbian language were selected as languages closely related to Macedonian. The performance of the tagger trained on the data set that included Orwell's novel "1984" had 88% accuracy.

After all, the authors of this paper are not aware that there exists publicly available part-of-speech tagger for Macedonian at the moment of writing.

III. LEARNING POS TAGGING

Currently, learning of automatic POS tagging is done with the powerful POS taggers' models: TnT tagger [7], averaged perceptron [8], guided learning for bidirectional classification [9], and Synatxnet – neural network that is part of Parsey McParseface [10]. Apart from these models, the bidirectional long short-term memory recurrent neural networks [11] were also examined. Although the last technique is the one of best for other languages, its accuracy for Macedonian language was rather poor, mainly due to extremely small learning corpus. Thus, it won't be presented in this paper nor the results will compared to the other techniques. In the rest of this section all the techniques that were used will be described without going into too much details.

TnT uses second order Markov model for part-of-speech tagging. The states of the model represent tags, outputs represent the words. Transition probabilities depend on the states, thus pair of tags. Output probabilities only depend on the underlying tag. Transition and output probabilities are estimated from a tagged corpus. TnT uses unigrams, bigrams and trigrams to calculate the transition probabilities [7]. One of the downsides of this model is that it can't handle unknown words. Thus, there are few different alternatives how unknown words should be handled. One option is to assign the most-frequent tag to all unknown words. Another possibility is to create a distribution of the frequencies of all tags and use it as a probability distribution to select the tag for the word. However, suffix trees created from the training corpus are one of the most common techniques that are currently used.

Collins has made great results in POS tagging for English language using the averaged perceptron [8]. Later this technique was adopted for many different languages. The pros of this technique are the simplicity, short training time and the great space for improvement by having larger training set [12]. The output depends on large number of binary values used as input vector that contains the features, and depends on the weights for the input vector. The weights are computed using iterative approach. The binary features describe the tag being predicted and its context. They can be derived from any information that is available about the text at the point of decision. In the averaged perceptron, the values of every coefficient are added up at each update, which happens (possibly) at each training sentence, and their arithmetic average is used instead. This makes the algorithm more resistant to weight oscillations during training and as a result, it substantially improves its performance. [12].

Guided learning framework described in [9], which has yielded state-of-the-art results for English and has been successfully applied to other morphologically complex languages such as Icelandic [13] and Bulgarian [14]. The advantage of this technique is the fact that it uses a beam-search to store the most probable tag sequences. Furthermore, the tagging is not unidirectional (left-to-right). For the Macedonian POS tagger presented in this paper, context features proposed by Ratnaparkhi [15] were used.

Syntaxnet is a framework for text processing created-by and open-sourced by Google [10]. It's mainly created for parsing sentences, but it contains a module for POS tagging. It's based on neural networks implemented in Tensorflow. The training time is very short, it gives excellent results and the number of features in the network and the number of nodes in the hidden layer can be easily configured. It uses greedy approach to tag the sequence of words left to right. It automatically creates a dictionary of all words that it has seen in the training set. It also creates prefix and suffix tables to tag words it hasn't seen before. The data should be in Conll format.

IV. DATASET AND DATA PREPROCESSING

In this research work the digitalized version of Orwell's novel "1984" was used as a corpus. The corpus is encoded in XML format according to the rules from Text Encoding Initiative, TEI P4 [16]. The novel is divided into three parts that contain several sections. Sentences are represented with <s> tag, every word is represented in <w> tag and every punctuation mark with <c> tag.

Macedonian corpus consists of 31.538 unique words, almost 38% of them with several MSDs (Fig. 1.). It comprises these 12 categories: nouns, verbs, adjectives, pronouns, articles, adverbs, conjunctions, numerals, particles, interjections, abbreviations and residuals. The frequency of word categories before the disambiguation is presented on Table I.

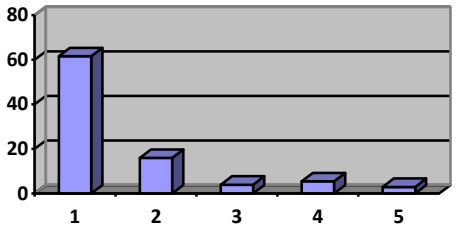


Fig. 1. Percentage of tokens assigned to one, two, up to five different classes

Initial preprocessing was extended with automatic annotation of 40 compound words, such as: "макар што, затоа што, кој било, сè уште" etc. After this step, the number of words that are associated with more than one POS tag was reduced down to 28%.

TABLE I. WORD CATEGORIES ASSIGNED TO WORDS

Word class	Frequency
Nouns	7.043
Verbs	5.015
Pronouns	4.103
Prepositions	4.269
Adjectives	3.740
Conjunctions	3.268
Adverbs	2.420
Particles	1.007
Numbers	513
Proper nouns	40
Abbreviations	28
Interjections	15
Other words	77
Punctuations	5.336

To obtain the possible POS tags for each word, the following three sources were used:

A. The first digital dictionary of Macedonian language (makedonski.info).

B. Aleksandar Petrovski's electronic lexicon for Macedonian language (http://www.ukim.edu.mk/dokumenti_m/2006-Predavanja.pdf#page=305)

C. Lists of the nouns, verbs and adjectives in Orwell's "1984", manually extracted to learn the rules for morphological analysis and synthesis [4]

A set that contains every word form from the corpus was made at the beginning. Then, all possible POS tags for every word were collected from the digital dictionary of Macedonian language (source A). Furthermore, for every word that had adjective as a possible tag, it was again checked in the electronic lexicon from (source B), since most adjectival forms in singular neuter could also be adverbs. Finally, every word in the set was cross-checked in the list of nouns, verbs and adjectives (source C) and in the possible tags that were detected with the previous two methods if these tags were already added as possible tags.

A. *Manual POS tagging of the corpus*

The creation of the system for automatic POS tagging of Macedonian Language was done in four independent stages:

- Creation of interactive tool (GUI) for manual annotation (Fig. 2 and Fig. 3)
- Annotating the text
- Building machine learning models
- Building interactive web application that processes unlabeled text and returns POS tag for each word in the text

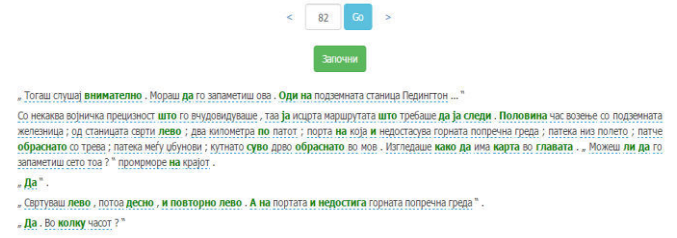


Fig. 2. Tool for manual annotation; ambigouos words are colored in red

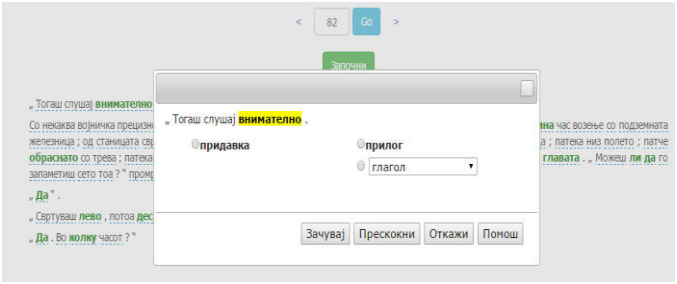


Fig. 3. Easy selection of possible POS tags

Manual POS tagging went through two stages:

- Parallel tagging of all ambiguous words done by two independent persons (the authors of this paper)
- Elimination of tags which were not identical

For both activities, two very interactive tools were prepared, enabling the selection of potential word category. To facilitate the manual tagging, apart from the direct connection to digital dictionary of Macedonian language, a file with explanation agreed among the authors of this paper of some utility and ambiguous words, such as: "како, каде, само, сами, исто, ниту" was also used. As a result of the first POS stage, still about 3% of the words remained without a unique value, due to different opinion of both taggers. They were eliminated in the second stage of the manual POS tagging.

Macedonian language is morphologically-rich language and together with Bulgarian language are the only Slavic languages that don't use cases. To express grammatical relations prepositions are used instead.

In the Macedonian language there are three type of disambiguates that were faced during this work that had to be disambiguated:

- Ambiguity between two word forms with the same lexeme
- Ambiguity between two word forms coming from different lexeme
- Ambiguity between a lexeme and a word form coming from other lexeme

This activity was the most time-consuming of all activities, but the quality of the manual tagging is essential for the results and from the learnt models.

V. RESULTS AND COMPARISON

The learning corpus was divided into two or three sets (training, development and test set), depending on the models that was built. It has undergone the following POS tagger models: TnT tagger, averaged perceptron, guided learning for bidirectional classification, and finally Synatxnet.

The quality of the tagging results are highly-dependent on the fact that Macedonian languages is a moderately inflective language. Also, there's almost free order of the words in a sentence leading to issue due to the small corpus.

For TnT model, two separate experiments were done. First, without any extra model that would tag unknown words, then Suffix Tagger of length 3 was built on the training set and was used to tag unknown words.

To evaluate the results using TnT, 10-fold cross-validation was used. The accuracy of the model is presented in Table II.

To assess the results obtained using the averaged perceptron, 10-fold cross-validation was used again. Various number of iterations to obtain the values of the weights were tested.

TABLE II. ACCURACY BASED ON TNT MODEL

Accuracy	TnT	TnT + Suffix
All words	81.34%	92.38%
Known words	81.34%	96.75%
Unkown words	0%	71.76%

The results are shown in Fig 4. According to Fig. 4 and Occam's razor the weights obtained after 3 iterations were selected. The accuracy with this parameters is 95.16%.

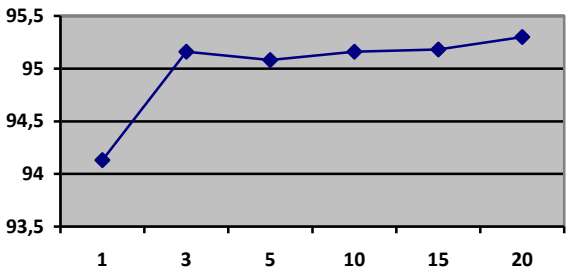


Fig. 4. Accuracy of POS tagging using the averaged perceptron

To be able to use this technique, further processing of the text had to be done for better generalization. Initially, the characters of the every word at the beginning of a sentence were converted to lowercase and replaced all digits with a special marker. They were derived from the template suggested from Ratnaparkhi [15]. The features are presented in Table III.

TABLE III. FEATURES USED FOR AVERAGED PERCEPTRON

Tags	Tag of previous word Tag of pre-previous word Combination of previous two tags
Word forms	Current word Previous word Pre-previous word Next word The word after the next word
Prefixes and Suffixes	Suffixes of the current word (length 1-5) Prefixes of the current word (length 1-3) Suffix of the previous word (length 3) Suffix of the next word (length 3)
Features of the current word	Capitalized

For Syntanxnet, the dataset was split in 3 sets using 70%-15%-15% sets due to the small corpus. Different learning rates were tried and various number of nodes in the hidden layer. The same features as described for the Averaged Perceptron were used. The results are presented on Fig. 3.

With the last technique, bidirectional guided learning framework the best results have been achieved. Beam size of 1, 2 and 3 was tested. The feature templates used in this model include features about the word, the left context, the right context and bidirectional features (Fig. 4.).

TABLE IV. RESULTS BASED ON SYNTAXNET NEURAL NETWORK

Nodes in the hidden layer	Learning rate	Accuracy
128	0.08	93.93%
128	0.1	93.72%
256	0.08	94.04%
256	0.1	93.93%
512	0.08	94.05%
512	0.1	94.57%

TABLE V. RESULTS BASED ON GUIDED LEARNING FOR BIDIRECTIONAL SEQUENCE CLASSIFICATION

Beam size	Accuracy per token	Accuracy per sentences
1	96.74%	52.83%
2	96.82%	50.94%
3	96.90%	52.83%

VI. WEB-BASED TAGGER О' ЗНАЧИ

In order to disseminate the created POS tagger, and to enable its evaluation, an interface leading towards the automatic POS tagger was created. It is available from the link http://bonchanoski.com/postagger/mk/tag. The interface of the system is presented on Fig. 5.

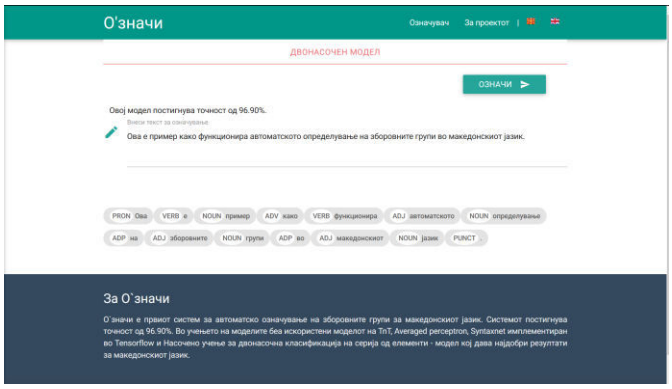


Fig. 5. О'значи – System for automatic POS tagging of Macedonian language

The attempt to tag the sentence "Ова е пример како функционира автоматското определување на зборовните групи во македонскиот јазик." resulted with: "PRON Ова, VERB е, NOUN пример, ADV како, VERB функционира, ADJ автоматското, NOUN определување, ADP на, ADJ зборовните, NOUN групи, ADP во, ADJ македонскиот, NOUN јазик, PUNCT ." The accuracy is actually 100%, proving the quality of the implemented approach.

VII. CONCLUSION AND FURTHER WORK

The efficiency and the accuracy of the POS tagger presented in this paper, although it was based on a very limited corpus is above many commercial taggers, proving the correctness of the approach implemented during its creation. It

encourages further development of the tool, and its extension towards a morpho-syntactic annotator.

The results (96.90% accuracy) in this paper are big improvement to the previous best-known results achieved by Nöemi Aepli (88%). The accuracy achieved is comparable to the state-of-the-arts results for English language (97.55%). Although some tests have been done on the text out of the domain and the results were very satisfactory, the fact that the corpus is very small should be taken into consideration.

The first improvements will be done by feeding the system with a larger corpus. To facilitate the manual POS tagging process, a professional linguist will probably be engaged, to fully annotate a new book from scratch using the extended version of the tool for manual tagging. Then, the most effective annotation algorithms will be implemented again to enable learning of morpho-syntactic annotation of Macedonian language. The second extension of the system will be directed towards automatic morpho-syntactic annotation.

REFERENCES

[1] Jurafsky D., and Martin J. H., "Speech and Language Processing, Chapter 10", November 2016.

[2] Erjavec, T.. "MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora." In *LREC*. 2004.

[3] Martin Frodl. "Part-of-Speech tagging using neural networks" , 2013.

[4] Ivanovska, A., Zdravkova, K., Erjavec, T. and Džeroski, S., "Learning rules for morphological analysis and synthesis of Macedonian nouns, adjectives and verbs" In *Proceedings of 5th Slovenian and 1st international Language Technologies Conference, Jozef Stefan Institute, Ljubljana*, pp. 140-145, 2006.

[5] Vojnovski, Viktor, Sašo Džeroski, and Tomaž Erjavec. "Learning PoS tagging from a tagged Macedonian text corpus." Proceedings of SIKDD 2005 (2005)

[6] Aepli, Noëmi, Ruprecht von Waldenfels, and Tanja Samardzic. "Part-of-Speech Tag Disambiguation by Cross-Linguistic Majority Vote."

[7] Brants, Thorsten. "TnT: a statistical part-of-speech tagger." Proceedings of the sixth conference on Applied natural language processing. Association for Computational Linguistics, 2000.

[8] Collins, Michael. "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[9] Shen, Libin, Giorgio Satta, and Aravind Joshi. "Guided learning for bidirectional sequence classification." ACL. Vol. 7. 2007.

[10] Petrov, Slav. "Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source", 2016., Retrieved on 1st February 2017, from https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

[11] Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991* (2015).

[12] Hajič, Jan, Jan Raab, and Miroslav Spousta. "Semi-supervised training for the averaged perceptron POS tagger." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

[13] Dredze, Mark, and Joel Wallenberg. "Icelandic data driven part of speech tagging." In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 33-36. Association for Computational Linguistics, 2008.

[14] Georgiev, Georgi, Valentin Zhikov, Petya Osenova, Kiril Simov, and Preslav Nakov. "Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 492-502. Association for Computational Linguistics, 2012.

[15] Ratnaparkhi, Adwait. "A maximum entropy model for part-of-speech tagging." Proceedings of the conference on empirical methods in natural language processing. Vol. 1. 1996.

[16] Sperberg-McQueen, C. M. and Burnard, L. (eds.). TEI P4: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Bergen, 2002

A Survey of Text Mining Techniques, Algorithms and Applications

Bojan Ilijoski and Zaneta Popeska
Faculty of Computer Science and Engineering
ss. Cyril and Methodius University
Skopje, Macedonia
bojan.ilijoski@finki.ukim.mk

***Abstract*— In this paper we will give a brief overview of text mining techniques, algorithms and its applications. We will also present some other surveys and recent works in the field of text mining. In the end we are showing some of the fields where text mining can find application.**

Keywords—text mining; survey; text mining techniques; text mining applications

I. INTRODUCTION

The goal of this paper is to present the recent development and improvements of text mining techniques, and to make comparison between them, if it is possible. Text mining is a part of data mining the goal of which is to discover some new information from a given text. The main idea is to collect several documents, combine them, and extract new facts, predications or assertions by using mining techniques. Text mining is interdisciplinary field which includes data mining, information retrieval, statistics, machine learning, computational linguistics, natural language processing (NLP), etc. According to Sumathy and Chidambaram [1] text mining incorporates four major areas shown on figure 1.

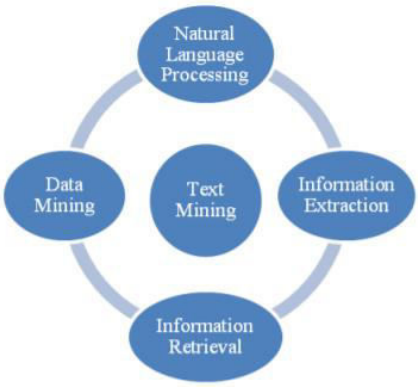


Fig. 1. Areas of text mining [1]

Information extraction (IE) is trying to automatically extract structured information from semi structured or unstructured data. The main purpose is identifying entities (as names of humans, cities, countries, companies, products, etc.), the attributes of the entities and relation between them. Information retrieval (IR) is a process of obtaining resources relevant to information. It provides textual documents, from a document data base, according to user query. The most known IR system are search engines. Natural language processing (NLP) is discipline which studies natural human language. The main idea is to make the human language understandable for computers. Natural Language Processing is concerned with Natural Language Generation (NLG) and Natural Language Understanding (NLU) [1]. Data mining is trying to find a patterns and statistical rules in huge amount of data so that can make predictions and provide understandable structure for further analysis.

We can say that the beginnings of text mining are somewhere in 1980s as text analytics that in the late 1990s emerged as text data mining or text mining. We can conclude that text mining is a relatively new field, but its potential is huge, especially with the rapid development of www which generates a huge amount of text every second. For example, there were 4.66 billion web pages online in March 2016 that contain a lot of text. The importance of this field is supported by the fact that more than 80% of the whole data is stored in text format [2]. Text is still the most common way for exchanging information between people, so a powerful tool for extracting information like text mining is very useful and can be applied to solve or improve solutions to many problems in different fields.

The task of text mining is not easy at all. Unlike other fields or techniques in data mining, which mostly work with structured data, text mining should handle unstructured and semi-structured data as well. The unstructured data might account for more than 70%–80% of all data in organizations [3]. So, this is another challenge for text mining. Before text mining most of the data mining techniques were made to work with structural data which is easier to be understood by computers. Unlike this structural data, natural language was developed to be understood by humans. So, in natural language there might be some slang, spelling variations, dialect that is easy for people to handle with contextual meaning, unlike computers for which it is still inconceivable to understand.

There is more than one challenge that should be solved. Starting from text splitting (tokenization, extraction of words,

etc.), through finding an appropriate representation of text, and then effective application of techniques suitable for text mining. All of those phases are very important. In the first phase of text preprocessing we must be very careful because there is a big possibility to lose data in this process of transformation. We should clean, format and extract meaningful features and all of this should be done by losing little or nothing of the meaning of the unstructured text. In the next phase we should find the right technique that should be applied on data in order to get the answers that we were looking for. According to the work that text mining techniques perform we can group them in several categories as Information Extraction, Topic Tracking, Categorization, Clustering, Concept Linkage, Information Visualization and Association Rule Mining [4].

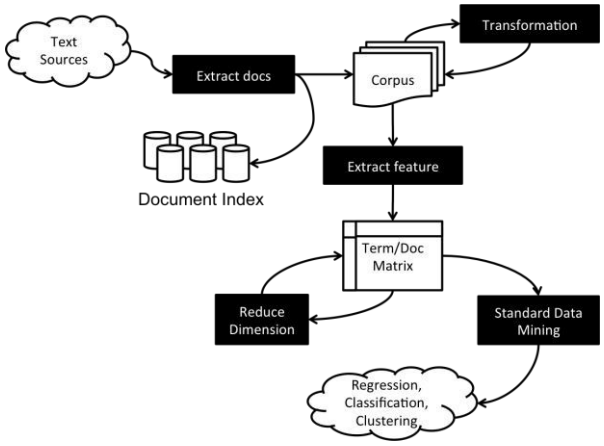


Fig. 2. Common text mining workflow (source https://dzone.com/articles/common-text-mining-workflow)

Because the amount of text data and the ability to detect hidden information in texts, which could previously be hard to spot, the text mining is one of the most popular field of data mining. There are many surveys for text mining techniques and their applications, but most of them explain what text mining is and what subcategories of text mining exist. Also, authors usually mention only the fields where those text mining techniques can be used, but not their purpose. For example publishing and media, telecommunications, energy and other services, information technology sector and Internet, banks, insurance and financial markets, pharmaceutical, research companies and healthcare.

This paper will present some of the most popular and effective techniques for text mining. We will also make a comparison between them (where it is possible) and we will provide information about what technique solves which kind of problems and where it can be used. At the end we will offer the real world applications of the techniques and possible future development of them.

II. RELATED WORK

In addition we will present some of the recent surveys for text mining. Briefly we will describe what they are talking about, what they present to us, which are trending's in text mining, comparisons, etc.

In the "Text Mining Research: A Survey" [5] the authors give us a short introduction in text mining, short description of the process of text mining and a techniques in text mining. In this paper they give a good overview of comparative analysis of information extraction techniques, information retrieval techniques, categorization techniques and clustering techniques. All analysis are presented well with short description for all of them. In the end the authors give very short review of the application of text mining as well as issues in text mining.

"A Comprehensive Study of Text Mining Approach" [6] contains brief introduction to text mining, also a categorization and explanation of text mining techniques (summarization, information extraction, categorization, visualization, clustering, topic tracking, question answering and sentiment analysis). For all techniques there is a description and steps visualization for the processes. Also in this paper are presented some text mining tools, with some information about them, as Text Analyst: natural language text analysis software (Megaputer Intelligence), Intelligent Miner for Text (IBM Software), Text Finder (Paracel Inc.), Text Finder (Paracel Inc.), Vantage Point, Fulcrum Search Server, Isaac and Amberfish, ISIS, AE1, WordSmith Tools and Harvest. This paper present some of applications of text mining as Competitive Intelligence, Detection of Junk Emails, Management of Human Resources, Customer Relationship Management, Multilingual Applications of Natural Language Processing, Classification of NEWS as Text, Classification of Scientific Documents and Sentiment Classification.

"A survey on classification techniques for text mining" [7] give comparison between text mining classification techniques and the authors describe the accuracy of different text classifications as Naive Bayesian, K-Nearest Neighbors, Support Vector Machine, Decision Tree and Regression.

A detail explanation of some text mining metrics is given in paper "A survey on similarity measures in text mining" [8]. The authors present a metrics for string - based similarity divided in two groups, character-based similarity measures (Longest Common Substring algorithm (LCS), Damerau-Levenshtein, Jaro, Needleman-Wunsch, Smith-Waterman) and Term-based Similarity Measures (Manhattan distance, Cosine Similarity, Dice's coefficient, Euclidean distance, Jaccard similarity, Matching Coefficient, Overlap coefficient). Also authors present measures for corpus-based similarity which are present on fig. 3.

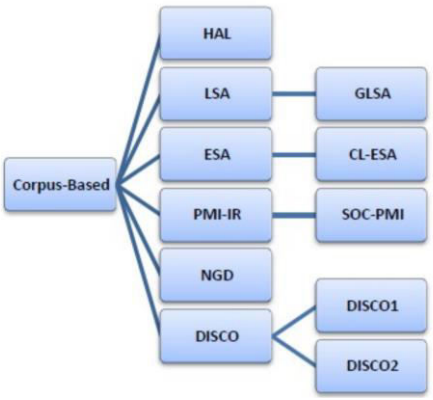


Fig. 3. Corpus-Based Similarity Measures [8]

In this paper are also present measures for Knowledge-Based Similarity which present the similarity between documents based on similarity between words in it and also use some information derived from semantic networks. Some of those measures are Leacock & Chodorow similarity, Lesk similarity, Wu and Palmer similarity, Resnik similarity, Lin metric, etc.

In the paper "Feature Extraction and Duplicate Detection for Text Mining: A Survey" [9] at first is given a good introduction in text mining. There are explained text mining models, applications and benefits of text mining, traffic based events in text mining as fetching and detecting tweets and classifying them. This paper gives a good overview of preprocessing of text mining, with all phases as feature selection and feature extraction. There is a good survey about discovering facets for queries from search result and about duplicate detection and data fusion.

III. Recent Work

In this section we will present the last achievements in text mining field. As we previously mention the text mining is raising field so there are a lot of researchers that do text mining and find a new ways to applicate its techniques.

In "KNN based Machine Learning Approach for Text and Document Mining" [10] the authors compare the techniques for text classification based on naïve bayes, term graph model and k-nearest neighbors. They use well known Reuters – 21579 dataset and get the results shown in table 1.

TABLE I. Accuracy for each method [10]

Category/Method	NAÏVE	Term Graph	KNN
EXCHANGE	74.68	97.41	98.00
ORGANIZATION	51.43	98.23	98.51
PEOPLE	33.19	99.61	99.70
TOPICS	81.80	99.19	99.29
PLACES	72.23	99.19	99.27

One field where text mining usage is increasing is marketing target and market prediction. The quantity of online text rapidly increases so processing of this information can give a good prediction of market gains or losses [11]. This is the reason why a lot researchers direct their work toward different aspects of this problem. This is interdisciplinary concept which includes Linguistics (to understand the nature of language), Machine-learning (to enable computational modelling and pattern recognition), Behavioral-economics (to establish economic sense). This filed includes topics as Efficient Market Hypothesis, Behavioural-economics, Adaptive Market Hypothesis, Markets' Predictability, Fundamental vs. Technical Analysis, Algorithmic Trading, Sentiment and Emotional Analysis. As all other text mining applications the data should go through feature selection, dimensionality reduction and future representation but in this problem there is one more step which make mapping between text procured from social media, blogs, forums, online news, etc. and market data.

The text mining gains special importance when we talk about analysis of user experiences. Because of online shopping this data is increasing and their analysis by human is almost impossible. So there are some linguistics-based methods for analysis of customer experience feedback and accuracy [12].

Also text mining can be used for analyzing and classifying code structures. One recent work [13] shows how text mining can be used for analyzing and classifying code structures in Android malware families. They use extraction of code structures and feature extraction, hieratical clustering, linage analysis and 1-nn classifier to find a malware family. The authors say that this method provides analysis of relationships among families, the existence of common ancestors, the prevalence and/or extinction of certain code features, etc.

The text mining can be also used for mining pharmacovigilance. In "Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art" [14] the authors present application of text mining methodologies in pharmacovigilance data for multiple sources fig 4.

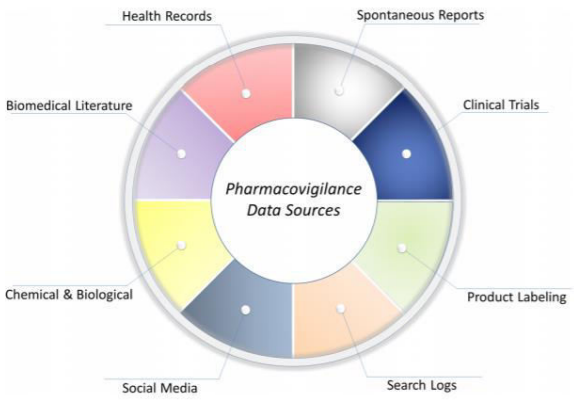


Fig. 4. Data sources currently used or researched to support holistic pharmacovigilance [13]

The flexibility of text mining alows to be applied different tasks in biology and medicine. Combined with other types od evidence it can be used for efficient dictionary-based tagger for named entity recognition of human genes and diseases [15]. Diseases software integrates evidence on disease-gene associations from automatic text mining, manually curated literature, cancer mutation data, and genome-wide association studies.

IV. Applications

Text mining has a lot of applications in different fields. Some of them are more specific and maybe not natural for text mining.

A. Sentiment analosys

Case of natural language processing for identifying the mood of the people about special product. [16][17].

B. Marketing

To study and analyze the needs and desire of the costumers in order to target them and to offer products or service [11][18].

C. Security

Monitoring and investigation resources such as news, blogs, comments, codes, etc. [13][19][20].

D. Medicine

Comparison of drugs, detection of anomalies, information extraction from medical examination, etc. [14][15][21].

E. Company resource planning

Analyzing the employees and resources, reports for activates, monitoring satisfaction levels, investigation of failures etc.[22].

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] K.L.Sumathy, M.Chidambaram, "Text mining: concepts, applications, tools and issues – an overview", International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

[2] V.Gupta, G.S.Lehal, "A survey of text mining techniques and applications", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009

[3] A.Holzinger, C.Stocker, B.Ofner, G.Prohaska, A.Brabenetz, R.Hofmann-Wellenhof, "Combining HCI, natural language processing, and knowledge discovery - potential of IBM content analytics as an assistive technology in the biomedical field", Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data pp 13-24, Volume 7947 2013

[4] K.Thilagavathi, V.Shanmuga Priya, "A survey on text mining techniques", INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS ISSN 2320-7345, Vol.2 Issue.10, Pg.: 41-50 October 2014

[5] R.Janani, S.Vijayarani, "Text mining research: a survey", International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online) 2320-9801, ISSN (Print) 2320-9798, Vol. 4, Issue 4, April 2016

[6] A.Kaushik, S.Naithani, "A comprehensive study of text mining approach", IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.2, February 2016

[7] S.Brindha ; K.Prabha ; S.Sukumaran, "A survey on classification techniques for text mining", Advanced Computing and Communication Systems (ICACCS), 2016 3rd International Conference, 2016

[8] M.K.Vijaymeena, K.Kavitha, "A survey on similarity measures in text mining", Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1, March 2016

[9] R.S.Ramya, K.R.Venugopal, S.S.Iyengar, L.Patnaik, "Feature extraction and duplicate detection for text mining: a survey", Global Journal of Computer Science and Technology: Software & Data Engineering , 2016

[10] V.Bijalwan, V.Kumar, P.Kumari, J.Pascual, "KNN based machine learning approach for text and document mining", International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.61-70

[11] A. K. Nassirtoussia, S. Aghabozorgia, T. Ying Waha, D. Chek Ling Ngob, "Text mining for market prediction: A systematic review", Expert Systems with Applications, Volume 41, Issue 16, 15 November 2014, pp. 7653–7670

[12] F.Villarroel Ordenes, B.Theodoulidis, J.Burton, T.Gruber, M.Zaki, "Analyzing customer experience feedback using text mining: a linguistics-based approach", Journal of Service, 2014

[13] G.Suarez-Tangil, J.E.Tapiador, P.Peris-Lopez, J.Blasco, "DENDROID: A text mining approach to analyzing and classifying code structures in Android malware families", Expert Systems with Applications, 2013

[14] R.Harpaz, A.Callahan, S.Tamang,Y.Low, D.Odgers, S.Finlayson, K.Jung, P.LePendu, N.H.Shah, "Text mining for adverse drug events: the promise, challenges, and state of the art", Drug Safety, October 2014, Volume 37, Issue 10, pp 777–790

[15] S.Pletscher-Frankilda, A.Pallejàa, K.Tsafoua, J.X.Binderc, L.Juhl Jensen, "DISEASES: Text mining and data integration of disease–gene associations", Methods, Volume 74, 1 March 2015, Pages 83–89

[16] Z.H.Khan, M.Atique, V.M.Thakare, "Combining lexicon-based and learning-based methods for twitter sentiment analysis", international Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 2015, pp.89-91

[17] W.Medhata, A.Hassanb, H.Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, Volume 5, Issue 4, December 2014, Pages 1093–1113

[18] K.Berezina, A.Bilgihan, C.Cobanoglu, F.Okumus, "Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews", Journal of Hospitality Marketing & Management, Volume 25, 2016

[19] R.Scandariato, J.Walden, A.Hovsepyan, W.Joosen, "Predicting vulnerable software components via text mining", IEEE Transactions on Software Engineering (Volume: 40, Issue: 10, Oct. 1 2014), pp. 993 – 1006

[20] H.Isah, P.Trundle, D.Neagu, "Social media analysis for product safety using text mining and sentiment analysis", Computational Intelligence (UKCI), 2014 14th UK Workshop, 2014

[21] I.Spasić, J.Livsey, J.A.Keane, G.Nenadićc, "Text mining of cancer-related information: Review of current status and future directions", International Journal of Medical Informatics, Volume 83, Issue 9, September 2014, pp. 605–623

[22] D.Dash Wua, C.Shu-Heng, D.L.Olson, "Business intelligence in risk management: Some recent progresses", Information Sciences Volume 256, 20 January 2014, pp. 1–7

Nutrient - Gene - Disease correlation through the understandings of 'omics'

Miodrag Cekikj , Slobodan Kalajdzhiski
"Ss. Cyril and Methodius" University in Skopje, Faculty of Computer Science and Engineering
"Rugjer Boshkovikj" 16, 1000 Skopje, Republic of Macedonia
cekicmiodrag@gmail.com slobodan.kalajdziski@finki.ukim.mk

***Abstract* - The human being is in the constant drive for learning and applying the acquired knowledge for development, improvement and optimization of life processes necessary to maintain civilization adaptation and survival. The dynamics of the modern way of living constantly highlight the benefits of increased quality and improvements visible within almost all living areas. However, this progressive trend opens a series of crucial topics that are closely related to the generic sense of the price paid by human civilization on account of these benefits. Today, one of the most discussed and popular topics of research interest is nutrients to human health relationship. In this paper we will make a systematic review of the beginnings of genetics, and the development of more modern sciences that allow a wider view of the interaction between genes and nutrients and improving health through diet.**

Keywords: Genomics; Nutrition; Nutrigenomics; Metabolomics; Bioinformatics; Genome; Health; Disease; System Biology; Diet; Nutrient - Gene

I. INTRODUCTION

A growing proportion of the world population is obsessed with healthy food and choice of nutrients that maintain physical and mental health. The results of the latest research very firmly stood in the way of the food industry and in general industrial revolution associated with the cultivation, processing and food production.

The relationship between food and human health is field of interest from the very first beginnings of the development of human civilization. Ayurveda or Ayurveda medicine is complete system of medicine with Indian historical roots. One of the main concepts of this alternative medicine is to provide a guide to diet and lifestyle of the people in order to stay healthy as well as people that can improve their health [1]. Ayurveda dates back to 5,000 years ago, taken from the ancient roots of the words *'ayus'* (life) and *'ved'* (knowledge) and nurture rich and overall appearance of a healthy lifestyle. This set of knowledge rests on the foundations of civilization of the Vedas in India, which implements the basics for life convenient and prescriptive systems [1] [2].

This fact is a clear indication that the people have been interested in the meaning and value of the food for a long time which de facto has been identified as a key element in maintaining a healthy mind and achieving proper life balance.

The human development has seen a number of different phases and periods that characterize the technological and industrial innovations. These findings can be thought of as factors that sustain the progressive trend of development and improving the quality of life. However, the consequences of such an accelerated pace that is more intense and severe in recent years, lead to a sacrifice that is often identified as a destructive phenomenon that disrupts life balance and therefore health.

The announcement of the creation of modern science dealing with the connection between food and health is happening with the definition of Mendelian laws scientifically described as inherited traits are passed from parent organisms to their offspring [3] [4]. Basically, Gregor Mendel, an Austrian priest, biologist, botanist and mathematician, was a scientist who laid the foundations of classical genetics which enables fertile ground for scientific research in the context of the characteristics and development of generations, or descendants. Its conclusions, in combination with medical research, are the first step in starting the process of analyzing the relationship between the genetic material of human factors that influence positive or destructive of its development and stability [5].

Basically, studying the inheritance of genes leads us to the field of genomics, which can be considered as one of the disciplines of genetics. Genomics, which refers to the study of the genome, applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes.

These leads to the field of nutrigenomics which in general is branch of nutritional genomics that research the effects of food on gene expression. The subject of interest is understanding the molecular - level interaction between food nutrients and other dietary bio actives with the human genome [6]. As a science that examines the response of individuals to dietary compounds, food and diets, nutrigenomics is using post - genomic and related technologies identified as genomics, transcriptomics, proteomics and metabolomics [7] [8].

The review will begin with the concept of interaction and interrelation between genes and nutritional values in the context of the meaning and impact of nutrients on metabolic processes in the body. In general, nutrients represent the most influential external factors that directly affect the balance of metabolism (*II. Nutrient - Gene Interactions*).

Then the focus will be placed on the specific field responsible for research of the genome, genomics and its disciplines as transcriptomics, proteomics, metabolomics, system biology. At the core of all these related disciplines is the impact of unbalanced diet propensity and the development of chronic diseases, especially modern ones which are unfortunately occurring more frequent lately. The main emphasis will be put on metabolomics, as a relatively new scientific discipline that focuses on the analysis of metabolites, the metabolome (*III. Genomics*).

In the next section, we will make an analysis and review of Gene - Gene and Gene - Disease interactions in the context of identifying the set of data that can be used for scientific - research work (*IV. Gene - Gene and Gene - Disease Interactions*).

In addition to this, we will present the most important projects, scientific journals and community representatives that are the basis for communication and collaboration of experts and scientists dedicated to the development and results of the field (*V. Projects, Scientific Journals, Community*).

Finally comes the conclusion and further guidance or directions on research that is based on any of the scientific disciplines related to the study of the relationship between nutrients and the potential development of chronic diseases (*VI. Conclusion and Further Guidance*).

II. Nutrient - Gene Interactions

The latest research identified and proved a very strong correlation between food and genes, in particular in the field of impact of new industrial food on the development of new civilization diseases [9]. The large number of publications enable the creation and access to datasets thus enabling visualization of unquestioned link between nutrients and genes.

As we previously mentioned, the genome evolves in response to many types of internal or external environmental factors, including nutrition [10]. According to this, the expression of genetic information can be highly dependent on, and regulated by, nutrients, micronutrients, and phytochemicals found in food [12].

Nutrient - Gene relation in the context of nutrition and lifestyle can be sequenced in positive or negative manner. Therefore, this subject of research can be considered applying the Darwinian model of carcinogenesis (for chronic diseases research purposes) as well as Mechanisms for the modulation of DNA adducts (ex. for the protective role of fruits and vegetables) [11]. As a result of all research papers covering this aspect stems the common complexity related to the interconnection between dietary patterns and individual susceptibility controlled by gene expression. One way of this complexity interpretation can be the fact that a big set of dietary components can alter organism`s genetic properties, events and thereby influence health. [12] In addition to the well known essential nutrients, there is a variety set of nonessential bioactive components that seem to significantly influence health [12]. In fact, nutrients through the cellular sensing mechanisms are considered to be signaling molecules that results in translation of the dietary signals into changes in gene, protein, and metabolite expression [12].

III. Genomics

DNA (Deoxyribonucleic acid), whose molecules are made of two twisted and paired strands, is the chemical compound that contains the instructions needed to develop and direct the activities of nearly all living organisms. In general, each DNA strand is made of four nucleotide bases, which comprise the genetic "alphabet". The order of these bases determines the meaning of the information encoded in that part of the DNA molecule. This means that with its four - letter language, DNA contains the information needed to build the entire human body [13].

The genome is the entire DNA sequence, genetic fingerprint of an organism, that contains all the nucleotide sequences including structural genes, regulatory sequences and non-coding DNA sequences [12]. Genes, as working subunits of DNA refer to the units that carry the instructions for making a specific protein or set of proteins. Each gene contains a particular set of instructions, usually coding for a particular protein or for a particular function [14]. The human genome is estimated to encode up to 30 000 genes, and to be responsible for generating more than 100 000 functionally distinct proteins. The most interesting part of the story is that functions of many thousands of genes are not yet known and strictly researched and identified [12].

Knowing all this information we can now focus on **genomics as the study of the genome, more specific an approach of mapping, sequencing, and analysis of all genes present in the genome**. This study focuses on resolving the variation in the genome between individuals [12]. In general, genomics is a concept that was first developed by Fred Sanger who first sequenced the complete genome of a virus and of a mitochondrion and initiated the practice of sequencing and genome mapping as well as developing bioinformatics and data storage in the 1970s and 1980s. The knowledge about genes that has so far been gathered has led to the emergence of **functional genomics, which is a field that tries to understand the pattern of gene expression, especially across different environmental conditions** [15] [16]. Functional genomics, often referred as system biology, is consisted of a few different methodologies that represents the basis of the biomics science. In general, these different scientific tools, which are followed by ‘omics’ suffix, defines specific science approaches that cover all aspects of the entire human genome. Omics refers to the collective technologies used to explore the roles, relationships, and actions of the various types of molecules that make up the cells of an organism [17].

We already mentioned the field of genomics - a science that focuses on genes and gene`s functions research. Respectively to the covering of DNA sequence, there is another area, **transcriptomics which is focused on the study of the transcriptome**. The transcriptome represents the complete set of RNA transcripts which under specific circumstances are produced by the human genome [18]. In addition to these omics areas, we define the appropriate science approach related to the human proteins. In fact, these are complex macromolecules consisting of one or more long chains of amino acid residues. As a basis of living tissues, they play a central role in biological processes and are required for the structure, function, and

regulation of the body's tissues and organs [19]. We can now introduce the term "proteome" that refers to the entire complement of proteins, including the modifications made to a particular set of proteins, produced by an organism or a cellular system [20]. In the context of proteome, there is new dedicated science area known as proteomics. **Proteomics is a large-scale comprehensive study of a specific proteome, including information on protein abundances, their variations and modifications, along with their interacting partners and networks** [20]. It attempts to characterize all proteins in a biological sample and addresses three categories of biological interest: protein expression, structure and function [12]. The relatively newest area that closes the circle of omics science tools is known as metabolomics. **Metabolomics refers to the systematic identification and quantification of the small molecule metabolic products (the metabolome) of a biological system (cell, tissue, organ, biological fluid, or organism) at a specific point in time** [21]. The metabolome consists of all the low-molecular-weight molecules or metabolites in a cell, tissue, or organism, thereby providing a functional readout of cellular biochemistry. Thousands of metabolites can now be measured quantitatively from relatively small amounts of biological material. In general, we can divide global and targeted metabolomics. On one hand, the global one enables new discoveries linking cellular pathways to biological mechanism in ways not previously suspected, while on the other hand the targeted metabolomics is defined by the identification and quantification of sets of structurally characterized and biochemically annotated metabolites, utilizing current knowledge of most biochemical pathways [7]. In short, metabolomics examines the whole metabolism, which ultimately reflects the behavior of different patterns of genes. It investigates metabolic regulation and fluxes in individual cells or tissues, in response to specific environmental changes [12].

Guided by the latest research in the field we can conclude that the one of the most influential environmental factors is the food and food nutrients. Also, there are a lot of project and research centers that concentrate on studies related to the interaction between modern chronic diseases and nutrients. Metabolomics is making great advances in this complex approach to nutrition research [12]. Metabolomics also has the advantages of offering more immediate information about our metabolism, which is not presented by changes in gene transcription or protein expression since both can occur without apparent metabolic consequences [12].

IV. Gene - Gene and Gene - Disease Interactions

We have to agree that one of the most challenging scientific topics is the gene - gene interaction analysis in the context of detecting the main indicators, triggers and consequences related to chronic human diseases, especially modern ones. Identifying genes with specific DNA sequence variations that increase or decrease susceptibility to disease depend largely on the genetic architecture of the disease, which can be defined as the (1) number of genes that impact disease susceptibility, (2) distribution of alleles and genotypes at those genes, and (3) manner in which the alleles and genotype impact disease susceptibility [22]. There is a big set of different statistical, computational and visualization methods for detecting and characterizing genes with effects that are dependent on other genes, but we are going to emphasize that the most important goal in human genetics and technology potential is to determine which of the multitude of genetic variants are useful for predicting who is at risk for common diseases [22]. Regardless of the technical detecting approach, all of this technique are based on well - known knowledge databases created to support the processes of identifying specific gene - gene interactions and their influence in different environment conditional systems. One interaction network is **GeneMANIA** which finds other genes that are related to a set of input genes, using a very large set of functional association data. In general, association data include protein and genetic interactions, pathways, co-expression, co - localization and protein domain similarity. GeneMANIA can be used to find new members of a pathway or complex, find additional genes you may have missed in your screen or find new genes with a specific function, such as protein kinases [23].

Another very useful tool in this context is **STRING database** of known and predicted protein - protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases [24].

In the processes of analysis and application of these data sets we can take in consideration that the both types of interactions are very tightly coupled. **This means that gene - gene and gene - disease interactions underlay on common concepts and goals, to understand the relation between genes properties and disease activation.** A lot of researchers work on ideas where main hypothesis is that the most central genes in an interaction network for a disease are likely to be related to the disease [25]. In light of scientific developments, one of the most well - known databases that stores gene - disease associations are **Online Mendelian Inheritance in Man (OMIM, 2007)** [26]. In fact, the content provides summaries of publications about gene - disease relationships. This Online Catalog of Human Genes and Genetic Disorders is another helpful tool in direction of achieving the challenges related to the fact that the most of the relevant information remains hidden in the unstructured text of the published papers [25].

In addition to all of this it is worth highlighting **DisGeNET** - a discovery platform integrating information on gene-disease associations (GDAs) from several public data sources and the literature [27]. Besides the large number of GDAs compiled in DisGeNET, the platform provides a score in order to rank the associations based on the supporting evidence. Moreover, DisGeNET can be queried through Search and Browse functionalities available from this web interface, or by a plugin created for Cytoscape to query and analyze a network representation of the data [27].

Another human disease database is **MalaCards** - an integrated database of human maladies and their annotations, modeled on the architecture and richness of the popular **GeneCards database** of human genes. The MalaCards disease and disorders database is organized into "disease cards", each integrating prioritized information, and listing numerous known aliases for each disease, along with a variety of annotations, as

well as inter - disease connections, empowered by the GeneCards relational database, searches, and GeneAnalytics set-analyses. Annotations include: symptoms, drugs, articles, genes, clinical trials, related diseases/disorders and more. An automatic computational information retrieval engine populates the disease cards, using remote data, as well as information gleaned using the GeneCards platform to compile the disease database. The MalaCards disease database integrates both specialized and general disease lists, including rare diseases, genetic diseases, complex disorders and more [28]. All of these tools are developed with just one purpose, to enrich the existing real world data set of gene - gene and gene - disease interactions and help the process of identifying the key factors related to the circumstances necessary for activating specific human disease or degeneration.

V. PROJECTS, SCIENTIFIC JOURNALS, COMMUNITY

Rapid technological development introduces the relatively new scientific areas which main research topic is related to the civilization global health. Using existing scientific knowledge in combination with existing modern technologies (as bioinformatics, data mining, big data, data visualization, database engines) leads us directly to the core fundamentals of the interactions between genes, nutrients and diseases. In the following section, we are going to present number of projects, organizations, scientific journals as well as online community focused on researching this area.

One of the most important project definitely is **The Human Genome Project (HGP)** which was led at the National Institutes of Health (NIH) by the National Human Genome Research Institute, produced a very high - quality version of the human genome sequence that is freely available in public databases. The sequence is not that of one person, but is a composite derived from several individuals. Therefore, it is a "representative" or generic sequence. To ensure anonymity of the DNA donors, more blood samples (nearly 100) were collected from volunteers than were used in this process. The Human Genome Project was designed to generate a resource that could be used for a broad range of biomedical studies. One such use is to look for the genetic variations that increase risk of specific diseases, such as cancer, or to look for the type of genetic mutations frequently seen in cancerous cells. More research can then be done to fully understand how the genome functions and to discover the genetic basis for health and disease [29]. This project was finished in 2003 with the publication of the complete human sequence [30].

The **BIOCLAIMS project** attempts to identify new "biomarkers" of the effects of food and food components on health, based on the new biological technologies, in particular, those of Nutrigenomics that will contribute scientific bases for the reshaping of the European Legislation on "Health claims made on food" (EU Regulation 1924/2006). This legislation, which is the subject of much controversy, entered into force in 2007 with the aim of solving the situation of anarchy, confusion and misleading around advertising statements made on food in relation to health. This significant project was coordinated by Professor Andreu Palou, from the University of the Balearic Islands and CIBER Physiopathology of Obesity and Nutrition (Spain), and involves scientists from 11 institutions from seven European countries. It started on 1st March 2010 and finished five years later 28th February 2015 [31].

NuGO - is an Association of Universities and Research Institutes focusing on the joint development of the research area of molecular nutrition, personalized nutrition, nutrigenomics and nutritional systems biology. NuGO evolved from an EU Sixth Framework Network of Excellence, and since 2010 the NuGO Association has taken over some of the activities of the Network of Excellence (NoE NuGO) NuGO Association is now expanding to global dimensions [32] [33].

The Joint Irish Nutrigenomics Organization (JINGO) Project is an Irish Government - funded initiative which has been running since 2007. By combining dietary, physical activity, body measurement and lifestyle data with cutting - edge nutrigenomics technology, a National Nutritional Phenotype Database of 7,000 people has been created [34].

The European Society of Human Genetics is a non-profit organization. Its aims are to promote research in basic and applied human and medical genetics, to ensure high standards in clinical practice and to facilitate contacts between all persons who share these aims, particularly those working in Europe. The Society will encourage and seek to integrate research and its translation into clinical benefits and professional and public education in all areas of human genetics. **The European Journal of Human Genetics** is the official publication of the European Society of Human Genetics, published monthly [35].

Karger - Medical and Scientific Publishers (Connecting the World of Biomedical Science) - the emerging field of nutrigenetics and nutria - genomics is rapidly gaining importance, and this new international journal has been established to meet the needs of the investigators for a high - quality platform for their research. Endorsed by the recently founded **'International Society of Nutrigenetics/Nutrigenomics' (ISNN)**, the Journal of Nutrigenetics and Nutrigenomics welcomes contributions not only investigating the role of genetic variation in response to diet and that of nutrients in the regulation of gene expression, but is also open for articles covering all aspects of gene - environment interactions in the determination of health and disease. Original papers and reviews cover the genetic basis for the variable responses to diet and lifestyle factors in chronic conditions (e.g. cardiovascular disease, obesity, diabetes, cancer), methods to assess gene - environment interactions and other related relevant topics, with research drawing from both human and animal studies [36] [37].

Nutrients (ISSN 2072-6643; CODEN: NUTRHU) is an open access journal of human nutrition published monthly online by MDPI. The Nutrition Society of New Zealand, Australasian Section - American Oil Chemists Society (AAOCS) and Asia Pacific Nutrigenomics Nutrigenetics Organization (APNNO) are affiliated with Nutrients [38].

Journal of Nutrition & Food Sciences - OMICS International through its Open Access Initiative is committed to make genuine and reliable contributions to the scientific community. OMICS International hosts over 700+ leading - edge peer reviewed Open Access Journals and organizes over 3000 International Conferences all over the world. OMICS

International journals have over 15 million readers and the fame and success of the same can be attributed to the strong editorial board which contains over 50000 eminent personalities that ensure a rapid, quality and quick review process. OMICS International signed an agreement with more than 1000 International Societies to make healthcare information Open Access. OMICS International Conferences make the perfect platform for global networking as it brings together renowned speakers and scientists across the globe to a most exciting and memorable scientific event filled with much enlightening interactive sessions, world class exhibitions and poster presentations. [39]

VI. CONCLUSION AND FURTHER GUIDANCE

In this article, we surveyed the most trending science areas that cover the interaction and dependencies between human genome and food nutrients. We presented nutrients as probably the most important external environment factor that is in direct correlation with genes processes.

Today, there are a lot of research centers, organizations, academic institutes as well as individual or groups of scientists that continuously explore all of the mentioned areas and related science methodologies. This means that each new day we are witnesses of new set of data and directions that lead us to even more topics of interest and research. According to this, the main goal is still the same, to find consistent and well controlled model for data processing and appropriate systematization.

This survey may serve as guide or initial reference point for all researchers involved in the process of interconnecting and visualizing all available sources of information in order to achieve their point of research. In general, the content is divided into separate sections describing the various disciplines that solve relevant complex issues and conditions which are very closely related to the human health and further human survival.Our future points of interest are the detailed analysis of nutrient - gene, gene - gene and gene - disease interactions and different techniques and approaches for data mining, visualization, processing and correlation in order to get understanding and better knowledge for nutrient - disease relationship and dependencies.

REFERENCES

[1] Steven D. Ehrlich, June 2015, ‘Ayurveda’, ‘Complementary and Alternative Medicine Guide’, Complementary and Alternative Medicine Guide, [http://umm.edu/health/medical/altmed/treatment/ayurveda] last visited: 17 August 2016

[2] Dr. John Douillard, ‘What is Ayurveda?’, LifeSpa Natural Health and Ayurveda, [http://lifespa.com/about-lifespa/ayurveda/what-is-ayurveda/], last visited: 12 October 2016

[3] H Ilona Miko, 2008, ‘Gregor Mendel and the Principles of Inheritance’, Nature Education, [http://www.nature.com/scitable/topicpage/gregor-mendel-and-the-principles-of-inheritance-593], last visited: 20 October 2016

[4] Dr. Dennis O'Neil, ‘Mendel's Genetics’, ‘Basic Principles of Genetics: An Introduction to Mendelian Genetics’, Behavioral Sciences Department, Palomar College, San Marcos, California, United States, [http://anthro.palomar.edu/mendel/mendel_1.htm], last visited: 22 October 2016

[5] Ph.D. Laura Rivard, ‘The Mendelian Concept of a Gene’, Mendelian Genetics, [http://knowgenetics.org/mendelian-genetics/], last visited: 25 October 2016

[6] Maxine Weinstein, James W. Vaupel, and Kenneth W. Wachter, Editors, 2007, ‘Biosocial surveys’, Committee on Advances in Collecting and Utilizing Biological Indicators and Genetic Information in Social Science Surveys, e-book, [https://www.nap.edu/catalog/11939/biosocial-surveys], last visited: 10 August 2016

[7] Lynnette R. Ferguson, 2013, ‘Nutrigenomics and Nutrigenetics in Functional Foods and Personalized Nutrition’, e-book, [https://www.crcpress.com/Nutrigenomics-and-Nutrigenetics-in Functional-Foods-and-Personalized-Nutrition/Ferguson/p/book/9781439876800], last visited: 20 August 2016

[8] Nutrigenomics New Zealand, Crown Research Institutes, [http://www.nutrigenomics.org.nz/], last visited: 6 August 2016

[9] Anita Shupe, 2014, ‘Truths and lies about food’, ISBN: 978-608-230-238-6

[10] Jung Kyoon Choi, Sang Cheol Kim, Apr 2007, ‘Environmental Effects on Gene Expression Phenotype Have Regional Biases in the Human Genome’, ‘Genetics - Genetics Society of America’ [http://www.genetics.org/content/175/4/1607]

[11] Vineis P, DNA adducts and the protective role of fruits and vegetables, Institute for Scientific Interchange, Torino, Italy, [https://www.iarc.fr/en/publications/pdfs-online/prev/sp156/sp156-ch9.pdf], last visited: 14 October 2016

[12] Riitta Törrönen, Marjukka Kolehmainen, Kaisa Poutanen, November 2006, ‘Nutrigenomics - new approaches for nutrition, food and health research’, Food and Health Research Centre, ETTK - Department of Clinical Nutrition/ Kuopio Centre of Expertise in Wellbeing, [http://kongzl.ind.ntou.edu.tw/Nutrigenomiikkaraportti.pdf], last visited: 25 October 2016

[13] National Human Genome Research Institute, [https://www.genome.gov/18016863/], last visited: 20 December 2016

[14] Dr Ananya Mandal, Oct 2014, ‘What are Genes?’, News Medical Life Sciences,[http://www.news-medical.net/life-sciences/What-are-Genes.aspx], last visited: 14 November 2016

[15] Dr Ananya Mandal, Jul 2014, ‘What is Genomics?’, News Medical Life Sciences,[http://www.news-medical.net/life-sciences/What-is-Genomics.aspx], last visited: 14 November 2016

[16] World Health Organization, 2005, Genetics, genomics and the patenting of DNA, Human Genetics Programme, [http://www.who.int/genomics/en/FullReport.pdf], last visited: 16 November 2016

[17] PhD. Sherry L. Ward, Jul 2014, ‘Omics, Bioinformatics, Computational Biology’, AltTox non-animal methods of toxicity testing - Emerging Technologies, [http://alttox.org/mapp/emerging-technologies/omics-bioinformatics-computational-biology/], last visited: 12 September 2016

[18] Zhong Wang, Mark Gerstein, Michael Snyder, Jan 2009, ‘RNA-Seq: a revolutionary tool for transcriptomics’, PMC - U.S. National Institutes of Health's National Library of Medicine (NIH/NLM), [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/]

[19] Genetics Home Reference, January 2017, ‘What are proteins and what do they do?’, How Genes Work, U.S. National Library of Medicine, [https://ghr.nlm.nih.gov/primer/howgeneswork/protein]

[20] National Cancer Institute - Office of Cancer Clinical Proteomics Research (OCCPR), ‘What is Cancer Proteomics?’, ‘What is Proteomics’, [https://proteomics.cancer.gov/whatisproteomics], last visited: 27 December 2016

[21] Nature Research, ‘Metabolomics’, ‘Metabolomics’, Nature Research, [http://www.nature.com/subjects/metabolomics], last visited: 5 January 2017

[22] Diane Gilbert-Diamond, Jason H. Moore, Jul 2014, ‘Analysis of Gene-Gene Interactions’, U.S. National Institutes of Health's National Library of Medicine, [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4086055/]

[23] GeneMANIA, Donnelly Centre for Cellular and Biomolecular Research - University of Toronto, [http://genemania.org/], last visited: 27 October 2016

[24] Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), Academic Consortium, [http://string-db.org/], last visited: 28 October 2016

[25] Arzucan Özgür, Thuy Vu, Güneş Erkan, Dragomir R. Radev, Jul 2008, 'Identifying gene-disease associations using centrality on a literature mined gene-interaction network', PMC - U.S. National Institutes of Health's National Library of Medicine (NIH/NLM), [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718658/]

[26] Online Mendelian Inheritance in Man (OMIM), Jan 2017, 'An Online Catalog of Human Genes and Genetic Disorders', 'Human Genetic Knowledge for The World', [https://www.omim.org/]

[27] DisGeNET, Jun 2016, Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics (GRIB) IMIM-UPF, [http://www.disgenet.org/web/DisGeNET/menu/home]

[28] MalaCards - The human disease database, Dec 2016, Weizmann Institute of Science, [http://www.malacards.org/]

[29] National Human Genome Research Institute, Aug 2015, 'DNA, Genes and Genomes', 'A Brief Guide to Genomics', [https://www.genome.gov/18016863/], last visited: 24 November 2016

[30] Human Genome Project Information Archive, 1990 - 2003, U.S. Department of Energy (DOE) and the National Institutes of Health, [http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml]

[31] Professor Andreu Palou, BIOCLAIMS project, 2010 - 2015, University of the Balearic Islands and CIBER Physiopathology of Obesity and Nutrition (Spain), [http://bioclaims.uib.eu/], last visited: 20 December 2016

[32] European nutrigenomics organisation - linking genomics, nutrition and health research(NuGO), Association of Universities and Research Institutes, [http://www.nugo.org/], last visited: 10 December 2016

[33] European nutrigenomics organisation - linking genomics, nutrition and health research(NuGO), Functional genomics in relation to food, nutrition and health, 2004 - 2010, Wageningen Universiteit Netherlands, [http://cordis.europa.eu/project/rcn/74154_en.html], last visited: 10 December 2016

[34] The Joint Irish Nutrigenomics Organisation (JINGO), UCD Institute of Food & Health, School of Agriculture, Food Science & Veterinary Medicine, College of Life Sciences, UCD, Belfield, Dublin 4, Ireland, [http://www.ucd.ie/jingo/], last visited: 15 December 2016

[35] The European Society of Human Genetics, [https://www.eshg.org/], last visited: 2 November 2016

[36] Karger - Medical and Scientific Publishers, Karger Publishers in Basel, Switzerland, [http://www.karger.com/], last visited: 4 December 2016

[37] PhD. Alfredo Martínez, International Society of Nutrigenetics / Nutrigenomics (ISNN), 2005, [http://www.nutritionandgenetics.org/], last visited: 17 October 2016

[38] Nutrients - Open Access Human Nutrition Journal, The Nutrition Society of New Zealand, Australasian Section - American Oil Chemists Society (AAOCS) and Asia Pacific Nutrigenomics Nutrigenetics Organisation (APNNO), [http://www.mdpi.com/journal/nutrients], last visited: 25 December 2016

[39] OMICS International Open Access Journals, [https://www.omicsonline.org/open-access-journals-list.php], last visited: 3 January 2017

Comparison of string matching based algorithms for plagiarism detection of source code

Tomche Delev*, Dejan Gjorgjevikj†
Faculty of Computer Science and Engineering
University of St. Cyrill and Methodius
Skopje, R. Macedonia
Email: *tdelev@finki.ukim.mk, *dejan@finki.ukim.mk

***Abstract*—In this paper we present the implementation and comparison of several string matching based algorithms for the purpose of comparing source code files written in the C programming language for similarity or potential plagiarism. We describe the complete pipeline of the algorithm with the tokenization phase, comparison and results representation. Then, we present the results of the comparison of the selected string matching algorithms in the task of comparing a selected dataset of source files submitted by students on the exercises and exams of the course in Structured programming. Finally, the differences, advantages and disadvantages of the selected algorithms are discussed.**

***Index Terms*—string matching, plagiarism, novice programmers.**

I. Introduction

Increased enrollment in introductory programming courses and the rise of popularity of massive open on-line courses (MOOC) introduced new challenges in handling large group of students and effectively assessing their learning. The solutions implemented by most educational institutions, including our faculty, uses some kind of automatic assessment tool or system. These systems are used to automatically assess the correctness of a submitted solution, without any human input. The system Code [1] is one example of such system, where students solve programming problems in different programming languages, and their solutions are automatically tested for correctness, by executing them and comparing their output to the expected output of a correct solution. However, this kind of assessment, although easy to implement and very helpful to the staff, alone is still not sufficient for final assessment and evaluation. One of the major problems it faces, is that it doesn't check for similarity and potential plagiarism of the submitted solutions. In environment such as laboratory exercises or homework, when students can not be restricted in their behavior and access to other resources on Internet or communication with other peers, they can submit someone else's solution as their own. This is one of the main form of plagiarism, serious academic dishonesty, where students use complete or partial work of other student as their own personal work. It is a serious problem that affects the regularity of the automatic assessment made by the system, but also the final assessment in general.

Integral part of most automated plagiarism detection systems is the comparison of two source files. In this paper several string matching based algorithms for calculation of similarity or distance between two strings are compared and evaluated. In order to compare source files with string matching algorithms effectively, the source code is first tokenized and this is described in the third part. The string matching algorithms are presented in the fourth part. The results of the comparison are presented and discussed in the fifth part, and finally a short conclusion of the work is presented in the final part.

II. Source code plagiarism

Plagiarism can be defined as the act of unacknowledged using or copying of portions or complete documents. This definition applies also in the context of source code, where the documents are in the form of source code written in some programming languages. Plagiarized program is either an exact copy of the original or a modified version using different types of modifications [2]. Source code modifications can be classified into two main categories: *lexical modifications* and *structural modifications* [3], [4]. Lexical modification do not require special programming knowledge and in most cases they are easily detected or ignored. Typical examples include modification of source code formatting (L1), comments alteration (L2), renaming identifiers (L3), split or merge of variable declaration (L4). Structural modifications, on the other side require higher level of programming knowledge and skills and their detection is more difficult. Typical examples are changing the order of variables in statements (S1), changing the order of statements within block (S2), reordering of code blocks (functions) (S3), addition of redundant statements or variables (S4), modification of control structures (S5), changing the data types and modifications of data structures (S6), redundancy (S7), temporary variables and subexpressions (S8). Other types of lexical and structural modifications are possible and are described in more details in [5]. For the purpose of our comparison we have chosen only the most relevant modifications for the context for novice programmers learning structured programming in C.

III. Plagiarism detection

Many tools for automated plagiarism detection have been developed and published. The tools can be categorized in two main categories: offline and online [6]. Offline (or desktop) tools can detect duplicates in a given collection of documents,

while the online tools search for potential duplicates on the web using through search engines or other tools. In this work, we are concerned with the offline detection tools and their approaches in plagiarism detection.

Plagiarism detection tools can employ different approaches that can be classified as text-based, attribute-oriented and structure-oriented. Text-based approaches rely on information retrieval techniques for ranking and fingerprinting methods and operate on the original contents of the source code files, which makes these approaches vulnerable to the lexical modifications. Attribute-oriented methods work by creating a fingerprint of a source file by using various numerical attributes such as average number of terms per line, unique terms, keywords, unique operands, the total number of operators and operands, and so on. Older systems [7], [8] are examples of implementation of these approaches and nowadays are mostly replaced with structure-oriented systems that consistently outperform them as presented in [9].

The approach that is most relevant to our work is the structure-oriented, which mostly uses tokenization and string matching algorithms to measure the code similarity [10], [11]. Systems that use this approach are relatively resilient to most common plagiarism modifications. Most known examples of such systems are JPlag [12] and MOSS [13].

IV. Tokenization

Source code of computer programs is usually enclosed in text files composed of keywords from the programing language, string and number literals and any natural language text in some form of comments. As is, this file can be read as a single string and compared for plagiarism using string matching algorithm. However, the same computer programs can easy be modified applying some of the presented modifications, producing semantically equivalent programs but different string contents. Standard approach used in most plagiarism detection systems is to convert the source file into list of tokens by tokenizing the source code. Tokenization is a complex process that is performed using a programming language lexer and parser. To perform the tokenization, we used the ANTLR (ANother Tool for Language Recognition) [14] tool, specifically its lexer and parser generation using a grammar for the C programming language. With the provided grammar file, the library generates lexer and parser for this grammar. Using the generated lexer and the parser we can parse each source file and convert it into list of tokens. The overall tokenization architecture is presented on fig. 1. An example of the tokenized representation of a simple source code sample in the C programming language is shown on fig. 2. Tokenization is very important phase and mostly used because it eliminates most lexical modifications to the code (L1, L2, L3). As a result, two source files with different formatting (L1), different comments (L2) and different identifier names (L3) will yield identical tokens list.

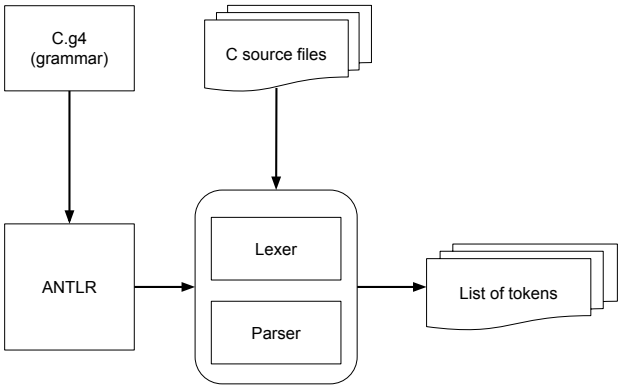


Fig. 1: Tokenization of C source files

```
int main() {
    float x;
    int y;
    double z = x + y;
    return 0;
}
```

(a) Source file in C

```
['int', Identifier, '(', ')', '{', 'float'
      , Identifier, ';', 'int', Identifier
      , ';', 'double', Identifier, '=',
      Identifier, '+', Identifier, ';', '
      return', 0, ';', '}']
```

(b) Generated list of tokens

Fig. 2: Example of tokenization

V. String similarity/distance algorithms

Once the source code is converted into list of tokens, we can compute the similarity or distance between two lists. For computing the similarity and distance we have used several string similarity algorithms. Since these algorithms operate on strings (arrays of characters), we adopted them to work on list of tokens, so each token is a substitute for a character in the string. If we represent the string S formally as an array of N characters c, $S = \{c_1, c_2, ..., c_n\}$, the list of tokens can be represented as $L = \{t_1, t_2, ..., t_n\}$. Each token is represented with the original token content from the source file, its position (line and column) in the source code file and its type.

The algorithms used to compute similarity or distance between programs are listed on table I. Some the string matching algorithms define similarity between strings (noted as *similarity* where 0 means that strings are completely different), and some define distance between strings (noted as *distance* where 0 means that strings are identical). Algorithms noted as normalized algorithms return the similarity or distance as a number in the range $[0, 1]$. Algorithms marked as metric, compute a metric distance that follows the triangle inequality 1.

TABLE I: String similarity algorithms

Algorithm	Similarity/Distance	Distance?	Metric?	Cost
Levenshtein	Distance	No	Yes	$O(m*n)$
Normalized Levenshtein	Similarity/Distance	Yes	No	$O(m*n)$
Weighted Levenshtein	Distance	No	No	$O(m*n)$
Damerau-Levenshtein	Distance	No	Yes	$O(m*n)$
Optimal String Alignment (OSA)	Distance	No	No	$O(m*n)$
Jaro-Winkler	Similarity/Distance	Yes	No	$O(m*n)$
Longest Common Subsequence (LCS)	Distance	No	No	$O(m*n)$
Metric Longest Common Subsequence (M-LCS)	Distance	Yes	Yes	$O(m*n)$
N-Gram	Distance	Yes	No	$O(m*n)$
Q-Gram	Distance	No	No	$O(m+n)$
Cosine similarity	Similarity/Distance	Yes	No	$O(m+n)$
Jaccard index	Similarity/Distance	Yes	No	$O(m+n)$
Sorensen-Dice coefficient	Similarity/Distance	Yes	No	$O(m+n)$

$$d(x,y) <= d(x,z) + d(y,z) \tag{1}$$

Levenshtein distance between two strings is the minimum number of single-character edits (inserts, deletions or substitutions) required to change one string to another.

In *Normalized Levenshtein* the distance is computed as Levenshtein distance divided by the length of the longest string.

Weighted Levenshtein is an implementation of the Levenshtein that allows different weights for different character substitution.

Damerau-Levenshtein, also called distance with transposition, is similar to Levenshtein represents the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters.

Optimal String Alignment is a variant of Damerau-Levenshtein that computes the number of edit operations needed to make the strings equal under the condition that no substring is edited more than once. The difference from the algorithm for Levenshtein distance is the addition of one recurrence for the transposition operations.

Jaro-Winkler is (roughly) a variation of Damerau-Levenshtein, where the substitution of two close (neighbor) characters is considered less important then the substitution of two characters that a far from each other.

Longest Common Subsequence (LCS) problem consists in finding the longest subsequence common to two (or more) sequences. It differs from problems of finding common substrings: unlike substrings, subsequences are not required to occupy consecutive positions within the original sequences. The LCS distance D between strings s_1 (of length n) and s_2 (of length m) is defined as 2.

$$D = n + m - 2|LCS(s_1, s_2)| \tag{2}$$

LCS distance is equivalent to Levenshtein distance when only insertion and deletion is allowed (no substitution), or when the cost of the substitution is the double of the cost of an insertion or deletion.

Metric Longest Common Subsequence is based on LCS [15] where distance is computed as 3.

$$1 - \frac{|LCS(s1, s2)|}{max(|s1|, |s2|)} \tag{3}$$

A. Shingle (N-gram) based algorithms

A few algorithms work by converting strings into sets of N-grams (sequences of N characters, also sometimes called k-shingles). The similarity or distance between the strings is then the similarity or distance between the sets. The cost for computing these similarities and distances is mainly dominated by k-shingling (converting the strings into sequences of k characters). Therefore, there are typically two use cases for these algorithms: directly compute the distance between strings, or, for large datasets, pre-compute profile or set representation of all strings, and then compute similarity between profiles (sets).

N-Gram [16] uses affixing with special character $\backslash$n to increase the weight of first characters. The normalization is achieved by dividing the total similarity score with the original length of the longest word.

Q-Gram [17] distance between two strings is defined as the $L1$ norm of the difference of their profiles (the number of occurrences of each N-gram) 4.

$$\sum_{i=1}^{n} V_{1i} - V_{2i} \tag{4}$$

Q-gram distance is a lower bound on Levenshtein distance, but can be computed in $O(m+n)$, whereas Levenshtein requires $O(m*n)$.

In *Cosine similarity* the similarity between the two strings is defined as the cosine of the angle between the two vectors representing the strings, and is computed as 5.

$$\frac{V_1 * V_2}{(|V_1| * |V_2|)} \tag{5}$$

For *Jaccard index* like Q-Gram distance, the input strings are first converted into sets of N-grams, but in this approach the cardinality of each N-gram is not taken into account. Each input string is simply a set of N-grams. The Jaccard index is then computed as 6.

$$\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \quad (6)$$

Sorensen-Dice coefficient is similar to Jaccard index, with the difference that the similarity is computed as 7.

$$\frac{2 * |V_1 \cap V_2|}{|V_1| + |V_2|} \quad (7)$$

VI. Evaluation

To compare the presented algorithms we prepared three datasets (denoted as A, B and C) with source code solutions of selected programming problems from the introductory course in Structured programming taught in C. The first dataset A was artificially created and contains solutions on a example problem from lab exercises on the topic of loops. This dataset contains total of 13 correct solutions of the selected problem. Single solution denoted as "original" was created from the authors. Four solutions (denoted L1 to L4) are different *lexical* (L) modifications of the original solution. Other four solutions (denoted S1 to S4) are different *structural* (S) modifications of the original solution. The last four solutions (denoted O1 to O4) are manually selected different *original* solutions submitted by students of the course. Summary of the datasets A is presented on table II. This is labeled dataset, so solutions with modifications (L, S) are plagiarism of the original solution, and the other four solutions (O) are not plagiarism.

Datasets B and C are composed from the submitted solutions on a selected problem from the first partial exam in academic year 2013/2014 on the topic of loops including simple arithmetic and modulo operations. Specifically, the dataset B contains the 49 solutions submitted on the partial exam which was set in controlled environment. The controlled environment means that student can not access the Internet, share ideas or contents of their solutions with other students, so the possibility of plagiarism is minimized. The other dataset C, contains all 697 solutions submitted when this same problem was given on one of the laboratory exercise in academic year 2015/2016. The laboratory exercises are set in computer laboratories at the faculty and the students' behavior is not restricted in terms of communication with other students or using auxiliary materials prepared beforehand. This makes possible for students to use solutions from their colleagues and engage in plagiarism. The dataset B contains total of 49 solutions of which only 2 were correct according to automatic assessment of the system Code. The other dataset C contains total of 697 solutions of which 480 (69%) were correct. The big difference in the ratio of correct solutions between two datasets raises the concern of possible plagiarism in the dataset C. The average number of tokens of the tokenized versions of the files is similar in both datasets (≈ 130). Summary of the datasets B and C is presented in table III.

TABLE II: Dataset A

File Name	Modification Description
original	The original solution from authors
L1	Source code formatting
L2	Comments alteration
L3	Renaming identifiers
L4	Split and merge variable declaration
S1	Changing the order of variables in statements
S2	Changing the order of statements within block
S4	Adding redundant statements or variables
S5	Control structures modification

TABLE III: Datasets B and C

Dataset	No. of solutions	Correct	Avg. number of tokens
B	49	2	133.0
C	697	480	131.0

VII. Comparison results

On the first dataset A we computed the pairwise distance and similarity between all solutions using all 13 presented algorithms. Because this dataset was labeled with plagiarism pairs and not-plagiarism pairs, we computed the average distance and similarity among these groups of files. We also computed the averages on pairs of plagiarism and original solutions From the results that are shown on table IV we can see that the average distance among plagiarism pairs is significantly smaller than the average distance among non-plagiarism pairs. For example for Levenshtein algorithm the average distance among plagiarism pairs is 7.86 and among non-plagiarism is 55.33. Biggest difference (91.67) was produced from the Q-Gram algorithm, where the average distance among non-plagiarism was 102.17 and among plagiarism pairs was 10.5. The distance among pairs of plagiarism and non-plagiarism solutions was very close to the distance among plagiarism pairs. This shows the original solutions selected for this dataset are relatively equally distant (different) among them and among them and other modified plagiarism solutions. Similar trend of significantly smaller average distances among plagiarism pairs $[0.3, 0.7]$ compared to the average distances among non-plagiarism pairs $[0.21, 0.64]$ is present also in the computed normalized distance. For the similarity measure, which is opposite of distance (larger similarity means larger probability of plagiarism), the average among plagiarism pairs $[0.93, 0.97]$ is significantly larger than the average among non-plagiarism $[0.36, 0.79]$ and mixed pairs $[0.35, 0.78]$.

Using the calculated pairwise distances we performed hierarchical clustering using complete-linkage method and the dendograms depicting the result clusters are shown on figure 3. From this figure we can notice that the original solution and all other solutions that were lexical and structural modification are forming single cluster, while other original solutions form separate clusters each.

For the other datasets B and C we have also computed the distance and similarity among all possible pairs using all algorithms. Since these datasets are not labeled with plagiarism and non-plagiarism pairs, we wanted to count the possible plagiarism pairs using some threshold for distance, normalized distance and similarity. We have handpicked the thresholds from the results for the dataset A. The threshold for distance

TABLE IV: Average distance, normalized distance and similarity among all plagiarism pairs (second column), pairs of original solutions (third column) and pairs between plagiarism and originals (fourth column)

Algorithm	Plagiarism <	Non-plagiarism	Plag/Orig
Distance			
Levenshtein	7.86	55.33	52.97
Weighted Levenshtein	7.86	55.33	52.97
Damerau-Levenshtein	7.86	55.33	52.56
OSA	7.86	55.33	52.56
LCS	8.36	73.50	69.19
Q-Gram	10.50	102.17	104.25
Normalized distance			
Levenshtein (N)	0.07	0.48	0.47
Jaro-Winkler (N)	0.05	0.21	0.22
Metric LCS (N)	0.06	0.36	0.34
N-Gram (N)	0.07	0.51	0.51
Cosine Similarity (N)	0.05	0.34	0.39
Jaccard index (N)	0.06	0.64	0.65
Sorensen-Dice (N)	0.03	0.47	0.49
Similarity			
Levenshtein	0.93	0.52	0.53
Jaro-Winkler	0.95	0.79	0.78
Cosine Similarity	0.95	0.66	0.61
Jaccard index	0.94	0.36	0.35
Sorensen-Dice	0.97	0.53	0.51

was selected a value of 11 which in the dataset A separates all plagiarism of non-plagiarism pairs. The value of the threshold for normalized distance was 0.1 and for similarity was 0.9, so all pairs with distance less than 0.1 or similarity greater than 0.9 will be counted as plagiarism. Using these values for thresholds we counted all the plagiarism pairs using the distances and similarities from all algorithms. On table V are shown the results for the dataset B and on table VI are shown the results for the dataset C. These results are showing that for the dataset B which was composed of solutions submitted on exams there are no plagiarism pairs detected for all of the algorithms. On the other side, for the dataset C, which was composed of solutions submitted on laboratory exercise, the percentage of solutions involved in plagiarism pair is almost for all algorithms above 50%.

VIII. Conclusion

In this paper we have presented the results of the implementation and evaluation of 13 different string matching algorithms for the task of computing distance and similarity between source files written in the programming language C. We have shown that the computed distances or similarities among the pairs of files using all of these algorithms can be used to detect possible plagiarism. To be useful for the task, the algorithms were adopted to work for list of tokens instead of strings, so they can be used to compare the tokenized versions of the source files. Tokenization proved to be successful method for preventing the most common lexical and structural

[0]Percentage of files that are included in at least one pair of files with distance or similarity above the threshold.

TABLE V: Plagiarism in dataset B

Algorithm	% Plagiarism [0]	Avg. distance
Distance (< 11)		
Levenshtein	0.00	97.14
Weighted Levenshtein	0.00	97.14
Damerau-Levenshtein	0.00	97.13
OSA	0.00	97.13
LCS	0.00	115.23
Q-Gram	0.00	201.03
Normalized distance $(distance < 0.1)$		
Levenshtein (N)	0.00	0.57
Jaro-Winkler	0.00	0.30
Metric LCS	0.00	0.53
N-Gram	0.00	0.62
Cosine Similarity	0.00	0.75
Jaccard index	0.00	0.85
Sorensen-Dice	0.00	0.74
Similarity $(similarity > 0.9)$		
Levenshtein	0.00	0.43
Jaro-Winkler	0.00	0.70
Cosine Similarity	0.00	0.25
Jaccard index	0.00	0.15
Sorensen-Dice	0.00	0.26

TABLE VI: Plagiarism in dataset C

Algorithm	% Plagiarism [0]	Avg. distance
Distance (< 11)		
Levenshtein	69.15	67.40
Weighted Levenshtein	69.15	67.40
Damerau-Levenshtein	69.15	67.40
OSA	69.15	67.40
LCS	66.14	81.11
Q-Gram	45.77	156.00
Normalized distance (< 0.1)		
Levenshtein (N)	71.59	0.43
Jaro-Winkler	77.62	0.23
Metric LCS	75.75	0.39
N-Gram?	100.00	0.47
Cosine Similarity	59.54	0.59
Jaccard index	50.65	0.73
Sorensen-Dice	57.82	0.60
Similarity $(similarity > 0.9)$		
Levenshtein	71.59	0.57
Jaro-Winkler	77.62	0.77
Cosine Similarity	59.54	0.41
Jaccard index	50.65	0.27
Sorensen-Dice	57.82	0.40

modifications of the code made for purpose of avoiding plagiarism detection. We have evaluated all the algorithms on the labeled dataset A, which was composed with a selected set of common lexical and structural modifications applied on a single original solution. For these selected modifications the results from all algorithms were very promising. All of the algorithms were able to produce very small distance and significant similarity among the plagiarism pairs in this dataset. On the contrary, the distances computed for the non-plagiarism pairs were significantly larger and the similarity computed was significantly lower comparing to the plagiarism

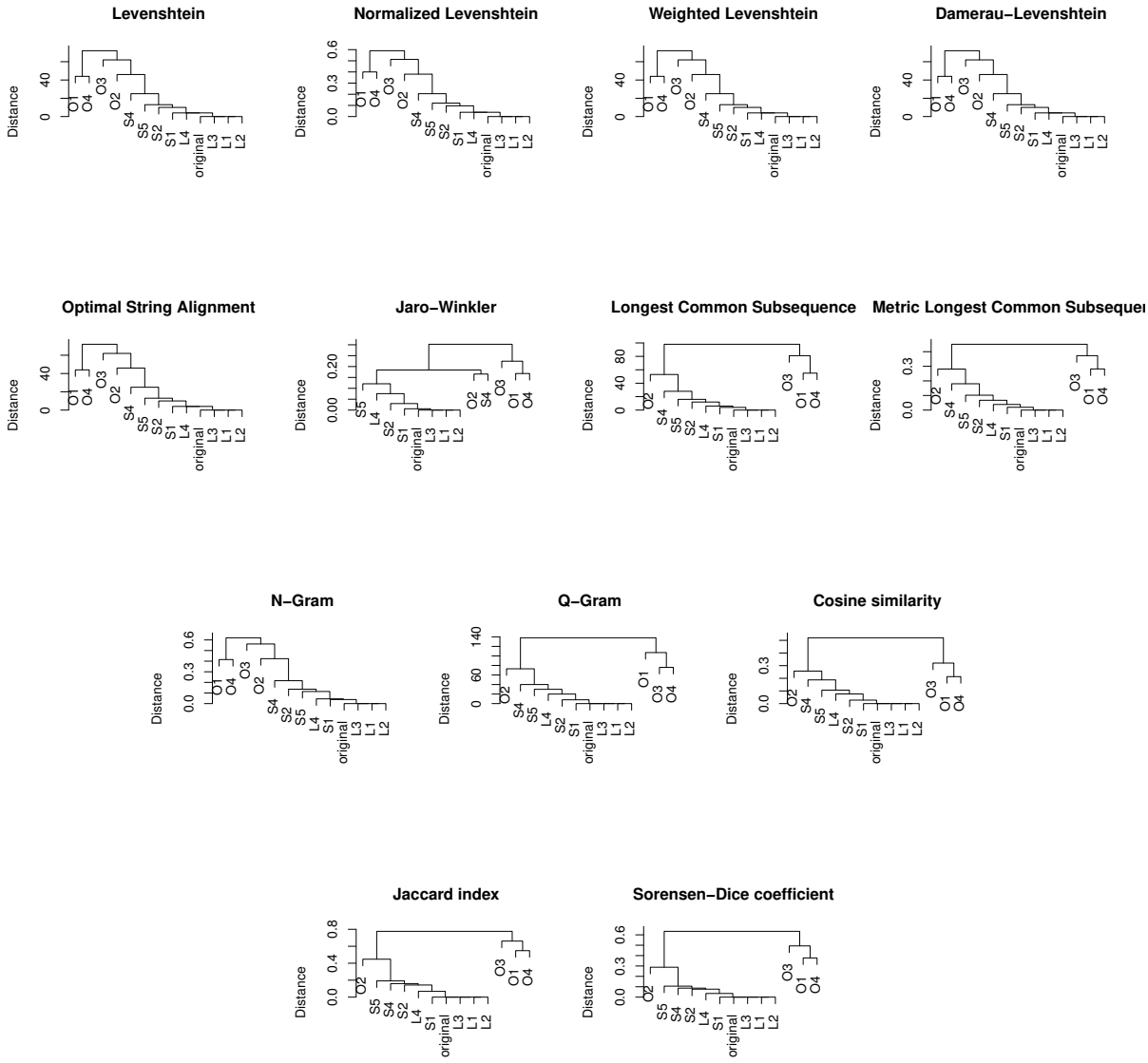


Fig. 3: Dendograms from hierarchical clustering based on distances computed from string matching algorithms

pairs. Using a selected threshold values for distances and similarities from the plagiarism pairs from the first dataset A we were able to detect significant percentage of plagiarism pairs in the dataset C and none in the dataset B. This was the expected result for these datasets, knowing the settings in which the students submitted solutions for the problem picked for these datasets. This also rediscovered the known problem of very big percentage of plagiarism involved in laboratory sessions.

It is important to mention that all of the tested algorithms were only used to compute distance or similarity among potential plagiarism pair of source files. To be able to confirm that a pair of solutions is plagiarism, still a human decision is needed. Some of these algorithms, along with the computed distance or similarity that can mark a pair for possible plagiarism, can also be used to mark the contents of files that were identical. That way, the value of distance or similarity between two source files will be accompanied with visualization of the matched segments of both source files. Also, it is worth mentioning that some of the algorithms are very sensitive to more advanced structural modifications which were not evaluated in this paper.

REFERENCES

[1] T. Delev and D. Gjorgjevikj, "E-lab: Web based system for automatic assessment of programming problems," *ICT Innovations 2012, Web Proceedings ISSN 1857-7288*, p. 75, 2012. [Online]. Available: delev2012lab.pdf

[2] E. L. Jones, "Metrics based plagarism monitoring," *Journal of Computing Sciences in Colleges*, vol. 16, no. 4, pp. 253–261, 2001.
[3] M. Joy and M. Luck, "Plagiarism in programming assignments," *IEEE Transactions on education*, vol. 42, no. 2, pp. 129–133, 1999.
[4] C. K. Roy and J. R. Cordy, "A survey on software clone detection research," *Queens School of Computing TR*, vol. 541, no. 115, pp. 64–68, 2007.
[5] Z. Djuric and D. Gavsevic, "A source code similarity system for plagiarism detection," *The Computer Journal*, p. bxs018, 2012.
[6] M. Mozgovoy *et al.*, "Desktop tools for offline plagiarism detection in computer programs," *Informatics in Education-An International Journal*, no. Vol 5_1, pp. 97–112, 2006.
[7] G. Whale, "Identification of program similarity in large populations," *The Computer Journal*, vol. 33, no. 2, pp. 140–146, 1990.
[8] J. A. Faidhi and S. K. Robinson, "An empirical approach for detecting program similarity and plagiarism within a university programming environment," *Computers & Education*, vol. 11, no. 1, pp. 11–19, 1987.
[9] K. L. Verco and M. J. Wise, "Plagiarism à la mode: a comparison of automated systems for detecting suspected plagiarism," *The Computer Journal*, vol. 39, no. 9, pp. 741–750, 1996.
[10] M. J. Wise, "Detection of similarities in student programs: Yap'ing may be preferable to plague'ing," in *ACM SIGCSE Bulletin*, vol. 24, no. 1. ACM, 1992, pp. 268–271.
[11] ——, "Yap3: Improved detection of similarities in computer program and other texts," *ACM SIGCSE Bulletin*, vol. 28, no. 1, pp. 130–134, 1996.
[12] L. Prechelt, G. Malpohl, and M. Philippsen, "Finding plagiarisms among a set of programs with jplag," *J. UCS*, vol. 8, no. 11, p. 1016, 2002.
[13] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: local algorithms for document fingerprinting," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. ACM, 2003, pp. 76–85.
[14] T. Parr, J. Lilly, P. Wells, R. Klaren, M. Illouz, J. Mitchell, S. Stanchfield, J. Coker, M. Zukowski, and C. Flack, "Antlr reference manual," *MageLang Institute, document version*, vol. 2, no. 0, 2000.
[15] D. Bakkelund, "An lcs-based string metric," *University of Oslo*, 2009.
[16] G. Kondrak, "N-gram similarity and distance," in *International Symposium on String Processing and Information Retrieval*. Springer, 2005, pp. 115–126.
[17] E. Ukkonen, "Approximate string-matching with q-grams and maximal matches," *Theoretical computer science*, vol. 92, no. 1, pp. 191–211, 1992.

A Survey of Models of Robotic Behavior for Emotional Robots

Vesna Kirandziska and Nevena Ackovska
Faculty of Computer Science and Engineering
ss. Cyril and Methodius University
Skopje, Macedonia
{vesna.kirandziska, nevena.ackovska}@finki.ukim.mk

Abstract – **Emotional robots are robots that can have emotions, sense emotions or act based on emotions. Programming emotional robots is a challenge due to the complexity and ambiguity of the emotion definition and the varieties of possible emotion representations. One of the most important part of programming such emotional robots is the model for robotic behavior that is based on emotions. In this paper a survey of the models that are being used in contemporary researches and also the models implemented in real emotional robots are presented. How the recognized or sensed emotion influences on the robotic behavior is investigated in more detail. The specifics and differences among existing models will be presented in this paper. The benefits and drawbacks of using the existing models will be analyzed in the end.**

Keywords—robotics; behavior models; emotions

I. INTRODUCTION

Personal service robotics is a wide research area whose goal is to create robots that assist people or perform useful services in numerous different ways. These robots are aimed for all variety of people among which are people with no special skills nor training to operate with robots. This raises the question on how these roots should behave and interact with humans. In [1] it is suggested that social interaction between human and robot, that is a specific to so called social robots, is needed for all personal service robots. In standard human-robot interaction a robot receives information, makes decision and actively changes its environment with its behavior. In social interaction in which social robots are engaged, robots posses histories and they explicitly communicate with and learn from all parties in their society of robots or humans [2]. Among the important factors of the design of social robots and especially social interaction are the emotions. But, what are the emotions?

A proper emotion definition does not exist, but many have tried to explain emotions better. For example, in [3] emotion is represented as the psychological energy that drives human behavior. Emotional state is the internal manifestation of this energy while human behavior is the external manifestation of emotions. Emotions influence several cognitive processes among which are problem solving, decision making, thinking, perception and learning [1,4]. For example a child uses perception of emotion expression of their parent as an encouragement or discouragement on his/hers behavior [5]. Another example is the influence the emotions have in human emotional experience that represents the accumulated knowledge gained by feedback from the environment [4]. Thus, emotional experiences are unique for each human and it is one reason humans act differently in similar situations. As seen in the examples above, the influence emotions have on cognitive processes in a human are crucial to human normal behavior and in fact is was proven in [6] that human intelligence is much impacted by emotions.

The great functionality of human emotions in human behavior suggest that emotions in robots could have great effect on their performance and especially on their social interaction ability. Many researches have supported this statement. For example, [7] suggests that emotions are valuable source of intelligence in robots that could provide autonomy of a robot. Even more some authors question the power of artificial intelligence in robots without emotions [4].

Emotions are modeled using mathematical language so that they could be represented in a robot. One emotion representation model is the categorization model of Ekman who have distinguished 6 basic emotions (anger, disgust, fear, happiness, sadness, and surprise) [8]. Robots that have implemented emotion model are called emotional robots. This paper focuses on the ways that emotions influence the emotional robots' behavior. We examine the special characteristics of these emotional robots, that others robot lack.

This paper is structured as follows: In the next section different possibilities for using emotions in robots are described. In Section III several examples of models of emotional robotic behavior are presented. Next, a discussion about the application of these models is given. In the end we conclude this paper.

II. USAGE OF EMOTIONS IN ROBOTS

Emotion usage in human behavior implies that there are varieties of ways emotions can be used in robots, since emotions in robots tend to simulate human emotions. Several researches have investigated the role emotion in artificial creatures and robots in particular and some of the features that correspond to those research findings are presented next.

Survivability – Emotion serve as one of the mechanisms for complete robot autonomy, but also provide support for a robot survival in a complex world [9].

Self model – Emotion can be used in the representation of robot's internal state. Short term emotional state is often called an emotion, while a longer term emotional state is called mood [5].

Environment perception – Emotions can play a role in receiving and processing information from its environment and thus in robot's external state representation [4]. For example, perceiving emotions in humans can be used as a feedback to a robot. From another point of view emotions can reflect how the robot is affected by different perceptual information it gathers in the world.

Behavior decisions – Emotion can influence on the specific actions robots make, but also on long term robotic behavior. On a long term, robots plan their sequel actions and also change their target goal. Information about emotions that are present in robots internal and external state can be used for planning [5,10].

Interaction – Emotions can facilitate believable human-robot interaction in the sense that robots could interact with humans in ways that humans are familiar and comfortable with [9].

Learning – Emotions can be included into the evaluation mechanism of the robots learning process [4]. This could enhance robot learning and trigger adaptation in robotic behavior. For example, emotional feedback can be used to adapt a robot in a non-expert interaction [9].

More roles of emotions include: motivation, alarming mechanism, strategic processing, memory control, emotional intelligence etc. [10]. Among all stated roles of emotions some directly influence robotic behavior: behavior decisions (directly affected by the model of robotic behavior) and learning (modifies the model of robotic behavior). In the next section several models of robotic behavior are presented.

III. Models of Robotic Behavior in Emotional Robots

Currently there is no single computational model for a robot that captures all the complexities and aspects of emotions. Studies have tried to create best fitting models for their own specific purpose. Since the application of emotional robots is huge and versatile there are varieties of models built for emotional robotic behavior. In this section some custom models that have been used in contemporary researches are given.

A. Traits, Attitudes, Moods and Emotions (TAME)

An affect-based behavior model named TAME (Traits, Attitudes, Moods and Emotions) was built by Moschkina and Arkin in 2003 [11]. This model aimed to present the basis for intelligent robotic behavior that improves human-robot interaction. Personality and affect module is designed to add affect in robotic behavior. This module consists of four separate components Traits, Attitudes, Moods ad Emotions that are interconnected and all together represent the robot internal state. Traits and attitudes determine the robot character. Moods and emotions represent long-term and short-term emotional states, accordingly. Robotic behavior is selected based on behavior parameters that are influenced by the robot internal state (presented in the Personality and Affect module) and external state (presented in the robot Perceptual module). The Personality and Affect module changes its inner parameters based on the perceptual model. In this way both robot character and emotional state are changing, thus the robot behavior changes.

B. EARL framework

EARL framework (Framework for systematic study between emotion, adaptation and reinforcement learning) was built by Broekens in 2007 [5] to model the relation between emotion and learning in emotional robots. EARL framework has four parts: Emotion recognition module, Reinforcement learning agent, Artificial emotion module slot and Expression module. Standard reinforcement learning techniques, as Q-learning, can be used by the learning agent. Here the value function Q for each possible action is approximated with a separate Multi-Layer Perceptron (MLP). Social reinforcement, coming from recognized emotion, can be added in the reward parameter in the action selection function. Beside using emotion as reinforcement, emotions are used as meta-parameters (parameters in the robot model) by the Artificial emotion module slot. At last, robotic emotions are used in emotion expression module where the robot's body expresses its current state.

Even though this framework is aimed for actual robots, it was first tested in a simulated environment. The results have shown that using social reinforcement and learning, as proposed in the framework, facilitates robot learning i.e. the robot would learn the optimal solution faster.

C. Emotion-triggered reinforcement learning model

The goal behind the model created in [7] was to model robot that adapts to its environment through reinforcement learning where emotions are used for relevant events detection. The emotion system is modeled by a recurrent network that mimics a simplified version of a human hormone system. It identifies the robot emotional state based on the perceived sensor information. From the emotional state the dominant emotion and robot's feelings are deduced. A dominant emotion is the emotion with the highest intensity that is above a given threshold. If there is no such emotion, the neutral emotion will be the dominant one.

The controller of the robotic behavior is an adaptive controller that has two separate modules: Associative Memory module and Behavior selection module. The first module calculates the expected evaluation for each possible behavior and the second selects the best behavior in a given moment. The Associative Memory module uses feed-forward networks to calculate the evaluation based on robot's feelings (feelings are defined as the cognition of the robot sensor data that also depend on the emotional state). The network is adapted using back-propagation where reinforcement is extracted from the dominant emotion. The Active controller is event-driven i.e. triggered by the event detector. For example, when there is a change in the dominant emotion with respect to previous detection, a new event is detected. To conclude, this whole

model incorporates emotions in three different model components: 1) reinforcement function, 2) robot feelings and 3) event detector.

D. Autonomous behavioral/emotional expression system

The main concept of the system proposed in [12] is creating behavioral/emotional expressions based on dopamine that is used to determine robot's motivation. Note that this substance is important in human emotion expression and here a concept of artificial dopamine is introduced. The proposed system consists on three parts: Recognition of the external situation, Cognition module and Emotion-selection module.

The Recognition system extracts valuable parameters, important for the determination of the dopamine, from the external situation. The Cognition module first determines the robot motivation based on the dopamine, and than finds the best candidate for behavior and emotion using two separate self-organizing maps (SOM). The network architecture of SOM enables multi-dimensional data to be mapped onto two-dimensional array of neurons that are all mapped to one specific behavior (Behavior map) and emotion (Emotional map), accordingly. The best matching neuron in the Behavior map determines the action the robot should perform. Based on the Emotional map, affective factors are calculated and imputed in the Emotion-selection module that outputs the emotion expression actions that the robot should perform. The Emotion-selection system is implemented as a Markov Process. The next robot emotion state is based on the previous emotion state and transition probabilities of change from one to another state that are incorporated in the topology of the Markovian emotion model. Emotion is presented as a random variable with 6 possible values that represent six emotions: neutral, sadness, fear, disgust, happiness and hope. By adding new data, the emotion model changes by the change of the transition probabilities. The novel property in the proposed model is that it predicts the next emotional state of a robot based on its previous emotion and this influences emotion expression actions.

E. Decision Making system that uses re-construction of emotions

A behavior selection system given in [13] is structured in four modules: Cognition, Emotion, Behavior-selection and Behavioral-making. The Markovian emotional model, the same used in [12], is used here. The only difference is that here only four emotional states are present: joy, sad, fear and anger. The input data of the system are sensor data extracted from the environment. The Cognition module calculates emotion-inducing factors (four parameters) given the environment sensor data. This module is implemented using SOM where online learning is enabled. In the next stage of the process of decision making, the Emotional module calculates the next emotional state based on its previous state. Here emotions are reconstructed by using emotional-inducing factors obtained during the task. Based on the current emotion the Behavior-selection module determines the probability distribution for all possible behaviours. This module is also implemented as a Markovian model. The final module, Behavioral-making module, generates the control vector for the robot based on which the robot is controlled.

More efficient association between emotion-including factors and input sensor data added additional learning in the model and that, in turn, has improved the robot's performance. This was shown in an experiment in a simulation environment [13].

F. Emotion expression based on HMM (Hidden Markov Model)

The emotional expression model presented in [3] focuses only on the problem of generating robotic expression behavior that can be used in human-robot interaction. The input of the model is the emotional features from the facial expression input from the human the robot interacts with. The output of the model is the emotional expression behaviour presented via facial and physical robotic behavior.

The model consists of two computational modules: Active field state space and HMM Transference module. Based on the expression features that are extracted from a human, a feature classification is done to get one of the 7 possible emotional states: anger, disgust, fear, happiness, sadness, surprise and calm state. The classified emotion is inputted in the Active field state space. In each point in the active field the strength for each emotion is adjusted based on the new data. Emotional intensity over time weakens and this is adjusted in the active field by using an attenuation function. Next, in the HMM Transference model dual stochastic processes for emotional state transference and behavior performance transference are implemented. Using HMM, the model produces the facial expression plan and physical behavior plan for the robot. These plans are used to control robot's hardware devices. The presented model is used in dynamic human-robot interaction. The double stochastic processes make the model more expressive and thus the interaction more natural [3].

There are many other models that are used and interleave robotics and emotions. One valuable example is a neural network architecture for the development of a desired phenotype from a hierarchically structured genome that has the ability to evolve to an emotion integrated neural leaning architecture. This architecture presented in [14,15] integrates learning and emotion in the sense that emotion based self-supervised learning is incorporated.

IV. Model usage and application

As elaborated in [2] social robots have various different applications. For example, social robots can be used for entertainment, as toys, as mobile social companions, as interactive tools, for therapy, for education and for anthropomorphic researches that aim to emulate natural human robot interaction. As stated before, one great feature of these social robots is for a robot to present emotional responses in its environment.

Based on the function of a robot, different behavior models can be implemented in an emotional robot. In this paper several models ware presented. These were all different, but still all have incorporated emotions in their decision making module.

The goal behind these models varies from more specific ones, like in the Emotion expression based on HMM that is specific for emotional expressions, to more general goals, like mimicking human interconnections between several emotion-based human characteristics that was done in TAME.

All presented models were divided in different modules that had a specific task and their interfaces were used to communicate with other modules. The input data in the models are sensory data collected form the robot environment, while the output is a robotic action that is chosen for the robot to perform. Most of the presented models implemented online learning that enable a robot to adapt to new situations in the environment. In the models for emotional robots, emotions have played a great part in the learning process. Some models have implemented reinforcement learning, some implemented neural network structures that have the property to be adjusted online and others used Hidden Markov Models.

Beside all similarities, the presented models are all different. It has been proven in their researches that all the proposed models have some benefits [3, 5, 7, 11, 12, 13]. Today a generic model for robotic behaviour is not known, so different approaches enriches the set of possible models that could be used for implementation in different robots and suggest ideas and direction in which the models can be improved.

The work we have done in our laboratories proves that emotion can be successfully detected using well chosen sound features [18], but even more when including facial data [19]. Experimental data have shown that the created robot emotion recognition system has similar power as human recognition on other people's emotions [18]. This implies on the possibility that human-robot interaction could be more believable with the inclusion of a module for perceiving emotions that influences robot behaviour. In [20] we have also discussed the ethical consequences on the involvement of an emotion perception module in a social robot on both humans and robots.

Some robots are created especially for research purposes where emotion based behaviour is investigated. One example of this kind of robots is the robot named Kismet. Robotic behaviour based on emotion could also be used in other general - purpose robots that have possibilities to implement emotional behaviour. Example include the commercial robots Nao and Pepper. In sequel a short description of the mentioned robots is given. Of course there are more robots that use or could use the robotic behaviour models that are based on emotions and some examples are: Plao, FACE, Kasper, Infanoid, Cog etc.

Kismet, a robot with a 15-degrees-of-freedom (DOF) face, was designed to encourage natural infant-caretaker interactions [2]. It can express seven different emotions: anger, disgust, fear, joy, sorrow, and surprise via facial expressions and via its voice. Kismet emotion state are implemented using active field states. Online learning is used to adjust the model with the new data from the environment and especially from the interaction. As a result, the same user input that triggers surprise, may soon trigger annoyance. The threshold for activation of Kismet's expressions vary over time.

Nao is a humanoid robot developed by Aldebaran Robotics, now SoftBank Robotics. Nao has tactical sensors, two speakers, two cameras, two lateral microphones, prehensile hands, and 25 DOF. Nao can potentially express emotions through both verbal and nonverbal means, but has limited facial expressions [16]. Many researches have used Nao so far as an experimental robot on which emotion robotic behavior is deduced [2,15]. Our work shows impressive results using NAO in interaction with children. This is true both in learning processes and in interaction with children with special needs [21,22].

Pepper is the novel humanoid robot developed by Aldebaran SoftBank Robotics. It is a friendly humanoid robot that "seems determined to make everyone smile" [17]. Pepper has an emotion engine that is designed to understand people's feelings. Based on these feelings the robot could act accordingly and for example talk, gesticulate, and zip by on its wheels.

V. Conclusion

Today emotional robots have many applications mostly in the areas where human-robot interaction tends to be more natural and human-like: nursing robot, teaching robot, companion robot etc. The trend of creating such robots can be observed with the appearance of commercial robots that are being upgraded with the ability to process emotions. One example is the robot Pepper, the emotional humanoid robot from Aldebaran SoftBank Robotics.

In this paper all aspects where emotions can be used in a robotic behaviour model are presented. We have also presented some existing models of emotional robotic behaviour. These all have some similarities and differences, and all have presented improvements for a concrete robotic application. We have also worked on robotics platforms and architectures that include and successfully process emotions.

In the future, there is a need for a complete behaviour model based on emotions that can be used for general purpose or some special specific purpose. The present models can be used as a starting point in creating a more general solution and shall be used in the emotional models that we are improving in order to obtain a more natural human – robot interaction.

References

[1] C. Bartneck, J. Forlizzi., A design-centered Framework for social-human interaction, In: Ro-Man2004, Kurashiki, 2004, pp. 591–594.

[2] T. Fong, I. Nourbakhsh, K. Dautenhahn, "A Survey of socially interactive robots: Concepts, Design and Applications", Technical Report CMU-RI-TR, No. 29, 2002.

[3] L. Xin, X, Lun, W. Zhi-Iang, F. Dong-mei, Robot emotion and performance regulation based on HMM, International Journal of Advanced Robotic System, Vol. 10, 2013.

[4] F. Yang, X. Zhen, Research on the Agent's behavior decision-making based on artificial emotion, Journal of Information & computational Science, Vol. 11, No. 8,2014, pp. 2723-2722.

[5] J. Broekens, Emotion and reinforcement: affective facial expressions facilitate robot learning, Human Computing, Springer-Verlag Berlin Heidelberg, 2007, pp. 113-132.

[6] A. R. Domasio, Descartes' Error: Emotion, Resech and Human Brain, Penguin, 2005.

[7] S.S. Gadanho, J. Hallam, Emotion-Triggered Learning for Autonomous Robots. In Workshop on Grounding Emotions in Adaptive Systems at the 5th International Conference of the Society for Adaptive Behavior (SAB'98) Zurich, 1998.

[8] P. Ekman, Basic emotions. Wiely, T. Dalgeleish and M. Power, eds., Handbook of Cognition and Emotion, New York, 1999.

[9] R.C. Arkin, "Moving Up the Food Chain: Motivation and Emotion in Behavior-based Robots" in Who Needs Emotions: The Brain Meets the Robot, Eds. J. Fellous and M.Arbib, Oxford University Press, 2005.

[10] M. Scheutz, Using roles of emotions in artificial agents: A case stady from artificial life, Proc. AAAI, San Jose, California, 2004, pp. 42-48.

[11] L. Moshkina, R.C. Arkin. On TAMEing Robots, Proc. IEEE International Conference on Systems, Man and Cybernetics, 2003.

[12] W. Jitviriya, M. Koike, E. Hayashi, Emotional model fr robotic system using a self-organizing map combined with Markovian model, Journal of Robotics and Mehcatronics, vol. 27, no. 5, 2015, pp. 563-570.

[13] S. Watada, M. Obayashi. Kuremoto, S. Mabu, A decision making system of robots introducing a re-construction of emotions based on their own experiences, Journal of Robotics, Networking and Artificial Life, Vol. 1, No. 1, pp. 27-32, 2014.

[14] N. Ackovska, S. Bozinovski, L. Bozinovska. "Evolving an emotion-based neural control architecture", Proc. IEEE Southeastcon, Huntsville, AL, 2008, pp. 404-408.

[15] N. Ackovska, System software in minimal biological systems, PhD Thesis, St. Cyril and Methodius University, Skopje, 2008 (in Macedonian) .

[16] V. Manohar, J.W. Crandall. Programming Robots to Express emotions: Interaction paradigms, communication modalities and context, IEEE transactions on human-machine systems, vol. 44, no. 3, June, 2014.

[17] E. Guizzo. Meet Pepper, Aldebaran's New Personal Robot With an "Emotion Engine". IEEE Spectrum , 5 June, 2014.

[18] V. Kirandziska, N. Ackovska. „A robot that perceives human emotions and implications in human-robot interaction". Proc. of IEEE RO-MAN '14 - The 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, 2014, pp. 495-498.

[19] V. Kirandziska, N. Ackovska and A. Madevska Bogdanova. „Comparing Emotion Recognition from Voice and Facial Data Using Time Invariant Features". International Journal of Computer, Electrical, Automation, Control and Information Engineering, World Academy of Science, Engineering and Technology, vol. 10, No. 5, 2016, pp.737-741.

[20] V. Kirandziska, N. Ackovska, "A concept for building more humanlike social robots and their ethical consequence", Proc. Of the International Conferences: ICT, Society and Human Beings 2014 (MCCSIS) 15-19 July, Lisbon, Portugal, 2014, pp. 37-44, (Best Paper).

[21] A. Tanevska, N. Ackovska, V. Kirandziska, "Robot-assisted therapy: considering the social and ethical aspects when working with autistic children", Proceedings of the 9th International Workshop on Human-Friendly Robotics - HFR 2016, Genova, 2016, pp. 57-60.

[22] A. Tanevska, N. Ackovska, V. Kirandziska, "Assistive robotics as therapy for autistic children", 13th International Conference for Electronics, Telecommunications, Automation and Informatics, Struga, Macedonia, 2016 (in print).

Comparative Analysis of Methods for Determination of Protein Binding Sites

Georgina Mirceva, Andreja Naumoski, 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, andreja.naumoski@finki.ukim.mk, andrea.kulakov@finki.ukim.mk

***Abstract*—Proteins are very important since they participate in many processes in the organisms. They effect these processes by performing particular functions. Therefore, it is very useful if we know the functions of proteins, so we can regulate the processes in the cells. The literature provides lots of methods for determinations of protein functions, but still there is a large gap between the known protein structures and those for which we know its functions. One way to determine the protein functions is to analyze the characteristics of protein binding sites where protein interaction happens. In this paper, we focus on determination of protein binding sites. We present an approach for protein binding sites prediction. Our approach considers the geometrical conformation of protein structures, and by using classification method, it identifies the atoms that are part of binding region. We present some results from the comparison with several approaches that are widely used.**

Keywords—protein molecule; protein function; protein binding site

I. INTRODUCTION

Protein molecules are very important in the processes in living organisms. Drug design is based on the knowledge about the functions of protein molecules. The improvements in technology provides fast determination of protein structures. However, the knowledge gathered with these technologies is not useful if it is not used to determine the protein functions. There are various methods for functional annotation of proteins. Some methods are based on identification of homologous proteins [1]. Another group of methods [2] identifies the stable regions of proteins that do not change during evolution, and then try to annotate protein structures based on the characteristics of these regions. Other methods [3] determine protein functions by using protein-protein interaction networks. Also, there are methods that annotate proteins by detecting binding sites and analyzing their characteristics [4]. In this work, we focus on identification of binding sites that could be used for determination of protein functions.

This paper is structured as follows. In section 2 we present our approach for predicting protein binding sites. Some experimental results are presented in section 3, while section 4 give some general conclusions.

II. OUR APPROACH

The prediction of protein binding sites is done in two steps. First, we extract the characteristics of the residues that constitute the protein structure. Then, in the second step, we use a classification method in order to build prediction model that would estimate whether a given test atom is part of binding region.

We consider the following characteristics of the atoms of the protein: Accessible Surface Area (ASA) [5], depth index (DPX) [6], protrusion index (CX) [7] and hydrophobicity [8]. We use the rolling ball algorithm [5] in order to estimate the ASA of the atoms. A probe sphere with predefined radius is rolled over the protein structure, and for each atom, we calculate the area of the atom that is touched by this sphere. The amino acid residues contain several atoms, so for each residue we calculate its ASA feature by summing these values from all atoms. The depth index (DPX) [6] is calculated as Euclidean distance between an atom and the closest atom that has ASA>0. The protrusion index (CX) [7] shows the density around the inspected atom, and it calculated as a ratio between the empty space and the filled space around the atom within a sphere with predefined volume. Because amino acid residues contain several atoms, therefore we aggregate the DPX and CX values and consider the average from these values. Hydrophobicity is a characteristic of the amino acids that shows the hydrophobic properties of the amino acids. In this work, we use the hydrophobicity scale given in [8].

After extraction of the characteristics of the amino acid residues, next, in the second step we induce prediction model by using classification method. In [9], we used various feature selection techniques in order to select the most relevant features and then we applied various classification methods for model induction. In this work, we consider the four classification methods that are best ranked according to the results presented in [9]. Namely, we consider the following classification methods: C4.5 Tree [10], Naïve Bayes Tree [11], Bayesian Network [12] and Functional Tree [13]. The classification model predicts whether a given residue is part of a binding region.

III. Experimental Results

Besides the four descriptive attributes described in the previous section, we also consider the class attribute that shows if the examined residue is part of binding region. For this propose, we use the LigASite v7.0 database [14], which contains biologically relevant binding sites in proteins with known apo-structures. This database contains both the redundant and non-redundant (<25% sequence similarity) sets. The test dataset is formed by the 105408 residues of the 542 chains that are present in the non-redundant set, while the 132773 residues from the remaining 703 chains are considered in the training set. Next, on the training dataset, we perform balancing by down sampling to 20%, and then both sets are normalized in the interval [0;1].

For comparison, we use the following approaches: PIADA [15], Atom nucleus distance [16], ASA change [17], Van der Walls distance [18]. We use their implementations in the PSAIA software [15]. The results for AUC-ROC obtained with our approach and the comparing approaches are shown in Table 1. For our approach, we use the four classification methods stated before. The results show that by using our approach best results are obtained by using the Naïve Bayes Tree and Functional Tree classifiers, while Bayesian Network achieves lowest results. Our approach showed better prediction performance on this dataset than the other approaches used for comparison. In [9], we already showed that the prediction power of the existing methods significantly varies on different datasets meaning that they are applicable only for some group of proteins. In this research, we confirm that our approach has better results for the proteins used in this research, while the other approaches are not applicable for this group of proteins.

TABLE I. The Results for AUC-ROC Obtained With Our Approach and the Comparing Approaches

Approach	**AUC-ROC**
Our approach - C4.5 Tree	0.613
Our approach - Naïve Bayes Tree	0.626
Our approach - Bayesian Network	0.596
Our approach - Functional Tree	0.625
PIADA	0.523
Atom nucleus distance	0.522
ASA change	0.530
Van der Waals distance	0.517

IV. Conclusion

In this paper, we presented an approach that could be used for determination of the binding sites of proteins. First, we extract several characteristics of the amino acid residues and then by using various classification methods we build prediction model.

The results show that our approach obtains best results with the Naïve Bayes Tree and Functional Tree classifiers. Our approach showed better prediction power on the examined dataset, while the existing approaches used for comparison are not applicable for this group of proteins.

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] R. Quinlan, "C4.5: Programs for Machine Learning," 1st ed., Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.

[11] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), Portland, Oregon, USA, August 2–4, 1996, E. Simoudis, J. Han, U. Fayyad, Eds., AAAI Press, Menlo Park, CA, USA, pp. 202–207, 1996.

[12] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," Mach. Learn., vol. 29, no. 2-3, pp. 131–163, 1997.

[13] J. Gama, "Functional Trees," Mach. Learn., vol. 55, no. 3, pp. 219–250, 2004.

[14] B. H. Dessailly, M. F. Lensink, C. A. Orengo, and S. J. Wodak, "LigASite a database of biologically relevant binding sites in proteins with known apo-structures," Nucleic Acids Res., vol. 36 (Database issue), pp. D667–D673, 2008.

[15] J. Mihel, M. Šikić, S. Tomić, B. Jeren, and K. Vlahoviček, "PSAIA – Protein Structure and Interaction Analyzer," BMC Struct. Biol., vol. 8, 21, 2008.

[16] Y. Ofran and B. Rost, "Predicted protein-protein interaction sites from local sequence information," FEBS Lett., 544, no. 1–3, pp. 236–239, 2003.

[17] S. Jones and J. M. Thornton, "Analysis of protein-protein interaction sites using surface patches," J. Mol. Biol., vol. 272, no. 1, pp. 121–132, 1997.

[18] A. S. Aytuna, A. Gursoy, and O. Keskin, "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," Bioinformatics, vol. 21, no. 12, pp. 2850-2855, 2005.

Combining LWE-Solving Algorithms

Dario Gjorgjevski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
gjorgjevski.dario@students.finki.ukim.mk

***Abstract*—The goal of this report is to provide a systematic study of LWE-solving algorithms and explore possibilities of combining and unifying current approaches. The first focus is on the Arora–Ge modeling, which we view as giving rise to another LWE instance with a new error distribution. Then, we adapt the BKW algorithm and provide complexity results for that particular error distribution. At the end, we analyze a combination of two strategies of improving the BKW algorithm: lazy modulus switching and coding. An improvement in both the running time and the memory is obtained over the individual approaches.**

***Keywords*—Learning with errors, algebraic cryptanalysis, multivariate cryptography, Blum–Kalai–Wasserman, modulus switching, lattice codes.**

I. Introduction

Definition 1 (Learning With Errors [1])**.** Let $n, m \geqslant 1$ be integers, q be an odd prime, $\mathcal{X}$ be a probability distribution over $\mathbb{Z}_q$, and $\mathbf{s} \in \mathbb{Z}_q^n$ be a secret vector. Denote by $L_{\mathbf{s},\mathcal{X}}^{(m)}$ the probability distribution over $\mathbb{Z}_q^{n\times m} \times \mathbb{Z}_q^m$ obtained by sampling a matrix $\mathbf{A} \hookleftarrow \mathcal{U}(\mathbb{Z}_q^{n\times m})$, a vector $\mathbf{e} \hookleftarrow \mathcal{X}^m$, and outputting $(\mathbf{A}, \mathbf{sA}+\mathbf{e}) =: (\mathbf{A}, \mathbf{c}) \in \mathbb{Z}_q^{n\times m} \times \mathbb{Z}_q^m$.

Definition 2 (Search-LWE)**.** Search-LWE is the problem of finding $\mathbf{s} \in \mathbb{Z}_q^n$ given $(\mathbf{A}, \mathbf{sA}+\mathbf{e}) \sim L_{\mathbf{s},\mathcal{X}}^{(m)}$.

Denote by $\mathcal{X}_{\alpha,q}$ the discrete Gaussian distribution over $\mathbb{Z}_q$ with mean 0 and standard deviation $\sigma := \alpha q$, which returns an integer $x \in \{\lceil -\frac{q}{2} \rceil, \dots, \lfloor \frac{q}{2} \rfloor\}$ (considered modulo q) with probability proportional to its mass, i.e.,

$$\frac{\exp(-\pi x^2 \,/\, s^2)}{\sum_{y=\lceil -\frac{q}{2} \rceil}^{\lfloor \frac{q}{2} \rfloor} \exp(-\pi y^2 \,/\, s^2)},$$

where $s := \sqrt{2\pi}\sigma$. Typically, $\alpha q = n^\epsilon$ with $0 \leqslant \epsilon \leqslant 1$. When $\epsilon > 1/2$, it has been shown that worst-case GapSVP reduces to average-case LWE [1], [2].

The LWE problem has been used in the design of many cryptographic primitives. Gentry, Peikert, and Vaikuntanathan [3] showed how to construct a trapdoor function based on LWE and created an identity-based cryptosystem. Applebaum *et al.* [4] used LWE to construct encryption scheme with strong security properties. The most important application of LWE, however, has been its use in the design of fully homomorphic encryption schemes (FHE) [5], [6].

II. Arora–Ge modeling

Arora and Ge [7] were the first to model LWE instances algebraically. Their approach reduces an LWE instance to the problem of finding the common root of a multivariate system of high-degree, error-free polynomials.

Namely, let $\left(\mathbf{A} =: \begin{bmatrix}\mathbf{a}_1^\top & \cdots & \mathbf{a}_m^\top\end{bmatrix}, \mathbf{c} =: \begin{bmatrix}c_1 & \cdots & c_m\end{bmatrix}^\top\right) \in \mathbb{Z}_q^{n\times m} \times \mathbb{Z}_q^m$ be an LWE instance with error $\mathbf{e} =: \begin{bmatrix}e_1 & \cdots & e_m\end{bmatrix}^\top \sim \mathcal{X}_{\alpha,q}^m$. Write

$$\mathcal{P}(\mathbf{e}) = \left\{ e_i \prod_{k=1}^{C_{\mathsf{AG}}\sigma} (e_i - k)(e_i + k) \right\} \forall 1 \leqslant i \leqslant m, \tag{1}$$

with $C_{\mathsf{AG}} \geqslant 0$ as a parameter, and note that $\mathcal{P}(\mathbf{e}) = \mathbf{0}$ if for all $1 \leqslant i \leqslant m$ it holds that $e_i \in \{-C_{\mathsf{AG}}\sigma, \dots, C_{\mathsf{AG}}\sigma\}$. $\mathcal{P}(\mathbf{e})$ is a polynomial system over $\mathbb{Z}_q[\mathbf{e}]$ of degree $D_{\mathsf{AG}} := 2C_{\mathsf{AG}}\sigma + 1$. Its purpose is to impose a certain structure on the problem, i.e., that the errors' magnitude should be restricted to a small interval. Using a well-known fact about the Gaussian distribution (Lemma 1), it can be seen that it fits the proposed structure with high probability.

Lemma 1. *Denote by $\mathcal{X}_\sigma$ the Gaussian distribution with mean 0 and standard deviation σ. Furthermore, for $x \geqslant 0$, denote by $Q(x) := \frac{1}{2}(1 - \operatorname{erf}(x \,/\, \sqrt{2}))$. Then, for all $C > 0$ it holds that*

$$\begin{aligned}\Pr[e \hookleftarrow \mathcal{X}_\sigma : |e| > C\sigma] &\approx 2Q(C) \\ &\leqslant \frac{2}{C\sqrt{2\pi}} \exp(-C^2 \,/\, 2) \in \exp(\Omega(-C^2)).\end{aligned}$$

Noting that $e_i = c_i - \langle \mathbf{a}_i, \mathbf{s} \rangle$, we can write Eq. (1) as a polynomial system over $\mathbb{Z}_q[\mathbf{x}]$:

$$\left\{ (c_i - \langle \mathbf{a}_i, \mathbf{x} \rangle) \prod_{k=1}^{C_{\mathsf{AG}}\sigma} (c_i - \langle \mathbf{a}_i, \mathbf{x} \rangle - k)(c_i - \langle \mathbf{a}_i, \mathbf{x} \rangle + k) \right\} \forall 1 \leqslant i \leqslant M_{\mathsf{AG}}. \tag{2}$$

From Lemma 1 a union bound on the probability of failure can be derived,

$$p_f = M_{\mathsf{AG}} \Pr[e \hookleftarrow \mathcal{X}_{\alpha,q} : |e| > C_{\mathsf{AG}}\sigma] \leqslant M_{\mathsf{AG}} \exp(\Omega(-C_{\mathsf{AG}}^2)).$$

Now, we have the following theorem due to Arora and Ge [7]:

Theorem 1 ([7, Theorems 3.1 and 3.3])**.** *Let $D_{\mathsf{AG}} < q$. Taking*

$$M_{\mathsf{AG}} = \mathcal{O}\left(\binom{n + D_{\mathsf{AG}}}{D_{\mathsf{AG}}} \sigma q \log q \right) = n^{\mathcal{O}(D_{\mathsf{AG}})} = 2^{\widetilde{\mathcal{O}}(D_{\mathsf{AG}})}$$

equations as in Eq. (2) *and linearizing the system gives at most one solution with high probability.*

Theorem 1 tells us is that having an additional factor of $\sigma q \log q$ equations allows us to retrieve the secret vector with high probability as long as $\mathcal{P}(\mathbf{e}) = \mathbf{0}$ holds. However, what happens if for some $1 \leqslant j \leqslant m$ we get an error $e_j \notin \{-C_{\mathsf{AG}}\sigma, \dots, C_{\mathsf{AG}}\sigma\}$? The Arora–Ge algorithm invariably fails. That is why we will attempt to consider the modeling as *a new* LWE instance; one where Eq. (1) always holds.

III. Arora–Ge as a new LWE instance

Consider the structure imposed by the Arora–Ge modeling in Eq. (1). As per the previous discussion, we can relax it by introducing a new error term $\hat{\mathbf{e}} \in \mathbb{Z}_q^m$ such that failure never occurs:

$$\mathcal{P}(\mathbf{e}) = \left\{ e_i \prod_{k=1}^{C_{\mathsf{AG}}\sigma} (e_i - k)(e_i + k) + \hat{e}_i \right\} \forall 1 \leqslant i \leqslant m. \tag{3}$$

More specifically, the components $\hat{e}_i$ of $\hat{\mathbf{e}}$ are simply

$$\hat{e}_i := -e_i \prod_{k=1}^{C_{\mathsf{AG}}\sigma} (e_i - k)(e_i + k).$$

Each $\hat{e}_i$ will be zero with probability $1 - \exp(\Omega(-C_{\mathsf{AG}}^2))$. We can expand the products in Eq. (3) to obtain

$$\sum_{k=0}^{C_{\mathsf{AG}}\sigma} t(C_{\mathsf{AG}}\sigma, k)(c_i - \langle \mathbf{a}_i, \mathbf{s}\rangle)^{2k+1} + \hat{e}_i, \tag{4}$$

where $t(n,k)$ denotes the *central factorial number* (see A008955 at OEIS for additional information on central factorial numbers).

Denote by $\widehat{\mathcal{X}}_{p,q}$ the distribution of the $\hat{e}_i$'s, where $p := \Pr[\hat{e} \hookleftarrow\!\!\$\ \widehat{\mathcal{X}}_{p,q} : \hat{e} = 0]$, and by $(\widehat{\mathbf{A}}, \hat{\mathbf{s}}\widehat{\mathbf{A}} + \hat{\mathbf{e}}) =: (\widehat{\mathbf{A}}, \hat{\mathbf{c}})$ the new LWE instance. It can be readily seen that $\widehat{\mathcal{X}}_{p,q}$ satisfies

$$\Pr[\hat{e} \hookleftarrow\!\!\$\ \widehat{\mathcal{X}}_{p,q} : \hat{e} = 0] =: p \in 1 - \exp(\Omega(-C_{\mathsf{AG}}^2))$$
$$\Pr[\hat{e} \hookleftarrow\!\!\$\ \widehat{\mathcal{X}}_{p,q} : \hat{e} \neq 0] = 1 - p \in \exp(\Omega(-C_{\mathsf{AG}}^2)).$$

We will make two heuristic assumptions: one regarding the distribution of the nonzero components of $\hat{\mathbf{e}}$, and the other regarding the distribution of $\widehat{\mathbf{A}}$.

Assumption 1. $\widehat{\mathcal{X}}_{p,q}$ is a probability distribution over $\mathbb{Z}_q$ which returns an integer $x \in \{\lceil -\frac{q}{2} \rceil, \ldots, \lfloor \frac{q}{2} \rfloor\}$ (considered modulo q) with probability

$$\begin{aligned} p &\in 1 - \exp(\Omega(-C_{\mathsf{AG}}^2)) && \text{if } x = 0, \\ \frac{1-p}{q-1} &\in \frac{\exp(\Omega(-C_{\mathsf{AG}}^2))}{q-1} && \text{if } x \neq 0. \end{aligned}$$

In other words, the probability mass of $\exp(\Omega(-C_{\mathsf{AG}}^2))$ is distributed uniformly among the $q-1$ nonzero values in the support.

Assumption 2. $\widehat{\mathbf{A}}$ is distributed uniformly at random, i.e., it follows Definition 1.

Assumption 1 allows us to prove Lemma 2 regarding the distribution of sums of i.i.d. random variables following $\widehat{\mathcal{X}}_{p,q}$.

Lemma 2. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_i \sim \widehat{\mathcal{X}}_{p,q}$. Define $X := \sum_{i=1}^{n} X_i \pmod q$, and recall that $p := \Pr[X_i = 0]$. X is distributed as*

$$\Pr[X = 0] = \frac{1}{q}\left(1 + (q-1)\left(\frac{pq-1}{q-1}\right)^n\right)$$
$$\Pr[X = j] = \frac{1 - \Pr[X=0]}{q-1} \text{ for } j \neq 0.$$

Proof. Let p_k denote the probability that k nonzero values drawn from $\widehat{\mathcal{X}}_{p,q}$ add up to 0. Note that these elements are assumed to be uniform. This probability satisfies the recurrence

$$p_{k+1} = \frac{1}{q-1}(1 - p_k),$$

as the only way ho have $k+1$ nonzero values add up to 0 is to have the previous k elements add up to some nonzero value j (probability of $1 - p_k$) and we pick $-j$ for the $(k+1)$-st value (probability of $1 / (q-1)$). The initial condition is $p_0 = 1$. Solving the recurrence yields

$$p_k = \frac{1 - (1-q)^{1-k}}{q}.$$

We add k nonzero elements with probability $\binom{n}{k}(1-p)^k p^{n-k}$, so the probability to get 0 is

$$\begin{aligned} \Pr[X = 0] &= \sum_{k=0}^{n} \binom{n}{k}(1-p)^k p^{n-k} \frac{1-(1-q)^{1-k}}{q} \\ &= \frac{1}{q}(1 - (1-q)^{1-n}(1 - p + p(1-q))^n) \\ &= \frac{1}{q}\left(1 + (q-1)\left(\frac{pq-1}{q-1}\right)^n\right). \end{aligned}$$

Since we're working in a group, every nonzero value can be reached from every other nonzero value in a unique way, so it can be argued by induction that the nonzero values receive the rest of the probability mass distributed uniformly among them.

Another way to arrive at the same result is to solve the recurrence

$$\Pr\left[\sum_{i=1}^{n+1} X_i = 0\right] = p \Pr\left[\sum_{i=1}^{n} X_i = 0\right] + \frac{1-p}{q-1}\left(1 - \Pr\left[\sum_{i=1}^{n} X_i = 0\right]\right).$$

□

IV. The BKW algorithm

The BKW algorithm due to Blum, Kalai, and Wasserman [8] was the first sub-exponential algorithm given for solving the Learning Parity With Noise (LPN) problem. Since LPN can be seen as a special case of LWE where $q = 2$, the BKW algorithm can be adapted to solve Search-LWE with general moduli.

One can view BKW as a variant of the standard Gaussian elimination which avoids multiplications and tries to eliminate entire blocks of elements in order to prevent the error from blowing up.

The algorithm takes a parameter $1 \leqslant b < n$, the *block size*, and defines the *addition depth* as $a := \lceil n/b \rceil$. It then repeatedly eliminates blocks of b elements per row addition over $a-1$ rounds and obtains samples for recovering blocks of $\mathbf{s}$.

Albrecht *et al.* [9] illustrate the BKW algorithm in three stages:

Stage 1. Sample reduction;
Stage 2. Hypothesis testing; and
Stage 3. Back substitution.

Below we describe the three stages and in particular adapt the *hypothesis testing* stage to the $\widehat{\mathcal{X}}_{p,q}$ distribution.

A. Sample reduction

Given an LWE oracle $L_{\mathbf{s},\mathcal{X}}$, the goal is to construct a series of oracles $B_{\mathbf{s},\mathcal{X},l}$, each of which produces samples $(\mathbf{a}, c)$ where the first lb elements of $\mathbf{a}$ are zero. To sample from $B_{\mathbf{a},\mathcal{X},1}$, we query $B_{\mathbf{a},\mathcal{X},0} := L_{\mathbf{a},\mathcal{X}}$ and store the samples $(\mathbf{a}, c)$ in a table T_1.

If T_1 contains a sample $(\mathbf{a}', c')$ s.t. $\mathbf{a}$ and $\pm\mathbf{a}'$ agree on their first b coordinates, we do not store $(\mathbf{a}, c)$, but output $(\mathbf{a} \mp \mathbf{a}', c \mp c')$ instead. If the sample from $B_{\mathbf{a},\mathcal{X},0}$ already has its first b elements be zero, we output it directly as a sample from $B_{\mathbf{a},\mathcal{X},1}$.

For $1 < l < a$, the algorithm proceeds recursively by populating a table T_l of samples from $B_{\mathbf{a},\mathcal{X},l-1}$ and outputting a sample as soon as a collision or a zero block is found.

Due to the symmetry of $\mathbb{Z}_q$, a table T_l contains at most $(q^b - 1)/2$ samples. Thus, to obtain m samples from $B_{\mathbf{a},\mathcal{X},l}$, one has to perform at most $m + (q^b - 1)/2$ queries to $B_{\mathbf{a},\mathcal{X},l-1}$.

The exact procedure is given in Algorithm 1. The complexity of obtaining m samples from the last oracle, $B_{\mathbf{s},\mathcal{X},a-1}$, is given by Lemma 3.

Lemma 3 ([9, Lemma 2])**.** *Let q be an odd prime, and $L_{\mathbf{s},\mathcal{X}}$ be an LWE oracle with $\mathbf{s} \in \mathbb{Z}_q^n$. Let $1 \leqslant b < n$ and $a := \lceil n / b \rceil$. Define $n' := n - (a-1)b$. The worst case complexity of obtaining m samples $(\mathbf{a}_i, c_i)$, where the $\mathbf{a}_i$'s contain at most n' nonzero components is upper bounded by*

$$\left(\frac{q^b - 1}{2}\right)\left(\frac{(a-1)(a-2)}{2}\right)\left((n+1) - \frac{ab}{3}\right) + m\left(\frac{a-1}{2}(n+2)\right) \in \mathcal{O}(q^b amn)$$

additions in $\mathbb{Z}_q$ and

$$(a-1)\frac{q^b - 1}{2} + m \in \mathcal{O}(\max(q^b a, m))$$

Algorithm 1 Oracle $B_{\mathbf{a},\mathcal{X},l}$ with parameters $1 \leqslant b < n$ and $a := \lceil n \,/\, b \rceil$ $(1 \leqslant l < a)$

State: Table T_l, initially empty
Output: An LWE sample $(\mathbf{a}, c)$ with $\mathbf{a}_{(l..lb)} = \mathbf{0}$

```
loop
    (a, c) ←$ B_{a,X,l-1}                ▷ a_{(1..(l-1)b)} = 0 by definition
    if a_{((l-1)b+1..lb)} = 0 then
        return (a, c)
    else if ∃(a', c') ∈ T_l . a_{((l-1)b+1..lb)} = ±a'_{((l-1)b+1..lb)} then
        return (a ∓ a', c ∓ c')
    else
        T_l ← T_l ∪ {(a, c)}
    end if
end loop
```

queries to $L_{\mathbf{s},\mathcal{X}}$.

The memory required for the tables T_1 *through* T_{a-1} *is upper bounded by*

$$\left(\frac{q^b-1}{2}\right)(a-1)\left(n+1-\frac{(a-2)b}{2}\right) \in \mathcal{O}(q^b a n)$$

elements of $\mathbb{Z}_q$.

B. Hypothesis testing

After the sample reduction step, we are left with an oracle $B_{\mathbf{s},\mathcal{X},a-1}$ outputting samples $(\mathbf{a}, c)$ with $\mathbf{a}$ having at most $n' \leqslant b$ nonzero components. Equivalently, we can view these as samples in $\mathbb{Z}_q^{n'} \times \mathbb{Z}_q$. Let $\mathbf{s}'$ denote the n' nonzero components of $\mathbf{s}$.

Since $\mathbf{a}$ was obtained by summing or subtracting up to 2^{a-1} samples of the original oracle, the error $c - \langle \mathbf{a}, \mathbf{s}' \rangle$ follows a distribution of the sum of up to 2^{a-1} original errors. Even though the analysis will be agnostic to the exact nature of this distribution, Lemma 2 showed that the distribution of the sum of $\widehat{\mathcal{X}}_{p,q}$-distributed i.i.d. random variables retains the same structure. This is not necessarily true for the discrete Gaussian case, even though existing literature, e.g. [9], [10], assumes it. The problem of recovering $\mathbf{s}'$ is equivalent to the problem of distinguishing between the error distributions for a guess $\mathbf{v}$ when $\mathbf{v} = \mathbf{s}'$ and $\mathbf{v} \neq \mathbf{s}'$.

Duc, Tramèr, and Vaudenay [11] proposed an improved version of the original BKW algorithm [8] which uses multidimensional Fourier transforms instead of maximum likelihood [9] to perform hypothesis testing. The result is an algorithm whose analysis we will adapt here.

Definition 3 (Multidimensional discrete Fourier transform)**.** Let $p_1, \ldots, p_b$ be integers, and let $\theta_{p_j} := \exp(2\pi i \,/\, p_j)$ for $1 \leqslant j \leqslant b$. Define the group $G := \prod_{i=1}^{b} \mathbb{Z}_{p_j}$. An element $\mathbf{x} \in G$ can be represented as $(x_1, \ldots, x_b)$ with $x_j \in \mathbb{Z}_{p_j}$. The discrete Fourier transform (DFT) of a function $f \colon G \to \mathbb{C}$ is a function $\mathcal{F}\{f\} \colon G \to \mathbb{C}$ defined as

$$\mathcal{F}\{f\}(\boldsymbol{\alpha}) := \sum_{\mathbf{x} \in G} f(\mathbf{x}) \theta_{p_1}^{-\alpha_1 x_1} \cdots \theta_{p_b}^{-\alpha_b x_b}.$$

The DFT can be computed in time $\mathcal{O}(|G| \log |G|) = C_{\mathsf{FFT}} |G| \log |G|$ for a small constant $C_{\mathsf{FFT}} > 0$.

Consider the function

$$f(\mathbf{x}) := \sum_{j=1}^{m} \mathbb{1}_{\{\mathbf{a}_j = \mathbf{x}\}} \theta_q^{c_j}$$

and its DFT

$$\begin{aligned}\mathcal{F}\{f\}(\boldsymbol{\alpha}) =: F(\boldsymbol{\alpha}) &= \sum_{\mathbf{x} \in \mathbb{Z}_q^{n'}} f(\mathbf{x}) \theta_q^{-\langle \mathbf{x}, \boldsymbol{\alpha} \rangle} = \sum_{\mathbf{x} \in \mathbb{Z}_q^{n'}} \sum_{j=1}^{m} \mathbb{1}_{\{\mathbf{a}_j = \mathbf{x}\}} \theta_q^{c_j} \theta_q^{-\langle \mathbf{x}, \boldsymbol{\alpha} \rangle} \\ &= \sum_{j=1}^{m} \theta_q^{-(\langle \mathbf{a}_j, \boldsymbol{\alpha} \rangle - c_j)}.\end{aligned}$$

Note that in particular

$$F(\mathbf{s}') = \sum_{j=1}^{m} \theta_q^{-(e_{j,1} \pm \cdots \pm e_{j,2^{a-1}})},$$

where the $e_{j,i}$'s are independent samples from $\widehat{\mathcal{X}}_{p,q}$.

Duc, Tramèr, and Vaudenay [11] showed that for appropriate values of m and a, $\arg\max_{\boldsymbol{\alpha} \in \mathbb{Z}_q^{n'}} \Re(F(\boldsymbol{\alpha})) = \mathbf{s}'$ with high probability. To adapt their result for the case where the errors follow a $\widehat{\mathcal{X}}_{p,q}$ distribution, we need the following lemma:

Lemma 4. *Let* $X \sim \widehat{\mathcal{X}}_{p,q}$ *and* $Y := 2\pi X \,/\, q$. *Then*

$$\begin{aligned}\mathbb{E}[\cos Y] &= \frac{pq-1}{q-1}, \text{ and} \\ \mathbb{E}[\sin Y] &= 0.\end{aligned}$$

Proof.

$$\begin{aligned}\mathbb{E}[\cos Y] &= \sum_{x=\lceil -\frac{q}{2} \rceil}^{\lfloor \frac{q}{2} \rfloor} \cos\left(\frac{2\pi x}{q}\right) \Pr[X = x] \\ &= Pr[X = 0] + 2 \sum_{x=1}^{\lfloor \frac{q}{2} \rfloor} \cos\left(\frac{2\pi x}{q}\right) \Pr[X = x] \\ &= p + 2\frac{1-p}{q-1} \sum_{x=1}^{\lfloor \frac{q}{2} \rfloor} \cos\left(\frac{2\pi x}{q}\right) = p - \frac{1-p}{q-1} = \frac{pq-1}{q-1}.\end{aligned}$$

$\mathbb{E}[\sin Y] = 0$ follows from the symmetry of both the distribution and the sine function, and the fact that $\sin 0 = 0$. □

Adapting the analysis from [11] to $\widehat{\mathcal{X}}_{p,q}$, we can state Theorem 2.

Theorem 2. *Let* q *be an odd prime,* $\mathcal{X} := \widehat{\mathcal{X}}_{p,q}$ *be the error distribution, and* $L_{\mathbf{s},\mathcal{X}}$ *be an LWE oracle with* $\mathbf{s} \in \mathbb{Z}_q^n$. *Let* $1 \leqslant b < n$ *and* $a := \lceil n \,/\, b \rceil$ *be parameters to the BKW algorithm. Define* $n' := n - (a-1)b$, *and let* $B_{\mathbf{s},\mathcal{X},a-1}$ *be an oracle constructed according to Algorithm 1 outputting samples* $(\mathbf{a}_i, c_i)$. *Denote the* n' *nonzero components of* $\mathbf{s}$ *as* $\mathbf{s}'$.

Fix an upper bound on the failure probability, $0 < \epsilon < 1$. *Then, the number of independent samples* m_ϵ *from* $B_{\mathbf{s},\mathcal{X},a-1}$ *required s.t. we fail to recover the secret block* $\mathbf{s}'$ *with probability at most* ϵ *satisfies*

$$m_\epsilon \geqslant 8n' \log\left(\frac{q}{\epsilon^{1/n'}}\right) \left(\frac{pq-1}{q-1}\right)^{-2^a}.$$

Proof. For a fixed upper bound $0 < \epsilon < 1$, we have

$$\Pr[\exists \boldsymbol{\alpha} \neq \mathbf{s}' . \Re(F(\boldsymbol{\alpha})) \geqslant \Re(F(\mathbf{s}'))] \leqslant q^{n'} \exp\left(-\frac{m_\epsilon}{8} \mathbb{E}[\cos Y]^{2^a}\right) \leqslant \epsilon.$$

Solving the last inequality for m_ϵ, we get the desired result. □

C. Back substitution

Once n' elements of $\mathbf{s}$ have been recovered, back substitution can be performed through the tables T_j, zeroing out n' elements in each sample. Note that the last oracle then becomes superfluous and can freely be discarded along with its table.

D. Complexity of the BKW algorithm with $\widehat{\mathcal{X}}_{p,q}$

Following the analysis from [11], we derive that the optimal value for the parameter a of the BKW algorithm is to set

$$a = W_0\left(\frac{-n\log q\log 2}{\log\left(p-\frac{1}{q}\right)}\right)/\log 2,$$

where W_0 is the principal branch of the Lambert-W function.

This choice of a does not yield sub-exponential complexity. This can be shown using the known result that for all $x \geqslant e$ it holds that $\log x \leqslant W_0(x)$.

V. Improvements to the BKW collision-finding procedure

As we saw, the cost of the BKW algorithm is largely determined by its collision-finding procedure. To this end, several improvements have been proposed in the literature. The two most notable are:

- Lazy modulus switching [12]; and
- Coded-BKW using lattice codes [10].

In what follows we would like to combine the two steps.

A. Lazy modulus switching

Lazy modulus switching [12] lets us perform collisions more "loosely." Namely, it searches for collisions modulo p rather than q, where $p < q$. However, arithmetic is still performed in $\mathbb{Z}_q$.

More formally, instead of checking whether $\mathbf{a}_{((l-1)b+1..lb)} = \pm\mathbf{a}'_{((l-1)b+1..lb)}$ when receiving a sample $(\mathbf{a}', c')$ at stage l, we check whether $\lceil \mathbf{a}_{((l-1)b+1..lb)} / D \rfloor = \pm\lceil \mathbf{a}'_{((l-1)b+1..lb)} / D \rfloor$, where $D := q/p$. This is viewed as a "looser" form of equality – equality modulo p rather than q. Of course, while this step practically reduces the problem from $\mathbb{Z}_q$ to $\mathbb{Z}_p$, it introduces additional error. Note that errors introduced early end up being added exponentially many times.

We can quantify the error introduced at stage l as

$$\overbrace{e \pm e'}^{\text{original error}} + \overbrace{\sum_{j=(l-1)b+1}^{lb} s_j(a_j \mp a'_j)}^{\text{new error due to mismatch}} \leqslant e \pm e' + D\sum_{j=(l-1)b+1}^{lb} s_j,$$

because

$$\left\lceil \frac{\mathbf{a}_{((l-1)b+1..b)}}{D} \right\rfloor = \pm\left\lceil \frac{\mathbf{a}'_{((l-1)b+1..b)}}{D} \right\rfloor \iff \left|\frac{\mathbf{a}}{D} \mp \frac{\mathbf{a}'}{D}\right| \leqslant 1 \iff |\mathbf{a} \mp \mathbf{a}'| \leqslant D.$$

In the worse case, the final error can be seen to include an additional

$$D\sum_{l=1}^{a-1} 2^{a-l+1}$$

number of error samples originating from the modulus switching performed at stage l. If secret error switching [13] is performed, then these errors also come from the same distribution as the original error and their distribution can be calculated exactly.

The variance of the error introduced by rounding, $\sigma^2_{\mathsf{round}}$, is equal to the variance of uniformly random elements in $\mathbb{Z}_{\lceil D \rfloor}$,

$$\sigma_{\mathsf{round}} = \frac{\lceil D \rfloor^2 - 1}{12}.$$

However, for the final variance, Albrecht *et al.* [12] showed that a bound of $n2^a\sigma^2\sigma^2_{\mathsf{round}}$ is not tight, and is instead given by

$$b(2^a - 1)\sigma^2\sigma^2_{\mathsf{round}}, \tag{5}$$

i.e., it is smaller by a factor of a.

Throughout the rest of the report we will be using the more pessimistic upper bound: the variance $\sigma^2_{\mathsf{round}}$ added up exponentially many times based on the number of stages. Adapting the analysis to use Eq. (5) is left as future work.

B. Lattice codes

A lattice Λ is a discrete additive subgroup of $\mathbb{R}^n$. More formally, Λ is a lattice if and only if there exists a basis $\mathbf{b}_1, \ldots, \mathbf{b}_m \in \mathbb{R}^n$ such that any $\mathbf{y} \in \Lambda$ can be written as an integer linear combination of the basis vectors, i.e., $\mathbf{y} = \sum_{i=1}^m \alpha_i \mathbf{b}_i$ with $\alpha_i \in \mathbb{Z}$.

The class of lattices we will be interested in is based on q-ary linear codes. If $\mathcal{C}$ is a linear $[N, k]$ code over $\mathbb{Z}_q$, then a lattice over $\mathcal{C}$ is

$$\Lambda(\mathcal{C}) = \{\boldsymbol{\lambda} \in \mathbb{R}^n : \boldsymbol{\lambda} \equiv \mathbf{c} \pmod q, \mathbf{c} \in \mathcal{C}\}.$$

A typical application is to use the lattice $\Lambda(\mathcal{C})$ as a codebook for quantization of sequences $\mathbf{x} \in \mathbb{R}^n$. Let $Q(\mathbf{x})$ be the lattice point closes to $\mathbf{x}$ if the squared error is used as a fidelity criterion. $Q(\mathbf{x})$ is then a MSE quantizer. The theory of lattice codes allows us to bound the variance of the quantization error. More specifically, it can be expressed as

$$\sigma^2_{\mathsf{code}} = G(\Lambda)\,\mathrm{Vol}(\mathcal{V})^{\frac{2}{n}},$$

where $G(\Lambda)$ is the normalized second moment of Λ, representing a figure of merit of a lattice quantizer with respect to the MSE measure; $\mathrm{Vol}(\cdot)$ is the volume of a closed set in $\mathbb{R}^n$; and $\mathcal{V}$ is the fundamental Voronoi region of Λ, i.e.,

$$\mathcal{V} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leqslant \|\mathbf{x} - \mathbf{w}\|, \forall \mathbf{w} \in \Lambda\},$$

where $\|\cdot\|$ is the L^2 norm.

If we let $G(\Lambda_n)$ be the minimum possible value of $G(\Lambda)$ over all lattices $\Lambda \subseteq \mathbb{R}^b$, it is known that

$$\frac{1}{2\pi e} < G(\Lambda_n) \leqslant \frac{1}{12}. \tag{6}$$

Moreover, the bounds given by Eq. (6) are tight. At stage l of the BKW algorithm, we fix a q-ary linear $[N_l, b]$ code, denoted $\mathcal{C}_l$. This code gives rise to a lattice code. Now, instead of colliding vectors $\mathbf{a}$ and $\mathbf{a}'$ if there is a match between their leading nonzero blocks of length b, we collide them when two codewords decode to the same block.

This is performed by looking at blocks (codewords) of length N_l, decoding them to blocks of length b, and then comparing those for equality. If they are equal, we eliminate *the entire block of length* N_l and put $\mathbf{0}$ in its place. The leftover due to inexactness is treated *as additional error*. σ^2_{code} quantifies the *variance of that error*.

The approach outlined above has the advantage of eliminating $N_l \geqslant b$ components per reduction step, at the cost of introducing additional error. Note that as in lazy modulus switching (Section V-A), errors introduced early in the reduction end up being added exponentially many times later in the algorithm.

Denote by $\Lambda_{[N,k]}$ the maximum possible value of $G(\Lambda)$ over all lattices Λ generated by an $[N, k]$ linear code. By definition $G(\Lambda_{[N,k]}) \geqslant G(\Lambda_N)$, so Eq. (6) tells us that

$$G(\Lambda_{[N,k]}) \leqslant \frac{1}{12}.$$

If a lattice is built from an $[N, k]$ linear code by *Construction A* [14], then $\mathrm{Vol}(\mathcal{V}) = q^{N-k}$. Thus,

$$\sigma_{\mathsf{code}} \approx \frac{q^{1-k/N}}{\sqrt{12}}.$$

The decoding procedure and the full details are described in [10]. It is simple syndrome decoding using the square error metric. Note that a single reduction results in *a pair* of blocks being erred with variance σ^2_{code}; therefore, the total error due to coding has variance $2\sigma^2_{\mathsf{code}}$.

C. Combining lazy modulus switching and coding

The good thing about the two improvements described in the previous sections is that they are *orthogonal* in a sense and can be used in conjunction. The overall algorithm will have the following structure:

Step 1. Perform secret error switching [13] to change the distribution of the secret to that of the error;
Step 2. Perform a_1 regular BKW reductions to eliminate $a_1 b$ components;
Step 3. Perform a_2 BKW reductions using lazy modulus switching and coding using $[N_l, b]$ linear codes to eliminate $n_{\mathsf{code}} = \sum_{l=1}^{a_2} N_l$ components;
Step 4. Guess the top n_{top} components;
Step 5. Perform *subspace hypothesis testing* using an $[n_{\mathsf{test}}, k]$ linear code and the multidimensional DFT.

To align with existing literature, we go back to the assumption that the errors of the LWE instance follow a discrete Gaussian distribution $\mathcal{X}_{\alpha,q}$ with standard deviation $\sigma := \alpha q$. It is intuitive that modulus switching is performed first as a smaller modulus would make the quantization less erroneous, too.

Note that in the final error expression there are 2^{a_2-l+1} error terms arising from the l-th coded BKW step coupled with modulus switching. We would like to match this error with the error introduced in the *subspace hypothesis testing* step, $\sigma^2_{\mathsf{pre\text{-}set}}$.

Thus, we get the following balance equation for stage l:

$$\begin{aligned} \sigma^2_{\mathsf{pre\text{-}set}} &= 2^{a_2-l+1}(\sigma^2_{\mathsf{code}} + \sigma^2_{\mathsf{round}}) \\ \implies \sigma^2_{\mathsf{pre\text{-}set}} &= 2^{a_2-l+1}\left(\frac{p^{2(1-b/N_l)} + \lceil D \rfloor^2 - 1}{12}\right), \end{aligned}$$

with a solution for N_l given by

$$N_l = \left\lfloor \frac{b}{1 - \frac{1}{2}\log_p\left(12\left(\frac{\sigma^2_{\mathsf{pre\text{-}set}}}{2^{a_2-l+1}} - \sigma^2_{\mathsf{round}}\right)\right)} \right\rfloor. \tag{7}$$

Note that the constraint

$$\frac{\sigma^2_{\mathsf{pre\text{-}set}}}{2^{a_2-l+1}} > \sigma^2_{\mathsf{round}} \tag{8}$$

which arises from Eq. (7) is a natural one – indeed, if it were not satisfied, then the error introduced just by modulus switching would end up being greater than $\sigma^2_{\mathsf{pre\text{-}set}}$.

D. Complexity of the algorithm

Our algorithm is parameterized by the following parameters:

- m and n: the number of samples and length of the secret respectively;
- a_1 and a_2: the number of blocks to eliminate using regular BKW reductions and coded BKW reductions coupled with modulus switching respectively;
- b: the "block size" of the BKW reductions;
- p: the new modulus, with $D := q \,/\, p$;
- n_{test} and k: parameters of the linear code used in the hypothesis testing step; and
- n_{top} and d: the number of components to guess and the range in which they should be guessed respectively.

In the following paragraphs, we adapt the complexity analysis from [10] to include lazy modulus switching.

a) Secret error switching: If $n' := n - a_1 b$ is the number of nonzero components after the initial BKW reductions, then the complexity of this step is upper bounded by

$$t_{\mathsf{switch}} = m(n+1)\left\lceil \frac{n'}{b-1} \right\rceil.$$

b) Regular BKW reductions: Since modulus switching is not yet performed at this stage, there are $(q^b - 2) \,/\, 2$ equivalence classes. The number of equivalence classes determined the number of samples lost per reduction stage, following the reasoning in Section IV-A, we obtain an upper bound of

$$t_{\mathsf{BKW}} = \sum_{k=1}^{a_1}(n + 1 - kb)\left(m - \frac{k(q^b-1)}{2}\right).$$

c) Coded-BKW reductions coupled with modulus switching: We are now working modulo p and using a p-ary $[N_l, b]$ linear code at each stage, where the N_l's are given by Eq. (7). Following [10], the decoding cost is upper bounded by

$$t_{\mathsf{decode}} = \sum_{k=1}^{a_2} 4\left(M + \frac{k(p^b-1)}{2}\right) N_k,$$

where M is the number of samples after the last reduction stage. Note that in the $(a_2 - k + 1)$-st stage, the number of processed samples is $M + k(p^b - 1) \,/\, 2$. Thus, the overall complexity is

$$t_{\mathsf{Coded\text{-}BKW}} = t_{\mathsf{decode}} + \sum_{k=1}^{a_2}\left(n_{\mathsf{top}} + n_{\mathsf{test}} + \sum_{l=1}^{k} N_l\right)\left(M + \frac{(k-1)(p^b-1)}{2}\right).$$

d) Partial guessing: Due to the secret error switching step, the secret follows a Gaussian distribution. We can take advantage of Lemma 1 and guess the top n_{top} components. More formally, given a parameter $d > 0$, we exhaust all possible entries with absolute value less or equal to than d. There are $(2d+1)^{n_{\mathsf{top}}}$ such entries, yielding a complexity of

$$t_{\mathsf{guess}} = M n_{\mathsf{top}}(2d+1)^{n_{\mathsf{top}}},$$

i.e., the complexity of guessing and updating the observed samples.

e) Subspace hypothesis testing: The subspace hypothesis testing step can be shown (see [10] for the details) to have a complexity of

$$t_{\mathsf{test}} = 4M n_{\mathsf{test}} + (2d+1)^{n_{\mathsf{top}}}(C_{\mathsf{FFT}} q^{k+1}(k+1)\log q + q^{k+1}).$$

f) Total complexity: Let $n_{\mathsf{total}} = n_{\mathsf{code}} + n_{\mathsf{test}}$ denote the total length affected by the coding and modulus switching. For a constant $C > 0$, we set a noise level $C^2\sigma^2\sigma^2_{\mathsf{pre\text{-}set}} n_{\mathsf{total}}$ to be the variance introduced by the coding and modulus switching, and then compute the required number of samples following literature on linear cryptanalysis ([15], [16]).

Denote by P_{test} the probability that the Euclidean length of $\mathbf{s}$ at the end of the reductions is less than or equal to $C\sigma\sqrt{n_{\mathsf{total}}}$. Using Lemma 1, we can set $C = 1.2$ for a probability of roughly 0.975. Let $P(d)$ denote the probability that the absolute value of one guessed entry of $\mathbf{s}$ is less than or equal to d.

Theorem 3. *Let n, q, σ be the parameters of the chosen LWE instance, and $a_1, a_2, b, p, n_{\mathsf{test}}, k, d$ be the parameters of the Coded-BKW algorithm coupled with lazy modulus switching. The complexity is then given by*

$$\frac{t_{\mathsf{switch}} + t_{\mathsf{BKW}} + t_{\mathsf{Coded\text{-}BKW}} + t_{\mathsf{guess}} + t_{\mathsf{test}}}{P(d)^{n_{\mathsf{top}}} P_{\mathsf{test}}}. \tag{9}$$

The required number of samples for testing, M, is set to be

$$M = \frac{4\log(2d+1)^{n_{\mathsf{top}}}q^k}{\Delta(\mathcal{X}_{\sigma_{\mathsf{final}}} \parallel \mathcal{U}(\mathbb{Z}_q))},$$

where $\mathcal{U}(\mathbb{Z}_q)$ is the uniform distribution over $\mathbb{Z}_q$, and $\mathcal{X}_{\sigma_{\mathsf{final}}}$ is the discrete Gaussian distribution over $\mathbb{Z}_q$ with standard deviation $\sigma_{\mathsf{final}} := 2^{a_1+a_2}\sigma^2 + C^2\sigma^2\sigma^2_{\mathsf{pre\text{-}set}}n_{\mathsf{total}}$. $\Delta(\mathcal{X}_{\sigma_{\mathsf{final}}} \parallel \mathcal{U}(\mathbb{Z}_q))$ denotes the divergence between the two probability distributions and is computed numerically.

The number of calls to the LWE oracle is

$$m = \frac{a_1(q^b-1) + a_2(p^b-1)}{2} + M.$$

Proof. The result is simply given by the cost of one iteration divided by the expected success probability. The number of required samples is as per [15], [16]. □

E. Experimental results

The complexity was determined experimentally for an LWE instance instantiated with security parameter $n = 128$ following [1]. The instance was generated by the open-source generator by Albrecht *et al.* [17]. Calculations were done by numerically evaluating Eq. (9).

We found that it is beneficial to decrease the new modulus, p, as low as possible without violating Eq. (8). In the case of the Regev instance with $n = 128$, for which $q = 16\,411$, the lowest such p is $p \approx 2q\,/\,3$.

The reported complexity using Coded-BKW with lazy modulus switching is $2^{83.6}$, as opposed to "vanilla" Coded-BKW whose estimated complexity is $2^{85.1}$. For the Regev instance with $n = 512$, for which $q = 262\,147$, the complexity was reduced to $2^{285.86}$ from $2^{287.77}$. Furthermore, modulus switching reduced the required memory in both cases by roughly a factor of 2.5.

VI. Conclusion and future work

In this report, we analyzed several algorithms for solving the LWE problem. We attempted to arrive at a sub-exponential algorithm by combining known approaches into a single procedure. Unfortunately, complexity analysis showed that all such BKW-like approaches are fundamentally limited by the trade-off of obtaining reduced samples and the addition of errors.

At the end, we presented an improvement to the Coded-BKW algorithm by combining it with lazy modulus switching. Experimental results showed improvement for the Regev LWE instance with $n = 128$, which is a widely-used "benchmark" LWE instance.

Even though current state-of-the-art approaches (e.g., [13]) are better (although not by much), they, too, are based off of the BKW algorithm and are not likely to achieve sub-exponential complexity. We leave it as future work to tighten the analysis and perhaps combine (lazy) modulus switching and coding in a smarter way.

References

[1] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," in *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, ser. STOC '05, ACM, 2005, pp. 84–93.

[2] Z. Brakerski *et al.*, "Classical hardness of learning with errors," in *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, ser. STOC '13, ACM, 2013, pp. 575–584.

[3] C. Gentry, C. Peikert, and V. Vaikuntanathan, "Trapdoors for hard lattices and new cryptographic constructions," in *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, ser. STOC '08, ACM, 2008, pp. 197–206.

[4] B. Applebaum *et al.*, "Fast cryptographic primitives and circular-secure encryption based on hard learning problems," in *Advances in Cryptology – CRYPTO 2009: 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16–20, 2009. Proceedings*, S. Halevi, Ed. Springer Berlin Heidelberg, 2009, pp. 595–618.

[5] Z. Brakerski and V. Vaikuntanathan, *Efficient fully homomorphic encryption from (standard) LWE*, Cryptology ePrint Archive, Report 2011/344, 2011.

[6] C. Gentry, A. Sahai, and B. Waters, *Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based*, Cryptology ePrint Archive, Report 2013/340, 2013.

[7] S. Arora and R. Ge, "New algorithms for learning in presence of errors," in *Proceedings of the 38th International Colloquim Conference on Automata, Languages and Programming*, ser. ICALP '11, vol. 1, Springer-Verlag, 2011, pp. 403–415.

[8] A. Blum, A. Kalai, and H. Wasserman, "Noise-tolerant learning, the parity problem, and the statistical query model," *J. ACM*, vol. 50, no. 4, pp. 506–519, Jul. 2003.

[9] M. R. Albrecht *et al.*, *On the complexity of the BKW algorithm on LWE*, Cryptology ePrint Archive, Report 2012/636, 2012.

[10] Q. Guo, T. Johansson, and P. Stankovski, "Coded-BKW: Solving LWE using lattice codes," in *Advances in Cryptology – CRYPTO 2015: 35th Annual Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2015, Proceedings, Part I*, R. Gennaro and M. Robshaw, Eds. Springer Berlin Heidelberg, 2015, pp. 23–42.

[11] A. Duc, F. Tramèr, and S. Vaudenay, *Better algorithms for LWE and LWR*, Cryptology ePrint Archive, Report 2015/056, 2015.

[12] M. R. Albrecht *et al.*, *Lazy modulus switching for the BKW algorithm on LWE*, Cryptology ePrint Archive, Report 2014/019, 2014.

[13] P. Kirchner and P.-A. Fouque, *An improved BKW algorithm for LWE with applications to cryptography and lattices*, Cryptology ePrint Archive, Report 2015/552, 2015.

[14] J. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed., ser. Grundlehren der mathematischen Wissenschaften. Springer-Verlag New York, 1999, vol. 290.

[15] T. Baignères, P. Junod, and S. Vaudenay, "How far can we go beyond linear cryptanalysis?" In *Advances in Cryptology - ASIACRYPT 2004: 10th International Conference on the Theory and Application of Cryptology and Information Security, Jeju Island, Korea, December 5-9, 2004. Proceedings*, P. J. Lee, Ed. Springer Berlin Heidelberg, 2004, pp. 432–450.

[16] A. A. Selçuk, "On probability of success in linear and differential cryptanalysis," *Journal of Cryptology*, vol. 21, no. 1, pp. 131–147, 2008.

[17] M. R. Albrecht *et al.* (2013). A generator for LWE and Ring-LWE instances, `https://www.iacr.org/news/files/2013-04-29lwe-generator.pdf` (visited on 09/13/2016).

Platform for data analysis obtained from a cultural exchange program

Ivana Maznevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje, Macedonia
ivana_maznevska@yahoo.com

Igor Mishkovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje, Macedonia
igor.mishkovski@finki.ukim.mk

Miroslav Mirchev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje, Macedonia
miroslav.mirchev@finki.ukim.mk

***Abstract*—The purpose of this work is to analyze data for students that participate in the Work and Travel program, supported by the US State department, which allows each student to bring the right decision for applying, according to his/her perspective. The main analysis is based on data extracted from a database gained from a Work and Travel agency in the program from several countries on the Balkan, for the period from 2011 until 2015. Furthermore, we have conducted a survey and obtained subjective analysis directly from the participants. Based on this data, we present full statistical analysis that resulted with a few conclusions regarding the conditions for issuing visas and the level of earnings. Using the machine learning tool WEKA, we predict the final result obtained from the analyzed data. As a final component in this paper we present a Web application that is based on Numbeo (the world's largest database of user contributed data about cities and countries worldwide). The application has a Facebook integration in order to get information about the current location and also the possibility to connect with students who are already logged on.**

Keywords— Data analysis; Machine learning; Work and Travel; Visualization

I. INTRODUCTION

Cultural exchange programs have high growth in recent years and they are an interesting topic for research. Thanks to the Mutual Educational and Cultural Exchange Act (Fulbright–Hays Act of 1961) Macedonian students have the opportunity to be part of those programs. This act opens a unique opportunity for young people to experience America from a different perspective, to visit and stay in it. Also, they have a great opportunity to continue their education or to gain work experience and training on short term, and then to implement the new skills when they return to their home country.

The diversity of the program allows applicants to become students of some of the American educational institutions. Moreover, there is a possibility to visit America without being part of an educational program, but to be part of an internship program in US companies, or to participate in the summer work and travel program where the intent is international students to work in America during their student holiday and then to visit different places. Basically, those programs are based on a working basis, allowing applicants to gain and learn new skills, and also to be part of the new cultural diversity on a daily basis without having to attend some educational institutions.

The purpose of the Work and Travel program is to enable international students to visit America and be part of the everyday life while they are employed, and then to travel around the US. The Work and Travel program is one of the most abundant cultural exchange programs in Macedonia and it is the subject of research in this work.

II. METHODOLOGY

A. Statistical Analysis

The research first started with developing a full statistical analysis using data from a work and travel agency starting from 2011. The analysis was carried out in two forms. One in terms of characteristics of students (age, academic year, faculty, etc.), their choice of job and location, method of finding work, working conditions and accommodation, their average earnings and their average period of time designated for travel. The second form in terms of their earnings during their stay in the United States. As an additional way to gain further information that will be covered on the Web platform, we have conducted a survey among former participants in the program, which included students from Macedonia, Bulgaria and Croatia and other surrounding countries. The survey was completed by 203 respondents, and the data were properly analyzed.

Using Excel as a platform for business intelligence, a complete analysis and systematization of the data was conducted, which was separated into two parts, namely an objective analysis obtained by processing data from a database of work and travel agency and a subjective analysis obtained from an anonymous survey [1].

From the results shown in table view (dashboard), we were able to pull detailed conclusions. From an objective analysis, as important arguments we define the following: for the male students, especially in the third year of studies, participation percentage is higher than the female whose biggest participation is mark in fourth year of their studies, as shown in Fig. 1.

Most students are opting to work in Maryland (20,30%), New Jersey (15,28%) and Colorado (13,38%), shown in Fig. 2. In Macedonia, most participants in the program are from Skopje, then from Bitola and Strumica. The percentage of approved visas for Macedonian students was 69.38% versus 30.62% rejected, as shown in Fig. 3, which indicates a problem

with a large number of students not being accepted into the program, unlike other countries on the Balkans.

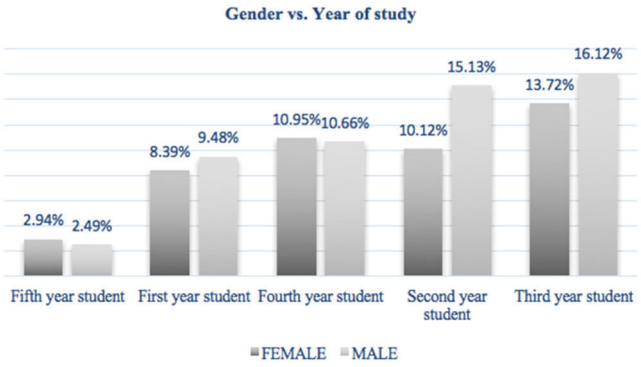


Fig. 1. Objective analysis of participation gender versus year of study

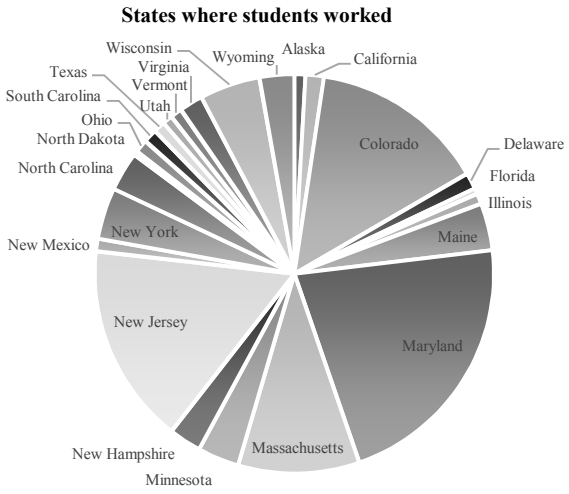


Fig. 2. Objective analysis of where students chose to work

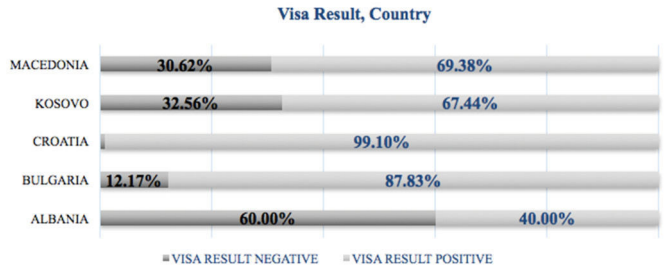


Fig. 3. Objective analysis of visa results compare to different countries

Next, we analyze the level of earnings using data extracted from a database using the service for reimbursement of paid tax during student stay and work period in the United States. Based on the processed data we manage to pull important conclusions, which can be used by the students to decide where they would like to work in the United States. Most students are earning between $5,000 and $10,000, and in nation comparison, Macedonians and Bulgarian students are earning more than Croatian, as shown in Fig. 4.

Best earnings are realized in Alaska, and students who previously took part in the program earn more than students who participate for the first time. In Macedonia, higher earnings are recorded in the seasons 2014 and 2015, especially among students who are fourth year of study. Students of economic studies have the highest percentage of participation in the program, and the highest average earnings have accomplished by students in management.

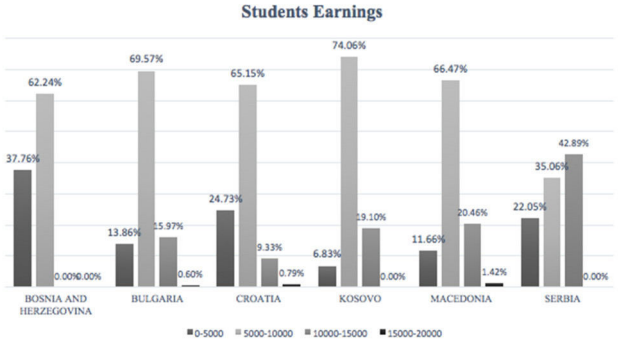


Fig. 4. Objective analyze for participation earnings

Based on the survey regarding the evaluation of satisfaction with the entire program the main conclusion is that respondents expressed positive opinion, which indicates that the program is highly appreciated by all students, results are shown in Fig. 5. However, students reported several issues related to instruction and the assistance by the agency or the sponsor and the interactions with other participants in the program. This was the main motive for building the Web platform that would increase the level of overall experience of the program. The platform will be used by the students and they will be able to verify all terms and ways of life in the location where they will choose to work without the need to bolster other data sources. Also through their Facebook login they will be able to check on other students who are in their vicinity, which would contribute to increase the interactivity and collaboration of students from different cultures.

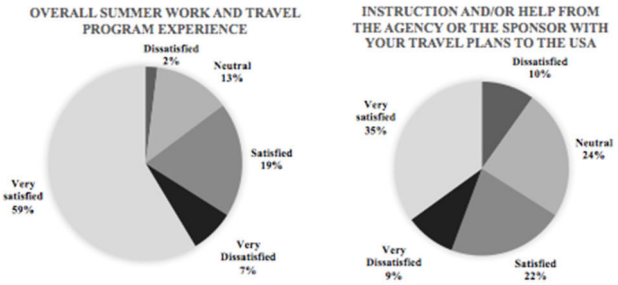


Fig. 5. Satisfaction analysis main results

B. *Machine learning with WEKA*

Waikato Environment for Knowledge Analyze (Weka) is a software employing machine learning algorithms written in Java, developed at the University of Waikato, New Zealand. It can be freely used as it is licensed under the GNU General Public License [2][6].

In the research using the open source machine learning WEKA program we developed a decision tree regarding participation in the program and obtaining a visa, and a predictive analysis for the level of earnings. The decision tree is easy to be used and the algorithms are easy to be understood. With the help of this tree, students based on personal preferences can predict if they would be approved for a visa or not, or whether they would earn above or below average earnings. In terms of pruning the tree, we came to the conclusion that online - pruning is useful for reducing the size of the decision tree, but does not always give accurate information in return. On the other hand, the reduction factor of trust not only that reduces the size of the tree, but also helps to

filter statistically irrelevant nodes, which otherwise would lead to errors in the classification. We can conclude that several values of the factor of trust should be tested while creating a decision tree to find the most appropriate value for the specified set [5].

Data was entered in Excel, and then formatted in CSV document defined by the specific format and structure. Furthermore, the data was converted into so-called ARFF (Attribute Relation File Format) or a special format that can be read by WEKA. ARFF format is an ASCII text file that contains a list of elements (cases) that share common attributes.

The total number of elements to be considered were 3075, divided and processed through 11 attributes such as age, nationality, year of study, visa result, etc.

Weka data can be analyzed using different techniques and data mining algorithms such as classification, clustering, association with rules on data mining, visualization and others. Fig. 6 shows a two-dimensional data visualization for participation in the program and obtaining a visa, all data are classified by class visa. Blue color image indicates approval of the visa, and red is negative outcome or refusal, so for example the second graph where data is stored for program participants by sex, graphically see the extent of obtaining a visa in 1429 females and 1655 male participants. In terms of graphic direction of study, we can see that the highest share of students of Economy and Finance (476), followed by computer science (279). Regarding the type of program in 1751 are premium versus 1324 students registered through the individual program, which visually shows a lesser degree of rejected visa unlike premium.

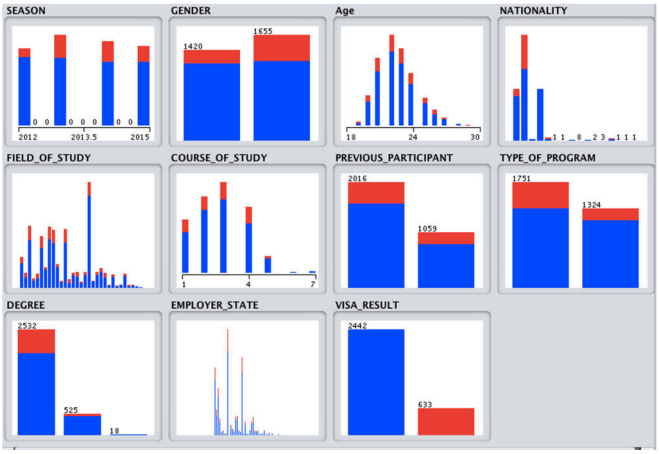


Fig. 6. Two-dimensional data visualization (Visa Score) in Weka

The main challenge when constructing a decision tree is to make the right choice of an attribute that should be further split to get the most accurate decision. The outcome gives us the so-called concept of getting information (information gain), which is actually a matter of entropy before and after making the decision. Entropy is a measure of uncertainty that contains some information. Entropy in decision tree is measured in bits, and one attribute that has the most bits apart until get a class. The formula for calculating the entropy is

$$Entropy = -pP \times log\,2(pP) - pN \times log\,2(pN) \quad (1)$$

Where pP is a proportion of positive examples while pN is the proportion of negative examples.

Eq. 1 is already implemented in the algorithm in WEKA J48, with a selection of the attribute that should be shared in order to reduce the tree to make decisions and to get a more relevant conclusion. The generation of the decision tree is done using the classification module J48, which is an implementation of the C4.5 algorithm that is an extension of the famous ID3 algorithm.

Particularly the Confusion matrix is

	Predicted class	
True class	TP	FN
	FP	TN

where

- TP = true positives: number of examples predicted positive that are actually positive
- FP = false positives: number of examples predicted positive that are actually negative
- TN = true negatives: number of examples predicted negative that are actually negative
- FN = false negatives: number of examples predicted negative that are actually positive

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.985	0.973	0.790	0.985	0.877	0.039	0.579	0.818	Visa_YES
	0.027	0.015	0.333	0.027	0.050	0.039	0.579	0.249	Visa_NO
Weighted Avg.	0.782	0.769	0.693	0.782	0.701	0.039	0.579	0.697	

In recognition of the model and information retrieval with binary classification accuracy (also called positive predictive value) is fraction, part of the retrieved cases that are relevant, while recall (recall) (also known as sensitivity) is a part of the relevant instances repeated. So both precision and recall are based on understanding and measures of importance.

Recall is actually TP (true positive) values divided by real positive values (1). Precision is TP (true positives) divided by the anticipated positive values (2).

$$Recall = \frac{tp}{tp+fn} \quad (2)$$

$$Precision = \frac{tp}{tp+fp} \quad (3)$$

where in our case, we have that *Recall* = 811/(811+12) = 0,98 and *Precision* = 811/(811+216) = 0,790.

F-Measure (Measure F) is a measure that combines precision and recall. It represents a harmonic environment of precision and withdrawal. Harmonic environment is one of the measures of central tendency, useful for quantitative data. It is represented by the following equation

$$F = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

where, in our case, we have F = (2 X 0,790 X 0,985) / (0,790 + 0,985) = 0,877.

According to results obtained with J48, where we used 66% of the data for training, and 34% for testing, we got that 78% of the items are accurately classified, while 21% were not.

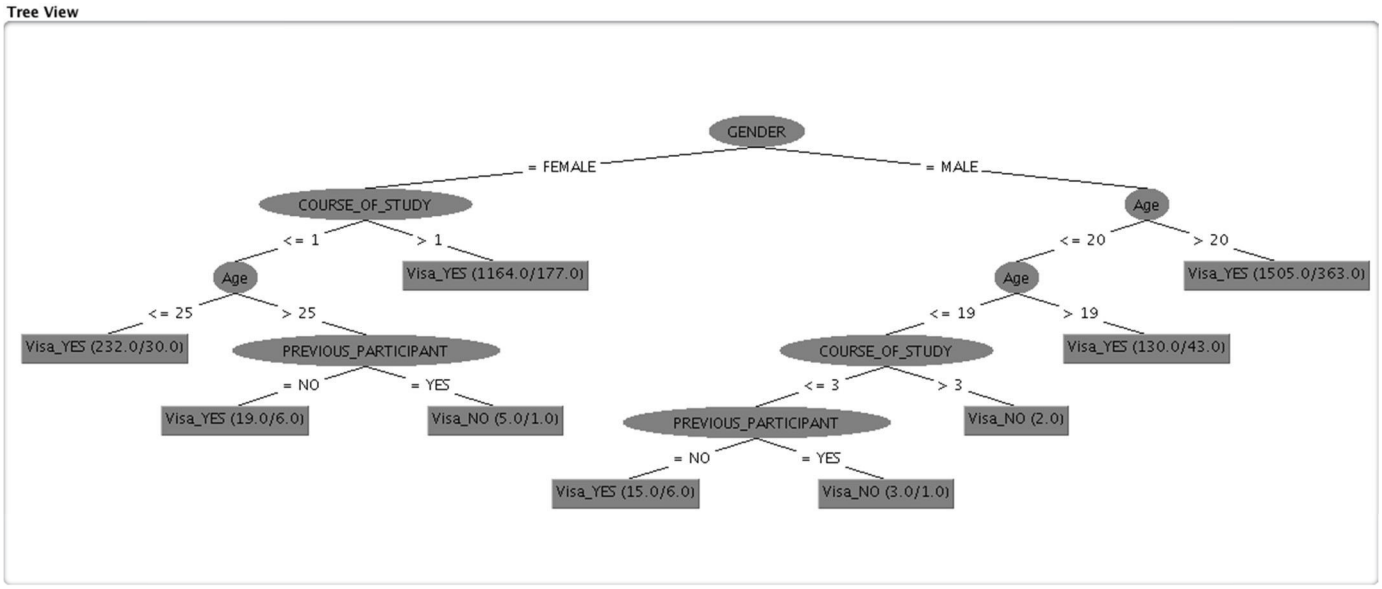


Fig. 7. Visualization of the decision Tree – Visa Result

Using visualization of the decision tree the first number is the total number of cases (weight of cases) that accurately covers that attribute. The second number is the number (weight) of those cases wrongly classified. The decision tree for determining the visa outcome is shown in Fig. 7.

C. *Web application*

As a final component in this work we present a Web application, prosetaj.com. The web application is made in order to allow students to be able to log in using their Facebook account and according to their current location to check basic information about where they are. The information will cover expenses for food and daily activities, expenses for accommodation, transportation, entertainment and other interesting facts that will be of great importance for students in their choice of location where they work. Also, they can connect with the other students who had already log in to the application. The interface of the application is shown in Fig. 8 [4].

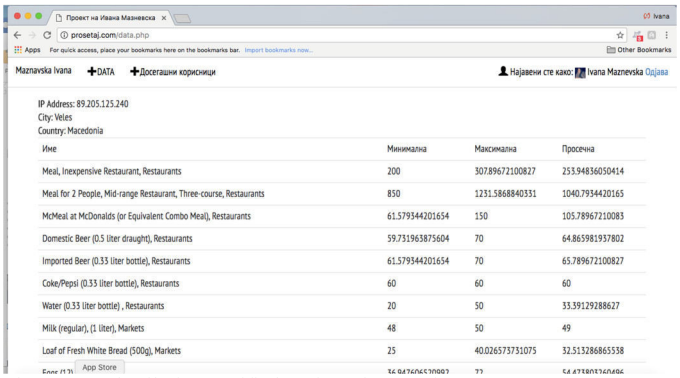


Fig. 8. Data details for specific location of prosetaj.com

The process of setting prosetaj.com consists of several steps:

1. Integration of Facebook API
2. Integration of Numbeo API

Numbeo (the world's largest database of user contributed data about cities and countries worldwide). Numbeo API is used for obtaining relevant information for basic expenses for food and daily activities, costs of accommodation, transport and everything else which can be of student's interest during their stay in the United States [7].

III. Conclusion

The results from the research can be used by multiple stakeholders, both by the students, US Embassy, the sponsors and agencies to obtain further information and details of the program and its participants at any time. With the development of this platform students and all interested parties will have the opportunity at any time to get all the information about Work and Travel program for cultural exchange, easily and without any problems to start and complete the process of the program, and certainly to facilitate communication between participants. With all the available data that the platform offers, students will be able to easily decide on whether to participate in the programs, employers will have the option of selecting students, sponsors to increase number of applicants, and certainly will have the positive impact on fulfilling a joint goal of all stakeholders, which is gaining a higher degree of exchange of cultural values, experiences and increasing friendships internationally.

References

[1] G. Business intelligence (BI) (2014, October). Retrieved from http://searchdatamanagement.techtarget.com/definition/business-intelligence

[2] Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. Retrieved 2011-01-19

[3] Data mining: What is Data mining? Retrieved from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[4] Tutorial: Integrate Database Based Facebook Connect To Your Website. Retrieved from http://artatm.com/2012/08/tutorial-integrate-database-based-facebook-connect-to-your-website/

[5] "Decision Tree Analyze using Weka" Machine Learning – Project II Sam Drazin and Matt Montag University of Miami

[6] G. Holmes; A. Donkin; I.H. Witten (1994). "Weka: A machine learning workbench" (PDF). Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[7] Numbeo. Retrieved from https://www.numbeo.com/cost-of-living/

Strategies for Network Reliability Evaluation and Estimation Based on Pivoting Method

Marija Mihova, Blagoj Mitrevski
Faculty of Computer Science and Engineering Ss. Cyril and Methodius University Skopje, Macedonia
marija.mihova@finki.ukim.mk, mitrevski.blagoj@students.finki.ukim.mk

***Abstract*— Most of the algorithms for calculation of network reliability are based on strategies that use minimal path or minimal cut vectors. Such approaches use two complex algorithms, algorithm for determining all such vectors and algorithm for reliability calculation given minimal cut or path vectors. The only technique that doesn't require the calculation of such vectors is the pivoting method. In this paper we proposed an algorithm which is modification of the pivoting method for computation of two terminal reliability between all pairs of nodes. Moreover, we proposed an algorithm for reliability estimation. This algorithm is based on the fact that the most probable path and the edges from the min cut have the biggest impact on the reliability of the system. Also, the results from the approximate algorithm are shown, analyzed and compared.**

***Index Terms*—reliability evaluation, pivoting method for reliability calculation, all pairs two terminal reliability, reliability estimation**

NOTATION

$e = \{i, j\}$	the link (edge) between the nodes i and j;
$p_e = p_{ij}$	probability that link $e = \{i, j\}$ is in the operative state;
S	set space (set of all possible states);
$\boldsymbol{X}$	state matrix;
$P(\boldsymbol{X})$	probability of observing state $\boldsymbol{X}$;
$R(G)$	reliability of graph G;
T	set of terminal nodes;
$G_{/\{i,j\}}$	graph obtained from G by deleting edge $\{i, j\}$;
$G_{\cdot\{i,j\}}$	graph obtained from G by merging edge $\{i, j\}$;

I. INTRODUCTION

A variety of real-life systems as transportation systems, communication networks, a variety of delivery and distribution systems [Lin], can be modeled as a communication capacity networks. Direct graph (digraphs) are more appropriate way for representation of such networks, where the vertices are communicating entities and the edges are communication paths, i.e. links. However in the real world the links are not perfect and they can fail with some probability. Therefore the invention of techniques for evaluation of the probability for communication of the nodes of such networks, i.e. reliability, is one of the most exploring fields in the theory of network communication analysis. Many techniques for computing *s-t* reliabilities and *k*-terminal reliabilities have been proposed [Mohamed 2, Page and Perry]. These techniques vary widely in computational efficiency. Although some of them might show impressive speed in some specific situations, the worst-case complexity function is still an exponential function of the network size. On the other hand the real network models could be very complex and large, provoking the reliability computation to become intractable. Therefore the need to invent some efficient approximate techniques that have a good precision [Younes, Mohamed1].

In the methods for reliability evaluation of network systems usually is used minimal path or cut sets and there are three general approaches: inclusion-exclusion [M. R. Hassan], sum of disjoint products [Chin-Chia] and pivoting. The most powerful technique between these is pivoting, also known as factoring, which is also the only one that can be applied directly to a network graph representation without first finding minimal path or cut sets [Page and Perry, 1991].

The primary network reliability measure for undirected probabilistic networks differ in the set of terminal nodes T for which the probability to communicate is computed [Network Reliability Optimization]. So we have two-terminal reliability, the probability that a selected node pair (a source node s and a sink node t) are connected, where $T = \{s, t\}$, all-terminal, the probability that every node can communicate with every other node, where $T = V$, and K-terminal, the probability that every node in a given set of nodes $K \subset V$ communicate with every other node in K, when $T = K$.

In this paper we extend the pivoting method on evaluation of all pairs two terminal reliability. Moreover, we present approximation strategy for estimation of the two-terminal reliability. The paper is organized as follows: In the next part we give basic definition, while the third part explains basic idea of the pivoting method for calculating network reliability. The next fourth chapter is devoted to the modified method for all pairs two terminal reliability, where we give the pseudo code for the algorithm and illustration with an example. The algorithm for estimation two terminal reliability is explain in the fifth chapter. The sixth chapter the results are presented and statistically analyzed.

II. BASIC DEFINITION

A network with unreliable components is represented as an undirected weighted graph $G = (E, V, P)$ with node set V and edge set E. The weights are interpreted as probabilities, so it is known as a probabilistic network. The elements in the E are such that each edge $e = \{i, j\}$, can be in either two states: operative or failed, with associated probabilities $p_e = p_{ij}$ and $1 - p_e = 1 - p_{ij}$, respectively. The probability p_{ij} is usually defined as the probability that link $e = \{i, j\}$ is in operative state at a random point in time. The common assumptions are that the nodes are perfectly reliable, while the arc failures are independent and no repair is allowed.

The set of all possible states of the links is called set space and it is usually denoted by S. We can define a function $x(i, j) = x(e)$ with $x(i, j) = 1$ if the link $\{i, j\}$ is in operative state and $x(i, j) = 0$ otherwise. Then each element of S can be represent with a binary matrix $\boldsymbol{X}$. Now the probability of observing particular state $\boldsymbol{X}$ is equal to

$$P(\boldsymbol{X}) = \prod_{\{i,j\} \in E} \left(1 - p_{ij} + x(i,j)\left(2p_{ij} - 1\right)\right) \quad (1)$$

Depending of its state, the whole network can also be in operation or failure state. Its state is represented by the structure function $\Phi(\boldsymbol{X})$, defined as follows: $\Phi(\boldsymbol{X}) = 1$, if all nodes in T are connected in state $\boldsymbol{X}$ and $\Phi(\boldsymbol{X}) = 0$ otherwise. Now the reliability of the network can be computed by

$$R(G) = \sum_{\boldsymbol{X} \in S} P(\boldsymbol{X}) \Phi(\boldsymbol{X}) \quad (2)$$

III. PIVOTING

The factoring theorem is based on the Bayesian formula. In fact, respective to the state of link $\{i, j\}$, the reliability of the network can be calculated by the formula

$$R(G) = p_{ij} R(G|x(i,j) = 1) + \left(1 - p_{ij}\right) R(G|x(i,j) = 0),$$

where $R(G|x(i,j) = 1)$ is the reliability of the network when the link $\{i, j\}$ is in operation, while $R(G|x(i,j) = 0)$ is the reliability of the network when the link $\{i, j\}$ is not in operation. Therefore we define two graphs, $G_{\cdot\{i,j\}}$, a graph corresponding to the network when the link $\{i, j\}$ is in operation and $G_{/\{i,j\}}$, a graph corresponding for the network when the link $\{i, j\}$ is not in operation.

The graph $G_{/\{i,j\}}$ is simply obtained by deleting the edge $\{i, j\}$ from G. The graph $G_{\cdot\{i,j\}}$ is obtained from G by merging nodes i and j into a new node and connecting each edge incident to at least one of them to this new node. Let us consider the situation when both i and j are connected with the same node k. The new node and the node k will be in direct communication if at least one of the edges $\{i, k\}$ and $\{j, k\}$ is in operation. Thus, the probability that there is no direct communication between the new node and the node k is $(1 - p_{ik})(1 - p_{jk})$, i.e. the probability of direct link between the new node and the node k is $1 - (1 - p_{ik})(1 - p_{jk})$. Therefore, for all nodes k incident with both i and j we replace the two links obtained by the merging procedure with one link that has probability $1 - (1 - p_{ik})(1 - p_{jk})$, Figure 1. Now the expression for $R(G)$ is given by

$$R(G) = p_{ij} R\left(G_{\cdot\{ij\}}\right) + \left(1 - p_{ij}\right) R\left(G_{/\{ij\}}\right). \quad (3)$$

We can also compute the probability that the system is not in operation with the expression:

$$\bar{R}(G) = 1 - R(G) = p_{ij} \bar{R}\left(G_{\cdot\{ij\}}\right) + \left(1 - p_{ij}\right) \bar{R}\left(G_{/\{ij\}}\right). \quad (4)$$

The reliability of the network can be computed by repeating this decomposition. Proper pivot selection and using some techniques for network reduction can improve computation time and reduce the computation steps [Satyanarayana and Chang, 1983, Ball et al., 1995].

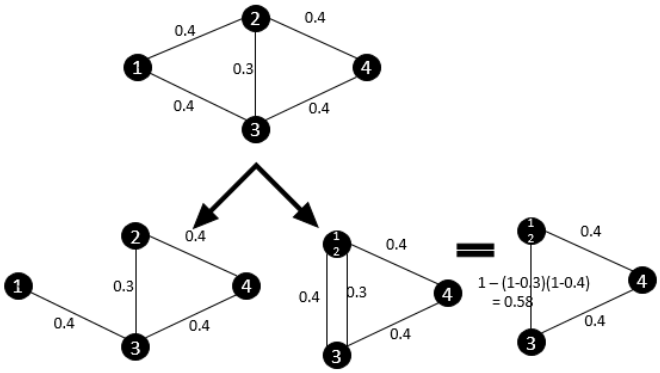


Figure 1. The first step of performing factoring decomposition

Termination step occurs in two situations:

- When the set T is divided between different connected components. This situation occurs after edge has been deleted. The reliability of such graph is equal to 0, Figure 2.
- When all of the nodes from T become one node. This situation occurs after edge has been merged. The reliability of such graph is equal to 1. We would have had such situation if in the graph in Figure 1 the set had been $T = \{1, 2\}$.

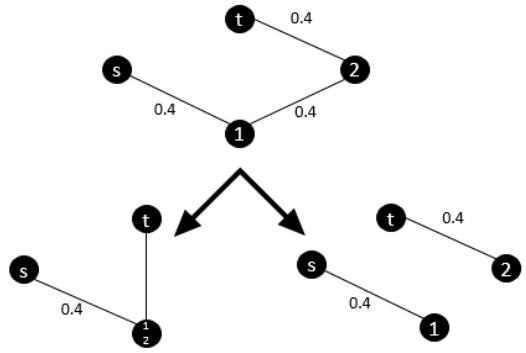


Figure 2. Termination step after reconnecting some nodes from T

Consider a graph obtained after performing number of factoring compositions. That graph is obtained by removing a set of edges either with merging procedure or with deleting procedure. Let G' be the graph obtained from G after merging nodes incident with edges $e_1, \ldots, e_k$ and deleting edges $e_{k+1}, \ldots, e_r$. Then, the reliability of G' is

$$R(G') = R(G|x(e_l) = 1, x(e_h) = 0, l = 1, k, h = k + 1, r).$$

Assume that after a number of steps applied on recursive formula (3), we get an expression in which the graph G' appears. The coefficient that multiplies this $R(G')$ is

$$\tilde{R}(G') = \prod_{m=1}^{k} p_{e_m} \prod_{m=k+1}^{r} (1 - p_{e_m}). \quad (5)$$

These coefficients will be used in the algorithms we explain further, since

$$\tilde{R}\left(G'_{\cdot\{ij\}}\right) = p_{ij}\tilde{R}(G') \quad (6)$$

$$\tilde{R}\left(G'_{/\{ij\}}\right) = (1 - p_{ij})\tilde{R}(G') \quad (7)$$

IV. Evaluation of All Pairs Two Terminal Reliability

In this part we give the generalization of the pivoting method for evaluation two terminal reliability for all pairs in the network. The proposed algorithm have the same time complexity as the pivoting procedure for one pair of terminals. In this part we will use $R_{ij}(G)$ to denote the reliability of the set $T = \{i, j\}$.

The basic algorithm is modified as follows:
Initialization: For all nodes i and j $R_{ij}(G) = 0$ at the beginning.
Reliability calculation: For each graph $G_{\cdot\{A,\,B\}}$ the probability $\tilde{R}\left(G_{\cdot\{A,B\}}\right)$ to that graph is added to the function $R_{ij}(G)$, for all $i \in A$ and $j \in B$.
Unreliability calculation: In addition we can also calculate the probability $\bar{R}(G)$. In fact, whenever a node i is disconnected from the rest of the graph, i.e. its degree is equal to 0, we can add $\tilde{R}(G)$ to $\bar{R}_{ik}(G)$, for all $k \neq i$. Thus, the values $R_{ik}(G)$ and $\bar{R}_{ik}(G)$ computed at each step are upper and lower boundaries for the reliability.

Given graph $G(V, E)$ and the matrix $\boldsymbol{P} = p_{ij}$, the basic recursion is given by the following algorithm:

Input: graph $G(V, E)$ and the matrix $\boldsymbol{P} = p_{ij}$.
Output: The reliability of the all pairs $\{i, j\}$.
Initialize:
$R_{ij} = 0, \bar{R}_{ik} = 0$ for all $\{i, j\}$.
Push graph G and $\tilde{R}(G)$, i.e. $(G, \tilde{R}(G) = 1)$ in the stack Q.
while $Q \neq \emptyset$ (or when time limit is exceed) do:
5 Take the element from the top of Q, $(G, \tilde{R}(G))$.
Choose an edge $\{A, B\}$, $A, B \subseteq V$.
Evaluate $G_{\cdot\{AB\}}$ and $G_{/\{AB\}}$.
Evaluate $\tilde{R}\left(G_{\cdot\{ij\}}\right) = p_{ij}\tilde{R}(G)$
$R_{ij} = R_{ij} + \tilde{R}\left(G_{\cdot\{ij\}}\right)$.
10 Evaluate $\tilde{R}\left(G_{/\{ij\}}\right) = (1 - p_{ij})\tilde{R}(G)$
For all nodes A with degree 0 from $G_{\cdot\{AB\}}$ and $G_{/\{AB\}}$ calculate
$\bar{R}_{ij} = \bar{R}_{ij} + \tilde{R}\left(G_{\cdot\{AB\}}\right), i \in A, j \notin A;$
$\bar{R}_{ij} = \bar{R}_{ij} + \tilde{R}\left(G_{/\{AB\}}\right), i \in A, j \notin A;$
Delete all nodes from $G_{\cdot\{AB\}}$ and $G_{/\{AB\}}$ with degree 0.
16 If $G_{\cdot\{ij\}}$ is not an empty graph push $(G_{\cdot\{ij\}}, \tilde{R}(\, G_{\cdot\{ij\}}\,))$ in Q
If $G_{/\{ij\}}$ is not an empty graph push $(G_{\cdot\{ij\}}, \tilde{R}(\, G_{/\{ij\}}\,))$ in Q
return R_{ij} or $(R_{ij} + \bar{R}_{ij})/2$, for all $\{i, j\}$.

Note that further optimization can be achieved in situations when the graph is a collection of simple cycles or lines. Such graphs are easy to detect, since degrees of all nodes are either 1 or 2. In fact, if two nodes i and j lie on the same simply cycle, then there are two simple paths from i to j, π_1 and π_2.

$$R_{ij} = 1 - (1 - \prod_{e \in \pi_1} p_e)(1 - \prod_{e \in \pi_2} p_e)$$

And when two nodes i and j lie on the same line, then there is only one simple path from i to j, π, and

$$R_{ij} = \prod_{e \in \pi} p_e.$$

V. Approximately Estimation of Two Terminal Reliability

Having in mind that the most influence on the reliability of a network have the paths with greatest probability and the cuts with smallest probability, we propose a method for estimation two terminal reliability based on these two elements.
Using Dijkstra's algorithm we can find the most probable path π, while using Ford Fulkerson algorithm we can find the cut C with the smallest sum of probabilities [MIT]. There must be an edge that lies on both the path and the cut. Choosing that edge in the factoring step, leads to graphs with following properties:

- The graph obtained by merging that edge increases the probability of merging s and t in the next steps
- The graph obtained by deleting that edge increases the probability of disconnecting s and t in the next steps.

We estimate the approximate proportion between R and $\bar{R}$ as follows:

$$R : \bar{R} = \left(\prod_{e \in \pi} p_e\right) : \left(\prod_{e \in C} (1 - p_e)\right).$$

Then the estimated reliability of the graph will be

$$\hat{R}(G) = \frac{\prod_{e \in \pi} p_e}{\prod_{e \in \pi} p_e + \prod_{e \in C} (1 - p_e)}. \quad (8)$$

Note that when there is no path between s and t, then $\prod_{e \in \pi} p_e = 0$, i.e. $\hat{R}(G) = 0$. Also when s and t are merged into a same node, then $\prod_{e \in C} (1 - p_e) = 0$, i.e. $\hat{R}(G) = 1$.

Using (8) we may evaluate the approximate reliability in each factoring step performed on the graph G, with the following expression:

$$R = R - \hat{R}(G)\tilde{R}(G) + \left(\hat{R}(G_{\cdot\{i,j\}})\tilde{R}(G_{\cdot\{i,j\}}) + \hat{R}(G_{/\{i,j\}})\tilde{R}(G_{/\{i,j\}})\right) \quad (9)$$

Another important observation is that the most probable path from s to t in $G_{\cdot\{i,j\}}$ is the same one as that in G (with removed $\{i, j\}$), and its probability is $(\prod_{e\in\pi} p_e)/p_{ij}$. Moreover, the most probable s - t cut in $G_{/\{i,j\}}$ is the same one as that in G (with removed $\{i, j\}$), having probability $(\prod_{e\in C}(1-p_e))/(1-p_{ij})$.

In order to improve speed in the case of two terminal reliability, we can use the following strategies:

- Removing all nodes different than s and t with degree equal to 0 or 1.
- Factoring immediately all graphs where s and t are connected with an edge, since the graph obtained by merging that edge has reliability 1, and there are no further factorizations of it.
- Factoring immediately all graphs where s or t have degree 1, since the graph obtained by deleting that edge has reliability 0 and there are no further factorizations of it.

Observe that the order of the graphs taken into factorization, play a key role on the convergence speed. In order to give a preference on graphs with greatest impact on the reliability, we should lead the following strategies:

- Graphs with the greater $\tilde{R}$ value are factorized earlier.
- Graphs with greater value of the most probable path, $\prod_{e\in\pi} p_e$, are factorized earlier.
- Graphs with smaller value of the most probable cut $\prod_{e\in C}(1-p_e)$ are factorized earlier.

Therefore, for each graph we care a number $I(G)$ that indicates its importance. Smaller $I(G)$ means greater importance. It is calculated by $I(G) = (n(G) + v(G))/\tilde{R}(G)$, where $n(G)$ is the length of the most probable path, while $v(G)$ is the number of edges on the most probable cut.

The algorithm is illustrated by the following example:

Example 1.

G_0: The most probable path and cut are shown in Figure 3. The reliability $\hat{R} = R_0 = \frac{0.9\cdot0.7\cdot0.9}{0.9\cdot0.7\cdot0.9+0.3\cdot0.3} = 0.86301$. The graphs G1 and G2 given in Figure 3 a) b) are obtained by the factorization on this graph

G1: $\tilde{R}(G1) = 0.7$. The most probable path and cut are shown in Fig.3 $\hat{R}(G1) = \frac{0.9\cdot0.9}{0.9\cdot0.9+0.1\cdot0.1} = 0.98780$, $I(G1) = 5/0.7 = 7.14$.

G2: $\tilde{R}(G2) = 0.3$. The most probable path and cut are shown in Fig.3. $\hat{R}(G2) = \frac{0.9^2\cdot0.7}{0.9^2\cdot0.7+0.3} = 0.65398$, $I(G2) = 4/0.3 = 13.33$.

Now the estimated value for R is

$$R_1 = 0.7 \cdot 0.98780 + 0.3 \cdot 0.65398 = 0.887654$$

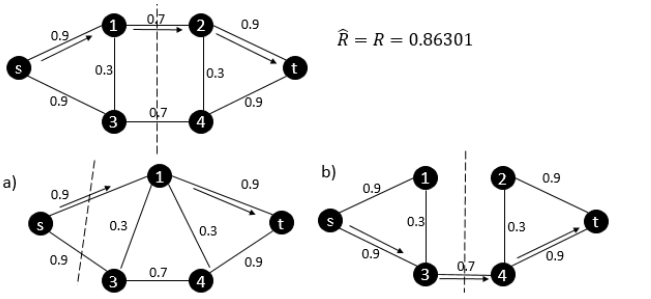


Figure 3. First step in the proposed algorithm. Graph a) is obtained by merging {1, 2}, while graph b) is obtained by deleting {1, 2}.

The graph $G1$ has smaller value for I, so he is the next one for factorization. Therefore $R_2 = 0.887654 - 0.69146 = 0.196194$.

$G3$: The graph $G3$ obtained by merging {s, 1} is given in Fig. 4 a). $\tilde{R}(G3) = 0.63$. Since {s, t} is in the graph we calculate R again, so now $R_2 = 0.196194+0.63\cdot0.9 = 0.763194$. Next, we delete the edge {s, t} and get the graph in Fig.4 b). For this graph $\tilde{R}(G3) = 0.63 \cdot 0.1 = 0.063$. Now, deg(t) = 1, so we merge {t, 4} and obtain the graph shown in Fig.4 c), which has $\tilde{R}(G3) = 0.063 \cdot 0.3 = 0.0189$. Again {s, t} is in the graph, so we add $0.3 \cdot 0.0189 = 0.00567$ to R, obtaining $R_2 = 0.768864$. By deleting this edge we obtained the graph from Fig. 4 d), that has $\tilde{R}(G3) = 0.0189 \cdot 0.7 = 0.01323$. The probability for merging s and t in the last graph is $0.93 \cdot 0.7$, which is obtained after one factorization where the edge {3, t} is deleted and one factorization where the edge {s, t} is merged. Now we have $R_2 = 0.768864+0.93 \cdot 0.7 \cdot 0.01323 = 0.777477$. Clearly, we are finished with this graph.

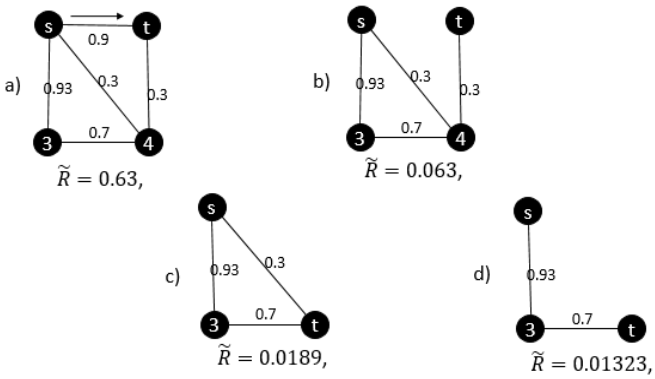


Figure 4. Factorized graph obtained from G1 merging {s, 1}.

$G4$: The graph $G4$ is obtained by deleting {s, 1} from G1, Fig. 5 a). $\tilde{R}(G4) = 0.07$. Here we only have one transformation, merging {s, 3}, Fig. 5 b). Again we calculate $\tilde{R}(G4) = 0.063$. The new $G4$ has $\hat{R}(G4) = \frac{0.9\cdot0.7}{0.9\cdot0.7+0.3\cdot0.7} = 0.75$ and $I(G4) = 4/0.063 = 63.49$. Adding corresponding value $0.75 \cdot 0.063$ to R we get

$$R2 = 0.777477 + 0.04725 = 0.824727.$$

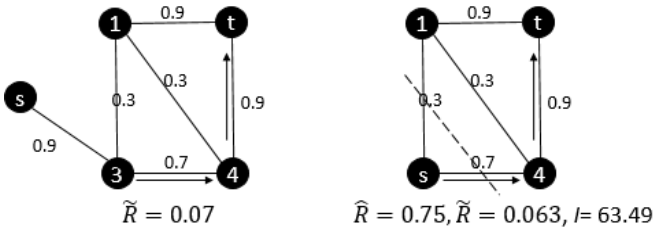


Figure 4. Factorized graph obtained from G1 by deleting {s, 1}

Next graph for factoring is G2, since 13.33 < 63.49. After the factoring we obtain $R3 = 0.835971$. Continuing, the approximated reliabilities are: $R4 = 0.802499$ and at the end $R5 = 0.802163$.

Given graph $G(V, E)$ and the matrix $\boldsymbol{P} = p_{ij}$, the algorithm for estimation *s-t* reliability is given by following pseudo code:

Input: graph $G(V, E)$, nodes s and t, and the matrix $\boldsymbol{P} = p_{ij}$.
Output: The *s-t* reliability
Evaluate the most probable *s* - *t* path and *s* - *t* cut in G
Compute $\hat{R}(G)$ by (8)
Initialize:
  $R = \hat{R}(G)$ for all $\{i, j\}$.
5 Push $(G, \tilde{R}(G) = 1, \hat{R}(G)$, $I(G)=\infty)$ in Q
while $Q \neq \emptyset$ (or when the number of iteration is exceed) **do**
  Take $(G, \tilde{R}(G), \hat{R}(G), I(G))$ from Q with the smallest $I(G)$,
  Evaluate most probable *s* - *t* path π and *s* - *t* cut C in G
  Choose the edge $\{i, j\}$ in $\pi \cap$ C
10Evaluate $R = R - \hat{R}(G)\tilde{R}(G)$
  Evaluate $G_{.\{i,j\}}$ and the new *s* - *t* cut of $G_{.\{i,j\}}$
  Evaluate $G_{/\{i,j\}}$ and the new *s* - *t* path of $G_{/\{i,j\}}$
  for $G'=G_{.\{i,j\}}$ and $G' = G_{/\{i,j\}}$ **do**
    if $(s, t) \notin G'$ and deg(t) > 1 and deg(s) > 1 **then**
      Compute $\hat{R}(G')$ using (8) and $I(G)$
      $R = R + \hat{R}(G')\tilde{R}(G')$.
      Push $(G', \tilde{R}(G'), \hat{R}(G'), I(G'))$ in Q
    else
      while $(s, t) \in G'$ or deg(t)=1 or deg(s)=1
20 **if** $(s, t) \in G'$,**then**
        $R = R + p_{st}\tilde{R}(G')$
        $G' = G_{/\{s,t\}}$
        $\tilde{R}(G') = (1 - p_{st})\tilde{R}(G')$.
        Evaluate the new *s* - *t* path of G'
      if deg(t)=1, $(t, i) \in G'$ (deg(s)=1, $(s, i) \in G'$),**then**
        $G' = G_{.\{t,i\}}$ ($G' = G_{.\{t,i\}}$)
        $\tilde{R}(G') = p_{si}\tilde{R}(G')$ ($\tilde{R}(G') = p_{ti}\tilde{R}(G')$)
        Evaluate the new *s* - *t* cut of G'
    Compute $\hat{R}(G')$ using (8) and $I(G)$
30 $R = R + \hat{R}(G')\tilde{R}(G')$.
    Push $(G', \tilde{R}(G'), \hat{R}(G'), I(G'))$ in Q
return R

VI. STATISTICAL RESULTS

In this chapter we will present the results. We have examined 20 different random generated graphs with different number of vertices and edges and with random generated probabilities for every edge. The Table 1 consist of details for every graph: |V| - number of vertices, |E| - number of edges, the number of iterations for the basic pivoting method to calculate the reliability between two random vertices, the number of iterations for the approximate algorithm (explained in chapter V) to calculate the **exact** reliability between the same two random vertices and the iterations needed for the approximate algorithm to compute the approximate probability with 0.05 precision.

TABLE 1
EXPERIMENTAL RESULTS

|V|	|E|	Iterations for basic pivoting algorithm	Iterations for approximate algorithm to finish	Iterations for the approximate algorithm to compute the probability with 0.05 precision
6	8	32	19	12
6	5	2	2	2
6	5	3	4	3
6	10	51	35	16
6	11	107	3	3
10	12	127	17	11
10	23	6381	322	242
10	26	12483	7	6
10	29	20679	653	427
10	14	736	49	37
15	33	1172377	32044	18765
15	32	4185307	7	6
15	30	404521	878	653
15	39	6327251	78951	51461
15	43	32491459	6852	4891
20	37	43168803	11054	8923
20	41	188224709	14645	11012
20	38	32769063	14036	9151
20	36	49534147	17578	2560
20	43	110644052	4296	3357

From the results we can conclude that the pivoting algorithm takes more iterations than the approximate algorithm to compute the exact result. But the approximate algorithm requires computation of the most probable path which can be computed using Dijkstra's algorithm (O(E + VLogV)) and the min – cut which can be computed with Ford-Fulkerson algorithm ($O(VE^2)$) both increasing the complexity of each iteration against the pivoting algorithm where each iteration uses only simple computations. Also the approximate algorithm needs only few steps to calculate the probability when the two vertices are connected with a small number of paths or are in two distinct connected components.

From observing the last column, the number of iterations for the approximate algorithm to approximate the probability with 0.05 precision, we can conclude that in average with 71% of the number of iterations to calculate the exact probability we get a solid approximation for the probability.

VII. Conclusion and Future work

In this paper we analyzed the pivot method, we proposed an algorithm for all pairs two terminal reliability and we proposed an estimation algorithm. The results showed that the estimation algorithm uses relatively small number of iterations and can further be optimized.

As a future work we want to make a cross estimation with the same idea with consideration of several parallel paths and parallel min cuts. Also we want to optimize and reduce the iterations of the pivot procedure and extend it for multistate systems.

References

[1] Konak, A. and Smith, A.E. (2005). "Network Reliability Optimization," in Handbook of Telecommunications, ed. P. M. Pardalos and M. G. C. Resende, Springer, 735-760.

[2] M.O. Ball, C.J. Colbourn, and J.S. Provan. Network reliability, volume 7 of Handbooks in Operations Research and Management Science, pages 673–672. Elsevier Science B.V., Amsterdam, 1995

[3] Satyanarayana and M.K. Chang. Network reliability and the factoring theorem. Networks, 13(1):107–20, 1983.

[4] Chin-Chia Janea, John Yuan A sum of disjoint products algorithm for reliability evaluation of flow networks, European Journal of Operational Research, Volume 131, Issue 3, 16 June 2001, Pages 664–675

[5] Mohamed–Larbi Rebaiaia, Daod Al-Kadi, Network Reliability Evaluation and Optimizations: Methods, Algorithms and software tools, CIRRELT, 2013

[6] Mohamed–Larbi Rebaiaia, Daod Al-Kadi,Anew algorithm for generating minimal cut sets in R-Networks, IFAC-PapersOnLine.net, Elsevier,1474-6670

[7] L.B. Page and J.E. Perry. A note on computing environments and network reliability. Microelectronics and Reliability, 31(1):185–6, 1991

[8] Picard, J.C., Maurice, Q., "On the structure of all minimum cuts in a network and application,"*Mathematical Programming Study*, vol.13, pp. 8-16, 1980

[9] M. R. Hassan, System Reliability Evaluation of a Stochastic - Flow Network using Spanning Trees, Indian Journal of Science and Technology, Vol 9(10), 2016

[10] Cormen, H.T., Leiserson, E.C., Rivest, R.L., Stein, C,"*Introduction to Algorithms*,"Third edition. The MIT Press , 2009

[11] Yi-Kuei Lin, Cheng-Ta Yeh, Cheng-Fu Huang, Reliability evaluation of a stochastic-flow distribution network with delivery spoilage, Computers & Industrial Engineering 66(2):352-359 · October 2013

[12] A. Younes, M. R. Hassan, A Genetic Algorithm for Reliability Evaluation of a StochasticFlow Network With Node Failure, International Journal of Computer Science and Security (IJCSS), Volume (4) : Issue (6), 2010

0-1 Knapsack problem parallelization using OpenMP and Nvidia CUDA

Nikola Trajanovski, Marjan Gusev, Vladimir Zdraveski
Faculty of Computer Science and Engineering, Ss. Cyril and Methodious University in Skopje, Macedonia

***Abstract*—The 0-1 knapsack problem is one of the most famous problems in computer science. The problem is defined for a set of items, each with predefined weight and value. A knapsack can carry a limited weight and the idea is to select items that will have a higher value. The problem is to determine the number of each particular item to include in a collection so that the total weight is less than or equal to the limit and the total value is as large as possible. We experiment with OpenMP parallel CPU solution and a CUDA GPU solution based on a simple brute force knapsack algorithm. This requires no additional input data transformation. The provided solutions demonstrate the high scalability and speedup in both the CPU and GPU implementations.**

I. Introduction

As one of the frequently used combinatorial optimization in computer science, the knapsack problem refers to finding a selection with a maximum value within constraints of limited weight capacity, The problem is an integer optimization problem well known in resource allocation and used in many related fields.

In this article, we analyze the 0/1 knapsack problem defined as follows. Let there be a knapsack that can hold a maximum weight W. Let there be n different items which have a weight w_i and value v_i, for $i = 1, 2, \dots n$. The 0/1 knapsack problem restricts to choose the number x_i of copies of each kind of item to 0 or 1. The goal is to choose the optimal subset of items to put in the knapsack to maximize the total value while not going over the maximum capacity. The mathematical presentation is to maximize the value presented by (1) due to constraints defined in (2).

$$\sum_{i=1}^{n} v_i x_i \tag{1}$$

$$\sum_{i=1}^{n} w_i x_i \leq W \quad and \quad x_i \in \{0, 1\} \tag{2}$$

The complexity of the algorithm is polynomial classifying it as NP-complete. There is no known algorithm that reduces the polynomial time and in the same time to be correct. There are several variants of a pseudo-polynomial time algorithm using dynamic programming, and this is why we present parallel solutions to exploit faster execution times.

In this paper we are explore parallelizing a purely brute force implementatation of the knapsack problem. We will design two different parallel algorithms. A CPU parallel algorithm using OpenMP, as well as a GPU parallel algorithm using Nvidia CUDA.

We are going to measure execution times for both algorithm as well as for other sequential algorithms and based to the results we are going to make a conclusion which algorithm is optimal for which type of input data.

The outline of the paper is as follows. Related work is presented in Section II. Section III specifies and compares the sequential and parallel CPU algorithms. Our nVidia CUDA algorithm is described in Section IV. Section V presents the results of the provided experiments using OpenMp CPU solution and the GPU algorithm. Finally, conclusions and future potentials for CPU and GPU parallelization are presented in in Section VI.

II. Related work

There are many different approaches to solve the knapsack problem.

The most well known is the sequential approach with dynamic programming. A simpler dynamic programming based parallel algorithm is introduced by Andonov et al. [1].

Over the years there has been a few efforts to parallelize this problem. Goldman and Trystram present an idea to parallelize the dynamic programming calculations by representing them in a precedence graph [2].

An interesting solution can also be found by exploiting the parallelism potential that exists during the backtracking steps of the branch-and-bound algorithm [3].

Another proposed solution by Li et al. [4] is based on a parallel algorithm where the optimal merging is adopted.

GPU computing is a relatively new concept, and there are no too many efforts to parallelize the knapsack problem using CUDA. Pospichal et al. [5] set focus on writing a generic parallel algorithm using nVidia CUDA.

The algorithm proposed in our solution is based on Lou and Chang's parallel algorithm [6] with a support of both the CPU and GPU.

III. CPU algorithms

In this section, we present the brute force algorithm, its parallel version and compare to the dynamic programming approaches.

A. Brute force algorithm

The implementation of the brute force algorithm is very simple. The idea is to loop through the every possible combination of items and choose which one is the optimal. Given$n$ total items there are a total of 2^n possible combinations of those

items for the 0/1 knapsack problem. An inner loop checks which items need to be added to form the value and weight for each combination. There are many ways to generate all combinations. The most intuitive is to do it recursively, but to parallelize the algorithm we present the iterative approach.

Sequential brute force algorithm

```
// number of items
int n = 10;
// capacity of the knapsack
int W = 100;
// array of all values
int* value = new int[n];
// array of all weights
int* weight = new int[n];
// =pow(2,n)
unsigned long long combinations = 1 << n;
// value in optimal combination
int maxValue = 0;
// temporary value
int tempValue;
// temporary weight
int tempWeight;
for (unsigned long long i = 1;
    i < combinations; ++i) {
  tempValue = 0;
  tempWeight = 0;
  for (int j = 0; j<n; ++j) {
      // BITWISE AND
    if (i & (1 << j)) {
      tempValue += value[j];
      tempWeight += weight[j];
    }
  }
  if (tempWeight <= W && tempValue > maxValue)
  {
    maxValue = tempValue;
  }
}
```

The main loop iterates from 1 to 2^n. The inner loop iterates from 1 to $n-1$. The inner loop checks digit by digit in the binary representation of the current number i and if it is equal to 1 adds the item whose index corresponds with the position of the digit that's being checked. For example, if $n = 6$ and the first loop is at iteration 37(100101 binary), the value of the combination will be: value[0] + value[3] + value[5], and the weight will be weight[0] + weight [3] + weight [5]. The time complexity of the algorithm is O(N * 2^N).

B. CPU parallel brute force Algorithm

Parallelizing the brute force algorithm is pretty straightforward. Because the loop iterations are independent from one another all loop iterations can be ran in parallel. All threads get part of the loop and calculate the optimal combination in that part and then compare it with all the other threads to find the optimal solution from all combinations.

Parallel CPU algorithm using openMP:

```
// number of items
int n = 10;
// capacity of the knapsack
int W = 100;
// array of all values
int* value = new int[n];
// array of weights
int* weight = new int[n];
// =pow(2, n))
unsigned long long combinations = 1 << n;
unsigned long long chunk = combinations / 8;
// value in optimal combination
int maxValue = 0;
#pragma omp parallel
{
  // optimal value for each thread
  int tempValue = 0;
  #pragma omp for schedule(static, chunk) nowait
  for (unsigned long long i=1; i<combinations; ++i) {
    int loopValue = 0;
    int loopWeight = 0;
    for (int j = 0; j<n; j++) {
      if ((i & (1 << j))) {
        loopValue += value[j];
        loopWeight += weight[j];
      }
    }
    if (loopWeight <= W &&
        loopValue > maxValue) {
      tempValue = loopValue;
    }
  }
  #pragma omp critical
  {
    if (tempValue > maxValue)
      maxValue = tempValue;
  }
}
```

IV. NVIDIA CUDA ALGORITHM

The general idea with the CUDA algorithm is very similar to the CPU parallel algorithm. The approach is still brute force, so we have to check all possible combinations and find the optimal one. That means the problem can be specified as finding the max element in an array, where each element is a sum of different values. A parallel reduction algorithm is used to find the max element. There are a 256 threads per block, so there the total number of blocks are $N/256$ blocks. The first iteration produces $N/256$ partial results, then the same kernel is invoked with the partial results. That persists until the array is reduced to one element, which is the largest and thus, the optimal combination. With this solution reduction is always done using shared memory which is much faster than global memory.

The kernel code used is a based on a modified version of the 'reduction' example provided by Nvidia [7], with the main difference being that out algorithm first processes the input into a single list of combinations before copying the list over to the GPU kernel for computation.

Nvidia CUDA kernel code:

```
__global__ void arrayReduce(int *g_idata,
    int *g_odata, unsigned int n)
{

    __shared__ int *sdata;

    int block_size = 256;
    unsigned int tid = threadIdx.x;
    unsigned int i = blockIdx.x * block_size
      * 2 + threadIdx.x;
    unsigned int gridSize =
      block_size * 2 * gridDim.x;
```

```
    unsigned int mySum = 0;

    while (i < n) {
        mySum = g_idata[i];
        mySum = max(mySum, g_idata[i+blockSize]);
        i += gridSize;
    }
    sdata[tid] = mySum;
    __syncthreads();
    if (tid < 128) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 128]);
    }
     __syncthreads();
    if (tid <  64) {
       sdata[tid]=mySum
          =max(mySum, sdata[tid + 64]);
    }
    __syncthreads();
    if (tid < 32) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 32]);
    }
    __syncthreads();
    if (tid < 16) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 16]);
    }
    __syncthreads();
    if (tid <  8) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 8]);
    }
    __syncthreads();
    if (tid <  4) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 4]);
    }
    __syncthreads();
    if (tid <  2) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 2]);
    }
    __syncthreads();
    if (tid <  1) {
        sdata[tid]=mySum
          =max(mySum, sdata[tid + 1]);
    }
    __syncthreads();

    // write result for this block to global mem
    if (tid == 0) g_odata[blockIdx.x] = mySum;
}
```

V. Evaluation of the experimental results

In this section, we describe the experiments and evaluate the obtained results. Further on, we evaluate and discuss the comparison between different solutions.

A. Experimental environment

The CPU used for testing is Intel i7-2670QM with 4 cores running at 2.2 GHz and turbo boost turned off for more consistent results.

The GPU used is Nvidia GeForce GT 635M with 144 CUDA cores, compute compatibility 2.1 and max threads per block: 256.

Each experiment was realized with a predefined set of n items with corresponding weights w_i and values v_i. We have realized experiments for different n values from 3 to 30.

	Sequential	2 Threads	4 Threads	GPU
n = 3	0.001	0.067	0.049	0.069688
n = 4	0.003	0.064	0.054	0.081019
n = 5	0.004	0.077	0.050	0.092828
n = 6	0.006	0.057	0.051	0.092315
n = 7	0.012	0.064	0.053	0.105812
n = 8	0.024	0.053	0.059	0.090433
n = 9	0.050	0.065	0.065	0.112036
n = 10	0.105	0.109	0.087	1.615185
n = 11	0.288	0.148	0.125	1.990603
n = 12	0.445	0.299	0.186	2.196509
n = 13	0.924	0.546	0.324	2.036372
n = 14	2.062	1.097	0.612	2.048851
n = 15	4.394	2.258	1.210	2.102739
n = 16	8.361	4.767	2.511	2.051978
n = 17	17.474	9.758	7.252	2.197807
n = 18	36.716	20.389	10.803	2.245127
n = 19	75.344	42.046	22.462	3.206844
n = 20	156.038	85.749	46.853	6.185677
n = 21	330.805	177.058	108.322	12.186106
n = 22	689.154	360.823	199.917	24.051862
n = 23	1472.200	791.977	447.171	48.486486
n = 24	2875.340	1541.770	880.449	92.180362
n = 25	5866.900	3316.490	1900.510	178.432452
n = 26	12221.000	6566.790	3712.010	250.186163
n = 27	25027.500	13392.000	7904.090	509.156108
n = 28	51127.600	27095.300	15902.600	1026.941358
n = 29	104953.000	56231.600	32270.500	2102.186102
n = 30	220269.000	114325.000	65681.200	3987.186103

Fig. 1. Execution times in ms for different algorithms and array sizes

Each test was performed 5 times and the average value of the measured time was used as the final result.

Fig. 1 presents the results obtained within the experiments.

We have calculated the speedup S_p as a ratio of the average measured execution time of the analyzed solution with the sequential solution, according to (3), where T_{seq} corresponds the sequential solution and the T_{par} the analyzed parallel solution.

$$S_p = \frac{T_{seq}}{T_{par}} \tag{3}$$

B. Comparing sequential and parallel CPU algorithms

Fig. 2 presents the obtained speedup when the parallel solutions were analyzed and compared to the sequential solution.

Because of the fact that each thread gets the same number of combinations to process and the small number of critical sections (1 per thread) we are able to get close to the linear speedup depending on the number of used processor cores.

However, the experiments have shown that this is only true if the number of items is larger than 12. When the number of items is relatively small the parallel algorithm performs worse than the sequential one. This is because the time it takes to go through all the combinations is relatively small compared to the time it takes to initialize, fork and then join the threads.

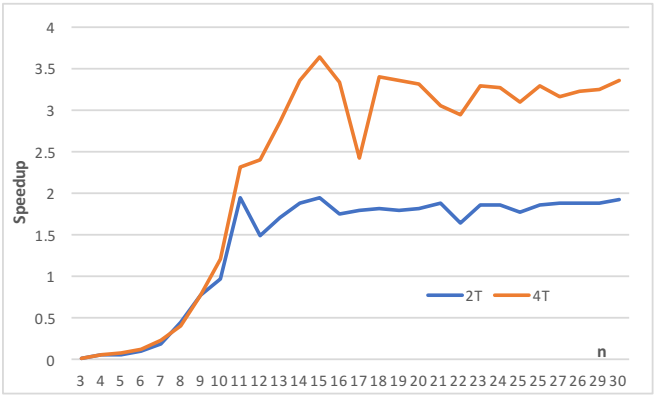


Fig. 2. Speedup of OpenMP solutions compared to the sequential

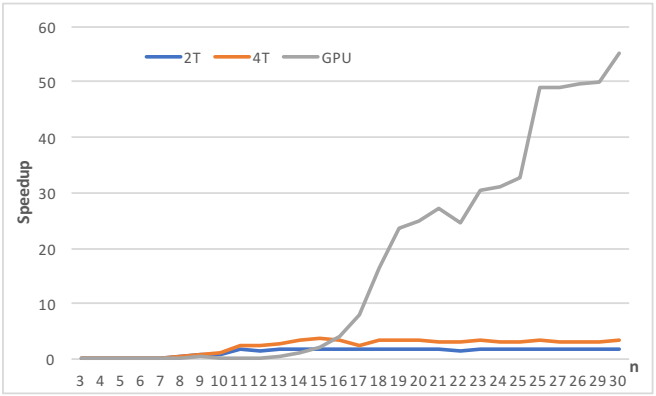


Fig. 3. Speedup GPU vs OpenMP parallel solutions

As the number of items increases that time becomes more and more negligible and we see the benefits of the parallelization. Interestingly, when the algorithm tasks are divided to two cores, the obtained speedup is approaching the value of 2 as n gets higher values.

In the case of 4 processor cores the speedup reaches only the value of 3.5, due to extended communication and data exchange.

C. Comparing CPU and GPU parallel algorithm results

Fig. 3 presents the obtained speedup when the parallel solutions were analyzed and compared to the sequential solution.

Due to the high initialization cost of the GPU algorithm and the data transfer between the CPU and GPU, the GPU algorithm is significantly slower on smaller values of n. Larger n values show the high potential of the GPU, since the data transfer and initialization time is less significant when compared to the computation time.

It is obvious that the GPU algorithm is superior when compared to the OpenMP parallel solutions. The expected speedup when n gets higher values is proportional to the number of processors in a streaming multiprocessor.

Another thing to take into account is the hardware used. The CPU used is a relatively high end mobile CPU and even the highest end desktop CPUs are 60-70% faster. On the other hand the GPU used for testing is a relatively low end mobile GPU with limited memory speed, CUDA cores and threads per block. The highest end GPUs can achieve performance over 50 times higher the GPU used. With that in mind, if we were to adjust the results based on the hardware used the results on larger arrays would be even more in favor for the GPU.

D. Comparing to dynamic programming algorithm

The most common way for solving 0-1 knapsack problem is with dynamic programming. In most cases it is the fastest algorithm for solving the knapsack problem. In essence it divides the problem in smaller sub-problems and uses the results from those sub-problems to build up to the solution. Unlike the brute force algorithm its complexity is influenced not only by the number of elements, but also by the capacity of the knapsack. Its complexity is O(N*W). If we compare that to the brute force complexity we can conclude that in order for the brute force algorithm to be faster the following statement has to be true $W > 2^n/numThreads$ The statement above is also assuming that that n is greater than 12 and trending towards infinity in order for the benefits of the parallelization to take effect.

VI. Conclusion and Future Improvements

The knapsack problem is one of the most famous and well known problems in computer science. This means that there have been a lot of different approaches to optimizing it. Parallelization, especially GPU parallelization is a relatively new concept that has gained significant momentum in recent years. Due to this fact, there have not been many articles regarding knapsack parallelization on GPUs.

Potential future speed improvement is drastic when comparing CPU and GPU parallelization. CPU clock speeds and overall performance, due to thermal and power limitations is growing slower then ever. On the other hand GPU parallelization is still in a very early stage, and is expected to grow at a higher pace. Based on that, we can expect an even bigger speed improvement on the GPU compared to the CPU on future hardware.

References

[1] P. Q. R. Andonov, F. Raimbault, "Dynamic programming parallel implementations for the knapsack problem," Tech. Rep. RR-2037, INRIA, 1993.

[2] D. T. A. Goldman, "An efficient parallel algorithm for solving the knapsack problem on hypercubes," in *Journal of Parallel and Distributed Computing*, pp. 1213–1222, March 2004.

[3] W. L. H. C. Smith, "A parallel algorithm for the 0/1 knapsack problem," in *International Journal of Parallel Programming*, pp. 349–362, October 1992.

[4] Q.-H. L. Ken-Li Li, Ren-Fa Li, "Optimal parallel algorithms for the Knapsack problem without memory conflicts," in *Journal of Computer Science and Technology*, pp. 760–768, December 2004.

[5] P. Pospichal, J. Schwarz, and J. Jaros, "Parallel genetic algorithm solving 0/1 knapsack problem running on the GPU," in *16th International Conference on Soft Computing MENDEL*, vol. 2010, pp. 64–70, 2010.

[6] C.-C. C. Der-Chyuan Lou, "A parallel two-list algorithm for the knapsack problem," in *Parallel Computing*, pp. 1985–1996, November 1997.

[7] H. Mark, "Optimizing parallel reduction in CUDA," *NVIDIA CUDA SDK*, vol. 2, 2008.

DAPS – A WEB BASED SYSTEM FOR SENSOR DATA PREDICTION

Ljubomir Ignov, Jordan Arsov
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering (FCSE)
1000 Skopje, Macedonia
{ignov.ljubomir, arsov.jordan}@students.finki.ukim.mk

***Abstract*—Since raw sensor measurements are voluminous and require high bandwidths to be sent to the data centers, data prediction is one useful approach for data reduction. Various algorithms based on different models for times series prediction can be used for this purpose. In this paper, we will present the development of a DAta Prediction System (DAPS), a web-based system that predicts future sensor measurements. Our online tool performs data prediction on one-dimensional and multidimensional sensor readings, using 3 different algorithms: Least Mean Square (LMS), Least Mean Square with variable step size (LMS-VSS), and Moving Average (MA) of different orders. Additionally, the visualization engine from DAPS visualizes the results obtained from the data prediction, by means of MSE and percentage of data reduction. This tool can also compare the performances provided by the three algorithms, since, depending on the nature of the sensor data, algorithms perform differently for different sensor measurements. We believe that our web-based system for sensor data prediction can be utilized by future developers of wireless sensor networks (WSN) and Internet of things (IoT) developers, who can choose the best technique for reducing sensor measurements.**

Keywords—prediction; sensor system; web-based system sensor data

I. INTRODUCTION

Today sensors are widely used, not only by scientists and researchers, but by ordinary people as well. Almost every person that owns some kind of technology has built in sensors in it and they are not even aware of it. All these sensors generate data, which means all this data has to go somewhere and to someone to be analyzed, so that conclusions can be drawn from it. We refer to this data as sensor data and in today's world it is widely used in many different fields.

The advancement of technology has brought changes to the way sensors operate and exchange data. So when we say now that a sensor sends its data somewhere, we automatically assume that it is being sent using wireless technology. That is why the focus of this paper is the Wireless Sensor Networks (WSNs). A Wireless Sensor Network is a network that consists of autonomous sensors that are spatially dispersed, that monitor some specific measurement like temperature or humidity, and communicate their data through the network to a main data gathering location. A single network can contain from a few to several hundreds of sensors. Some of the areas where WSNs are being used are: health care monitoring, environmental sensing, natural disaster prevention etc. This shows how widely spread and important the WSNs are.

A new, fast growing concept that has sensor data and their usage at its core is the Internet of Thing (IoT). The Internet of Things is a system of interconnected computing devices, machines, vehicles, buildings, people etc. that are provided with unique identifiers to transfer data over a network without the need of human to human or human to computer interaction. We intentionally mentioned people in the description of the IoT explanation above, because a human can also be a "thing" in the IoT. If a person has a heart monitor implant for example, he/she can become a part of the IoT. Even animals can be a part of it, since many of them have biochip transponders. This creates opportunities for better and easier integration of the physical world into computer systems, which results in improved efficiency, accuracy and eventually economic benefit. There are some predictions that by 2020 IoT will consist of more than 50 billion of objects [1].

Having in mind the volume and velocity of data produced in potential use case scenarios of sensors the problem of data size reduction becomes crucial if one wants to optimize the costs for the potential solution and reduce the data latency. In this paper, we use data prediction as a very efficient strategy to reaching the above-mentioned goals.

This paper is organized as follows. The Data prediction algorithms are presented in Section II, in Section III we describe the development of the DAPS , and a case study is shown in Section IV. Finally, the conclusion is in Section V.

II. DATA PREDICTION ALGORITHMS

In this section we are going to explain the basics of data prediction. Then, we will briefly introduce the algorithms used in our DAPS.

The term prediction has very clear meaning. A prediction is a statement about an uncertain event. It is often based on experience or knowledge, but this is not always the case. Among many different types of predictions, data prediction is becoming the most interesting one in many fields. When we talk about data prediction in this paper, we refer to time series prediction. Time series is a series of data points listed in time order. Usually the data in the sequence is taken at successive equally spaced points in time. The data is usually collected

from sensors that take certain measurements. Time series analysis and prediction is important in many different fields (studying wildlife, environment, households, analytics in multimillion companies, etc.) [2]. This can help to make a better decision about some repeatable event, which would reduce costs in long term.

Different models are used in the literature for data prediction [3]. These models can have many forms and represent different stochastic processes. There are three broad classes of models: the Autoregressive (AR), the Integrated (I) and the Moving Average (MA) [4]. All of these classes rely on previous data points. Combinations of these classes are possible, such are Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) etc.

In our web-based system, we implemented different types of Moving Average (MA) and two types of Least Mean Square (LMS).

A. Simple Moving Average

Simple Moving Average (SMA) is the unweighted mean of the previous n data. The calculated value is the prediction for the following one in the time series [5].

$$SMA = \frac{p_M + p_{M-1} + \dots + p_{M-(n-1)}}{n} \quad (1)$$

P_M is the last value we have had until the moment of prediction, P_{M-1} is the previous value etc. until we reach the $P_{M-(n-1)}$ value, where n is the number of previous measurements in the data series. Usually, we choose the n value.

In this system after the next value is predicted, we compare it with the real value and decide if the prediction is range [SMA - Ɛ, SMA + Ɛ], where Ɛ is a constant that is user defined. If the prediction is within this range then we take it as correct, if not, then it is incorrect and the actual value from the time series is being taken and used later for the next prediction.

There are different approaches to using the SMA technique depending on the number of previous data points used in the prediction.

Moving Average 1 (MA1) takes the previous value as the next prediction. It is insufficient to say that this is the simplest predicting technique possible and very inaccurate. Although for some types of data, like temperature, it is adequate.

The Moving Average 2 (MA2) calculates the mean of the previous two values in the series. This is slightly better than the MA1.

Finally Moving Average 3 (MA3) calculates the average of the previous 3 measurements in the series.

Additionally, there are other types of Moving Average algorithms. One of them is Weighted Moving Average (WMA). This is an average that adds multiplying factors to give different weights to data [6].

B. Least Mean Square

Least Mean Square (LMS) are a class of adaptive filters used to mimic a desired filter. LMS gives the least mean square of the error signal, that is the difference between the desired and the actual signal [7].

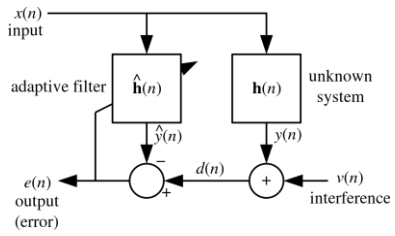


Fig. 1. Least Mean Square scheme

There is an unknown system marked above as **h(*n*)** and an adaptive filter **ĥ (*n*)** which tries to adapt to the system and be as close as possible to it. The input to the system **x(*n*)** is the data from the time series. The variables **y(*n*)** and **ŷ(*n*)** are output of the system and the filter and they are compared to give the error **e(*n*)**. **ŷ(*n*)** is calculated as a dot product with the filter weights.

$$\hat{y}(n) = W(n) * X(n) \quad (2)$$

The main goal of the LMS is to adapt the filter weights, so the filter gives prediction as close as possible to the actual value. If the Mean Square Error (MSE) gradient is positive, it implies that the error would keep increasing positively. This means we need to reduce the weights, and vice versa.

$$W(n+1) = W(n) + 2\mu e(n) X(n) \quad (3)$$

X(n) is the input signal vector of adaptive filter at n times, W(n) is the estimate value of weights vector of filter, e(n) is the error signal, μ is the step factor, which used to control the stability and convergence rate of algorithm. The mean-square error, as a function of filter weights is a quadratic function which means it has only one extremum, that minimizes the mean-square error, which is the optimal weight. The LMS thus, approaches towards this optimal weights by ascending/descending down the mean-square-error vs filter weight curve.

There is another algorithm we implemented, that is a slight variation of the LMS. It is called Least Mean Square – Variable Step Size (LMS-VSS) [8]. The only difference to the LMS is that here the step μ is changing. We set the initial step, the minimum step, maximum step and incremental step. The incremental step is increment by which the step size changes from iteration to iteration.

$$\mu_0 = \mu + (\text{IncStep})\,(gg_{prev}) \quad (4)$$

μ_0 is the new step size, μ is the previous step size, IncStep is the incremental step. g is denoted with the following equation:

$$g = Xe^* \quad (5)$$

where g_{prev} is the analogous expression from the previous iteration, and the * operator denotes the complex conjugate. X is the vector of all inputs and *e* is the error.

Then the new step size is given by

- μ_0, if it is between MinStep and MaxStep
- MinStep, if μ_0 < MinStep
- MaxStep, if μ_0 > MaxStep

Additionally, we will explain what Root Mean Square Error (RMSE) is. We use it to show the error of the prediction in our system. RMSE is a frequently used method to measure the difference between the predicted and actual values in a time series. The RMSE is a measurement of accuracy.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{n}} \quad (6)$$

In the equation $\hat{y}$ and y_i is the corresponding sample from the time series, *n* is the total number of samples. So the closer the RMSE is to zero, the more correct the prediction is.

III. Development of DAPS

In this Section, we will explain the process of developing DAPS, the web part of the system, the implementation of the algorithms and the graph drawing part.

A. Design and implementation of DAPS

In order to visualize the importance of data prediction and show the results of its algorithms, we created a web-based system that shows the calculations in a more comprehensible way for the user. Since it is a system on the web, the client-server architecture is the most common and appropriate. It is designed in a way that the user sends all the necessary data to the server. The server analyses and processes the request, and later visualizes the results.

The system is implemented with different technologies for different parts of it. The main part is making this web-based. That was written in Spring Boot, which is Java based framework. The back end was made with it. Views are made with HTML and interactions with it are made possible with JavaScript language. Data transfer between the controllers in Spring Boot and the HTML is managed with server-side Java template Thymeleaf.

Fig. 2 shows the flow of the actions in the system, step by step.

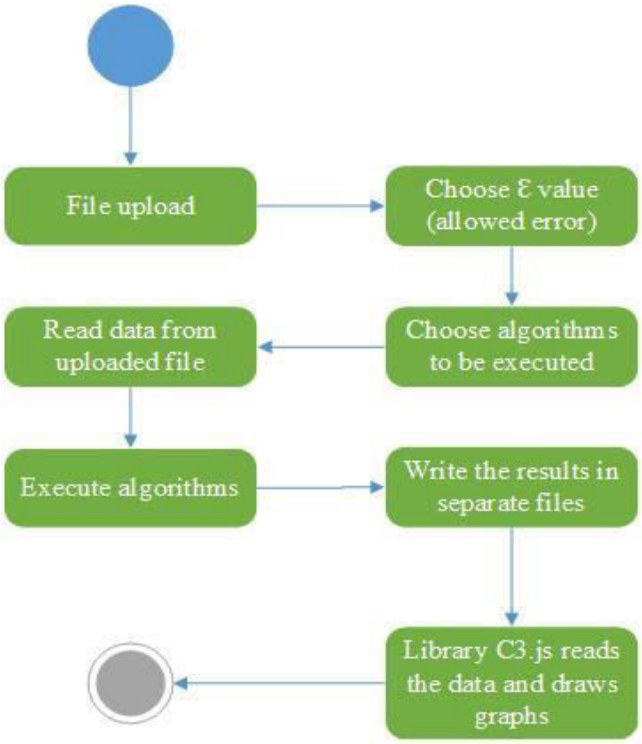


Fig. 2. Activity diagram for the DAPS

B. Functionalities of DAPS

The first part is uploading a file with the time series data. Here it is important to say that the file must be in CSV format and that the data inside is one dimensional. The first row (cell) should contain the name of the data and in the rest is written the data from the time series.

Next, there is choosing the Ɛ range, which sets the range in which a prediction is considered correct. This is set by the user along with the interval in that range. The user also selects the algorithms to be executed on the data. One or more can be chosen and run.

After the user clicks the "Upload" button, the system reads the data, runs the chosen algorithms and writes the results in separate files that are created in the user's Downloads folder.

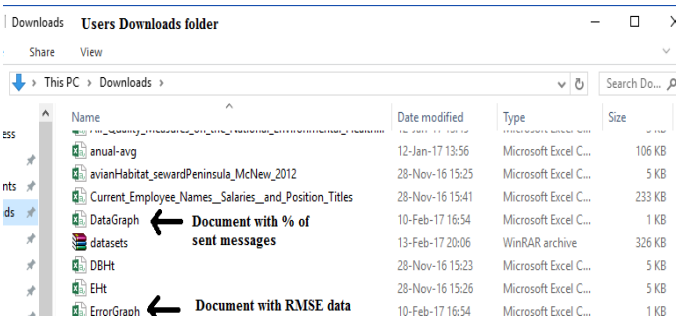


Fig. 3. User's Donwloads folder and the two files created in it.

Finally, the c3.js library reads the results and draws two graphs with certain parameters. The first one shows the percent of sent messages with all of the selected algorithms. The second presents the Root Mean Square Error (RMSE) for each of them.

As mentioned before this system runs prediction algorithms and then presents the results on graphs. Here we are going to explain what those "results" represent. There are 2 graphs, each of them representing one measurement. Those are percentage data reduction and Root Mean Square Error (RMSE).

The percent of data reduction is self-descriptive. Clearly, it shows the relation of the number of sent messages to the number of total possible messages in a sensor system. What that means is that each of the samples in the time series is considered as one message and if there are no prediction algorithms every single one of them must be transmitted in the system. But when an algorithm is run on the data we get some predictions and those predictions are compared to the actual data. If the prediction is good enough (the prediction is in [val. - Ɛ, val. + Ɛ] interval) then we consider that the message is not sent since the prediction is correct. The division of sent messages and total messages gives us the desired result.

IV. Case Study

In this section, we will show the actual output of the system and see how it performs on different datasets. We took two datasets in order to do this and compare the results. The first one is a dataset that has measurements for the air quality in the United States [9], and the second has data that shows the percentage of readmissions in the US [10]. Readmission is when a patient comes back in the same hospital after initially being released.

In the percentage of sent messages graphs that are produced, the X axis represents the allowed error that the user sets. The Y axis is the percent of sent messages for each of the values for the allowed error. In the RMSE graphs, the X axis is also the allowed error and the Y axis represents the RMSE values for the allowed errors.

A. *Air quality measurements*

In the first case, input DAPS uses data to measure air quality, in particular annual averages of the presence of PM 2.5 particles in micrograms per cubic meter. The data contains measurements from approximately 4,000 monitoring stations around the US, mainly in urban areas. Regarding the frequency of the sampling and taking measurements, it is different for every station. The data here is annual from all stations. The output of the system is shown in the following graphs:

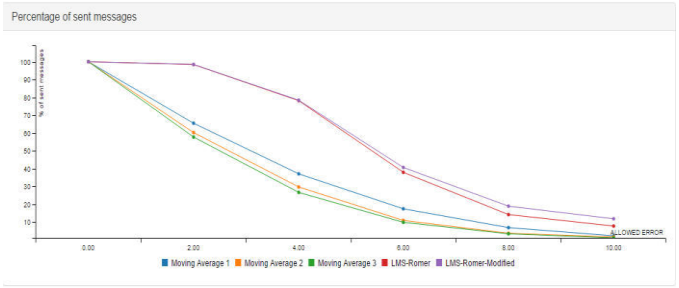


Fig. 4. Percentage of sent messages for air quality measurements

The graph shows that Moving Average 3 is the best, with the lowest number of sent messages, but the difference compared to other Moving Average algorithms is low.

The following chart shows Root Mean Square error of the same data-set:

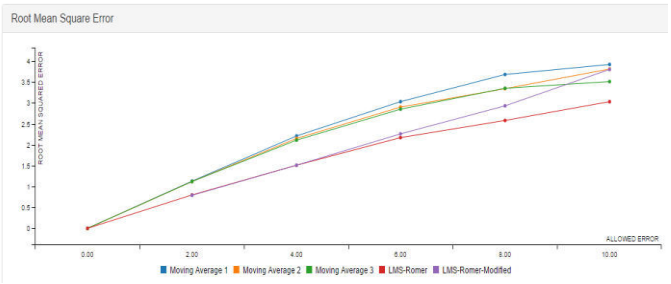


Fig. 5. Root Mean Square Error (RMSE) for air quality measurements

This graph shows that the LMS has the smallest error, but by increasing the allowed values for the Ɛ, the results of all algorithms are getting closer. As explained in one of the sections above, Ɛ is user defined and presents a constant that helps specifying the range in which a prediction is considered correct.

B. *Readmissions in hospital*

This data set refers to the percentage of people who had been received back again after their first discharge from hospital for treatment. A readmission is when a patient comes back to the hospital in a time span of 30 days after being released from there. Readmission rates have been increasingly used as an outcome measure in health services research and as a quality benchmark for health systems. This is annual data from the US Department of Health.

The following graphs are produced as results:

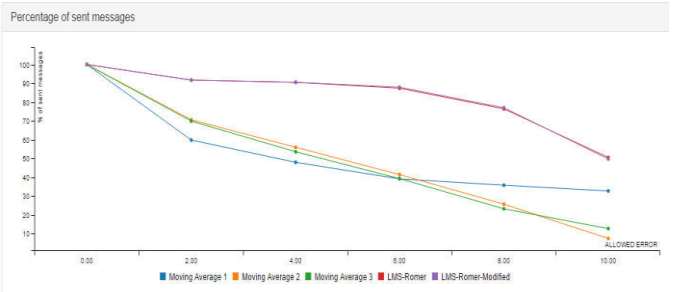


Fig. 6. Percentage of sent messages for people received back

For small Ɛ values, least sent messages are with Moving Average 1, but for bigger Ɛ values, Moving Average 2 and Moving Average 3 are better.

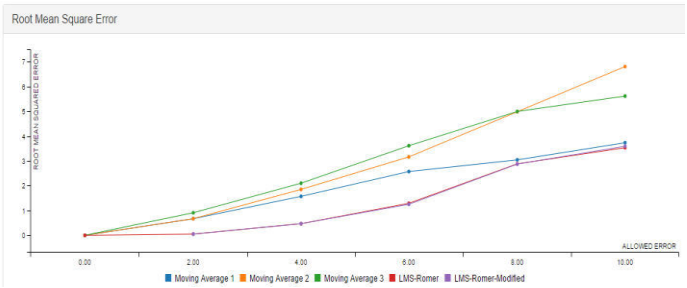


Fig. 7. Root Mean Square Error(RMSE) for people received back

This graph, which is similar to the previous case, confirms that LMS algorithms have the smallest error.

Most of the analyzed datasets, as the two previous cases, show the following:

- for the number of sent messages, better results indicate Moving Average algorithms, but they have a greater Root Mean Square Error.
- LMS sends more messages, but they have lower Root Mean Square Error.

Of course, there are exceptions in this rule, confirmed in fewer processed datasets, which means that finally the results depend on the data set itself.

V. Conclusion

Seeing the results of the case study we can confirm the importance of the data prediction, as it allows us to spend less resources and use them more efficiently.

The main achievement of this system is that it allows the users to see which algorithm is the best fit for their data very easily, and make decisions about further analysis of their data. Further development of such systems is needed as more complex and better prediction algorithms exist. We believe that DAPS can hugely help future developers of wireless sensor networks (WSN) and Internet of things (IoT) solutions, by providing the best sensor measurements reduction regardless of the specific case.

As time passes, the data keeps getting bigger and that is why the importance of systems like these will become greater.

References

[1] D. Evans, “The Internet of Things: How the next evolution of the internet is changing everything”, Cisco, 15 February 2016.

[2] D. Zissis, E. Xidias, D. Lekkas, "Real-time vessel behavior prediction". *Evolving Systems*, 2015

[3] D. J. Sheskin, “*Handbook of parametric and nonparametric statistical procedures*”, Boca Raton, FL, April 27, 2011 CRC Press. p. 109.

[4] N. Gershenfeld, *“The nature of mathematical modeling”,* New York: Cambridge University Press, 1999, pp. 205–208.

[5] B. Stojkoska Risteska, A. Popovska Avramova, P. Chatzimisios, "Application of wireless sensor networks for indoor temperature regulation." *International Journal of Distributed Sensor Networks*, 2014

[6] J. Devcic, “Weighted Moving Averages”, [Online] Available: http://www.investopedia.com/articles/technical/060401.asp

[7] S. Santini, K. Romer. "An adaptive strategy for quality-based data reduction in wireless sensor networks." In *Proceedings of the 3rd international conference on networked sensing systems (INSS 2006)*, pp. 29-36. 2006

[8] B. Stojkoska, D. Solev, and D. Davcev. "Data prediction in WSN using variable step size LMS algorithm." In *Proceedings of the 5th International Conference on Sensor Technologies and Applications*. 2011

[9] U..S. Department of Health & Human Services, Centers for Disease Control and Prevention, “Air quality measures on the national environmental health tracking network”, [Online], Available: https://catalog.data.gov/dataset/air-quality-measures-on-the-national-environmental-health-tracking-network

[10] U.S. Department of Health & Human Services, Centers for Medicare and Medicaid Services, “Readmissions and Deaths - Hospital”, [Online], Available: https://catalog.data.gov/dataset/readmissions-and-deaths-hospital

Random Walks on Protein Interaction Networks

Martin Milenkoski, Kire Trivodaliev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
milenkoski.martin@students.finki.ukim.mk, kire.trivodaliev@finki.ukim.mk

Abstract— **Random walk is a stochastic process that describes a path which consists of a succession of random steps on some mathematical space. In terms of graphs, a random walk that starts from a given query node results in a stationary vector (distribution) which can be interpreted as the affinity of the query node being "connected" with other nodes in the graph. In this paper random walks are used for protein interaction networks analysis. More specifically, different types of random walks, such as simple random walk and random walk with restart are simulated on a weighted graph representation of a human protein interaction network. The performances of the variants of the random walk process are evaluated for topologically different query nodes in the graph, in terms of their convergence time and node coverage. Further comparison is done by analyzing the overlap in the stationary vectors of the random walks. An application of random walks in solving a biologically motivated problem is shown in predicting members of a partially known protein complex (complex membership problem). The input for this is a core set of proteins (i.e. the queries) making up a protein complex. The problem in terms of biology is whether this core set is complete or not. In this paper, potential members of protein complexes are found by ranking proteins according to their random walk based affinity to the query proteins in the partially known protein complex. Evaluation is done by using a leave-one-out method for experimentally confirmed protein complexes.**

Keywords— random walk; protein interaction network; complex membership problem

I. INTRODUCTION

Many processes in the real world are random by nature, or at least appear random because of the numerous factors which can affect their outcome. Because of this, a lot of effort has been put in developing mathematical models which represent such processes. One such model is the random walk, which describes the path taken in some mathematical space, when each successive step is random. There are various types of random walk*s*, which can differ in several ways. One difference between the variants of random walks is the mathematical space in which they are implemented. The simplest example is the random walk on the integer line, but many other possibilities exist such as a plane [1], a higher dimensional vector space [2], a graph [3, 4], groups and random transformations [5] and many others. Another difference is the size of the steps taken. In a simple random walk on a lattice, the path can only move to some of the neighboring positions, while, in the Gaussian random walk the step size varies according to a normal distribution. Due to the inherent stochastic nature of many natural phenomena, the random walk is the algorithm of choice for analysis in a plethora of fields and applications, ranging from economy [6] and social networks [7], to physics [8] and biology [9].

Research utilizing random walks has been prominently focused on graphs and problems that can be transformed in a graph-like metric space. The usual scenario in such usage is for a random walker to start from a "query" node and move along the existing edges of the graph with some transition probability. Such method has been very successfully used in finding similarities between movies in a relational database [10]. In [11] a random walk over a document-level context graph is used in improving the ranking of videos for a given query phrase. A more localized random walk around the query node is achieved using a random walk with restart in which an artificial edge is added between the query node and every other node in the graph, thus allowing the random walk a probability of going back to the query node at each step. Such version of random walk has been used in image annotation [12]. Some common properties of real graphs like linear correlations and block-wise community-like structures have been used to improve the performance of this method [4].

In this paper the focus is on using random walks on protein interaction networks (PINs) which are fundamental to almost all biological processes [13] and are the richest source of information about proteins and their properties. Protein interaction networks (PINs) can be represented using a graph having each node represent a protein with the graph edges corresponding to an interaction between the connected nodes (proteins). The advancement of high-throughput technologies such as yeast-two-hybrid [14], mass spectrometry [15] and protein chip technologies [16] has enabled the construction of large interaction networks [17]. Since interactions between proteins are found experimentally there are a lot of false positive interactions [18] and a lot of effort is put into assessing their reliability [19].

With the growing size of protein interaction networks, various graph analysis techniques have been proposed for extracting potential new knowledge. Several studies have shown that random walks give superior results in the analysis of protein interaction networks [20]. Random walks on protein interaction networks for a given query protein result in a

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

stationary vector which contains an entry for every other protein in the network. These entries can be interpreted as an estimate of the probability of that protein being connected with the query protein. When the stationary random walk vector is sorted, the "top" proteins can be used for predicting various functions of the query protein [21]. Modification of this technique augmented with random walk on two other types of networks has successfully been used in protein function prediction [22].

In addition to function prediction, another biologically motivated problem in protein interaction networks is the detection of cluster structure in the graph and the prediction of new members of a partially known protein complex. Regarding the clustering of the PIN many algorithms and techniques have been proposed and used [23, 24]. In terms of random walks both stated problems have been treated. In [20] a modification of random walk with restart in which at each step, the walker can go back to any of the proteins comprising the starting complex is used as a technique for predicting complex membership. A technique known as repeated random walk in which random walk with restart is iteratively executed has been proposed for local cluster discovery [25].

In this paper the performance of different random walks in a human protein interaction network is explored. Topologically different proteins are considered as query nodes and results are compared both in terms of the stationary vector and the convergence time. For the complex membership problem, random walks are evaluated in terms of their predictive ability, and also in terms of their sensitivity to the topology of the complex i.e. the average "connectivity" measure of the proteins in a given complex, as well as the effect of the within complex "closeness" measure. The rest of the paper is organized as follows. Section 2 presents the data used in the paper. Additionally, it gives technical details for the algorithms used in the paper. In the third section a detailed description of the performed experiments and the corresponding results is given. Discussion for the results is also provided. Finally, the paper is concluded in the fourth section.

II. MATERIALS AND METHODS

This section provides details on the data used in the research, also the algorithms and problem for which they are employed.

A. Protein Interaction Data

The HIPPIE database [26] integrates several different expert curated databases of protein-protein interactions and assigns a confidence score to every interaction in the range of [0, 1]. This scoring scheme is optimized by human experts along with a computer algorithm and reflects the amount of experimental evidence for the given interaction. In order to build the protein interaction network used in this paper, PPI data is downloaded from the HIPPIE database. Additionally, all self-interactions, duplicate interactions and interactions with confidence score 0 are removed and the data is modeled as a weighted undirected graph. In this graph each vertex represents a protein named by the protein primary structure descriptor, and each edge denotes an interaction weighted by the HIPPIE confidence score. After all the preprocessing the final network is comprised of 16769 proteins and 277055 interactions.

For the work in this paper, we represent the protein interaction network as a graph $G = (V, E, W)$ where nodes $i, j \in V$ represent the proteins, edges $(i, j) \in E$ correspond to interactions between proteins *i* and *j*, and *W* is a matrix whose element w_{ij} is the weight associated to the (i, j) edge.

For the complex membership problem, the data is downloaded from the CORUM database [27]. This database provides a resource of manually annotated protein complexes from mammalian organisms, mainly human (64%), mouse (16%) and rat (12%). Each protein complex is described by protein complex name, subunit composition, function annotation along with mapping to Gene Ontology terms, Entrez Gene IDs of the subunits as well as the literature reference that characterizes the complex.

B. Random Walks on Graph

When modeling a random walk on a graph the aim is to simulate the trajectory of a random walker that starts from a query protein and moves to its neighbors with a probability proportional to the weight of each connecting edge. Random walk method simulates a random walker that starts on a query node, *q* (or a set of query nodes simultaneously). At every time tick, the walker chooses randomly among the available edges (based on edge weights). Let $p_i^{(t)}$ be the probability of finding the random walker at node *i* at time *t*. Then, the probability of the walker moving from node i to its neighboring node *j* at time $t + 1$ is $f_{ij}^{(t+1)} = q_i^{(t)} w_{ij}$. Thus, the probability of finding the random walker at node *j* at time $t+1$ is found by summing over the probabilities of moving to the node *j* from all its neighboring nodes:

$$p_j^{(t+1)} = \sum_i f_{ij}^{(t+1)} = \sum_i p_i^{(t)} w_{ij} \tag{1}$$

Let $p^{(t)}$ be the vector of all the values $p_i^{(t)}$. From equation (1), the following expression is derived:

$$p^{(t+1)} = W\, p^{(t)} \tag{2}$$

The simple random walk represents the iterative calculation of this matrix equation with the starting vector $p^{(0)}$ comprised of all zeros, except the value corresponding to the starting query node which is one. This value of 1 corresponds to the walker being at the query node with probability 1 when the process starts. In the case where the random walk starts from a set of nodes the cumulative starting probability of 1 is distributed among the nodes in the query set, with the default distribution being uniform. The goal of the random walk algorithm is finding the stationary vector $p_s = \lim_{t\to\infty} p^{(t)}$ whose elements $p_{s,i}$ represent the stationary probability of finding the walker at the node i.

For the random walk with restart algorithm, additional parameter c is added which represents the probability of the walker jumping back to the starting node (or set of nodes) at each point in time. This parameter c enforces a restriction on how far the random walk goes from the starting node. A value close to 1, will result in a stationary vector which reflects the local structure around the starting node, and a value close to 0 will give a more global view of the graph. The random walk with restart algorithm can be thought of as a simple random walk on a modified version of the original graph, in which an additional link is added from the starting node to all other nodes in the graph with a weight c, and the weights of all the other links are scaled by a factor of 1-c. From this representation the following equation can be derived:

$$p^{(t+1)} = (1-c)\ W\ p^{(t)} + c\ p^{(0)} \tag{3}$$

The number of iterations to converge is closely related to the restart probability *c*. As *c* gets smaller the diameter of the observed neighborhood increases, thus the number of iterations to converge gets larger.

C. Complex Membership Problem

The complex membership problem is defined in the following way: Given a core set of proteins in a protein complex C_1, and another set of proteins C_2, the goal is to rank the proteins in the set C_2 by their probability of being a member of the protein complex comprised of the proteins in the C_2. In this way the set of potential members of the complex can be greatly reduced and the experimental research can be focused on the proteins with the highest probability of being members of the complex. Since random walks produce a stationary probability distribution it is a very intuitive approach to solving this problem.

III. Result and Discussion

First, we want to investigate how the different random walks behave in the same experimental scenarios. Namely, we want to test how similar are the stationary distributions of different random walks when the walks are started from topologically different nodes in the graph. In order to do this, we rank the proteins by the average value of 4 node properties: local clustering coefficient, degree centrality, betweenness centrality and eigenvector centrality.

The local clustering coefficient of a node *i* is the proportion of possible edges between its immediate neighbors that actually exist in the graph. We define $N_i = \{ j \mid (i,j) \in E \}$ as the neighborhood of vertex *i*, which is a set of its immediately connected neighbors. Let $k_i = |N_i|$ be the number of vertices in the neighborhood of *i*. Then, the maximum number of possible undirected edges between the vertices in N_i is $m_i = k_i(k_i - 1)/2$. The local clustering coefficient of node *i* is defined as follows:

$$C_i = \frac{|\{(j,k) \mid j,k \in N_i \wedge (j,k) \in E\}|}{m_i} \tag{4}$$

Degree centrality of a node *i* is defined as the number of edges with one end in *i*: $D_i = \deg(i) = \sum_j w_{ij}$. In this paper the degree centralities of the nodes are normalized, so their values are in the [0,1] range.

Betweenness centrality of a node is a measure that quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of all shortest paths from node *j* to node *k* and let $\sigma_{jk}(i)$ be the number of shortest paths from *j* to *k* that pass through *i*. Then, the betweenness centrality of the node *i* is defined as:

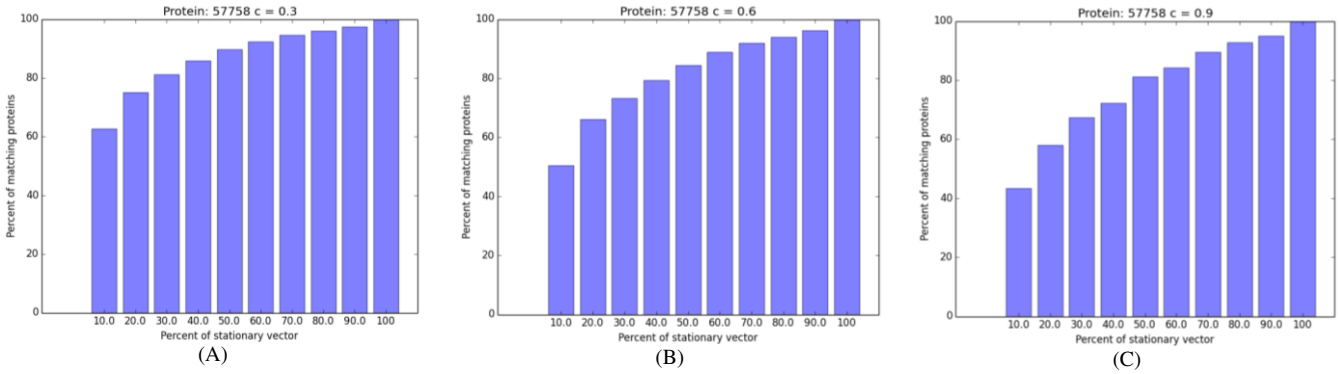
$$B_i = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{5}$$

Eigenvector centrality is a measure that assigns relative scores to all nodes in the graph based on the concept that connections to high-scoring nodes contribute more to the score of the node than connections to low-scoring nodes. Let *A* be the adjacency matrix of *G* such as its elements $a_{ij} = 1$ if $(i,j) \in E$, and $a_{ij} = 0$ otherwise. Then, the eigenvector centrality of the node *i* is defined as:

$$E_i = \frac{1}{\lambda} \sum_{j \in G} a_{ij} E_j \tag{6}$$

where λ is a constant. In matrix form Eq. (6) can be rewriten as $\lambda E = EA$. Hence the centrality vector *E* is the left-hand eigenvector of the adjacency matrix *A* associated with the eigenvalue λ. Usually, λ is chosen as the largest eigenvalue in absolute value of matrix *A*.

Finally, a measure $R_i = (C_i + D_i + B_i + E_i)/4$ is defined for every node, and the nodes are sorted in ascending order by this value. Then, the first (EntrezID = '57758'), middle (EntrezID = '6285'), and last element (EntrezID = '351') of this sorted arrangement of nodes are used as query nodes in the subsequent tests.



(A) (B) (C)

Fig. 1. Proportion of overlap between simple random walk and random walk with restart for query node 57758 with (A) c = 0.3, (B) c = 0.6, (C) c= 0.9

For every one of these three proteins, the stationary vector from the simple random walk is compared with the stationary vector from the random walk with restart algorithm. Namely, we test the proportion of overlapping proteins in the first x percent of the stationary vectors for different values of x in order to observe the difference between the two variations of random walk. Additionally, these tests are done for different values of the restart probability parameter *c* in order to examine the effects of restraining the walker to the local structure around the query node. Fig. 1 shows the results for the protein with EntrezID = '57758'.

We can see from the results what we previously stated about the effect of the parameter c on the stationary vector. Namely, as c increases the random walk is restricted to smaller subset of nodes in the local neighborhood of the query node as opposed to the simple random walk which gives a global picture about the graph. This results in a decrease of overlap between their stationary vectors which is most noticeable in the first 30-40% of the vectors. For comparison, Fig. 2(A) and Fig. 2(B), depict the overlap results for proteins with EntrezIDs '6258' and '351', accordingly, for values of the restart probability *c* of 0.3 (top) and 0.9 (bottom).

From these results it is clear that the previously stated effect from changing the parameter c remains the same even for topologically different nodes in the graph. Additionally, we can observe that the difference in the stationary vectors is most noticeable for the protein '351' which has the highest average measure of connectedness in the graph. This is because in this case the random walk is even more restricted to the local structure around the query protein because of its high connectivity.

TABLE I. EXECUTION TIMES IN SEC FOR DIFFERENT RANDOM WALKS WITH TOPOLOGICALLY DIFFERENT QUERY NODES

Protein ID	Simple random walk	Random walk with restart		
		c = 0.3	*c = 0.6*	*c = 0.9*
57758	0.66974051	0.245688658	0.136837883	0.095624431
6285	0.661865497	0.251660547	0.139837249	0.100178746
351	0.663240436	0.253963956	0.145360299	0.097390575

Additionally, the execution time in terms of different starting nodes and different walks is measured. The results are shown in Table 1.

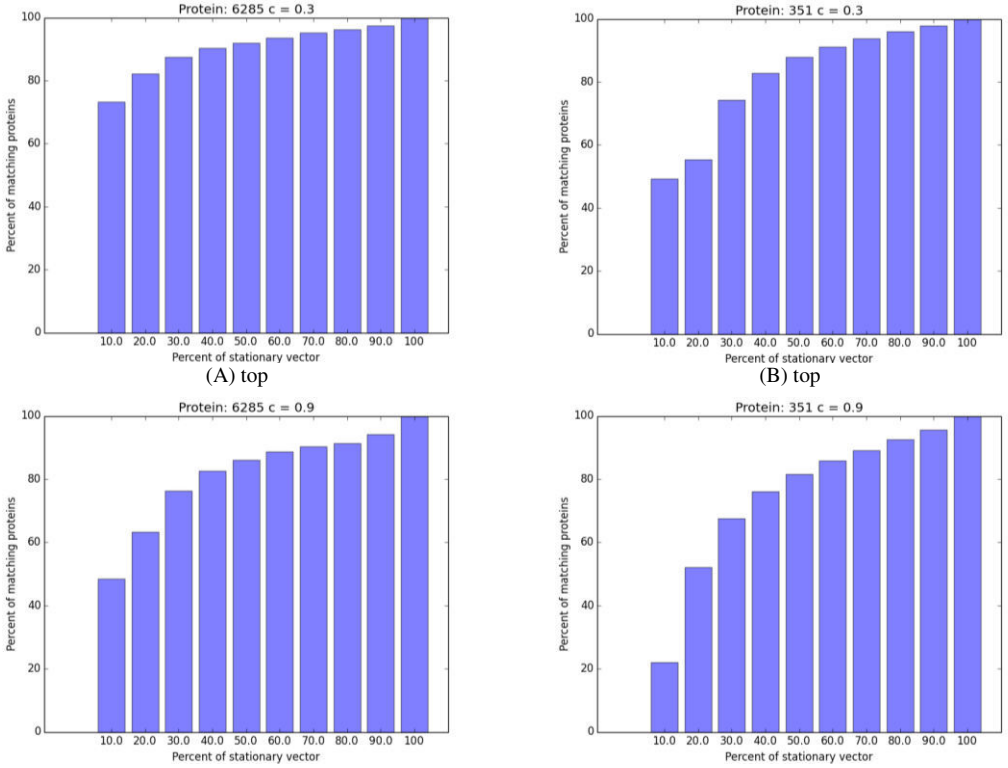
From these results we can conclude that the execution time of the random walks is independent of the topological characteristics of the query node. On the other hand, we can see that the execution time of the random walk with restart is smaller than that of the simple random walk, i.e. as the value of the parameter c increases the execution time of the random walk decreases.

For the complex membership problem, we consider only the 7 largest available complexes from the CORUM database whose constituents are all present in the PIN and these complexes are then ranked by the average R_i of their consisting proteins. The results are shown in Table 2.

TABLE II. PROTEIN COMPLEXES CONSIDERED FOR EXPERIMENTS ORDERED IN ASCENDING ORDER BY THE AVERAGE R_i OF THEIR CONSTITUENTS

Complex name	No. of proteins	Average Ri
55S ribosome, mitochondrial	77	0,082755567
Spliceosome	141	0,085403171
39S ribosomal subunit, mitochondrial	48	0,087629698
C complex spliceosome	79	0,100414267
Nop56p-associated pre-rRNA complex	104	0,154006929
60S ribosomal subunit, cytoplasmic	47	0,168434288
Ribosome, cytoplasmic	81	0,169347943

In order to observe the effect of the average value R_i, the complexes with highest and lowest average value are compared in terms of the complex membership problem experiments. These experiments are done using a leave-one-out technique. Namely, for every protein in a given complex, the position of the protein in the stationary vector of a simple random walk is calculated (also its position in the stationary vector resulting from the random walk with restart) the query set of nodes for the random walk composed of the remaining proteins in the complex. In the end, the results are averaged over all the proteins in the complex. This is done for three different values of the parameter c and the results are presented in Table 3.



(A) top (B) top

(A) bottom (B) bottom

Fig. 2. Proportion of overlap between simple random walk and random walk with restart for query node (A) 6285, (B) 351, with c = 0.3 (top) and c= 0.9 (bottom)

It is evident that the random walk with restart ranks the left out protein much higher in the stationary vector and is therefore a better approach for the complex membership problem. The results also show that the complex with higher average value R_i has a higher average position in the simple random walk, because its proteins have a higher connectivity with the rest of the graph. However, the rate of improvement of the random walk with restart is independent of the average value R_i of the different complexes, i.e. the average "connectivity" has no effect on the random walk performance.

TABLE III. COMPLEX MEMBERSHIP PREDICTION RESULTS USING DIFFERENT RANDOM WALKS ON COMPLEXES WITH HIGHEST AND LOWEST CONNECTIVITY

Complex name	Simple random walk	Random walk with restart		
		c = 0.3	*c = 0.6*	*c = 0.9*
55S ribosome, mitochondrial	4576,5	646	587,9	563
Ribosome, cytoplasmic	1176,69	246,08	217,7	217,1

Additionally, clusters are differentiated in terms of within complex "closeness". The measure for this feature of the complexes is defined as follows: A matrix D is defined such that for every pair of proteins i, j in the set of proteins comprising the protein complex $D(i, j) = D(j, i)$ is the length of the shortest path from i to j in the unweighted version of the graph G. We then define the average value of the matrix D as the measure of "closeness" between proteins in the given complex. The sorted results are shown in Table 4.

TABLE IV. PROTEIN COMPLEXES CONSIDERED FOR EXPERIMENTS ORDERED IN ASCENDING ORDER BY THEIR CLOSENESS MEASURE

Complex name	No. of proteins	Closeness
60S ribosomal subunit, cytoplasmic	47	1,257582617
Ribosome, cytoplasmic	81	1,274196007
Nop56p-associated pre-rRNA complex	104	1,550850592
39S ribosomal subunit, mitochondrial	48	1,661458333
C complex spliceosome	79	1,764460824
55S ribosome, mitochondrial	77	1,87890032
Spliceosome	141	1,891856546

In order to observe the effects of the within complex "closeness" measure on the performance of the random walks the lowest and highest ranking complexes are considered for the complex membership problem. The results are given in Table 5.

These results confirm the hypothesis that the random walk with restart algorithm improves the results of the simple random walk, and additionally higher values of the parameter c result in even bigger improvements. Again, as in the previous experiment, the rate of improvement is independent of the "closeness" measure of the protein complex.

TABLE V. COMPLEX MEMBERSHIP PREDICTION RESULTS USING DIFFERENT RANDOM WALKS ON COMPLEXES WITH HIGHEST AND LOWEST CLOSENESS

Complex name	Simple random walk	Random walk with restart		
		c = 0.3	*c = 0.6*	*c = 0.9*
60S ribosomal subunit, cytoplasmic	1144,4	314,7	278,5	273,1
Spliceosome	2468,1	647,5	512,3	486,7

From Tables 3 and 5 we can see that the random walk gives better results when working with a complex that is more central and has higher within cluster closeness, with the average centrality being a good indicator for the appropriateness of the random walk approach.

IV. CONCLUSION

This paper evaluates the usage of simple random walk and random walk with restart on a weighted graph representation of a human protein interaction network. Within the graph nodes are topologically differentiated based on an average connectedness (centrality) measure. Results for topologically different nodes suggest that the degree of change in stationary distributions of the different random walks with the increase in restart probability is proportional to the centrality of the node. Additionally, the results show that the speed of the algorithm increases with the increase of the restart probability. These two results imply that one can utilize faster random walks when a node centrality is lower with very little loss of information. This is important, especially when dealing with massive networks.

The paper also presents an application of random walks in solving the protein complex membership problem. The results show that the random walk performs better with complexes that have a high average centrality and within cluster closeness. Also, a high restart probability yields better results which is expected since protein complexes are usually locally organized clique-like structures in the protein interaction network.

REFERENCES

[1] J. M. Borwein, D. Nuyens, A. Straub, and J. Wan, "Random walks in the plane," in *22nd International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2010)*, 2010, pp. 191-202.

[2] A. J. D'Aristotile, "The nearest neighbor random walk on subspaces of a vector space and rate of convergence," *Journal of Theoretical Probability,* vol. 8, pp. 321-346, 1995.

[3] R. Burioni and D. Cassi, "Random walks on graphs: ideas, techniques and results," *Journal of Physics A: Mathematical and General,* vol. 38, p. R45, 2005.

[4] H. Tong, C. Faloutsos, and J.-y. Pan, "Fast Random Walk with Restart and Its Applications," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*, 2006, pp. 613-622.

[5] A. Furman, "Random walks on groups and random transformations," *Handbook of dynamical systems,* vol. 1, pp. 931-1014, 2002.

[6] S. Schmitt-Grohé and M. Uribe, "Closing small open economy models," *Journal of international Economics,* vol. 61, pp. 163-185, 2003.

[7] 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*, 2011, pp. 635-644.

[8] J. Kempe, "Quantum random walks: an introductory overview," *Contemporary Physics,* vol. 44, pp. 307-327, 2003.

[9] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology," *Journal of the Royal Society Interface,* vol. 5, pp. 813-834, 2008.

[10] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on knowledge and data engineering,* vol. 19, 2007.

[11] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 971-980.

[12] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Image annotation refinement using random walk with restarts," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 647-650.

[13] L. Hakes, S. C. Lovell, S. G. Oliver, and D. L. Robertson, "Specificity in protein interactions and its relationship with sequence diversity and coevolution," *Proceedings of the National Academy of Sciences,* vol. 104, pp. 7999-8004, 2007.

[14] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 4569-4574, 2001.

[15] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams*, et al.*, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature,* vol. 415, pp. 180-183, 2002.

[16] H. Zhu and M. Snyder, "Protein chip technology," *Current opinion in chemical biology,* vol. 7, pp. 55-63, 2003.

[17] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature,* vol. 402, pp. C47-C52, 1999.

[18] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields*, et al.*, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature,* vol. 417, pp. 399-403, 2002.

[19] C. M. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg, "Protein interactions two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics,* vol. 1, pp. 349-356, 2002.

[20] T. Can, O. Çamoğlu, and A. K. Singh, "Analysis of protein-protein interaction networks using random walks," in *Proceedings of the 5th international workshop on Bioinformatics*, 2005, pp. 61-68.

[21] K. Trivodaliev, I. Cingovska, S. Kalajdziski, and D. Davcev, "Protein function prediction based on neighborhood profiles," in *ICT Innovations 2009*, ed: Springer Berlin Heidelberg, 2010, pp. 125-134.

[22] W. Peng, M. Li, L. Chen, and L. Wang, "Predicting protein functions by using unbalanced random walk algorithm on three biological networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 2015.

[23] K. Trivodaliev, A. Bogojeska, and L. Kocarev, "Exploring function prediction in protein interaction networks via clustering methods," *PloS one,* vol. 9, p. e99755, 2014.

[24] K. Trivodaliev, S. Kalajdziski, I. Ivanoska, B. R. Stojkoska, and L. Kocarev, "Chapter Nine-SHOPIN: Semantic Homogeneity Optimization in Protein Interaction Networks," *Advances in protein chemistry and structural biology,* vol. 101, pp. 323-349, 2015.

[25] K. Macropol, T. Can, and A. K. Singh, "RRW: repeated random walks on genome-scale protein networks for local cluster discovery," *BMC bioinformatics,* vol. 10, p. 283, 2009.

[26] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, "HIPPIE: Integrating protein interaction networks with experiment based quality scores," *PloS one,* vol. 7, p. e31826, 2012.

[27] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo*, et al.*, "CORUM: the comprehensive resource of mammalian protein complexes—2009," *Nucleic acids research,* vol. 38, pp. D497-D501, 2010.

TRANSACTION PROCESSING APPLICATIONS IN CLOUD COMPUTING

Filip Mitrevski
St. Kliment Ohridski University
Faculty of Information and Communication Technologies
Bitola, Macedonia

Darko Pajkovski
St. Kliment Ohridski University
Faculty of Information and Communication Technologies
Bitola, Macedonia

Tome Dimovski
St. Kliment Ohridski University
Faculty of Information and Communication Technologies
Bitola, Macedonia

ABSTRACT

Cloud computing is a recent attractive term in the IT world. The term "Cloud Computing" comes out of the idea for centralizing the storage and computation in distributed data. Its long term goals are to provide a flexible, on – demand package to the cloud user, giving him much more freedom, flexibility and reliability at the same time, achieving all of the above by using a simple "utility computing model". It promises to bring on-demand pricing, less IT overhead and an ability to scale IT up and down quickly.

The focus of this work falls down on transaction processing applications which work in multi – processing and cloud environments. All major vendors have adopted a different architecture for their cloud services. As a result, in this paper we will be reviewing some of them and their fundamental approaches on improving Cloud Transactions.

I. INTRODUCTION

Transaction processing has been an important software technology in the last five decades. The government, telecommunication sectors, finance, transportation and military are all dependent on the transaction processing applications for their services, namely order processing, banking, electronic reservations, telephone switching, etc. Transaction processing systems are used by many large hardware and software vendors such as IBM, Microsoft, Google, Amazon, Oracle, Dell and their revenue for transaction processing products and services is in the tens of billions of dollars per year.

Cloud computing is a computing service offered over the Internet, in which the software is seen as a service and the applications and data are stored on multiple servers (locations).

In the current cloud computing architecture (Fig.1) there are data centers which are able to provide services to all of the clients participating in the same cloud.

Cloud computing allows us to move the processing effort from the local devices such as laptops, personal computers from various locations, to the data center facilities. For example, in such a way, any device could be able to solve some complex differential equations by simply passing the specific arguments to a data center service which will be capable to give back the result in a very short time. However, the security of data and applications becomes a very major issue.

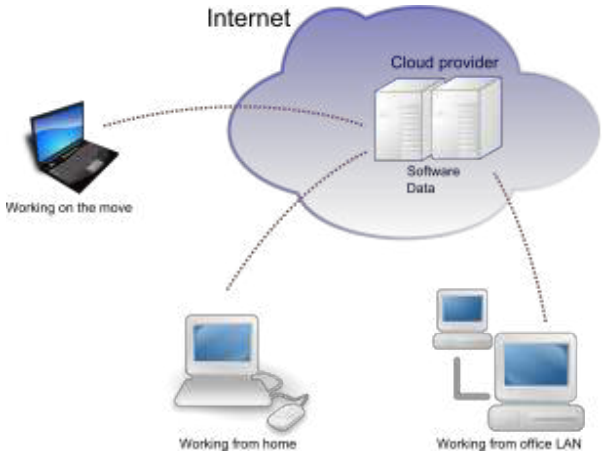


Figure 1: Cloud Computing Architecture

The main benefits of the cloud computing are the following:

- Flexibility – If our needs increase over time, it is very easy to scale up the cloud capacity, likewise if our needs scale down again, the flexibility is baked into the service.
- Disaster recovery – There are several solutions implemented as a "cloud-based" backup and recovery that save cut down the time penalty.
- Automatic software updates – Eliminate the need to spend time maintaining the system manually.
- Increased collaboration – Cloud based workflow and file sharing applications help to provide the updates in real time, gives them full visibility of their collaborations.
- Work from anywhere – Any device connected to the internet is able to do the job.
- Security – If the client computer crushes, there is almost nothing lost because everything is stored into the cloud in real time.

One of the main advantages of cloud computing is the promise of (virtually) infinite scalability so that IT administrators needn't worry about peak workloads. Finally, cloud computing provides flexibility and reliability at the same time in the utilization and management of both hardware and

software, which translates into savings in both "production time" and cost.

The major players in this field of cloud computing are Google, Amazon, Yahoo, Microsoft and some hardware manufacturers like IBM, Dell, HP, Intel. In this paper we will be reviewing some of them and their services, applications and fundamental approaches on improving cloud transactions.

II. RELATED WORK

With the emergence of cloud computing, some studies have evaluated the performance, scalability and reliability of cloud computing infrastructures. Some of them compared the performance of Hadoop, like an open source Java-based programming framework versus the more traditional (SQL-based) database systems [1]. The results of related studies on cost-consistency trade-off for OLTP workloads in the cloud have been reported [2].

OLTP (online transaction processing) is a special "class" consisted of software programs, capable of supporting transaction-oriented applications on the Internet, which is a much better approach compared to the traditional "distributed transaction" model – companies have many challenges regarding the performance when dealing with multiple nodes and databases required for mission-critical transactions. The second challenge comes from a case when the companies are highly distributed and they are communicating with their business partners who might be located all around the world. These are the main reasons why the traditional "distributed transaction" model is slow and unreliable.

Managing data in highly distributed OLTP environments means containing customer and product data which must be read from and written to constantly and in real-time in order to support the quality of each transaction. For example, this type of transaction takes place when we take out money at an ATM machine. Once our card is validated, a debit transaction takes place against our current balance to reflect the amount of cash that is being withdrawn. This type of transaction also takes place when we deposit money into our accounts and the balance gets an update.

III. DISTRIBUTED DATABASE ARCHITECTURES

Recently, distributed database architectures have found a place in cloud computing. First, the classic multi-tier database application architecture is described and then, two other variations of this architecture are also described such as replication, partitioning [12].

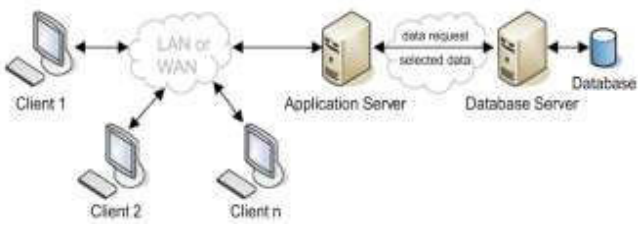


Figure 2: Distributed Database Architecture

A. Classic

Requests from the clients are dispatched to an available machine which runs a web and an application server. Afterwards, a web server handles the HTTP request from the clients and the application server executes the code specified in some program language with embedded SQL, which is shipped to the database server and interprets this request, returns a result and updates the database. The interface between the database server and the database itself includes shipping physical blocks of data (64K blocks), by using GET and PUT requests.

B. Partitioning

The difference between classic database architecture and partitioning is simple: The usage of a separate database server for controlling each partition in a database which is logically partitioned. In any database literature there are many examples for partitioning schemes: vertical partitioning vs. horizontal partitioning, round-robin vs. hashing vs. range partitioning [3].

For cloud computing the database architecture for partitioning was first founded by Force.com, by the platform that runs the "Salesforce" application. In Force.com, the partitioning includes a whole server-side application stack, with web and application servers. Here, all the requests to the same tenant are handled by the same app, web and database server.

C. Replication

In replication as with partitioning, there are several database servers and each of them saves copy of the whole database. The most important characteristic of replication is the mechanism to keep the replicas consistent. The replication is recommended to be transparent, therefore the requests are routed automatically to the Master. If the replication is not transparent, certain applications direct all update requests to the database server, which controls the Master copy and the Master server propagates all committed updates when these have been successfully committed.

IV. CLOUD SERVICES

In this section we will describe the alternative services offered by several big players of cloud computing world: namely Amazon (AWS), Google, Microsoft and Oracle. Their services are different in many aspects like the business model which is in use, software components used at all tiers and by the programming model. Only the Google App engine provides a service which can do automatic scalability and persistent data storage whilst fully automated hardware resources for all tiers. Amazon provides a service called "AutoScaling" which is used to automatically scale-out and scale-down EC2 machines for web tier.

A. Google App-Engine

Google App-Engine [5] is a platform for development and deployment of web applications which provides a whole platform as a service ("PaaS") to the cloud users. Several languages are supported in Google App-Engine like Python, Java and any extension of JVM languages, both with

embedded SQL for access to the database. The main advantage of the Google App-Engine is the automatic scaling of the resources consumed by an application, depending on the workload, so the cloud user doesn't have to worry about the spikes in the traffic or data. These applications [4] are intended for social networking start-ups, event-based websites and public institutions (school, universities, governments) etc.

Google App-Engine provides live migration on engine's instances to nearby hosts while active-even while under extreme load (up to 1.5 TB with their working SSD storage). Also, Google App-Engine offers optimal pricing for per-minute-billing, sustained-use, and special pricing for particular use.

B. Amazon EC2

One of the leading web based services that Amazon provides to the public, is Amazon's Elastic Compute Cloud, or simply called EC2[7]. EC2 is a web service, which provides resizable and secure compute capacity in cloud environments. At the same time, it has been specifically designed for the developers in order to make web-scale cloud computing easier. One of the main advantages that EC2 provides is the simple web service interface, which allows the user to control and obtain the desired capacity with minimal effort, providing total control over the computing resources. EC2 also reduces the time required to boot server instances to just a few minutes. What makes EC2 "unique" is the fact that it allows the user to pay only for the capacity that he actually uses – cuts down unnecessary cost and space usage to the server side as well.

Amazon EC2 Features:

- Elastic Web-Scale Computing: Increasing and decreasing capacity in a few minutes, commissioning many server instances simultaneously.
- Completely Controlled: Complete control of the instances together with root access, remote control using web service APIs
- Flexible Cloud Hosting Services: Multiple instance types, software packages and operating systems.
- Integrated Design: EC2 is integrated in most AWS services
- Reliability: Proven infrastructure, highly reliable environment.
- Security: EC2 works in pair with Amazon VPC, providing robust networking.
- Pre-defined instances: Allows the in-experienced user pre-built packages according to his needs, ranging from General Purpose Instances, which provide a certain "baseline" level of performance, up to GPU Graphics Instances, featuring up to several NVIDIA GRID GPU's.
- Elastic Load Balancing: Achieving greater fault tolerance, dynamically providing the correct amount of load balancing needed to response the incoming traffic, detection of unhealthy instances.

C. Gnubila – GPaaS (G Platform as a Service)

One of the products that *gnubila* provides is G Platform as a Service, or GPaaS [8]. GPaas is based on the G Platform, which evolved from a "simple" application server into a native cloud platform which includes DBMS, BPM and middleware technologies. All of this is coupled together with a kit of tools for maintenance for applications and services. A development framework which supports a user interface which provides infrastructure capabilities, session management (authentication, authorization, protection) and most importantly, transaction integrity, scalability and reliability are being supplied from the integrated development and deployment tools.

GPaas includes an environment dedicated for development, named GDeveloper, which can customize the User Interface based on HTML, CSS and JavaScript, making it easy for people to work together with social networks.

GPaaS advantages for developers:

- Faster time to market: Using Metadata as an approach instead of code makes changes to the application easy
- Simplified Deployment: All focus is put towards development and innovation
- Scalability: Use of hardware on demand, eliminates the need to write code for scalability
- Simplified application architecture: Fast data access, eliminates bottlenecks found in web applications, completely configurable
- Flexible data model
- Hybrid architecture: Providing transitions from local machines and server to a cloud model.
- Data management from multiple databases

D. Microsoft Azure

Microsoft Azure [6] represents a set of cloud services using .NET and SQL Server. The focus of this service falls in the category of Platform as a Service (PaaS). It is actually between a complete application framework like Google App-Engine and hardware virtual machines solutuon like Amazon EC2.

By using Windows Azure, customers can run applications and stored data on internet accessible machines owned by Microsoft. Applications in Microsoft Azure run only in user mode - no administrative access is allowed here. The primary goal of Windows Azure platform is to support a large number of simultaneous users compared to the Google App-Engine, which is more interested in small applications with light workloads. Also, Windows Azure differs from Amazon EC2 and Google App-Engine with regard to the pricing. Windows Azure charges a monthly or hourly flat fee depending on the database size with unlimited connectivity and virtual machine size.

E. CloudTran & Oracle Coherence

CloudTran [9] is specifically made to enable developers easy use of the IMDG (In-memory data grids) architecture. It provides ACID-property transactions across nodes. CloudTran is also "Built for Scale" – enables the usage of industry

standard components in order to quickly and easily build scalable, transactional applications. This gives us transactions with ACID-property without using phase commits, therefore the applications scale and run at grid speed all the time. CloudTran is suitable for use in environments where there is a need for scalable and fast transactions, coupled with easy recoverability of data.

CloudTran represents a middleware solution that works alongside, and also uses the features of Oracle's Coherence. Data which is bigger than the memory cache size is partitioned into multiple sets. All of the data sets are stored across nodes in the cluster – stores all the data which is required for an application in IMDG.

Oracle Coherence represents an IMDG tool which can take care of data objects stored in the RAM across servers organized in a cluster. The number of servers [10] can be modulated easily, together with the amount of RAM that is available. IMDGs use distributed caching in order to increase the performance and reduce the latency to existing databases.

Oracle Coherence provides several core functions and benefits:

- Caching – Consistent view of cached data, making data analysis easy for applications – maximizing the parallel capabilities of the data grid.
- Analytics – Aggregating and sorting of data, parallelizing operations across an entire data grid, ensuring that server failures do not affect calculation results.
- Transactions – Guaranteed data consistency in extreme transaction procession workloads.
- Events – Event handling mechanisms, capable of dealing with intense event rates, such as stream processing and continuous query for desktop applications.

When compared to other leading products, such as EC2 for example, Coherence provides scalability and flexibility in a similar way, enabling applications to scale linearly and dynamically in order to reduce and make the cost more predictable and to maximize the resource utilization. By offloading the processing from back end systems [11], Coherence supports continually growing application loads with minimal risk of data loss.

V. Conclusion

Cloud computing is a recent attractive term in the IT world that provides a service which is offered over the Internet, where the applications and the data are stored on multiple servers (locations). Corporations like Google, Amazon, Microsoft, Oracle are already providing cloud services. Their products in the likes Google App-Engine, Amazon EC2, Microsoft Azure, CloudTran and Gnubila are one of the best in the market with their ease of use, availability aspects, reliability and flexibility. At the same time they are achieving all of the above by using a simple utility computing model.

In this paper we have evaluated the benefits from each product as a package, enabling users to choose what is best for them, depending on their specific needs.

VI. References

[1] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker. A Comparison of Approaches to Large-Scale Data Analysis. In Proc. of SIGMOD, pages 165–178, 2009 *(references)*

[2] T. Kraska, M. Hentschel, G. Alonso, and D. Kossmann. Consistency Rationing in the Cloud: Pay Only when it Matters. In Proc. of VLDB, volume 2, pages 253–264, 2009

[3] S. Ceri and G. Pelagatti. Distributed databases principles and systems. McGraw-Hill, Inc., 1984.

[4] Persistent Systems. Google app engine. http://www.persistentsys.com

[5] Intoducing Google app engine https://cloud.google.com/

[6] D. Chappell. Introducing windows azure. http://go.microsoft.com/ December 2009

[7] Amazon Elastic Compute Cloid: https://aws.amazon.com/ec2/

[8] G Platform as a Service: http://www.gnubila.com/geas/gpaas

[9] CloudTran Overview: http://www.cloudtran.com/productOverview.php

[10] Indu Arora, Dr. Anu Gupta. Modeling and Designing Integrated Framework for Data Management of Transactional Applications in Cloud. http://www.ijser.org/researchpaper%5CModeling-and-Designing-Integrated-Framework-for-Data-Management.pdf

[11] Features and benefits of Oracle Coherence: http://www.mythics.com/about/blog/the-features-and-benefits-of-oracle-coherence

[12] Donald Kossmann, Tim Kraska and Simon Loesing. An Evaluation of Alternative Architectures for Transaction Processing in the Cloud, pages 2-3, 2010

Instance Based Learning in Protein Interaction Networks

Martin Josifoski, Kire Trivodaliev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
josifoski.martin@students.finki.ukim.mk, kire.trivodaliev@finki.ukim.mk

Abstract— **One of the essential challenges in proteomics is the computational function prediction. Protein interaction networks (PINs), as the richest source of information, can be utilized to solve this problem. PINs are modeled via graphs where proteins correspond to nodes, interactions to edges, and protein function to labels associated to nodes. The problem of computational function prediction now becomes one for proper labeling of its corresponding node. In this paper the PIN graph is used to derive a continuous vector representation of its nodes using semi-supervised learning. The approach used employs a biased random walk procedure with a flexible notion of a nodes neighborhood, which efficiently explores diverse neighborhoods. The produced vectors maximize the likelihood of preservation of the graph topology locally and globally. Once the vector representations of the nodes are produced learning is modeled as a set of binary classifications where a single classification corresponds to a single label (from a set of possible labels). In this single label classification the objective is to determine the existence (or non-existence) of a label to which aim the node vector representations are noted positive (if node has the label) or negative (node does not have label). Experiments are performed using a highly reliable human protein interaction network. Classification is done using two well known algorithms, namely, k-nearest neighbors and support vector machines. Results prove that this approach can be used to successfully tackle protein function prediction, but also provide insight to improvements that can make it comparable to state-of-the-art methods.**

Keywords— protein interaction networks; instance based learning; protein function prediction

I. INTRODUCTION

Proteins are the essential components of every living cell. Knowing the function/s of a protein elucidates many modern age priorities like drug development, disease understanding, synthetic biochemicals design, etc. Initial solutions of computational protein function prediction [1] were low-throughput since the time and resources needed to analyze a specific gene or protein were substantial. However, currently, with the advent of high-throughput technologies, vast amounts of useful data is being produced, ranging from simple protein sequences to complex proteomic data, such as gene expression data sets and protein interaction networks (PINs). The increased quantity of data comes at a price of incompleteness especially in view of protein functions. A major challenge now arises in the development of cost effective and precise procedures for such big data analysis with the aim of uncovering their intrinsic functional meaning [2].

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

The concept of protein function is highly context-sensitive and is not clearly defined. It can be interpreted as an umbrella term for all the cellular, molecular or physiological activities that a protein is involved in. Protein functions are usually viewed in terms of notational schemes that are organized as ontology, the most prominent one being the Gene Ontology (GO) [3, 4]. GO characterizes proteins in three major aspects stored as separate ontologies within the GO: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). Each ontology consists of a set of terms (GO terms), connected to each other in a directed acyclic graph. At this moment, GO can be recognized as the most applied functional annotation scheme across a wide variety of biological data [5, 6] and as such is the scheme considered in our research.

Protein-protein interaction (PPI) data produced by high-throughput techniques are fundamental to most biological processes [7] and as such are the best choice of single source for biological data used in protein function prediction. The PPI data has the nature of networks having proteins as nodes and interactions between proteins as edges between the nodes. These networks are referred to as Protein Interaction Networks (PINs). In the PIN graph representation functions associated with a given protein are modeled as labels associated to the node corresponding to the protein. Thus, the problem of computational function prediction of a protein becomes one for proper labeling of its corresponding node in its PIN graph representation. Based on the approach being used function prediction in the PIN can be categorized as: (1) *neighborhood-based* [8, 9, 10], where protein annotations are transferred from the most "dominant" annotations among the protein's neighbors, (2) *global optimization-based* [11, 12, 13], where the neighboring proteins may not contain enough information, so the functions of the query protein are inferred from the indirectly connected proteins, (3) *clustering-based* [14, 15, 16, 17], where protein's functions are taken from the functional majority of the module (cluster) where the query protein belongs, (4) *association-based* [18], similar to clustering approaches, but here functional modules are hypothesized from frequently occurring sets of interactions in PINs of protein complexes.

Recently, a lot of research has been focused on the problem of producing network embeddings. The DeepWalk [19] method builds on the fact that short random walk sequences are similar to natural language sentences and uses Skip-Gram, a model for representing words [20], to learn a vector representation of the nodes in a graph. The LINE [21] method

first learns node representations that capture first- and second-order proximities, and in the next step concatenates the two in a final representations. Similarly, GraRep [22] defines different loss functions to capture different k-order proximities and combines the representations learned from each function. The TADW method [23] first proves that DeepWalk is equivalent to matrix factorization. Next, it uses the fact to incorporate node level rich text information for network representation learning based on DeepWalk-derived matrix factorization. In order to capture the non-linear network structure, [24] proposes a deep model with non-linear functions, also, the first- and second-order proximities are preserved. The node2vec [25] method uses a biased random walk procedure with a flexible notion of a nodes neighborhood, which efficiently explores diverse neighborhoods. As with DeepWalk, node2vec treats these walks as the equivalent of sentences and produces vectors that maximize the likelihood of preservation of the graph topology and semantics locally and globally.

In this paper the graph representation of the PIN is used to derive a continuous vector representation. The procedure proposed in node2vec is adopted, since with certain parameter settings it becomes equivalent to DeepWalk, and by transitivity incorporates the key features of other previously proposed methods. Once the vector representations of the nodes are produced any one of the well known instance based learning algorithms can be used to achieve the goal of function prediction. To that aim the learning is modeled as a set of binary classifications where a single classification corresponds to a single label (from a set of possible labels). In this single label classification the objective is to determine the existence (or non-existence) of a label to which aim the node vector representations are noted positive (if node has the label) or negative (node does not have label). In this paper experiments are performed using a highly reliable human protein interaction network. A highly reliable set of protein functions is used as the set of labels for which the single label classifications are performed. For each label the information for positive instances is apriori known and we make certain assumptions concerning the negative instances. Classification is done using two well known algorithms, namely, k-nearest neighbors and support vector machines. Comparison is done between the performances of different vector extraction settings and different classifiers used. Results are also used to draw conclusions on certain improvements that can boost the overall performance of this approach.

The rest of the paper is organized as follows. Section 2 presents the steps in acquiring and building the data for the research and the technical details on the methods used. In the third section a detailed description of the performed experiments and the corresponding results is given. Discussion for the results and possible improvements of the method is also provided. Finally, the paper is concluded in the fourth section.

II. MATERIALS AND METHODS

This section presents the steps in acquiring and building the data for the research. Additionally, technical details on the methods used are also provided.

A. Research Data

1) Construction of the Protein Interaction Network

The construction of the PIN is based on PPI data from HIPPIE (v2.0) [26], which is a human PPI dataset with a normalized scoring scheme optimized to reflect the amount and quality of evidence for a given interaction. It is an integration of multiple experimental datasets. After this data is processed in a way that removes all the self-interactions, as well as zero-confidence interactions and removing all but the interactions with highest confidence score that are duplicated, it is used to construct an undirected weighted graph, where the proteins are represented as nodes and the interactions as edges, with weights equal to their confidence scores. We will focus on the largest connected component of this graph, which consists of 16,769 proteins and 277,055 protein-protein interactions.

2) Collection of protein functions

In order to tackle the task at hand, the PIN needs to be enriched i.e. describe every protein with all of its known functional annotations. To that aim the Gene Ontology (GO) annotations and protein-function data available at the STRING [27] database, are used. STRING is a biological database that contains information from numerous sources, including experimental data, computational prediction methods and literature collections. Annotations are associated to proteins with a confidence score in the range [1,5]. We consider only the annotations with confidence score higher or equal than 3.

The proteins in the previously constructed PIN are identified using the Entrez Gene IDs, hence a mapping is made for every protein using its Etrez-Id with its corresponding String-Id. According to this mapping and the data from the STRING database the PIN is augmented with its respectable function annotations and the annotations are propagated. Propagating annotations refers to transferring annotations from a more specific GO term to its broader parent terms. When a gene is annotated to a term, associations between the gene and the terms' parents are implicitly inferred. Because GO annotations to a term inherit all the properties of the ancestors of those terms, every path from any term back to its root(s) must be biologically accurate or the ontology must be revised [28].

There are 2041 protein that can not be mapped to a valid String-Id and 1148 nodes that are not annotated with any label, and these are removed from consideration.

B. Learning the Vector Representation of the PIN

In order to use instance based learning in networks, before using any supervised learning algorithm one first needs to construct a feature representations of the nodes/edges that are highly informative, discriminating and mutually independent. The node2vec algorithm provides a semi-supervised method for learning continuous vector representation for nodes in a network, that map every node to a d-dimensional feature space, while aiming at maximizing the likelihood of preserving network neighborhood of nodes. Formally, given a network $G = (V, E)$. Let $f : V \rightarrow R^d$ be the mapping to the feature vectors. For every protein node $u \in V$, a neighborhood of node u is defined with $N(u) \subset V$. Now the problem is formulated as

a maximum likelihood optimization problem with the following objective function:

$$\max \sum_{u \in V} \log \Pr(N(u) \mid f(u)) \quad (1)$$

Equation 1 maximizes the log-probability of observing a neighborhood node for a node *u*, conditioned on its vector representation, given by *f*. The advantage of this approach over other algorithms arises in its scalability, as well as flexibility to easily custom-fit the representation for detecting node dependencies based on communities they belong to, structural equivalences based on the nodes role in the network, or a mixture of both.

Let *G = (V, E)* be the graph representation for the PIN. The process of learning the representations starts by generating *r* random walks from every protein node *u* as a source, with fixed length *l*. Let c_i be the *i*-th protein in the walk and $c_0 = u$. The generation of a sequence of proteins is done using the distribution

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & (u,v) \in E \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The random walks are biased to provide a representation that captures the right mix of equivalences from the graph using the parameters *p* and *q*. Suppose we are at a protein *v*, and we have traversed there through edge (*u*, *v*) from protein *u*, the next protein *t* in the walk is decided using the transition probability π_{vt} (Fig. 1). Let w_{vt} be the weight and π_{vt} the transition probability on the edge *(v, t)* directed from *v*, than the transition probability is set to $\pi_{vt} = \alpha_{pq}(u,t) * w_{vt}$, such that

$$\alpha_{pq}(u,t) = \begin{cases} 1/p & if\ d_{ut} = 0 \\ 1 & if\ d_{ut} = 1 \\ 1/q & if\ d_{ut} = 2 \end{cases} \quad (3)$$

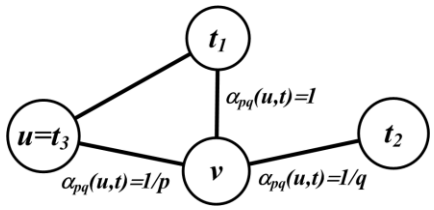


Fig. 1. Random walk transition probability ilustration

Through the parameters p, q we bias the walk in a way that enables us to control the notion of a neighborhood for a given protein node, by controlling the distance traversed from the source node in a random walk. Intuitively, having high values of p (> (max(1,q)) reduces the probability of returning to an already visited node, and therefore encourages exploration, conversely, low values of p (<(min(1,q)) result in going back a step, hence keeping the walk in a short radius. On the other hand, having a higher value for q (q > 1) makes the walk focused on nodes closer to the source and results in sample proteins within a small locality, while having lower values for q (q < 1) tend to explore interactions that lead to more distant protein nodes from the source protein.

The whole representation learning process can be summarized in three phases: preprocessing to compute transition probabilities, random walk simulations and optimization of the vector representations using stochastic gradient descent.

C. Learning the Protein Functions

The problem of learning the protein function prediction i.e. learning the appropriate labels for the PIN graph nodes is modeled as a binary classification problem, that is, whether a given protein should be labeled with a specific functional label or not. A protein that has been annotated with a GO label, is referred to as a positive sample. Although some research is being done on specifically defining negative annotations for proteins such resource is still not standardized or even available. For the purpose of this research all of the proteins that are not annotated with a GO label are referred to as negative samples for that label. Such an assumption is very general and will introduce a huge amount of error in the results because the annotation data itself can have a lot of false positive or even missing information. Additionally, the annotation propagation process introduces huge positive sets for GO labels high in the hierarchy which are very diverse. In such a scenario there is an overrepresentation of positive instances which has a negative effect on the classification. In order to showcase the performances of the proposed approach in this paper the worst case scenarios are considered. Figure 2 shows the labels for which the binary classification is performed.

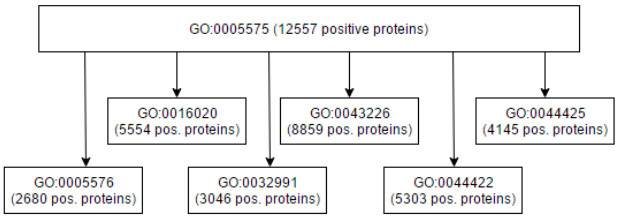


Fig. 2. The excerpt of the Gene Ontology hierarchy considered with each label associated the number of positive instances

First a classifier is built in reference to label GO:0005575. Next, the non-positive instances for the label are discarded. The former means that only the positive are further taken into consideration when building classifiers for the child labels of GO:0005575. In order to keep the classification unbiased, only labels with a positive sample ratio in the range of [0.2, 0.8] are considered, which leads to six child labels out of nine (Fig. 2). It is worth mentioning that the three child labels discarded due to the previous restriction give results similar to the ones produced for GO:0005575. This process can be continued iteratively for the child labels, until a child label that has a positive sample ratio in the targeted range no longer exists.

III. Results and Discussion

In the experiments conducted two feature representations learned from the PIN are used. For the first one the algorithm is run with parameters q = 1, p=1. The representations produced in this manner are equivalent to the DeepWalk algorithm. For the second representation q = 2, p = 1 are used, which semantically should imply that nodes with similar structural roles in the network will have similar representations. Both representations are set to produce vectors of length 128. Once the vector representations are generated, the datasets for the classification are generated. In the classification dataset a class attribute is added that is equal to 1 if the corresponding protein belongs in the positive samples for the given label, and 0 otherwise. This results in 1 dataset with 13580 instances and 6 datasets with 12557 instances, for the parent and child labels, respectively, corresponding to Fig. 2. All of the instances are 129 dimensional. The classification is done using two different supervised learning algorithms.

The first is the k-nearest neighbors algorithm (KNN) which is a non-parametric method used for classification and regression, that when used for classification assigns the object to the class most common among its k nearest neighbors. The second is the support vector machines (SVM) that are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible and new objects are classified based on which side of the gap they belong to. The algorithms were implemented in python using the scikit-learn library, where the optimal parameters where found using grid search for $k \in [10,25)$ in KNN, and for $C \in \{2^{-15}, 2^{-13}, \ldots, 2^{15}, 2^{17}\}$ in SVM, where the other parameters were left with the default values.

In the classification process we are interested in the following events: True Positive events (TP) - when a functional label is assigned to a query protein and is part of its true annotation label set, True Negative events (TN) - when a functional label is not assigned to a query protein and is not part of its true annotation label set, False Positive event (FP) - when a functional label is assigned to a query protein but is not part of its true annotation label set, False Negative event (FN) - when a functional label is not assigned to a query protein but is part of its true annotation label set. Using the counts for these events the following statistical measures can be calculated:

$$Sensitivity\ (TruePositiveRate) = \frac{TP}{TP + FN} \tag{4}$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{5}$$

Graphed as coordinate pairs, the Sensitivity and the FalsePositiveRate form the Receiver Operating Characteristic curve (or ROC curve). The ROC curve describes the performance of a model across the entire range of classification thresholds. 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 [29]. Figures 3-9 show the ROC curves for the binary classifiers for each of the 7 GO labels under consideration.

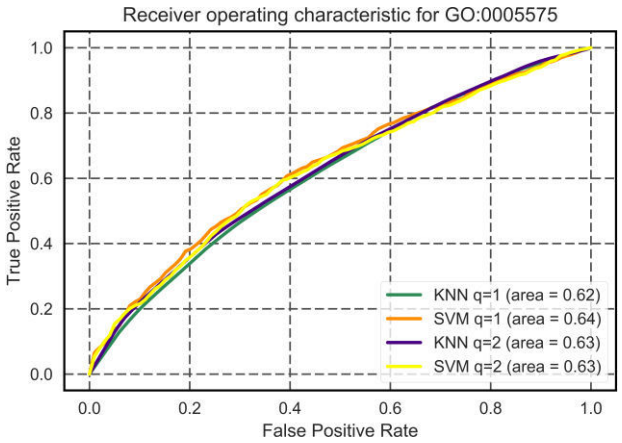


Fig. 3. ROC curve for the parent GO label GO:0005575

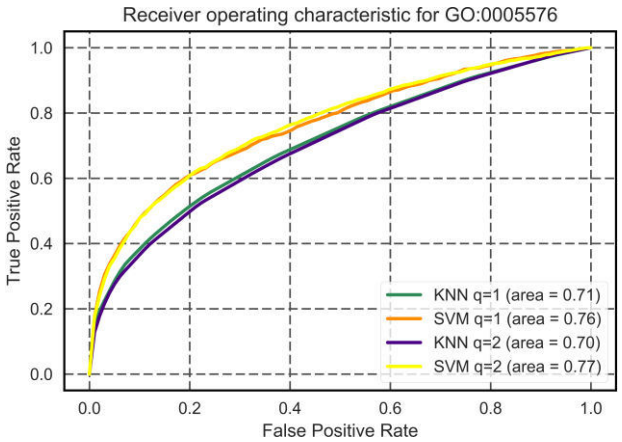


Fig. 4. ROC curve for the child GO label GO:0005576

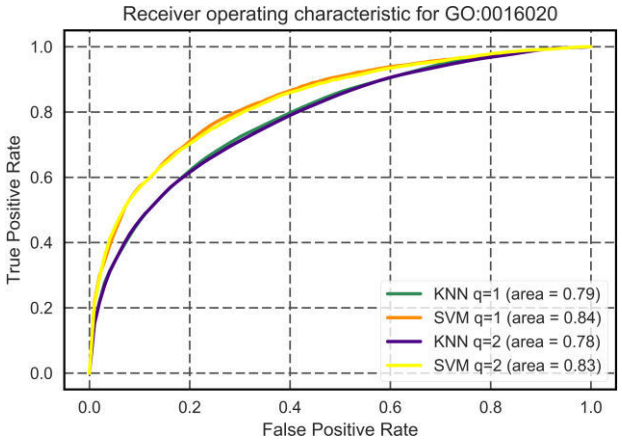


Fig. 5. ROC curve for the child GO label GO:0016020

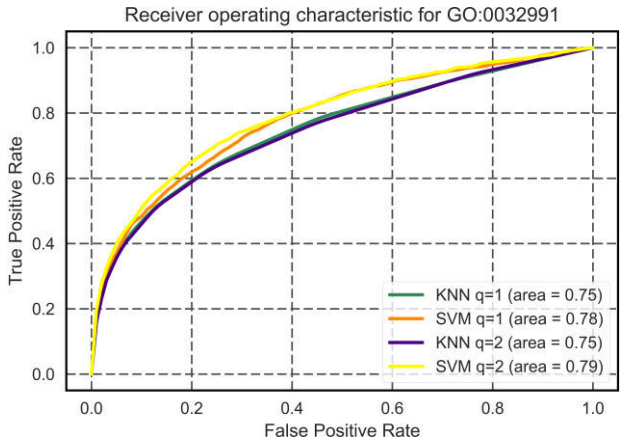


Fig. 6. ROC curve for the child GO label GO:0032991

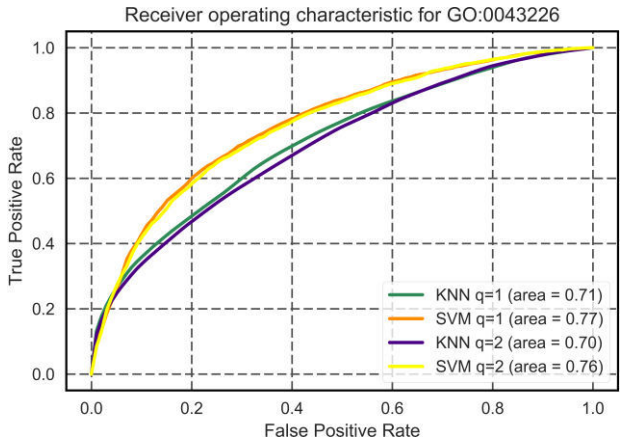


Fig. 7. ROC curve for the child GO label GO:0043226

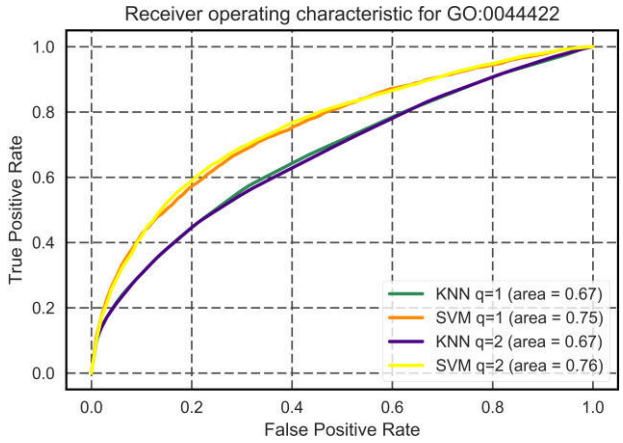


Fig. 8. ROC curve for the child GO label GO:0044422

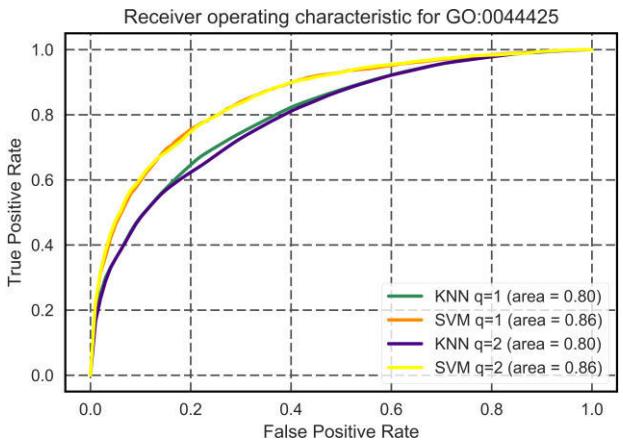


Fig. 9. ROC curve for the child GO label GO:0044425

As can be seen from the results, the SVM classifier always outperforms the KNN classifier for the protein function predictions under considerations in this research. The results are quite satisfying considering that as afore mentioned assumption that all of the proteins that have not been annotated with a particular label are considered as negative instances, which is a reason that greatly contributes for the misclassified samples. Nevertheless these experiments can serve as a proof of concept that this approach has potential to be competitive and maybe even outperform the currently used techniques in protein function prediction.

This research has shed light on the possible improvements to be made, out of which most important are the following: precise definition of negative instances in the training phase of the classification will greatly improve performance; also usage of more sophisticated classifier will contribute to improvement in the function prediction. In addition to the classification improvements, some modifications and augmentations need to be done in the vector representation learning. Namely, the presented approach takes into consideration only the topological features of the PIN graph ignoring the functional labels distribution on the nodes. The inclusion of these semantic contexts in the representation learning should improve the veracity of the produced vectors.

IV. CONCLUSSION

This paper introduces instance based learning to the problem of protein function prediction based on protein interaction networks. The graph representation of PINs is used to derive a continuous vector representation of the graph's nodes using semi-supervised learning. The graph to vector mapping uses a biased random walk procedure with a flexible notion of a nodes neighborhood, which efficiently explores diverse neighborhoods. Sequences of nodes in such walks are considered as the equivalent of sentences, thus a skip-gram learning approach is used. The procedure results in vectors that maximize the likelihood of preservation of the graph topology and semantics locally and globally. Once the vector representations of the nodes are produced instance based learning algorithms are used for function prediction. To that aim the learning is modeled as a set of binary classifications

where a single classification corresponds to a single label (from a set of possible labels). In this single label classification the objective is to determine the existence (or non-existence) of a label to which aim the node vector representations are noted positive (if node has the label) or negative (node does not have label). Experiments are performed using a highly reliable human protein interaction network. Assumptions are made for the negative instances in the classification phase, taking all proteins that are missing a functional label as being negative per that label. Classification is done using two well known algorithms, KNN and SVM. The results are in the range of expectations and are a good proof of concept that the presented approach can be very efficient in solving the computational function prediction problem. The results also indicate the potential improvements of the approach, including, but not limited to: different and more precise definition of negative instances for the classification phase, a more sophisticated learning algorithm for the binary classification, and alterations in the vector representation learning that will include not only the topological, but also the semantic context of the PIN graph.

REFERENCES

[1] R. F. Weaver, *Molecular Biology*: McGraw-Hill, 2002.

[2] I. Friedberg, "Automated protein function prediction—the genomic challenge," *Briefings in bioinformatics,* vol. 7, pp. 225-242, 2006.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry*, et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics,* vol. 25, pp. 25-29, 2000.

[4] G. O. Consortium, "Expansion of the Gene Ontology knowledgebase and resources," *Nucleic acids research,* vol. 45, pp. D331-D338, 2017.

[5] L. J. Jensen, R. Gupta, H.-H. Staerfeldt, and S. Brunak, "Prediction of human protein function according to Gene Ontology categories," *Bioinformatics,* vol. 19, pp. 635-642, 2003.

[6] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences,* vol. 100, pp. 8348-8353, 2003.

[7] L. Hakes, S. C. Lovell, S. G. Oliver, and D. L. Robertson, "Specificity in protein interactions and its relationship with sequence diversity and coevolution," *Proceedings of the National Academy of Sciences,* vol. 104, pp. 7999-8004, 2007.

[8] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein–protein interactions in yeast," *Nature biotechnology,* vol. 18, pp. 1257-1261, 2000.

[9] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi, "Assessment of prediction accuracy of protein function from protein–protein interaction data," *Yeast,* vol. 18, pp. 523-531, 2001.

[10] J. McDermott, R. Bumgarner, and R. Samudrala, "Functional annotation from predicted protein interaction networks," *Bioinformatics,* vol. 21, pp. 3217-3226, 2005.

[11] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein–protein interaction data," *Journal of Computational Biology,* vol. 10, pp. 947-960, 2003.

[12] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics,* vol. 21, pp. i302-i310, 2005.

[13] K. Trivodaliev, I. Cingovska, S. Kalajdziski, and D. Davcev, "Protein function prediction based on neighborhood profiles," in *ICT Innovations 2009*, ed: Springer Berlin Heidelberg, 2010, pp. 125-134.

[14] A. Mukhopadhyay, S. Ray, and M. De, "Detecting protein complexes in a PPI network: a gene ontology based multi-objective evolutionary approach," *Molecular BioSystems,* vol. 8, pp. 3036-3048, 2012.

[15] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Li, and B. Xu, "Protein complex prediction in large ontology attributed protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 10, pp. 729-741, 2013.

[16] K. Trivodaliev, A. Bogojeska, and L. Kocarev, "Exploring function prediction in protein interaction networks via clustering methods," *PloS one,* vol. 9, p. e99755, 2014.

[17] K. Trivodaliev, S. Kalajdziski, I. Ivanoska, B. R. Stojkoska, and L. Kocarev, "Chapter Nine-SHOPIN: Semantic Homogeneity Optimization in Protein Interaction Networks," *Advances in protein chemistry and structural biology,* vol. 101, pp. 323-349, 2015.

[18] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou, "Mining coherent dense subgraphs across massive biological networks for functional discovery," *Bioinformatics,* vol. 21, pp. i213-i221, 2005.

[19] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701-710.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[21] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1067-1077.

[22] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 891-900.

[23] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang, *"Network Representation Learning with Rich Text Information," in *IJCAI*, 2015, pp. 2111-2117.

[24] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1225-1234.

[25] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855-864.

[26] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, "HIPPIE: Integrating protein interaction networks with experiment based quality scores," *PloS one,* vol. 7, p. e31826, 2012.

[27] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic*, et al.*, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Research,* p. gkw937, 2016.

[28] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nature Reviews Genetics,* vol. 9, pp. 509-515, 2008.

[29] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, pp. 861-874, 2006.

User-Friendly Admin Panel Solution For SOHO Environments

Sasho Najdov*, Atanas Kostovski†, Elvedin Selimoski‡, Ivona Micevska§ and Pance Ribarski¶
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
Email: *najdov.sasho@students.finki.ukim.mk, †kostovski.atanas@students.finki.ukim.mk,
‡micevska.ivona@students.finki.ukim.mk, §elveselimoski@outlook.com, ¶pance.ribarski@finki.ukim.mk

***Abstract*—Administrating the tasks in a small environment can be a real problem considering that hardware and software solutions can be expensive and confusing to work with, or just overwhelming for that line of work. The need to have a certain set of tools to ease the everyday work and increase productivity becomes a real concern. Research has shown that there are not many administration panels that provide all the required services a small environment might use, as well as providing it in an user-friendly way. The purpose of this paper is to come up with an easy-to-use solution that uses limited system resources for administrating daily office tasks. The anticipated outcome is to provide non-technical people with free, open-source software to administer their daily office tasks, all the while running on non-expensive hardware.**

***Index Terms*—SOHO; Server; Administration panel; Linux; Services; Python; Flask; Angular;**

I. Introduction

With the appearance of personal computers and certain breakthroughs in communication channels, office work was possible to decentralize. This decentralization brought great productivity to office employees, and also helped with lowering the company costs in office expenses. However, the need for administration of the technology equipment is still present, or it can even be greater than before. The decentralization of office work brought burden to the IT administration because of its extended responsibilities. The burden to individuals is also present, now they have to take care of their own IT infrastructure, which is almost always out of their knowledge domain.

Technology has created a demand for more individuals to be employed from home, and it allowed some companies to outsource their work to individuals, globally. The rise of new technologies, like email, video conferencing, remote desktop software and file sync, helped for a better decentralized co-working experience. Even though these new technologies create greater accessibility for (theoretically) everyone, there is a knowledge barrier between people. Anyone that is not included in the tech industry may have problems setting-up a good working environment. In the classical workspace, there are employees that have a role to administrate and setup the environments for every employee.

The goal is to make the setup and maintenance of the required technology easy and accessible for everyone, through an administration panel that requires minute technical knowledge. The implemented solution can be installed on a Linux server, and used through a web browser in order to administrate the Linux services on the server.

A. Related work

A research was conducted on existing technologies which were in the field of administration panels. Recognized were the following products which are related work to this research:

Ajenti [1], ***Virtualmin/Webmin*** [2], ***InterWorx*** [3], ***BlueOnyx*** [4], ***Froxlor*** [5].

These products are administration panels designed to ease the managing of certain Linux services which are appropriate for SOHO environments. The research has also found out that these products are targeted at experienced users with good technical knowledge, with their main purpose focused on making their work faster, ignoring accessibility for unexperienced users. This solution differs in the basic orientation towards inexperienced users in the Linux service domain. The solution focuses on office-related services, with user-friendly control forms allowing the end-users to use professional Linux services without the need to learn the detailed management tasks. By using this administration panel, the end-users can continue their domain work and still use professional Linux services to support their everyday jobs.

B. Paper organization

The paper is organized in five chapters. This first chapter is introduction to the topic and related work. The second chapter explains the need for services that will support the Small Office/Home Office (from now on SOHO) concept. In the third chapter, the architecture of this proposed system is explained, divided in back-end and front-end sections. The fourth chapter talks about the UX/UI of the system, especially about the front-end component. Finally, the last chapter is conclusion of the paper, and the developed system.

II. The need for services

A. Which services are needed for SOHO

This environment prototype is using services handpicked by the authors, that can help with productivity. These chosen services include a few file sharing and synchronization technologies, a VPN service, and a backup service. The services were picked, as they are the most commonly used tools for any workplace. Using this environment, the user will get SMB and NFS file server, as well as, desktop and mobile synchronization

server, VPN access for road-warriors, torrent manager and media server, and incremental backup software.

B. List of services and their administration

The services chosen for this administration panel are implementing the roles mentioned in section II-A. The list of implemented services and their description is given in the following list:

- Seafile: a file synchronization system. Files are stored on a central server and can be synchronized with personal computers and mobile devices through applications. The user can:
  - Modify the server name,
  - Modify the domain,
  - Insert user email and password.
- Monitoring and logging your system can help you prevent system crashing, and increase the server availability:
  - Insert email for reporting.
- Transmission: a BitTorrent client which features a variety of user interfaces on top of a cross-platform back-end. The user can:
  - Insert database username and password.
- Transmission RSS: adds torrents from RSS feeds. The user can:
  - Insert username and password,
  - Set the update interval for transmission to check for RSS feeds,
  - Insert a list of RSS urls.
- Borg: deduplicating backup program with compression and authenticated encryption. The user can:
  - Set the name of the backup process,
  - Choose where to save the backup (locally, ftp, nfs, samba, ssh),
  - Select the files/directories to include in the backup,
  - Set the name of the Borg repository,
  - Select the allowed user to do the backups,
  - Insert the destination host, username and password, if backup is not saved locally.
- OpenVPN: an open-source software application that implements virtual private network techniques for creating secure point-to-point or site-to-site connections in routed or bridged configurations and remote access facilities. The user can:
  - Insert the username of the OpenVPN user
- Samba: a free software re-implementation of the SMB/CIFS networking protocol. The user can:
  - Insert a list of usernames and passwords,
  - Insert a list of samba shares (a full path to the share),
  - Insert a list of mappings username -¿ sharename and access rights.
- NFS: a distributed file system protocol originally developed by Sun Microsystems in 1984, allowing a user on a client computer to access files over a computer network much like local storage is accessed. The user can:
  - Insert a list of access paths.

III. Architecture of the administration panel

A. The back-end

For the back-end API, Python with Flask Framework [6] is used, and SQLAlchemy [7] for management of the simple database. The passwords are stored in the database as a bcrypt hash.

The API is configured with a rules for the endpoints, that are consistent throughout the API. Each service is represented as a Model class, and its more specific settings are represented as a Submodel classes, that are kept in a list in the main model.

Each model is kept in the database, with columns and table name as defined using the SQLAlchemy package. For each model, there is a Schema class, that is representing the JSON version of the model that will be eventually sent through the API.

The models are singletons, and the sub-models are lists of objects. Each class has a simple database table representation, where the singletons are tables with only one row, and each sub-model is a new entry.

The idea is that the backend framework is mostly used as an API provider, and its getting all the data and informations from other python and bash scripts that compute and get or set the data needed for every service. For example, in the Transmission RSS service, it would mean that the API functions will call few scripts that interact with the services provided API to use its functions.

A RESTful API provider was created, following the JSON API 1.0 protocol. The idea is that things are kept modular, so its easier to add and remove new features and services later.

The API links are defined as followed, all the API links have `/API/v1.0/` as prefix:

- `POST "/service_name/<string:action>"`
  - Do some action to service_name
- `GET "/service_name/<string:action>"`
  - Get the state of the action in service_name
- `GET "/service_name/"`
  - Get the settings and informations about the service_name
- `PUT "/service_name/"`
  - Edit the settings and informations for the service_name
- `GET "/service_name/submodel/"`
  - Get the list of submodels for the service name
- `POST "/service_name/submodel/"`
  - Insert new submodel object (instance) in service_name
- `GET "/service_name/submodel/<int:id>"`
  - Get the submodel from service_name by id (submodel.id)
- `PUT "/service_name/submodel/<int:id>"`
  - Edit the submodel by id in service_name (submodel.id)
- `DELETE "/service_name/submodel/<int:id>"`
  - Delete the submodel by id from service_name

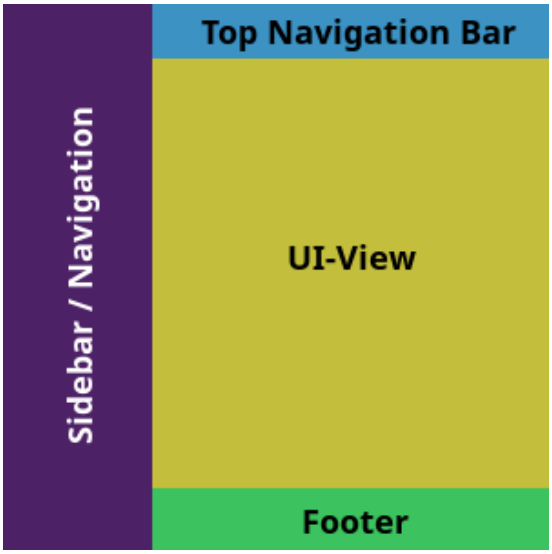


Fig. 1. Application structure

B. The front-end

The visual part of the application, gives the user the ability to manipulate with services in a familiar (point and click) manner, discarding the CLI approach where thorough knowledge of the inner workings of the services may be required. In this section, the different components used are explained in designing the front-end interface and how they work together to deliver the final experience (see Fig. 1).

The focal point of this section will be AngularJS [8] and how the application is structured with it. Although many other JavaScript libraries and plug-ins are used, they do not contribute to the functionality of the application, but to user experience only, and as such, they will be mentioned only briefly.

The backbone of the application is consisted of four components:

- Top Navigation Bar,
- Sidebar / Navigation,
- Footer,
- UI-View.

The first three parts (views) are static templates with dynamically injected content and the last is a dynamically loaded view. Their purpose is described in the following sections:

1) Top Navigation Bar: This view is more of a user-oriented tool bar, with helper tools to better find yourself in the application. It always resides on top of the window and consists of:

- A button to toggle the navigation view,
- A search input box for searching services and settings,
- Notification area where alerts/messages from services are gathered,
- A login/logout button for starting/ending a user session.

2) Sidebar / Navigation: No application is useful unless you can quickly and easily navigate through it. The sidebar gives you the ability to switch between different services just by clicking on the displayed service name. The possible locations the user can navigate to are:

- The Dashboard - A view with accumulated information from all services along with info about the hosting system(s). Information can be displayed in a tabular form, charts, status icons or just plain text,
- The Services - A per-service view is designed, where the list of services are displayed as a second-level navigation link of the Services section. Each service link leads to a state transition to the corresponding service state, with its controller, service and view,
- Other - This section can is used for all other settings and options that can be done in the administration panel such as account management, theme management and social links.

3) Footer: Of course every application is not complete without a footer that displays about, contact and other useful information and links.

4) The UI-View: The fourth part of the application is the UI-View and is of different kind that the other three parts. This is an AngularJS directive where a view is dynamically injected, for the corresponding service. When the application is loaded at first, it shows the dashboard as a default view, then by using the navigation, different services can be loaded in the view.

The content of this view depends on the state that the user is currently in. Not only does the view depend on the state, but its Angular controller and implicitly the used services do, too.

One controller is an exception to the rule, and that is the MainController. This controller is always present, it controls the index.html which, as previously mentioned, glued all components together, and is explicitly injected into the view. Being an always-present controller, it has other special functions, such as: controlling the user session (saving/clearing user credentials), authenticating a user on restricted pages, displaying user data and more.

All other controllers, services, and view are injected into the UI-View component and comply to the following nomenclature:

The Controller:

- Name: ServiceNameController
- Dependant services: serviceNameServices
- Mandatory properties: serviceName, serviceDescription, serviceIsOn, serviceIsEnabled, serviceSettings
- Mandatory functions: serviceTurnOn, serviceTurnOff, serviceEnable, serviceDisable, getServiceSettings, toggleSettings

The Service:

- Name: serviceNameService
- Dependencies: $resource
- Mandatory properties: req (*requests service settings from the back-end API*)
- Mandatory functions: getServiceSettings

The View:

- Name: servicename.view.html
- Header: serviceName, serviceDescription, serviceActions, breadcrumbs
- Body: service dependent details, service dependent settings

The State:

- Name: service.servicename
- Controller: ServiceNameController
- ControllerAs: vm
- TemplateUrl: servicename.view.html
- LoginRequired: state dependent, can be on of: *USER_ROLES.*{*all, guest, user, editor, admin*}

C. Other Controllers, Services, Views & States

This subsection covers the components that are left and did not belong in the previous sections. Some of these only serve as helper tools such as the *AuthenticationService* and the *ServiceController*.

- LoginController
  - Dependant services: *AuthenticationService, UserService*
  - View: *login.view.html*
  - State: *login*
- RegisterController
  - Dependant services: *AuthenticationService, UserService*
  - View: *register.view.html*
  - State: *register*
- ServiceController
  - Dependant services: *SystemService*
  - View: *service.view.html*
  - State: *service* $\langle abstract \rangle$
- authService
  - Used to authenticate a user's credentials
  - Authorizes a user's permission to certain resources
  - It uses the *session service* check the state of the current user
- userService
  - Returns information about the currently active user
- dataService
  - Returns service settings about the requested service name
- sessionService
  - Stores information about the current user, using a cookie
  - Provides information such as: current user, is the user authenticated, is the user authorized
  - Creates and destroys the user session on the front-end
- user-authInterceptor
  - Intercepts an HTTP request, and if a user is logged in i.e. authenticated, it appends the token in the "Authorization" header of the request
- OTHER
  - Broadcast events are used to handle user logging in and logging out
  - Broadcast events are used to handle the authorization of a user

D. Application Deployment

Concerning deployment, *NPM* [9] is used, along with *Bower* [10] to install required packages, while *Gulp* [11] is used to deploy the application to the server. Using *Gulp*, all required application resources are concatenated into: application scripts (*app.js*), application styles (*app.css*), used libraries and plugin scripts (*lib.js*), used library styles (*lib.css*).

Another useful option that comes with *Gulp* is *livereload*, where any change in the source code is immediately detected, "compiled" and deployed, such that the application is always up-to-date.

IV. UX/UI

This section will cover the general work-flow of the administration panel, concerning how to access the panel, how to login, how to view service details and change them. After that the results of a questionnaire will shown from users after using the application. The dashboard is displayed in Fig. 2.

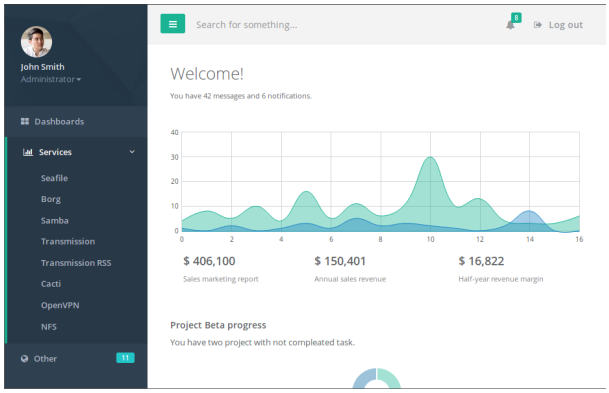


Fig. 2. Screenshot from the application dashboard

A. Application Work-flow

When accessing the application through a browser, the first page that will be displayed is *The Dashboard (no login is required to access this)*. On the left-hand side is *The Navigation* where the user can choose to go to one of the available services. Unauthenticated users, trying to access a service, will first be redirected to *The Login View*, and after *successful* authentication, they will be redirected back to the selected *Service View*.

The Login View is consisted of a login form, where *username* and *password* are submitted. Unregistered users can easily create an account by clicking on *Create an account*, which will lead them to *The Register View*. From there they will be able to create an account using the register form.

What will be displayed in *The Service View* is service-dependent. The general structure is:

- Service title
- Service description
- Breadcrumbs
- Service actions area - start/stop, enable/disable, settings
- A tab to display corresponding service information in tabular form, charts or status icons
- A tab to display/edit the settings of that service

B. User Questionnaire

A questionnaire was done on a small group of people, both from technical and non-technical background. All of the technical persons had learned the panel's structure in a fast pace, and were quickly able to manipulate with the offered services, managing them in all of the available ways from the purposed panel. On the other hand, non-technical persons had a little struggle figuring out the service's manipulation, but were satisfied of the intuitive control forms, to which they were already familiar. As seen in Fig. 3 below, the gathered information shows that on average, the users are satisfied with the administration panel. The most notable dislike was that this panel is still not complete, and that is why the next steps would be expanding the possibilities of this panel, adding new features and services.

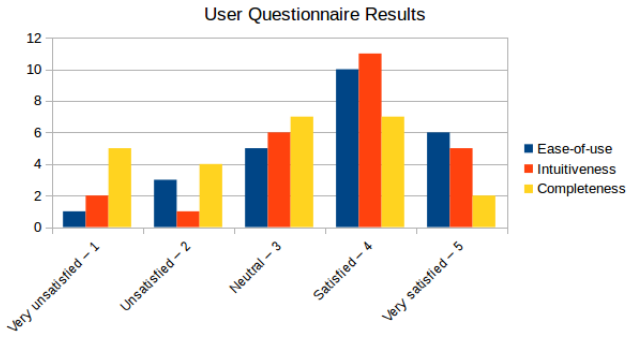


Fig. 3. Results of the questionnaire

V. Conclusion

With the increasing number of people that work from home or in small offices, there is a need for an environment that will be accessible, easy to setup and configure. The tools included, are great for boosting productivity of anyone, but are not easy to install for less tech experienced people.

An administration panel was designed that is targeted for non-technical users. More experienced users might find it useful too, but they might want even more customizations and control.

The future of this project is focused on making it possible to pick the initial services and tools, and install new ones. That way this environment would be easily customizable and adaptable for every users needs. It would be implemented with usage of repository with the pre-packed installation packages, so the simplicity of administration is kept from the users viewpoint. Additional features such as multi-language support and scaling for larger organizations are the future of this project.

References

[1] "Ajenti - admin panel," http://ajenti.org/, accessed: 2017-03-07.
[2] "Virtualmin - a webmin module," http://www.webmin.com/virtualmin.html, accessed: 2017-03-07.
[3] "Interworx - a smart, scalable and reliable web hosting panel," http://www.interworx.com/, accessed: 2017-03-07.
[4] "Blueonyx - internet hosting platform," https://www.blueonyx.it/, accessed: 2017-03-07.
[5] "Froxlor - server management panel," https://www.froxlor.org/, accessed: 2017-03-07.
[6] "The flask framework," http://flask.pocoo.org/, accessed: 2017-03-07.
[7] "Sqlalchemy - the python sql toolkit and object relational mapper," https://www.sqlalchemy.org/, accessed: 2017-03-07.
[8] "Angularjs - javascript library," https://angularjs.org/, accessed: 2017-03-07.
[9] "Npm - package manager for javascript," https://www.npmjs.com/, accessed: 2017-03-07.
[10] "Bower - a package manager for the web," https://bower.io/, accessed: 2017-03-07.
[11] "Gulp - automation toolkit," http://gulpjs.com/, accessed: 2017-03-07.

Affordable Server Solution For SOHO Environments

Atanas Kostovski[‡], Sasho Najdov[§], Ivona Micevska[*], Elve Selimoski[†], and Pance Ribarski[¶]
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
Email: [‡]kostovski.atanas@students.finki.ukim.mk, [§]najdov.sasho@students.finki.ukim.mk,
[*]micevska.ivona@students.finki.ukim.mk, [†]elveselimoski@outlook.com, [¶]pance.ribarski@finki.ukim.mk

***Abstract*—SOHO LANs have evolved today into inexpensive home and office networks that support a variety of home and office applications. Home applications include the ubiquitous disk and printer sharing as well as sharing high-speed Internet access. Business applications add Web serving and telephony applications. This paper studies a solution for a distributed system with infrastructure office services using ARM architectures. The provided services will include: file synchronization/sharing, logging, backup, RSS, VPN as well as infrastructure system services. The expected outcome of this paper is to provide an overview of an affordable server solution for Small Office/Home Office (SOHO) Environments.**

***Index Terms*—SOHO, Server, Small Bussiness, ARM, Linux, Services**

I. Introduction

The number of people who work in SOHO(Small Office Home Office) environments is rising. Along with it, the need for more productivity tools rises too. Teams that work on same projects like to share files securely, managers want access to the employees work, employees have the need to print and share documents, and most important everyone wants to be connected on high-speed network or stay anonymous on the internet. Our goal is to bring the productivity tools available by using a home server to the regular user, with each service needed one click away.

A. Related work

There are commercial products that offer this level of service and all those systems have their advantages and disadvantages. We want to make cheap and yet powerful system that is easy to use. We have decided to make SOHO network because we want to support people who work or have businesses in this environment. In the end, the users should be able to store, sync, access, collaborate and backup their data securely and with high performance and reliability, and at the same time the system will provide monitoring of application logs, log files, event logs, service logs, and system logs. The current tools that are being used for management of servers, are mostly, if not all, targeted for the advanced user. The last referred project, Daplie, is the closest to our goals, but that project provides only cloud as it's main feature.

- cPanel - https://cpanel.com/
- Plesk - https://www.plesk.cloud/
- Ajenti - http://ajenti.org/
- Daplie - https://daplie.com/

What our project is doing differently is that we are targeting the regular user, and we want to provide many features and add-ons that the user will be able to pick and install or remove. For the users that aren't sure what services they want to pick and install, few handpicked services are provided by default. In the next chapter we will explain the services and architecture of the system.

B. Organization of the paper

The paper is divided in five sections. This first section gives introduction to the topic and related work. The second section gives a list of services that fit on a typical SOHO server. In the third section we give intro to ARM SoC's and the Banana Pi board that we use for a server. The installation and usage of the chosen services is explained in section four. Finally, in section five, we present the conclusion of our SOHO server paper.

II. SOHO list of services

Every SOHO network needs various services. Network sharing, file sharing, printer sharing, data synchronization, virtual private network for staying anonymous online and protection, ability to backup data, all of these services and others are necessary for a small office to work properly.

As we mentioned earlier, there are default handpicked services that include:

- Seafile
- Monit
- Borg
- Transmission+rss
- OpenVPN
- Samba

A. File sync & sharing

The software Seafile [1] is a self-hosted file sync and share solution with high performance and reliability. We chose Seafile as the cloud storage solution, as the project is maintained, and user friendly. Everyone that has used Dropbox or similar cloud services, will find the features by Seafile familiar.

The project provides web interface, multi-platform file sync tools and applications that include even mobile platforms.

B. Logging/monitoring

Monit [2] is a very useful tool for managing and monitoring processes, files, directories, permissions and filesystems on a UNIX system. It will conduct automatic maintenance, repair and execute meaningful casual actions in error situations.

We chose Monit as our default monitoring tool. Every system needs proper monitoring, and Monit covers pretty much everything that is needed in this project. It can provide us an easy implementation of basic monitoring for the user, and notification system when the system reaches critical moments.

C. Incremental backup

BorgBackup (short: Borg) [3] is a deduplicating backup program. Optionally, it supports compression and authenticated encryption. The main goal of Borg is to provide an efficient and secure way to backup data.

This provides the best features we can ask for. Borg is using less space for storing the backups because of the incremental way of storing data, and it provides encryption of the storage, so the user's data is stored securely.

We chose backup as one of the preinstalled services, as it might be the single most important one. Because our product offers storage, the last thing that users want to happen is data loss.

Our default setup provides automatic daily backup. In the first release, we don't provide a rollback to backup feature, but the user will be able to mount the older backup and take whatever is needed from the backup. All of the actions and configurations are done from the administration panel.

D. Torrent client

Transmission [4] is a light-weight and cross-platform BitTorrent client. Transmission has CLI component and web component. Transmission-rss [5] is a useful addition to Transmission, it has the ability to monitor RSS feeds and automatically add torrent links.

We mainly chose this two features, to show that our product might be used for more purposes than only data storage. With few services like this two, it's possible to turn the product in a media box too.

E. VPN for road-warriors

OpenVPN [6] is an open-source software application that implements virtual private network (VPN) techniques for creating secure point-to-point or site-to-site connections in routed or bridged configurations and remote access facilities.

Since this is offered for professional businesses, VPN is required by many offices, it's included in the default setup. OpenVPN is a free, well maintained and supported package.

F. Windows file sharing

Samba is a software re-implementation of the SMB/CIFS networking protocol, that provides file and print services for various Microsoft Windows clients and can integrate with a Microsoft Windows Server domain.

We included Samba as a service, because most of the people are using Windows as their office desktop OS, and Samba provides nice integration of our product with a Windows client.

III. ARM Architecture and servers' inter-connectivity

Our product is based on Banana Pi [7] board, with Arch ARM as an operating system [8]. The Banana Pi is a series of credit card-sized single-board computers. Its hardware design was influenced by Raspberry Pi in 2013. Banana Pi software is compatible with Raspberry Pi boards.

It uses the Allwinner SoC (system on chip) and as such is mostly covered by the linux-sunxi port. Banana Pi is the open source hardware and software platform which is designed to assist lemaker.org [9].

Arch Linux ARM is a port of Arch Linux for ARM processors. Its design philosophy is "simplicity and full control to the end user", and like its parent operating system Arch Linux, aims to be very Unix-like. This goal of minimalism and complete user control, however, can make Arch Linux difficult for Linux beginners as it requires more knowledge and responsibility for the operating system.

Arch Linux was picked as the platform, because of the simplicity it offers. One of the main positive things it offers is that it comes streamlined, and the user has to install everything needed. This way, there is less redundant software that makes everything run slower on the low performance board.

IV. Implementation

For the implementation, we chose to use Docker [10]. Docker provides a way to implement and test the services, with ability to extend with new containers with ease.

Each service is contained in one Docker container. The implementation and configuration is manually done and tested for the project's platform, so the problems are minimized.

This way, each service can be abstracted as a package, and it's possible to implement a web store which provides the user a freedom to pick and download new extensions.

V. Conclusion

Small Office / Home Office LANs are an essential part of every small business that aims to achieve and maintain its business objectives. In the context of this paper, we introduced an expressive server solution that will, at the same time, be affordable and accessible.

Having a local server in the office is a great productivity boost. But not all offices, especially the smaller ones, can afford buying and paying for an expert to configure and maintain it.

The goal is to provide a solution, that is both affordable and accessible at the same time. Using this project, a small office or a freelancer working from home, can just plug in and pick the needed features to get the productivity boost.

The huge difference between using a local server and online cloud solution like Dropbox or Google Drive for document sharing, is that this way all the data is stored on the local

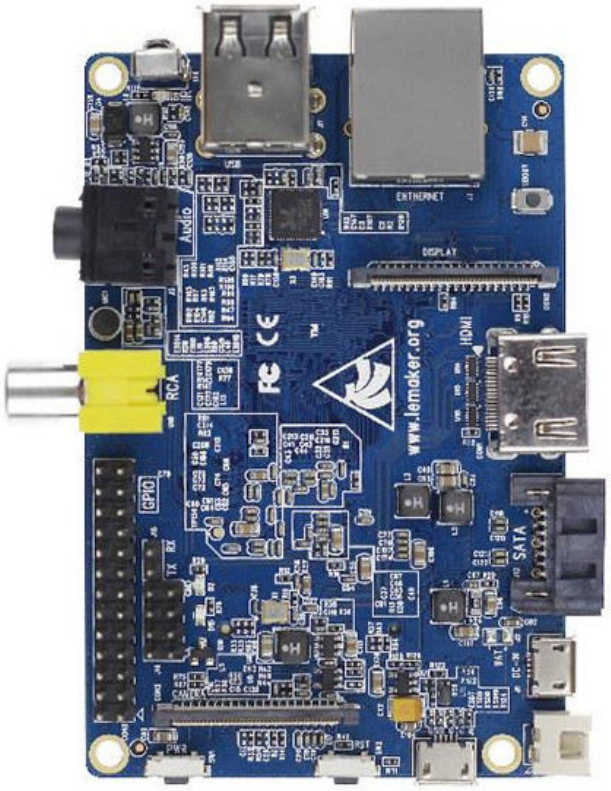


Fig. 1. The Banana Pi ARM SoC board. source[1]

machines, and never leaves the local network providing more security and privacy.

Also, we demonstrated various services and their applications, which are necessary for a small office to work properly.

All of the currently installed services are selected carefully and with a particular purpose in mind. There is a possible extensibility using the repository where the user will be able to download and install other services.

This solution can be immensely applicable in the future, because of the affordable price and the goal of user friendliness.

REFERENCES

[1] "Enterprise file sync and share platform with high reliability and performance," https://www.seafile.com/en/home/, accessed: 2017-03-07.
[2] "Monit - open source utility for managing and monitoring unix systems," https://mmonit.com/monit/, accessed: 2017-03-07.
[3] "Deduplicating backup program with compression and authenticated encryption," https://github.com/borgbackup/borg, accessed: 2017-03-07.
[4] "A fast, easy, and free bittorrent client," https://transmissionbt.com/, accessed: 2017-03-07.
[5] "Adds torrents from rss feeds to transmission web frontend," https://github.com/nning/transmission-rss, accessed: 2017-03-07.
[6] "Openvpn - open-source vpn over tcp or udp," https://openvpn.net/index.php/open-source.html, accessed: 2017-03-07.
[7] "A20 arm soc board with arm cortex a7 dual core cpu," http://www.lemaker.org/product-bananapi-index.html, accessed: 2017-03-07.
[8] "A port of arch linux, which aims for simplicity and full control to the end user," https://archlinuxarm.org/, accessed: 2017-03-07.
[9] "Lemaker arm board," http://www.lemaker.org/, accessed: 2017-03-07.
[10] "Docker is the world's leading software container platform," https://www.docker.com/, accessed: 2017-03-07.

[1]Source: LeMaker http://www.lemaker.org/product-bananapi-index.html

Implementing Easy Prepaid Card Programme Infrastructure

Darko Gjorgjiev* and Pance Ribarski§
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
Email: *gjorgjiev.darko@gmail.com, §pance.ribarski@finki.ukim.mk

***Abstract*—Debit cards or prepaid cards are an easy way to lower the complexity of payments in everyday bussineses. Besides payments, they can give the customer conveniency of keeping knowledge about spent and remaining credits, very similar to credit card reports. To the bussiness owner, the debit cards are means to securing loyal customers base and statistics into what is being purchased. This paper explains the implementation of a web and mobile application which enables easy debit card programme for bussinesses and their clients. The infrastructure is consisted of Python backend API and web application and Android frontend clients. The bussiness owners register Android devices as terminals which can process prepaid cards in the form of QR code card or rfid card. All the operations from the Android terminal are funneled to the Python backend where the central database resides. The clients can use the web application to check the balance on their prepaid cards. The implemented infrastructure gives easy tool for bussiness owners to use the power of prepaid cards.**

***Index Terms*—prepaid card, python, flask, api, android, qr code, rfid**

I. Introduction

Debit cards are plastic payment cards which banks issue to be used instead of cash. This is very convenient for the customers, especially when every store offers the possibility to use the debit card instead of cash. This paper introduces bussines-personal debit card programme which can be used with specific services that each bussines owner can have. The implemented system allows bussines owners to issue their own debit cards to their customers and define sets of services that they offer. When the customer uses a service from the bussines owners, they can easily pay with the issued debit card. The bussines owners have full insight into the selling statistics of their services, and the customers also can view the history of their payments and the services they purchased. This platform is very easy to use from the bussines owner point of view because the Point-Of-Sale (POS) terminal is an Android application which can be run on any Android phone or tablet. The debit card can be made of any material on which QR code can be printed. The debit card can also be an RFID card which will be read by an RFID reader on the POS device.

A. Related work

The debit card programme we propose in this paper is very similar to prepaid club cards with prepaid funds or gift cards. In Macedonia, cards of this type that we found are: Verna, M-Karta, Cineplexx Bonus, Skopska and SunWireless. Euro-Standard Bank offers prepaid gift card and many companies offer loyalty cards for gathering points without offering the prepaid funds possibility. The Verna prepaid card [1] is issued by the company Makpetrol and is used on their gas stations in Macedonia. The M-Karta [2] is a prepaid smart card issued by the Skopje municipality centar and used for the parkomats on their parking lots. The Cineplexx Bonus Card [3] is issued by the Cineplexx cinema and doubles as prepaid card and loyalty card. The city of Skopje's pulic transport company issued the Skopska card [4] which is a prepaid card for the public bus transport in Skopje. The EuroStandard Gift Card [5] is a debit card from EuroStandard Bank and is intended as money transfer in form of a gift. The common ground of all these cards is that they are purchased at the company stores, and they are spent at the company stores or company kiosks. They can't be recharged online are isolated and work only for the specified stores. The solution we are proposing is a generic debit card which can be used across bussineses. The card can be recharged online and can be reused across bussineses that are using the same platform. Our solution can even accept cards (2D-barcode, QR-code or RFID) from other companies.

B. Paper organization

The paper is organized as followed...

II. Architecture

The proposed system is divided in two parts: the backend together with the website and the POS terminal which is a mobile application. The API backend and the website are actually a python application which uses the Flask framework [6]. For the abstraction of the data model we are using SQLAlchemy [7] which is Object Relational Mapping library.

A. The website

Th web application is designed to support three roles of users:

- Administrator of the system,
- Bussines owners,
- Client that uses debit cards.

1) Administrator: This role is designated to the System Administrator and is only used for administration of the system. The user with this role has all the privileges in the system and can manage other users in the database (see Fig. 1).

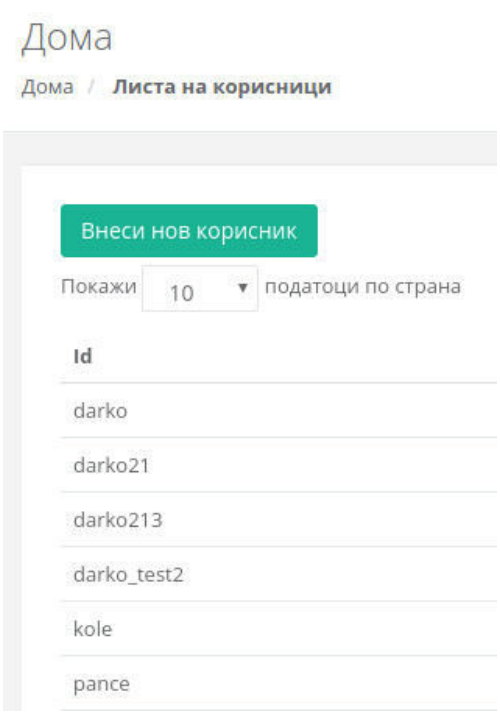


Fig. 1. The administration page for editing users of the system

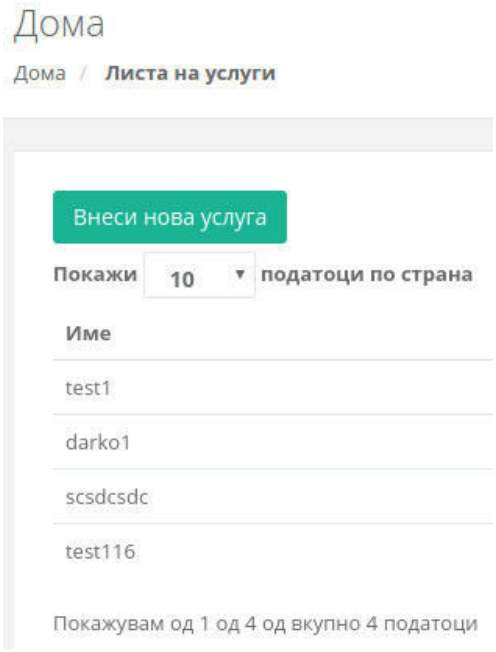


Fig. 2. The page with services for Business owners

2) Business owner: This role is for users which want to create Business profile and start using debit cards on the system. Users can register for this role on the website and wait for the Administrator to approve their registration. Once approved, the Business owner can start creating services for the created business (see Fig. **??**). They can also register mobile applications as point-of-sale terminals with the registered business and through them they can associate debit cards with their business. On the website, the Bussines owner can check the debit card transactions for their services. The Business owner can also manually recharge debit cards that are associated to their business.

3) Debit card client: The Debit card clients are the end-users of the system. They are the owners of the debit cards and can register their debit card with businesses. Once associated, the Debit card clients can spend funds from the debit card on the services of the business on the POS terminals. The clients can recharge their debit cards online on the website, or the cards can be manually recharged through the Business owners.

B. The API backend

The API backend in the Flask application is the backbone of the mobile applications for the POS terminals. It is consisted of a set of RESTful endpoints which support the mobile application. All the endpoints have a prefix `/api/v1.0/` and follow the JSON API specification [8].

The list of RESTful endpoints is:

- `GET "/services/"`
  - Get list of services
- `POST "/transactions/"`
  - Make transaction for a service with client's card
- `PUT "/transactions/<int:id>"`
  - Edit commited transaction
- `PUT "/transactions/<int:id>"`
  - Delete commited transaction
- `POST "/recharge/"`
  - Manually recharge client's card
- `POST "/terminals/"`
  - Register POS terminal

C. The POS terminal

The proposed system uses mobile application for the Android OS which acts as POS terminal. This allows transforming any Android device, be it phone or tablet, into a POS terminal. The mobile application allows login to registered Business owner accounts (see Fig. 3).

After the login, the Business owner can register the POS terminal to their account on the system. When registered, the application shows the Home screen (see Fig. 4). From here, the Business owner can start using the application as POS terminal for the system.

When the Business owner wants to process a payment, they have to read-out the Client's card and proceed with the desired service from the business (see Fig. 6).

After the service is chosen and the Client's card is read, the Business owner can confirm the payment **??**.

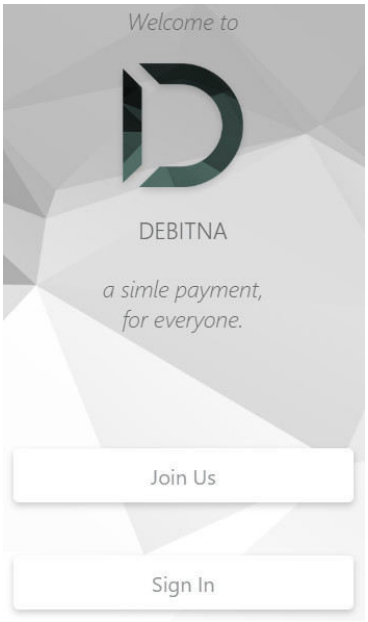


Fig. 3. The welcome screen of the mobile application

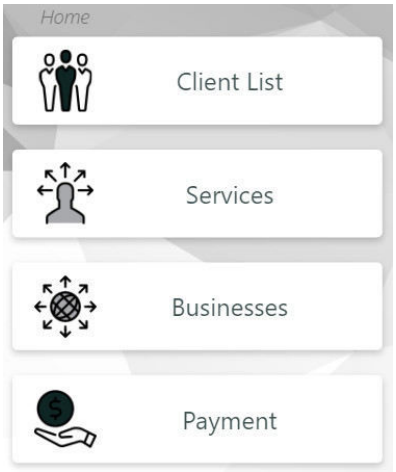


Fig. 4. The home screen of the mobile application

III. Deployment

The proposed system is deployed on a Linux server using Nginx as reverse proxy server and uWSGI as Python server for the Flask application. Currently it is bound with the domain http://debitna.mk and it is in testing phase. For the database we are using PostgreSQL which is managed by the SQLAlchemy ORM library. The system is designed as multi-tenant, meaning that many businesses can use the same deployment of the application. If needed, the application can be deployed on specific customer's premises for single-tenant scenario. In this situation, the mobile application should be informed with the new DNS name or the IP address of the Flask application server.

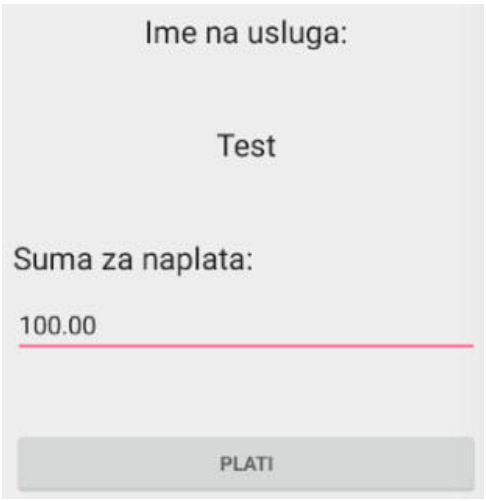


Fig. 5. The payment screen of the mobile application

Fig. 6. The payment confirmation screen of the mobile application

IV. Conclusion

The proposed system for debit card infrastructure solves the problem of incorporating prepaid or loyalty cards in everyday businesses. This system enables easy integration of debit cards for the end-user clients of the businesses. The system is multi-tenant, enabling many businesses to use the application, create their services and register POS terminals. With the POS terminals

References

[1] "Verna loyalty card," http://www.verna.mk, accessed: 2017-03-07.

[2] “M-karta prepaid smart card,” http://www.poc.mk/index.php?option=com_content&view=article&id=216&Itemid=334, accessed: 2017-03-07.
[3] “Cineplexx bonus card,” http://www.cineplexx.mk/klub/, accessed: 2017-03-07.
[4] “Skopska city bus card,” http://www.skopska.mk, accessed: 2017-03-07.
[5] “Eurostandard gift card (e-pay),” https://e-pay.mk/, accessed: 2017-03-07.
[6] “The flask framework,” http://flask.pocoo.org/, accessed: 2017-03-07.
[7] “Sqlalchemy - the python sql toolkit and object relational mapper,” https://www.sqlalchemy.org/, accessed: 2017-03-07.
[8] “Json api specification 1.0,” http://jsonapi.org/format/, accessed: 2017-03-07.

Robot Tracker Based On Semantic Segmentation Computer Vision Algorithms

Aleksandar Jovanov, student
Faculty of Computer Science and Engineering
University of Sts. Cyril and Methodius,
Skopje, Macedonia
jovanov.aleksandar.1@students.finki.ukim.mk

***Abstract*—This project presents an approach to human tracking based on a semantic segmentation algorithm which uses convolutional neural networks and optical flow based on Farneback's method. Tracking is implemented via person discrimination using Neural Networks and Speeded Up Robust Features as an initial step and optical flow tracking in next steps if movement is present. Actual robot actuation, localization and central planning are done by the move_base package of the Robot Operating System navigation stack and ROSARIA package with an intermediary node that transforms image data to goal coordinates. Microsoft Kinect and p3dx robot are the hardware components used in the realization of the project.**

Keywords: *tracking; semantic segmentation; convolutional networks; machine vision; robotics; robot operating system*

I. INTRODUCTION

There has been a lot of work in human detection and tracking, including both human body/face detection (Histogram of Oriented Gradients[1] and Haar-Viola Cascade Classifiers[2]), in optical flow methods (Horn-Schunck[3], Lucas-Kanade[4], Farneback[5]) and in feature descriptors like Speeded Up Robust Features[6]. Additionally convolutional neural networks like the ones studied in LeCun[7] and FCN[8] are of great importance in this work. Of central importance is the work from CRFasRNN[9].

Histogram of oriented gradients (HOG) is a method of detecting people with a multiscale sliding window (rectangle) approach under which HOG features are extracted and which are classified as human/not human by a Support Vector Machine.

Haar-Viola is a method similar to HOG sliding window that uses a different set of features for classification and comprises a more complicated classification procedure.

Optical flow based on brightness constancy equation (as used in this project) can not be solved without extra assumptions and constraints. Horn-Schunck method is an algorithm used to estimate the optical flow by adding an extra constraints for smoothness which which tries to minimize the distortions in flow in the whole image.

Lucas - Kanade is a very popular method for approximating optical flow that uses the concept of a k by k window around the pixel of interest that adds additional constraint which claims that optical flow is constant in that window. Such an assumption allows the system of equations for optical flow to be solved by the least squares criterion. Lucas - Kanade with pyramids is an extension that provides better results for larger images which calculates the optical flow for every pyramid level as a function of the next pyramid level and a correction step.

Farneback optical flow is a method developed by Gunar Farneback that is based on polynomial expansion of a pixel neighborhood and parametrized displacement fields.

Speeded Up Robust Features (SURF) is a key point detector and feature descriptor that uses an approximation to the hessian matrix for detection and Haar-wavelet responses for description.

Convolutional Neural Networks are a special kind of Artificial Neural Networks that combine a set of convolution, pooling and fully connected layers. Fully connected layers are characterized by the fact that every neuron in layer n is connected to every other neuron in layer n + 1. Pooling layers serve as a fast method of reducing the number of neurons. Convolutional layers differ from fully connected ones by not having a connection between each neuron pair in layers n and n + 1, instead focusing on a tiling approach to connection distribution that in the end appears to be a filter that can be learned, i.e. a convolution kernel.

Recurrent Neural Networks (RNN) are a special kind of Artificial Neural Networks that model temporal dependencies with a system of cells that meld input at time t, output from time t-1 and memory of past events and produce some output.

Conditional Random Fields (CRF) are an extension to dynamic Bayesian networks that opt to model the conditional $p(\mathbf{y}|\mathbf{x})$ distribution instead of the joint $p(\mathbf{y}, \mathbf{x})$. Inference in a CRF is a process which establishes the most likely **y** given **x**.

Conditional Random Field as Recurrent Neural Network (CRFasRNN) is a neural network based on convolution, pooling and softmax layers and its special CRFasRNN layer.

Softmax layer applies the softmax function which squashes an input vector so that all its components become members of the [0, 1] interval and all components sum to 1.

Pooling layers reduce the dimensionality of the input image by allowing only some pixels to proceed further into the network based on some criterion.

CRFasRNN layer is a type of neural network layer that models the CRF inference process as a recurrent neural network. It does so in two interlinked parts. The first part is the mean field iteration cell that consists of five steps based on convolutional and softmax layers plus an unary potential addition layer (result from previous part of CRFasRNN network). The second part is the recurrent cell based on the mean field iteration that unrolls in T steps (in [9] the authors claim that T=5 usually gives a good enough answer).

Fully Convolutional Network (FCN) is a Convolutional Neural Network that generates an image at output of same dimensions as provided in input where the output image is a segmentation mask. It does so by an up sampling process based on a deconvolution layer.

Robot operating system (ROS) is a system for robot control. ROS Node is an application that performs some kind of service. ROS Topic is a channel over which nodes communicate via messages. TF transform is a linear transform that relates two coordinate frames and is integrated inside ROS.

Segmentation mask is a binary image that contains 1 for pixels that are part of humans and 0 for everything else.

Tools used in this project are: Robot Operating System[10], ROSARIA[11], OpenCV[12], Caffe[13].

II. System Architecture

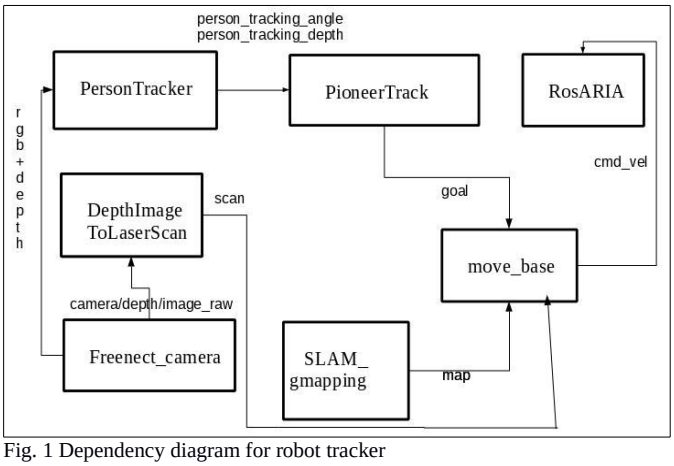


Fig. 1 Dependency diagram for robot tracker

The system consists of three main parts, the remote server that hosts the CRFasRNN neural network (not pictured here), the tracker component (PersonTracker and PioneerTrack node) and other supporting ROS nodes. The tracker is also explained as a separate entity.

A. Remote server

This is a computing device that has a NVIDIA Titan X graphics card that allows it to do fast neural network feed forward operations that generate the segmentation mask.

The CRFasRNN network hosted here takes a 500x500 BGR image and provides a 3D tensor with 20 entries per pixel that signify the probabilities of class membership. The argmax operation then generates a 2D image with values 0,1,..19 where 0 is background. That matrix is transformed so as to keep only the pixels with value 15 (class human) and set them to 255 and set all other pixels to value 0.

This new image is the semantic segmentation mask that is sent to the robot for further processing after a lossless compression step. Access to this computer was via ssh for this project. The network used in this project is previously trained by the original authors of CRFasRNN[9] and comes as a free download in form of caffe layer definition file and caffe model weights file. The network is not specifically trained for human detection, but that did not stop it from giving excellent results which is why a variant trained specifically for such a task was not considered.

B. Tracker

This component starts tracking when there is a single person in an image, continues doing so as long as there are one or more persons in the image and stops tracking when all humans are lost in an image. This is formalized in the following pseudo code (Fig. 2).

```
Input: rgb_image, is_tracking, global
tracking descriptors

m <- getMask rgb_image, is_tracking
c <- getClustersWithDescriptors m

if (c.clusters.length == 0)
      is_tracking <- false
else if (c.clusters.length == 1 and
is_tracking == true)
      is_tracking <- true
      tracking descriptors <-
c.descriptors[0]
else if (c.clusters.length > 1)
      ms <- findMostSimilar c, tracking
descriptors
      if (ms is not null)
            is_tracking <- true
            tracking descriptors <- ms
      else
            is_tracking <- false
            tracking descriptors <- null
else
      is_tracking <- false
      tracking descriptors <- null

return: mask that coincides with
tracking descriptors
```

Fig. 2 Pseudo code for main tracking algorithm

getMask is a method that decides between tracking via dense optical flow or reaching out to remote server to get the mask from the neural network. The general idea is to track via optical flow as long as you can because it is faster. Unfortunately that fails when there is no movement or only slight movement.

getClustersWithDescriptors partitions the input mask into clusters using the connected components algorithm and assigns a matrix of feature descriptors to each using the SURF algorithm.

findMostSimilar takes the clusters and their descriptors, takes the tracking descriptors and finds the one that is the closest to tracking descriptors where the closeness criterion is sum of distances of key point matches. It can return null as a result if the minimum distance found is above some threshold. This is necessary because slight movement is insufficient so there is a need of knowing when to stop using it and go get more accurate data from remote server.

C. Data publisher node

The tracker algorithm gives as a result a binary image, from now on labeled t_m. This image is used to extract two pieces of information which are then sent to the robot controller, the aforementioned move base component of Robot Operating System. The information extracted from t_m is relative orientation of tracked person w.r.t to Kinect camera and distance to it.

The formulas used for angle (α) are:

$$\lambda = 54.0 * (\pi / 180.0) \quad (1)$$

$$\alpha = \lambda * ((c_x - m_x) / w) \quad (2)$$

m_x is the half width of the input image, w is the width of the input image, c_x is the center x of tracked person contour generated from t_m.
The number 54 is the horizontal field of view of the Kinect device used in this project.

The formula for depth is:

I = depth image provided by kinect. (4)
I' = image and of I and t_m (5)
d = mean of all pixel values in I' that are not zero. (6)

These values are then published on person_tracking_angle and person_tracking_depth topics as Float64 values. Not a Number values are published if there is no person in an image.

D. PioneerTrack helper node

The following formulas are used to provide the data which is used for actuation:

$$g_x = p_x + d * \cos(\alpha) \quad (7)$$
$$g_y = p_y + d * \sin(\alpha) \quad (8)$$
$$g_\theta = p_\theta + \alpha \quad (9)$$

d is depth, α is angle, **p** is current pose (x, y, θ) shared via map -> odom tf transform.

The values (g_x, g_y, g_θ) are generated by PioneerTrack node and are sent on the goal topic to move_base, which then plans the trajectory. **NB**: d is actually shortened a little amount because people are not points and have a radius of volume. move_base should handle collisions with tracked people in and by itself but this is an extra layer of safety.

The tracker robot's behavior can be succinctly compacted to the following definition. Continuously build a map around yourself, localize in it and plans trajectories to goal (supplied by the new custom developed PioneerTrack node).

III. CONCLUSION

The main drawback of this approach to tracking is the amount of time it takes to provide results. It takes around 1 second to do inference with CRFasRNN network on a high end NVIDIA Titan X. It takes around 2 more seconds to transfer this data from remote server to robot (dependent on network speed). Actual tracking via SURF descriptor matching takes up to a 1 second depending on the number of clusters (i.e people returned in mask).Thus getting tracking information takes around 4 seconds, if the remote server is contacted. It works considerably faster when the remote server is not contacted and instead dense optical flow is used, around 0.3 seconds per frame on an Intel Core i3 4005U processor.

Quality wise this method works really well because the CRFasRNN neural network provides near pixel perfect image segmentation results and errors are very rare.

Compared to local methods of human detection tried previously this method is slower, but provides higher quality results. The methods tried before CRFasRNN are the HOG algorithm and Haar-Viola algorithm. Lucas-Kanade method with key points tracking was also tried but quickly dismissed because of inability to separate tracked person from other persons in image.

ACKNOWLEDGMENT

This project was realized with help from the OpenCV, ROS & Caffe communities and the robots provided by FCSE, Skopje and with aid from professors Andrea Kulakov and Petre Lameski.

REFERENCES

[1] Navneet Dalal and Bill Triggs. Histogram of oriented gradients for human detection. 2005.
[2] Paul Viola and Michael J. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. IEEE CVPR, 2001.
[3] Berthold K.P. Horn and Brian G. Schunck. Determining Optical Flow. Artificial Intelligence, 17, pp. 185-203, 1981
[4] Lucas, B., and Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679.
[5] Farneback, G. "Two-Frame Motion Estimation Based on Polynomial Expansion." Proceedings of the 13th Scandinavian Conference on Image Analysis. Gothenburg, Sweden, 2003.
[6] Bay, H. and Tuytelaars, T. and Van Gool, L. "SURF: Speeded Up Robust Features", 9th European Conference on Computer Vision, 2006
[7] Yann LeCun, Léon Bottou, Yoshua Bengio, Pattrick Haffner "Gradient-Based Learning Applied to Document Recognition", PROC. OF THE IEEE, November 1998

[8] Jonathan Long, Evan Shelhamer, Trevor Darrell “Fully Convolutional Networks for Semantic Segmentation”, CVPR, 2015
[9] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, Philip H. S. Torr, “Conditional Random Fields as Recurrent Neural Networks”, IEEE ICCV, 2015
[10] http://www.ros.org/
[11] http://wiki.ros.org/ROSARIA
[12] http://opencv.org/
[13] http://caffe.berkeleyvision.org/

Analysis and Comparison of School Me & Sitos Six, two Learning Management Platforms in Secondary Education

Berat Jashari[1], Florinda Imeri[2], Agon Memeti[3]
Department of Informatics, Faculty of Math and Natural Sciences[1,2,3]
University of Tetova
Tetova, Macedonia
berat-m@msn.com[1], {florinda.imeri, agon.memeti}@unite.edu.mk[2,3]

***Abstract*—To meet the need for flexible and interactive learning process, the government of Kosovo gave instructions to develop and implement online learning platforms, as tools that will facilitate content creation as well to support the teacher-student, student-student, and teacher-teacher interactions. The purpose of this paper is to report a study on the evaluation of two online learning platforms which are currently used at two secondary school in Kosovo. In order to evaluate their use and the impact in the education process we have conducted a survey, where we have analyzed the teachers' and students' viewpoint in terms of difficulties and benefits while using them. The platforms are names School Me and Sitos Six. This study describes and includes concrete recommendations to guide users of these platforms.)**

Keywords—lms; platform; sitos six; school me.

I. INTRODUCTION

Online learning delivery tools are designed to support a full range of teaching and learning activities conducted by educational institutions. With the rapid advances in technology, several online learning tools come onto the stage. The use of electronic platforms in our country as part of a secondary education process has also gone through different stages.

Research shows that using LMSs possesses numerous benefits for teaching and learning. It enables schools to shift the focus from content-based learning to process-based learning[1] and helps to "facilitate change from passive to active learning" [2]. Using LMSs also has the potential to increase student enrollment [3] and to promote interaction between students and school members [4].

The focus of this paper are two learning management platforms, School Me and Sitos Six platforms, which are implemented and being used in two secondary schools in Kosovo. Schools were faced with the challenge of the support and in deciding which tool to use, but it is seemed that once a school has implemented a LMS, there has been little or no significant research conducted to see the effect on students, teachers and schools (either positive or negative) of the system after implementation.

This research has evolved out of a desire to understand what issues exist for teachers, students and administrators of two secondary schools, when implementing and using a LMS. By researching a school that has recently implemented and is currently using a LMS, this research aims to provide some insight into the factors which work well and those which limit LMS use in schools.

Having introduced the research paper in section. I, section. II reviews the literature relevant to this study, with an overview of Information and Communication Technology ICT and Learning Management Systems (LMS). Sec III presents the research methodology used, sec IV presents the study's findings organized under each of the research questions. These findings are analyzed and discussed in section V, where the conclusions drawn from the findings are presented as well.

II. ICT AND LMS

As instruction and learning gain new dimensions in today's world due to the proliferation of information and communication technologies (ICTs) - multimedia, the Internet or the Web, as a medium to enhance instruction or as a replacement for other media-, education becomes independent of time and space. Consequentially, learners and instructors can utilize new modes of learning and communication due to the proliferation of information and communication technologies (ICTs).

Many governments have, developed plans to intensify their investments regarding ICT in education. The quick rise of the Internet and worldwide web (WWW) have led to the adoption of objectives to equip all schools with access to these facilities in a relatively short period of time

Use ICT has brought a fundamental turn in the teaching culture, by making knowledge accessible for every student. The appliance of ICT in teaching, provides great benefits for students, it increases their conceiving and perceiving abilities/skills during the class while facilitating the teaching process. ICT brings dynamism in the teaching/learning process and puts students in control of their teaching, allowing independent development progress. Despite the apparent benefits of the use of ICT for educational purpose, studies

showed that in many cases, the learning potential of ICT is deprived as many teachers are still not fully ICT literate and do not use it in their teaching. Studies on teachers' readiness for ICT generally, suggest that there is still a long way to go before schools in the region will be able to take full advantage of the opportunities provided by 21st century technology [5].

Education in the past was usually a matter of uni-directional transfer of information: from the teacher to the student. The main pedagogical approach was whole class teaching. Many argue that new pedagogical models need to be explored in order to prepare future citizens for lifelong learning

As Brown, 2004; Kortecamp & Croninger, 1996; Tearle, 2003, refer, in education, commercial technology is being integrated with promises of transforming learning, often without specific information on the effects of each of the technologies once implemented. One such technology currently being implemented into schools are LMS.

LMS (referred to as Virtual Learning Environments, Digital Learning Environments, Course Management Systems or Electronic Learning Environments) are web based applications, running on a server and accessible with a web browser from any location with an Internet connection. LMS presents educators with the following functionalities: tools for the administrative support of learning processes; the facilitation of communication processes between school board, teachers, students and parents; electronic support of learning processes (knowledge collaboration, contact sessions, feedback module) and the design and implementation of course material [6].

Although LMS originated in the late nineties of the previous century and despite their high adoption rate in higher education [7] and later in secondary education [8], little is known about the technology acceptance of LMS [9]; about how LMS influence learning [10]; how the use of LMS is related with teachers' and students' perceptions about teaching and learning [11]; learning outcomes resulting from the use of an LMS, and about teachers' motivation and training for using the LMS [12].

It is known that educational innovations usually do not succeed if teachers are not provided with the skills and knowledge needed to carry them out. Training teachers is a very expensive activity and hence, often much neglected in that there are big differences between schools, since they depend on the level of their budget and/or investments, as well as on the way of managing the income.

There are many research done within this field of study and from their conclusion it is seemed that if LMS use is professionally managed, the benefits are numerous in order to raise the level of capabilities at different level of education.

III. Empirical Analysis

The purpose of this study is to report a study on the evaluation of two online learning platforms which are currently used at two secondary schools; we have prepared a questionnaire which is designed with predefined answers and some open questions for general reflection. We have used mixed study that integrates both qualitative and quantitative studies. In order to have clear understanding at some points some of the data are transformed from qualitative into quantitative presentations.

As regarding the sampling frame and size of our survey, the data presented in this paper are answers taken only from students, teaching staff, and support staff (administrators and managers).

Below are presented research questions of our survey.

- RQ1. Which is the usage level of the School Me and Sitos Six platforms from teachers and students in secondary schools?
- RQ2. Have School Me and Sitos Six platforms affected the improvement of teaching and learning*?*

3.1. Presentation of results

Below we will present the results of the survey, the statistical data, based on answers, which we consider are relevant to our research topics. Data are analyzed using Statistical Package for the Social Sciences (SPSS).

We have done a correlation among teachers and students comparing their answers by showing the positive and negative aspects of the learning platform they use.

The result are separated in different tables, working based on the description between different factors that influence toward learning and teaching through their online learning platforms.

A. Teacher's results

Firstly, the idea was whether both (students and teachers) are using the platforms, especially in the schools or mostly at home, and the results are shown below in the table 1.

TABLE I. Platform Usage

		Sum of Squares	df	Mean Square	F	Sig
Do you use the platforms at School?	**Between Groups**	.292	1	.292	.312	.587
	Within Groups	11.208	12	.934		
	Total	11.500	13			
Do you use the platforms at home?	**Between Groups**	.054	1	0.54	.048	.830
	Within Groups	13.375	12	1.115		
	Total	13.429	13			

Based on the analysis of the first question whether they you use the platform in your house or school, results shown that their usage is not on the proper level and it is statistically not significant.

In order to analyze the impact of these platforms toward learning and teaching, especially the difference between them, the next step as looking for positive and negative aspects of their usage toward learning process.

TABLE II. PLATFORM'S POSITIVE/NEGATIVE VIEWPOINTS

		Sum of Squares	df	Mean Square	F	Sig
Positive aspects: easy to use	**Between Groups**	.292	1	.292	.265	.616
	Within Groups	13.208	12	1.101		
	Total	13.500	13			
Positive aspects: can be accessed from home	**Between Groups**	6.428	1	6.428	2.561	.136
	Within Groups	30.375	12	2.531		
	Total	36.857	13			
Positive aspects: it is convenient	**Between Groups**	1.339	1	1.339	1.627	.226
	Within Groups	9.875	12	.823		
	Total	11.214	13			
Positive aspects: enables me to learn more	**Between Groups**	.482	1	.482	.735	.408
	Within Groups	7.875	12	.656		
	Total	8.357	13			
Negative aspects: does not function well	**Between Groups**	36.214	1	36.214	1.046	.327
	Within Groups	415.500	12	34.825		
	Total	451.714	13			
Negative aspects: has problems with regis.	**Between Groups**	.149	1	.149	.095	.763
	Within Groups	18.708	12	1.559		
	Total	18.857	13			
Negative aspects: takes too much time	**Between Groups**	1.524	1	1.524	1.055	.325
	Within Groups	17.333	12	1.444		
	Total	18.857	13			
Negative aspects: it is not convenient	**Between Groups**	1.929	1	1.929	1.714	.215
	Within Groups	13.500	12	1.125		
	Total	15.429	13			

Results shows that generally, they do not positively see the use of platform "Sitos Six" toward learning and teaching process based on the fact that as there are a quite enough number of the respondents, which address the different platform problems that affects negatively into learning process.

B. Student's results

The same analysis like done with teachers, are done also with the students and below are presented separately the results.

TABLE III. PLATFORM USAGE

		Sum of Squares	df	Mean Square	F	Sig
Do you use the platforms at School?	**Between Groups**	.084	1	.084	.420	.521
	Within Groups	7.818	39	.200		
	Total	7.902	40			
Do you use the platforms at home?	**Between Groups**	.146	1	.146	.196	.660
	Within Groups	28.976	39	.743		
	Total	29.122	40			

Student analysis shows that, they are more prone and interested using both the platforms.

TABLE IV. PLATFORM'S POSITIVE/NEGATIVE VIEWPOINTS

		Sum of Squares	df	Mean Square	F	Sig
Positive aspects: easy to use	**Between Groups**	.113	1	.113	.053	.820
	Within Groups	83.789	39	2.148		
	Total	83.902	40			
Positive aspects: can be accessed from home	**Between Groups**	1.726	1	1.726	1.113	.298
	Within Groups	60.469	39	1.550		
	Total	62.195	40			
Positive aspects: it is convenient	**Between Groups**	3.446	1	3.446	2.022	.163
	Within Groups	66.457	39	1.704		
	Total	69.902	40			
Positive aspects: enables me to learn more	**Between Groups**	3.618	1	3.618	1.952	.170
	Within Groups	72.285	39	1.853		
	Total	75.902	40			
Negative aspects: does not function well	**Between Groups**	29.166	1	29.166	20.083	.000
	Within Groups	56.639	39	1.452		
	Total	85.805	40			
Negative aspects: has problems with regis.	**Between Groups**	9.618	1	9.618	5.326	.026
	Within Groups	70.431	39	1.806		
	Total	80.049	40			
Negative aspects: takes too much time	**Between Groups**	1.032	1	1.032	.464	.500
	Within Groups	86.773	39	2.225		
	Total	87.805	40			
Negative aspects: it is not convenient	**Between Groups**	18.038	1	18.038	7.609	.009
	Within Groups	92.450	39	2.371		
	Total	11.488	40			

Based on the results (positive and negative viewpoints), students are interested on their usage but also they respond that platforms are inappropriate for usage still they are facing with different problems (such: registration and lack of mobility). They respond that the registration point as well as the time it takes time and the platforms do not have conditions for everyday usage.

IV. CONCLUSION

We believe that this study has been very useful to view the correlation between the usages of both learning platforms in secondary schools. In general the empirical data collected in this study represents an overview of the current situation among LM Platform usage in secondary schools.

Based on the given statistical (empirical) results taken from teachers and students we can conclude that platforms need to be efficient in a manner of having a comfortable environment; offering sufficient internet connectivity: access online courses simultaneously, considering infrastructural bandwidth upgrades to increase network access speed; supplying computer accessories and communications technologies; considering a range of communications options; and especially offering adequate technical support (being adequate to keep the computers functioning properly and to solve problems that students and facilitators cannot).

In addition, we recommend that there must be done an additional training for all parties of the school, such as, staff, teachers, students, in order to use these the two platforms in an appropriate manner. Training should be focused on the way of platforms appliance, to fit with the curriculum, and to ensure access for all, or thinking about using additional online learning platforms (such as: Google Classrooms or Edmodo) as they are designed to help users save time, keep classes organized, and improve communication with students.

REFERENCES

[1] Vogel, D. And Klassen, J, "Technology-supported learning: status, issues and trends", Journal of Computer Assisted Learning, 17, 2, 104–114, 2001.

[2] Herse, P. and Lee, A, "Optometry and WebCT: a student survey of the value of web-based learning environments in optometric education", Clinical and Experimental Optometry, 88, 1, 46–52, 2005.

[3] Nunes, M. B. and McPherson, M, "Action research in continuing professional distance education", Journal of Computer Assisted Learning, 19, 4, 429–437, 2003.

[4] Lonn, S. and Teasley, S. D, "Saving time or innovating practice: investigating perceptions and uses of Learning Management Systems", Computers & Education, 53, 3, 686–694, 2009.

[5] Ya'acob et al, "Implementation of the Malaysian Smart School: An Investigation of Teaching-Learning Practices and Teacher-Student Readiness", Internet Journal of e-Language Learning & Teaching, 2(2), pp. 16-25, 2005.

[6] De Smet, C., and Schellens T, "ELO's in het Vlaams secundair onderwijs: Nieuw of alweer achterhaald", [LMS in Flemish secondary education: New or outdated], Advies & Educatie, 26(5), 12–14, 2009.

[7] Kember, D. et al, "Understanding the ways in which design features of educational websites impact upon student learning outcomes in blended learning environments", Computers & Education, 2010.

[8] Pynoo, B. Et al, "Predicting secondary school teachers' acceptance and use of a digital learning environment: A cross-sectional study. Computers in Human Behavior", 27(1), 568-575, 2011.

[9] Sánchez, R. A. and Hueros, A. D, "Motivational factors that influence the acceptance of Moodle using TAM", Computers in human behavior, 26(6), 1632-1640, 2010.

[10] Koszalka, T. A. and Ganesan, R, "Designing online courses: A taxonomy to guide strategic use of features available in course management systems (CMS) in distance education. Distance Education", 25, 243–256, 2004.

[11] Joo, Y. J et al, "Online university students' satisfaction and persistence: Examining perceived level of presence, usefulness and ease of use as predictors in a structural model", Computers & education, 57(2), 2011.

[12] Keramati, A., et al, "The role of readiness factors in elearning outcomes: An empirical study", Computers & Education, 57(3), 1919-1929, 2011.

Author Index

ISBN 978-608-4699-07-1