ISSN 2029-249X
ISSN 2029-4824 (elektroninis leidinys)

VYTAUTO DIDŽIOJO UNIVERSITETAS
KAUNO TECHNOLOGIJOS UNIVERSITETAS
VILNIAUS UNIVERSITETO KAUNO HUMANITARINIS FAKULTETAS

INFORMACINĖS TECHNOLOGIJOS

XXIV TARPUNIVERSITETINĖS TARPTAUTINĖS MAGISTRANTŲ IR DOKTORANTŲ KONFERENCIJOS „INFORMACINĖ VISUOMENĖ IR UNIVERSITETINĖS STUDIJOS" (IVUS 2019) MEDŽIAGA

2019 M. BALANDŽIO 25 D., KAUNAS, LIETUVA

2019
VYTAUTO DIDŽIOJO UNIVERSITETAS

ISSN 2029-249X
ISSN 2029-4824 (electronic issue)

VYTAUTAS MAGNUS UNIVERSITY
KAUNAS UNIVERSITY OF TECHNOLOGY
VILNIUS UNIVERSITY KAUNAS FACULTY OF HUMANITIES

INFORMATION TECHNOLOGY

Proceedings of the XXIV international master and phd students conference "Information Society and University Studies" (IVUS 2019)

25 April 2019, Kaunas, Lithuania

2019
VYTAUTAS MAGNUS UNIVERSITY

TURINYS / CONTENT

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Composition of the Information Security Methods for a Smart Environment and the Research

Nerijus Šatkauskas
Department of Computer Sciences
Kaunas University of Technology
Kaunas, Lithuania
nerijus.satkauskas@ktu.edu

***Abstract*—Smart devices and the smart environment itself is getting more and more popular. A big part of smart devices uses the Android operating system. Since any information on these devices can become available to the third parties on the basis of granted permissions, it is very important to consider it properly before granting them. A permission monitoring system prototype has been proposed for this purpose.**

Keywords—dangerous permission group, dangerous permission, information leakage, android operating system, smart environment, smart device, information value, information sensitivity, Android permissions, permission monitoring

I. INTRODUCTION

Smart environment is rather an abstract conception and it may refer to a number of more specific areas in question. If we referred to one of many definitions for the smart environment, it would sound like [1] "ordinary environments equipped with visual and audio sensing systems, pervasive devices, sensors and networks that can perceive and react to people…". It is expected that the number of such devices will only increase in the future.

One of the smart devices which makes a big part of the smart environment is a smartphone. A dominating operating system currently is Android [2]. This operating system has been created by Google on the basis of Linux. The operating system due to its nature of being an open source one has to be well controlled and maintained in order to keep it as safe as possible.

The purpose of this research is to analyze security issues the Android operating system faces. It assesses the security of the smart environment information storage in the Android operating system. It attempts to detect whether any unauthorized parties can get an access to this information. The methods which may strengthen the security are considered.

A prototype has been proposed for this purpose. This prototype shall classify the tested applications based on their permissions which suggest any potential information leakage. The results will be compared with some other applications which are currently available on the Play Store for the same purpose.

II. SMART ENVIRONMENT THREATS

Mobile devices once were considered as safe ones but everything has changed as soon as operating systems were introduced. Installing an application is not only an additional comfort. It can be an additional concern as well. Especially if it is a malware which can leak any information.

IoT environment or the smart environment in this particular case since the issues are rather common can be divided into three main levels [3]: application level, transportation level and perception level. All these three levels bear threats which are typical to them.

TABLE I. SMART ENVIRONMENT THREAT LEVELS

Layer	Main Threats
Application level – provides customer requested services like air temperature	Data leakage: stealing data
	DoS attacks: making services unavailable
	Malicious code injection: exploiting known vulnerabilities
Transportation level – transmits and receives any collected data	Routing attacks: intermediate malicious nodes
	DoS attacks: making nodes unavailable
	Data transit attacks: attacks in networks
Perception level – physical sensors to collect any data and to process it	Physical attacks: node tempering, replacing
	Impersonation: fake identity for attacks
	DoS attacks: making nodes unavailable
	Routing attacks: intermediate malicious nodes
	Data transit attacks: sniffing, man-in-the-middle

This research focuses on the application level. The operating system Android is picked due to its leading positions in the market.

III. ANALYSIS OF THE CURRENTLY AVAILABLE ANDROID DATA LEAKAGE MONITORING TOOLS

Data availability to third parties in the Android operating system relies on the permission model [4]. Permissions are such labels which should be assigned by developers to their application. The application must define in the manifest file which sensitive resources it needs to have an access to. The user during the installation has a chance either to grant these permissions or not.

A. Preinstalled permission manager

As Android 6.0 "Marshmallow" has been introduced in 2015, the ability was provided to toggle any granted dangerous permission groups for any specific application [5]. The accessibility of this tool may vary depending on the manufacturer of a device, but it can be accessed in general via Settings > Apps / Application Manager > Permissions.

A screenshot is provided below of the operating system Android 8.1.0. It gives an access to the list of all the installed applications. Dangerous permission groups can be reviewed, granted or revoked at any time.

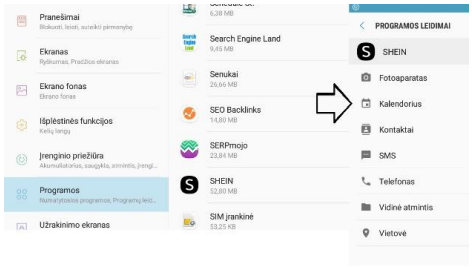


Fig. 1. Preinstalled permission manager

B. *Application Inspector*

A good alternative which is available on Play Store for the permission management is Application Inspector. This is a third-party application which is developed by UBQSoft.

The tool once it is launched provides a list of all the installed applications. One can see more details after picking any particular application within that list concerning libraries, last update time etc. Involved permissions are described as well as their level they belong to is provided: dangerous, normal, signature. The status of granted or not granted is available which can be changed after tapping and being directed to relevant Settings submenus.

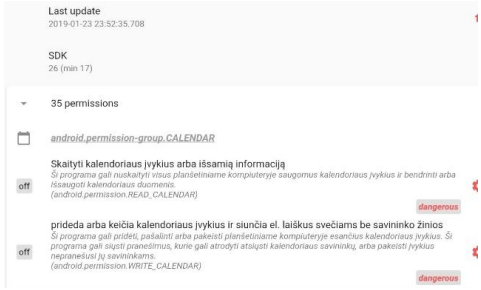


Fig. 2. Application Inspector

C. *Apk Analyzer*

It is a very extensive analyzer and it provides an access to different statistical data after a specific application is picked within a general scanned applications list. There is a tab for used permissions. These permissions are listed after tapping the tab, but the information resources are very limited. There are no descriptions about these permissions. Which level they belong to is undefined. There is no information if any of these permissions in the manifest file are granted or not.

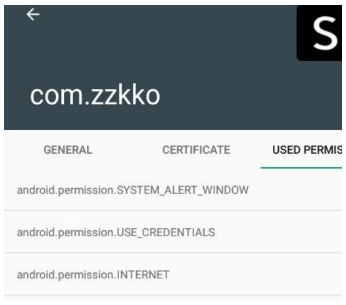


Fig. 3. Apk Analyzer

D. *PackageInfo*

Another application which can be helpful for scanning any installed applications on the device is PackageInfo. It gives a list of applications after scanning which are available for a more detailed review after picking any of them. It gives some package information, including the list of permissions. There are no detailed descriptions of these permissions. The state whether they are granted or not is unidentified.

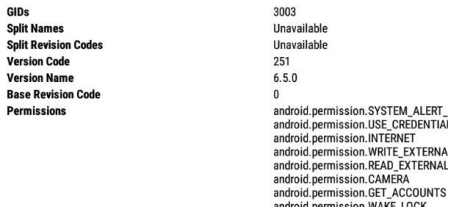


Fig. 4. PackageInfo

It becomes obvious after the analysis of some currently available tools for permission scanning and monitoring that the focus given on the permissions may not be enough for a regular user. A regular user may not want to search for any explanatory information about the granted permissions in external sources. It may lead the user to underestimating any potential threat due to personal information leakage.

IV. V-S AXIS INFORMATION SENSITIVITY ASSESSMENT

Different data classification methods were taken into consideration but V-S method [6] was chosen as the most appropriate one in this case. This method classifies any available information based on 2 axis which stand for information value and sensitivity. As the authors suggest who have introduced this method it is possible to assign the data to different information classes while implementing different security measures.

A. *V-S axis method in the prototype*

In order to able to use the proposed V-S axis method for the data on Android device, we first need to define the value of the vertical axis for information **sensitivity (Y)**. Sensitivity axis has tree levels: low (0), middle (1), and high (2).

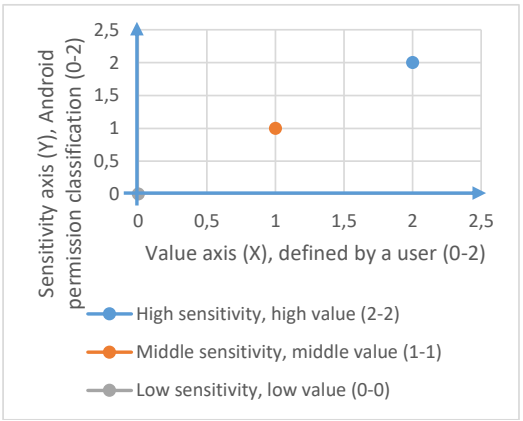


Fig. 5. V-S axis chart

The horizontal axis for information **value (X)** has also three levels. The levels are correspondingly: low (0), middle (1), and high (2).

The official classification of permissions available on Android developers portal was used for that purpose [7]. Permissions are classified there into four groups: normal, dangerous, signature, and special ones. This official classification reflects different information sensitivity levels to any potential information leakage. These permissions were assigned to the **sensitivity (Y) axis** in the following manner:

1) *Low (0):* Normal permissions are assigned to this level due their low potential threat. These permissions are granted to any installed application on a smart device without any intervention on the user side.
2) *Middle (1):* Some normal permissions are assigned to this level. Applications with these permissions may cause some inconvenience to users like CHANGE_NETWORK_STATE which allow to change the connectivity to wireless networks.
3) *High (2):* Dangerous permissions groups were assigned to this level. It is officially confirmed and classified as having negative impact once the information which belongs to the above class is unintentionally exposed to any third part parties.

Signature permissions and special permissions were not further considered in this research. Therefore, they were not assigned to any axis level.

The **horizontal (X) axis** for information value is used for a personal assessment of the information stored on the smart device. The values for this axis are selected by default in the prototype but a user can change them any time.

1) *Low (0):* This information is not valuable to the user or the user will not have any significant issues upon losing it. Permissions of low sensitivity (Y) axis level are matched to this value (X) axis level by default which results in 0 as a score.
2) *Middle (1):* This information might have some value to the user or the user might have some issues upon losing it. Permissions of middle sensitivity (Y) axis level are matched to this value (X) axis level by default which results in 1 as a score.
3) *High (2):* This information is valuable to the user. Losing it might cause considerable issues or financial losses. Permissions of high sensitivity (Y) axis level are matched to this value (X) axis level by default which results in 4 as a score.

B. Proposed prototype based on V-S classification

The proposed prototype Permission Monitoring System gives a quick review of the installed applications. It consists of 2 main lists. The first one is made of applications which are sorted in the order of the highest danger point score to the lowest one. A total danger point score for a specific application is compared to the maximum possible danger point score (maximum point score is 134). It gives that way a quick review of the data leakage potential.

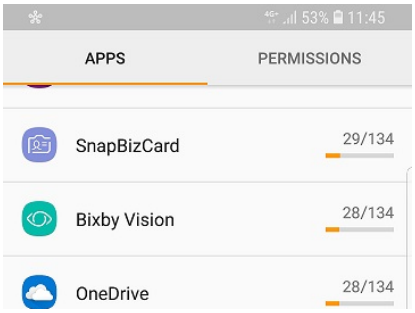


Fig. 6. Permission Monitoring System

V-S axis classification method is used both for the maximum danger point score calculation and for the current danger point score calculation of any specific application. As mentioned above, permission classification and default information values which a user can adjust to his / her own priority any time are taken into consideration.

If a user taps any application in the application list provided by the prototype, further options are available. The user can see the package name, version number, last update time etc. It also provides the number of dangerous, potentially dangerous and normal permissions. These permissions can be further explored after tapping their titles in this submenu.

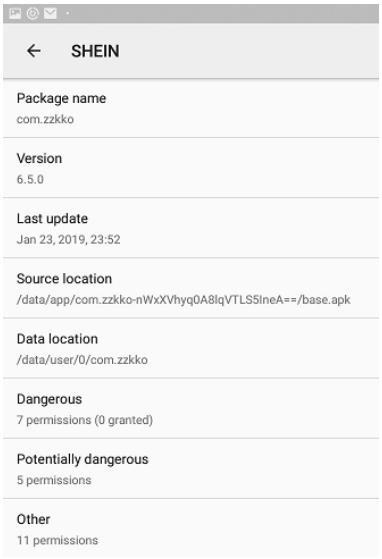


Fig. 7. Specific application data

As soon as a submenu option for permissions is tapped, one can see the dangerous permissions which of them are namely granted. If these dangerous permission groups are not granted but they are still in the list, it means that the manifest file contains that permission group, and as the application is used, sooner or later this permission group will be requested by the corresponding application. It is also the place where a user can change the value (horizontal one) axis level to a preferred one if he / she thinks that the default value does not meet his / her expectations. E.g. if a user feels that there is no important information in his / her contacts book and exposing it unintentionally to any third parties is not a big concern, it is

possible to change the value axis level to the middle one or the low one. Danger point score will be recalculated accordingly.

C. V-S classification of permissions

The following default values were used to calculate the score for any permission used within an application.

TABLE II. MAX. OF POINTS FOR A SPECIFIC PERMISSION

Dangerous permissions	2	0	2	4
Potentially dangerous	1	0	1	2
Normal permissions	0	0	0	0
		0	1	2
		Low value	Average value	High value

The maximum amount of points for a **dangerous** permission is 4. Meanwhile, the maximum amount of points for **potentially dangerous** permissions is 2.

The following formula was used to calculate the danger point score:

$(Y_D * X_D) + (Y_{PD} * X_{PD})$

All the permission groups which belong to the dangerous protection level are used for this prototype. As it was mentioned above, they belong to level High (2) on the sensitivity (Y) axis. Further details are provided below.

TABLE III. PERMISSION GROUPS AND MAX. POINT SCORE

Permission group	**Permissions and max. score on both axis**			
	Permissions	Y_D	X_D	***Multiplication***
CALENDAR	READ_CALENDAR	2	2	4
	WRITE_CALENDAR	2	2	4
CALL_LOG	READ_CALL_LOG	2	2	4
	WRITE_CALL_LOG	2	2	4
	PROCESS_OUTGOING_CALLS	2	2	4
CAMERA	CAMERA	2	2	4
CONTACTS	READ_CONTACTS	2	2	4
	WRITE_CONTACTS	2	2	4
	GET_ACCOUNTS	2	2	4
LOCATION	ACCESS_FINE_LOCATION	2	2	4
	ACCESS_COARS_LOCATION	2	2	4
MICROPHONE	RECORD_AUDIO	2	2	4
PHONE	READ_PHONE_STATE	2	2	4
	READ_PHONE_NUMBERS	2	2	4
	CALL_PHONE	2	2	4
	ANSWER_PHONE_CALLS	2	2	4
	ADD_VOICEMAIL	2	2	4
	USE_SIP	2	2	4
SENSORS	BODY_SENSORS	2	2	4
SMS	SEND_SMS	2	2	4
	RECEIVE_SMS	2	2	4
	READ_SMS	2	2	4
	RECEIVE_WAP_PUSH	2	2	4
	RECEIVE_MMS	2	2	4
STORAGE	READ_EXTERNAL_STORAGE	2	2	4
	WRITE_EXTERNAL_STORAGE	2	2	4
Maximum point score for dangerous permissions				104

Some normal protection level permissions are used for the sensitivity (Y) axis with the default value set to Middle. These values officially are considered as not dangerous, but a user may find it uncomfortable if their status becomes uncontrollable. Therefore, the level on the sensitivity axis (Y) is 1, and the level on the value axis (X) which can be changed by a user is also 1 by default. However, this default value is considered as 2 when calculating the maximum danger point score. The following table provides further calculation details.

TABLE IV. MAX. SCORE FOR POTENTIALLY DANGEROUS

Permissions	Y_{PD}	X_{PD}	**Multiplication**
CHANGE_NETWORK_STATE	1	2	2
CHANGE_WIFI_STATE	1	2	2
MODIFY_AUDIO_SETTINGS	1	2	2
REQUEST_DELETE_PACKAGES	1	2	2
NFC	1	2	2
REORDER_TASKS	1	2	2
REQUEST_INSTALL_PACKAGES	1	2	2
FLASHLIGHT	1	2	2
GET_TASKS	1	2	2
BILLING	1	2	2
SET_ALARM	1	2	2
DISABLE_KEYGUARD	1	2	2
SET_WALLPAPER	1	2	2
SYSTEM_ALERT_WINDOW	1	2	2
WRITE_SETTINGS	1	2	2
Maximum point score for potentially dangerous permissions			30

TABLE V. TOTAL MAXIMUM SCORE

Maximum danger point score	**Max. score**
Maximum point score for dangerous permissions	104
Maximum point score for potentially dangerous permissions	30

Maximum danger point score	Max. score
Maximum point score for dangerous permissions + potentially dangerous permissions	**134**

The maximum danger point score therefore is 134. If a user changes the level on the value (X) axis for any dangerous permission group or a potentially dangerous permission to low, it means that this permission will be multiplied by 0 which leads this permission to be unconsidered in the total danger point score for applications.

V. EXPERIMENTAL FINDINGS

The purposes of completing information leakage experiments based on permissions were the following ones:

1) Which categories do pose the highest risk of an information leakage among the tested ones?
2) Which applications do pose the highest risk of an information leakage among the tested ones?
3) Which permissions are requested the most frequently?

The following devices were used in one or other way in order to download the applications for testing them with the prototype.

TABLE VI. USED DEVICES

Device	Basic specifications
Lenovo Yoga 530	Windows Pro 10 Intel® Core™ i3-8130U CPU @ 2,20 Ghz 16,0 GB RAM
Samsung Galaxy S8	Android 8.0.0 Octa-core (2.3GHz Quad + 1.7GHz Quad), 64 bit, 10nm processor 4 GB RAM (LPDDR4)
Samsung Tab A (SM-T585)	Android 8.1.0 Octa-core (4x1.6 GHz Cortex-A53 & 4x1.0 GHz Cortex-A53) 3 GB RAM

Applications were downloaded based on different categories. Applications within the categories were picked while using the most popular application list since these applications are the most relevant ones to the biggest number of users.

The most popular 20 applications from the categories below were downloaded and installed.

- Shopping
- Finance
- Communication
- Education
- Business

Tested categories according to the results of the information leakage risk are distributed on the chart in the following way.

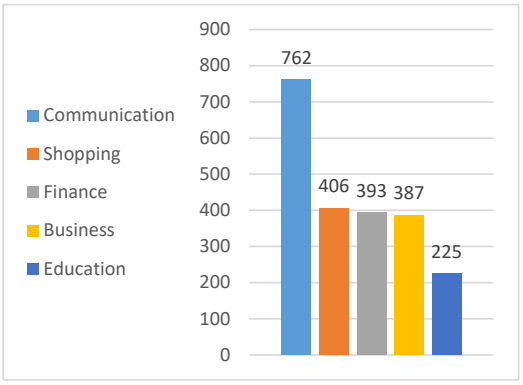


Fig. 8. Distribution of the tested categories

There results were calculated by summing up the danger point score of all the tested applications within that category. It was 20 top applications in it based on their popularity.

The following applications pose the highest risk of an information leakage among the tested ones.

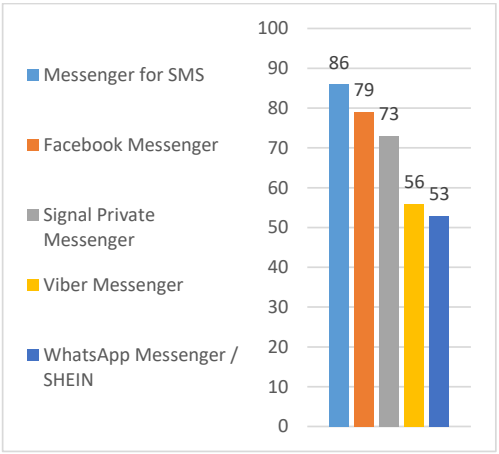


Fig. 9. The most dangerous applications

These applications were picked by looking for the highest danger point score among all the tested applications. The number of the tested applications is 100 at the moment.

The following dangerous permissions are requested the most frequently by the downloaded applications which were used for this research.

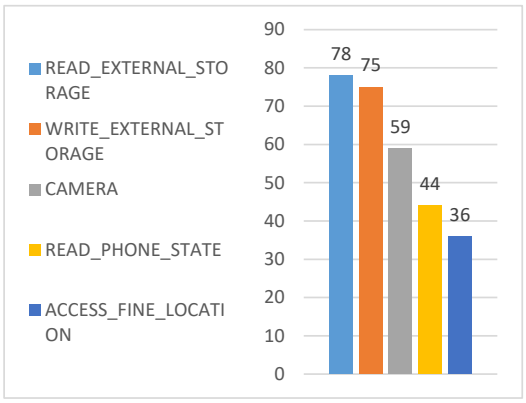


Fig. 10. The most frequent permissions

Numbers of the usage of different dangerous permissions were calculated in this test. As 100 Android applications were currently tested in this research, the chart numbers suggest the amount of instances the corresponding permission was requested or was to be requested. It means in this case that the permission READ_EXTERNAL_STORAGE was requested by 78 applications out of 100 tested applications. Top 5 permissions with the highest usage number were picked.

VI. CONCLUSIONS

Android OS security is based on the permission model. However, granting the permissions can be underestimated by a regular user due to a lack of available information or interest in his/her personal security.

A prototype has been offered which provides a simple risk assessment of any information leakage. A user does not need to have any awareness of permissions to understand the results.

100 applications in total from 5 different categories were tested. The results are provided in charts for a comparative purpose.

REFERENCES

[1] (2006) ACM DIGITAL LIBRARY, “State of the art smart spaces application models and software infrastructure”. [Online]. Available: http://ubiquity.acm.org/article.cfm?id=1167869

[2] (2017) IEEEXplore, “Critical Review of Static Taint Analysis of Android Applications for Detecting Information Leakages”, 8th International Conference on Information Technology (ICIT). [Online]. Available: http://ieeexplore.ieee.org/document/8080041/

[3] (2017) IEEEXplore, “Evaluating critical security issues of the IoT world: Present and Future challenges”. [Online]. Available: http://ieeexplore.ieee.org/document/8086136/

[4] (2017) IEEEXplore, “Android Permissions Unleashed”. [Online]. Available: https://ieeexplore.ieee.org/document/7243742

[5] Google Play Help, “Control your app permissions on Android 6.0 and up”, [Online]. Available: https://support.google.com/googleplay/answer/6270602?hl=en-GB

[6] (2007) IEEEXplore, “Research on Supply Chain Information Classification Based on Information Value and Information Sensitivity”. [Online]. Available: http://ieeexplore.ieee.org/document/4280248/

[7] (2018) “Protection levels”. [Online]. Available: https://developer.android.com/guide/topics/permissions/overview#normal-dangerous

INFORMACINËS TECHNOLOGIJOS • IVUS 2019 • ISSN 2029-249X • eISSN 2029-4824

Markov chains as a simulation technique for epidemic growth

Karolina Kęsik
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: karola.ksk@gmail.com

***Abstract*—Modeling various phenomena occurring in nature allows us to predict future effects in reality. One of such example is modeling the growth of infection in a given population depending on various parameters. In this article, we show the use of discrete Markov chains in order to model the epidemic with distinction into four states in which individuals in the population may be. More accurately – healthy, infected, sick and recovered. The article presents a mathematical model describing the phenomenon together with a calculation example and simulations. The obtained results were described and discussed.**

I. Introduction

Stochastic processes are a family of random variables defined in a certain probabilistic space. It is the field of the probability theory, which in today's science has become one of the most dynamically developed fields due to the numerous applications in optimization techniques or artificial intelligence. Such processes enable defining various phenomena based on a certain probability of its occurrence, as well as its components.

The field of stochastics found its place in the theory of optimization, where Lévy and Poisson processes were used in modeling various types of phenomena occurring in nature. Optimization theory has gained the most, where techniques inspired by nature have been created to this day. An example is the cuckoo algorithm intended originally to search for extremes of function. The movement of these birds was modeled using the Lévy flights. In [1], [2], the author presented the idea of searching key-points on images which can be used for feature extractions. Proposed idea is important in security because using stochastic algorithms almost guarantees finding other features due to the random placement of individuals in the population at the beginning of the algorithm. With similar motives in [3], the author showed the use of these algorithms in two-dimensional games, where such an algorithm can be used to model the opponent's movement. Additionally, in [4], [5], heuristic approach was used in different games. Similar algorithms are genetic ones based on chromosomes. Their use is visible in route planning programs [6], [7], whether the deployment of service stations in some area due to existing roads and their traffic. [8].

Another application is artificial neural networks, where there is a specific variation thereof with a stochastic note. In [9], [10], the authors model such networks and analyze the flow of information. Hybrids are also created, such as the connection of heuristics with neural networks [11]. Not only neural networks, but other artificial intelligence algorithms are used in practical terms, an example of which is optimization green computing awareness [12]. Another interesting idea is modeling of contaminant transport in porous media and monitoring of water quality [13], [14].

In this paper, we describe one of the classic stochastic tools such as Markov chains. We show their use in modeling the epidemic phenomenon and pay attention to their use in forecasting future phenomena.

II. Markov chains

The process $\{X_n, n \geq 0\}$ of state space S is called the discrete Markov chain, if for each $n \in \{0, 1, \dots\}$, the following equation occurs

$$\begin{aligned} P\{X_{n+1} = i_{n+1} | X_n = i_n, \dots, X_0 = i_0\} \\ = P\{X_{n+1} = i_{n+1} | X_n = i_n\} \end{aligned} \tag{1}$$

for all possible states $i_0, \dots, i_{n+1} \in S$.

It is a mathematical model of a random phenomenon evolving over time in such a way that the past affects the future only through the present. This model has state space S, where we can give the following properties

1) S it is a finite or at most a countable set of states,
2) $S = \{0, 1, 2, 3, \dots\}$,
3) $X_n = i$, what means, that at time n, the process is in the state of i.

Let us define the initial distribution of states, i.e. the distribution of the variable X_0. A probability vector as $\pi = [p_0, p_1, \dots]$. For each state $i \in S$, we have

$$p_i = P\{X_0 = i\}. \tag{2}$$

In addition, we define the probability of transition between two specific states. It is determined using a matrix $\mathbf{P} = (p_{ij})$. If S is a set defined as $(1, 2, \dots, m)$, than a matrix $\mathbf{P}$ has dimension equal $m \times m$. As p_{ij}, we can define the probability of going from one state i to another j in one step. Generalizing this, we have

$$p_{ij} = P\{X_{n+1} = j | X_n = i\}. \tag{3}$$

Using the initial vector π and the probability matrix $\mathbf{P}$, we define the Markov chain. And its behavior is defined using a stochastic matrix $\mathbf{P} = (p_{ij})$.

We assume that every element of such a matrix must be greater than 0 and the sum of elements of each row must be 1. This matrix has the following form

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix} \tag{4}$$

If the Markov chain is in the state i at n, then the probability of being in the state j after k periods is defined as follows

$$P(X_{n+k} = j | X_n = i) = P(X_k = j | X_0 = i) = p_{ij}(k), \tag{5}$$

where probabilities are independent of n, $p_{ij}(k)$ is an element at position (i, j) in $\mathbf{P}^k$.

III. Markov chains for epidemic simulations

The spread of infection can be represented using a mathematical model – a discrete Markov chain. In this model, we divide the population into four groups marked as states S=$\{0, 1, 2, 3\}$, i.e.

- 0 – susceptible sus,
- 1 – infected inf,
- 2 – sick $sick$,
- 3 – recovered rec.

Unfortunately, there is no way to create a perfect model, that is why we create several assumptions that simplify this model

- a healthy person can become infected with an infection,
- an individual in the infected group may go only to the disease state,
- a sick subject can leave a group of patients only through complete recovery,
- recovery guarantees immunity,
- immunity is not inherited
- age, sex and social status do not affect the probability of infection,
- climate or demographic changes do not affect the epidemic.

For such assumptions, the stochastic matrix can be defined as

$$\begin{pmatrix} p_{sus,sus} & p_{sus,inf} & p_{sus,sick} & p_{sus,rec} \\ p_{inf,sus} & p_{inf,inf} & p_{inf,sick} & p_{inf,rec} \\ p_{sick,sus} & p_{sick,inf} & p_{sic,sick} & p_{sick,rec} \\ p_{rec,sus} & p_{rec,inf} & p_{rec,sick} & p_{rec,rec} \end{pmatrix} \tag{6}$$

where p is the probability of changing states at t to $t+1$. An initial vector can be represented as

$$\mathbf{P}_0 = \begin{pmatrix} n_{sus} & n_{inf} & n_{sick} & n_{rec} \end{pmatrix}, \tag{7}$$

where n is the number of individuals in specific group at the initial stage.

IV. Experiments

A. Numerical example

Let us assume that at the beginning of our experiment, the population is made up of a thousand people, where 495 people are healthy, 3 are infected and 2 are sick. Hence, the initial vector is

$$\mathbf{P}_0 = \begin{pmatrix} 495 & 3 & 2 & 0 \end{pmatrix}. \tag{8}$$

The stochastic matrix is defined as follows

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.35 & 0.4 & 0.25 \\ 0 & 0 & 0.85 & 0.15 \\ 0 & 0.05 & 0.02 & 0.93 \end{pmatrix}. \tag{9}$$

After the first iteration, there is

$$\mathbf{P}_0\mathbf{P} = \begin{pmatrix} 495 & 3 & 2 & 0 \end{pmatrix} \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.35 & 0.4 & 0.25 \\ 0 & 0 & 0.85 & 0.15 \\ 0 & 0.05 & 0.02 & 0.93 \end{pmatrix} = \\ = \begin{pmatrix} 396 & 51 & 52 & 1 \end{pmatrix}. \tag{10}$$

After the second iteration, there is

$$\begin{pmatrix} 396 & 51 & 52 & 1 \end{pmatrix} \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.35 & 0.4 & 0.25 \\ 0 & 0 & 0.85 & 0.15 \\ 0 & 0.05 & 0.02 & 0.93 \end{pmatrix} = \\ = \begin{pmatrix} 317 & 57 & 104 & 21 \end{pmatrix}. \tag{11}$$

B. Simulation experiments

Simulations were conducted for two stochastic matrices. The first of these was described by the Eq. (9) and the second one has the following form

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.59 & 0.4 & 0.01 \\ 0 & 0 & 0.99 & 0.01 \\ 0 & 0.05 & 0.02 & 0.93 \end{pmatrix}. \tag{12}$$

Both matrices are similar but differ in the selected values. In the second matrix, the probability of transition from the infected to recovered and sick to the recovered state drastically changed, which means minimal chance of recovery. In all simulations, we assumed that the population is composed of 500 individuals. The effect of the difference in people needed for the spread of the disease on the entire population was examined. Tests were performed for the population composed of 1000000 individuals in population and 1 sick for two different step (time) values – $\{25, 150\}$. The results are shown in the form of diagrams in the Fig. 1–4. It is easy to see that the charts are identical regardless of the initial vector parameters (for the same probability matrix). Hence, the simple conclusion that Markov models allow simulation of the epidemic phenomenon, but the initial vector has little effect. Mainly stochastic matrices play a role, thus estimating the probability of transition between selected states.

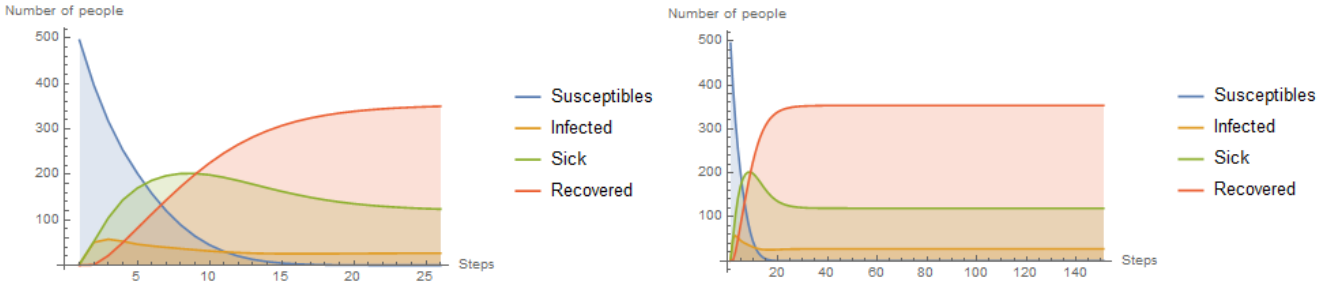


Figure 1: Measurements for the first matrix and population $\{495, 3, 2, 0\}$.

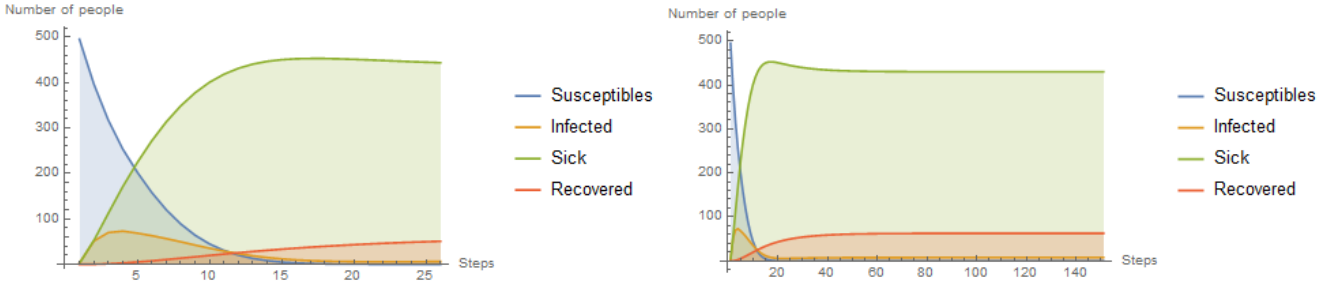


Figure 2: Measurements for the second matrix and population $\{495, 3, 2, 0\}$.

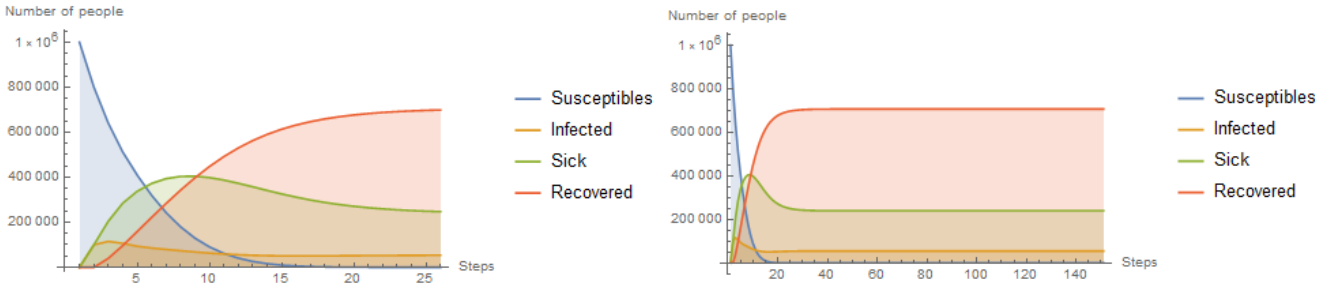


Figure 3: Measurements for the first matrix and population $\{1000000, 0, 1, 0\}$.

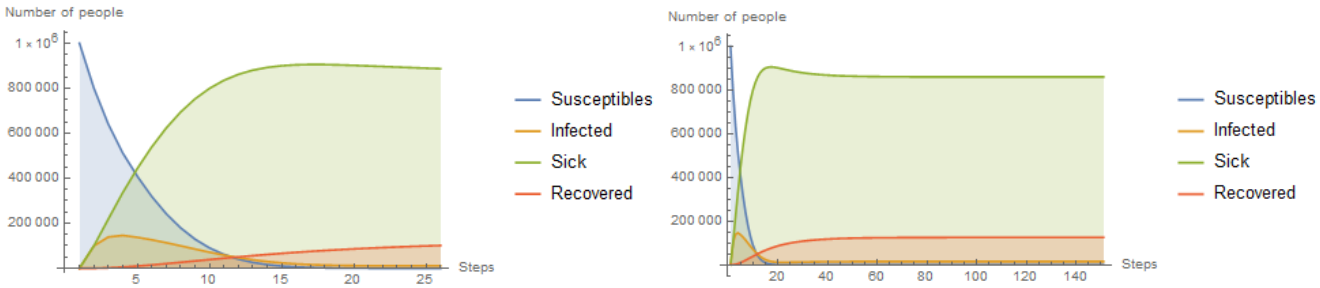


Figure 4: Measurements for the second matrix and population $\{1000000, 0, 1, 0\}$.

Note that in the probability of transition from infected and sick to healed state are 25% and 15%. In the population the average number of healthy is around 80% people and it stays independent of the step. Large differences in jumps between the population can be seen in the first 20 steps, where the population is infected and gets ill. The further step, the more stable the chart is, which may be due to the lack of changes in the stochastic matrix. From a practical point of view, the infection can evolve, and then these matrices can change. Unfortunately, the classic approach to Markov's chains does not modify the matrix during operation, although there are models that can do it. Impact on these matrices will result in a much better realignment of the model.

For the experiments we have carried out, we have performed

	Statistic	p-value
Anderson-darling	12173.87	0
Cramer-von Mises	34.27	0
Kolmogorov-Smirnov	0.84	$1.59 \cdot 10^{-95}$
Kuiper	0.92	$6.11 \cdot 10^{-114}$
Pearson χ^2	1573.91	$5.65 \cdot 10^{-328}$
Watson U^2	11.52	0

Table I: Statistical tests for susceptible table.

	Statistic	p-value
Anderson-darling	406.59	0
Cramer-von Mises	45.77	$1.22 \cdot 10^{-15}$
Kolmogorov-Smirnov	0.93	$8.45 \cdot 10^{-117}$
Kuiper	0.93	$3.19 \cdot 10^{-115}$
Pearson χ^2	2084.21	$4.73 \cdot 10^{-438}$
Watson U^2	11.69	0

Table II: Statistical tests for infected table.

statistical tests. On the significance level $\alpha = 0.1$, we also verify the hypothesis about the compatibility of the data distribution with the Gamma distribution with the shape parameter equal to 1 and the scale parameter equal to 2, which results are presented in Tab. I–IV. According to the tests carried out, in the case of each table presenting the process of healthy, infected, sick, recovered, at the significance level $\alpha = 0.1$, we can reject the hypothesis about the compatibility of the distribution of data with the Gamma distribution with the parameter equal to 1 and scale parameter 2.

V. Conclusion

Discrete Markov chains allow us to model phenomena occurring in nature. Each step depends on the previous one, although there is no change in probabilities during operation. This is a quite serious shortcoming in predicting the future effects of the model. In the analyzed case of epidemic spread, such action resulted in the absence of a possible mutation of the disease. However, it is worth noting that if the disease started to be fatal after a certain amount of time, in the case of the second matrix, it could kill almost the entire population, as opposed to the first one.

	Statistic	p-value
Anderson-darling	22104.23	0
Cramer-von Mises	49.74	0
Kolmogorov-Smirnov	0.99	$3.04 \cdot 10^{-132}$
Kuiper	0.99	$9.11 \cdot 10^{-131}$
Pearson χ^2	147.19	$2.61 \cdot 10^{-24}$
Watson U^2	12.35	0

Table III: Statistical tests for sick table.

	Statistic	p-value
Anderson-darling	3601.29	0
Cramer-von Mises	46.94	$6.66 \cdot 10^{-16}$
Kolmogorov-Smirnov	0.95	$3.19 \cdot 10^{-122}$
Kuiper	0.95	$2.19 \cdot 10^{-122}$
Pearson χ^2	1967.18	$8.61 \cdot 10^{-413}$
Watson U^2	12.04	0

Table IV: Statistical tests for recovered table.

This type of modeling of phenomena may allow us to improve the predictions of some phenomena that depend only on selected states and the table of probabilities. The number of states can be huge, but then the problem arises to create such a matrix and find its coefficients. The simulations showed that there is a large impact of even a small change in the main matrix.

References

[1] D. Połap and M. Woźniak, "Voice recognition by neuro-heuristic method," *Tsinghua Science and Technology*, vol. 24, no. 1, pp. 9–17, 2019.

[2] D. Połap, "Model of identity verification support system based on voice and image samples," *j jucs*, vol. 24, no. 4, pp. 460–474, 2018.

[3] A. Winnicka, "The opponent's movement mechanism in simple games using heuristic method," *Symposium for Young Scientists in Technology, Engineering and Mathematics (SYSTEM 2018)*.

[4] V. V. Zakharov and V. A. Shirokikh, "Heuristic evaluation of the characteristic function in the cooperative inventory routing game," *Journal on Vehicle Routing Algorithms*, vol. 1, no. 1, pp. 19–32, 2018.

[5] A. Iordan, "A comparative study of the a* heuristic search algorithm used to solve efficiently a puzzle game," in *IOP Conference Series: Materials Science and Engineering*, vol. 294, no. 1. IOP Publishing, 2018, p. 012049.

[6] V. Roberge, M. Tarbouchi, and G. Labonté, "Fast genetic algorithm path planner for fixed-wing military uav using gpu," *IEEE Transactions on Aerospace and Electronic Systems*, 2018.

[7] S. Yang, H. Wang, Z. Ren, S. Mu, J. Li, and J. Zhao, "Distribution network inspection route planning and the application of inspection automatic control technique," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. IEEE, 2018, pp. 1312–1315.

[8] T. H. Tran, G. Nagy, T. B. T. Nguyen, and N. A. Wassan, "An efficient heuristic algorithm for the alternative-fuel station location problem," *European Journal of Operational Research*, vol. 269, no. 1, pp. 159–170, 2018.

[9] P. Baskar, S. Padmanabhan, and M. Syed Ali, "Novel delay-dependent stability condition for mixed delayed stochastic neural networks with leakage delay signals," *International Journal of Computer Mathematics*, pp. 1–14, 2018.

[10] Z. Liu, "Design of nonlinear optimal control for chaotic synchronization of coupled stochastic neural networks via hamilton–jacobi–bellman equation," *Neural Networks*, vol. 99, pp. 166–177, 2018.

[11] S. Chatterjee, S. Sarkar, N. Dey, A. S. Ashour, and S. Sen, "Hybrid non-dominated sorting genetic algorithm: Ii-neural network approach," in *Advancements in Applied Metaheuristic Computing*. IGI Global, 2018, pp. 264–286.

[12] E. Okewu, S. Misra, R. Maskeliūnas, R. Damaševičius, and L. Fernandez-Sanz, "Optimizing green computing awareness for environmental sustainability and economic security as a stochastic optimization problem," *Sustainability*, vol. 9, no. 10, p. 1857, 2017.

[13] V. Nourani, S. Mousavi, D. Dabrowska, and F. Sadikoglu, "Conjunction of radial basis function interpolator and artificial intelligence models for time-space modeling of contaminant transport in porous media," *Journal of hydrology*, vol. 548, pp. 569–587, 2017.

[14] D. Dąbrowska, A. Witkowski, and M. Sołtysiak, "Representativeness of the groundwater monitoring results in the context of its methodology: case study of a municipal landfill complex in poland," *Environmental Earth Sciences*, vol. 77, no. 7, p. 266, 2018.

INFORMACINËS TECHNOLOGIJOS • IVUS 2019 • ISSN 2029-249X • eISSN 2029-4824

Impact of manipulation on initial population in heuristics

Alicja Winnicka
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: Alicja.Lidia.Winnicka@gmail.com

***Abstract*—The optimization problem is important due to practical use in various areas of our lives. Unfortunately, quite often there is a situation that we need to find the minimum values with certain criteria. Finding a solution using a classic approach is impossible to apply, which is due to an infinite number of solutions. That is why heuristic algorithms are used to find the optimal solution in a finite time. In this paper we propose the idea of manipulating the initial population due to better distribution of individuals in the whole space. The proposed solution has been described, tested and discussed.**

I. Introduction

Optimization algorithms find great practical use in various areas of our lives. Designing a construction such as a house requires an ideal weight distribution on the columns. From a mathematical point of view, the pressure, length and width of walls or columns can be described with the help of several equations, where the unknown value will depend on all elements. However, finding the best values to make this construction also as cheap as possible is quite a complicated problem. For this purpose, various algorithms are used that allow finding solutions that meet all criteria.

However, the problem is to search for these values. Quite often the set of values is an infinite one. And this prevents the use of iterative algorithms. These types of problems have contributed to the creation of heuristics, ie methods that do not guarantee ideal (and sometimes even correct) solutions in a finite period of time. Although, a large number of optimization problems are possible to solve using them.

The development of heuristics is driven by practical use. This is particularly important in the methods of artificial intelligence, where training the classifier occurs by minimizing weights in order to get the least error [1], [2] or the optimization used in the operation of various systems [3]–[6]. Another important element is the optimization in smart microgrid [7]. In this paper, we want to show how the manipulation of the initial distribution of individuals in the population depends on two classic heuristic algorithms.

II. Optimization problem

The optimization problem is understood as finding global extremes for a specific function $f(x) : \mathbb{R}^n \to \mathbb{R}$ for a given points $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$, $x_i \in \langle L_i, R_i \rangle$ for $i = 0, 1, \ldots, n-1$. In the case of minimization problem, it can be formulated as

$$\begin{aligned} &\text{Minimize} && f(\mathbf{x}) \\ &\text{subject to} && g(\mathbf{x}) \geq 0 \\ &&& L_i \leq x_i \leq R_i \qquad i = 0, 1, \ldots, n-1, \end{aligned} \tag{1}$$

where $g(\cdot)$ is inequality constraint.

III. Cuckoo Search Algorithm

Cuckoo Serch Algorithm (CSA) is one of the heuristic algorithms based on living creatures – in this case on cuckoos [8]. After observation of these birds, people noticed, that cuckoos have specific way to hatch their eggs. They actually do not hatch them by themselves, but search another bird's nests to leave eggs there. In that case, they need to find the nest whose host does not recgnize unfamiliar egg.

The behaviour of these birds inspired to extend minimization algorithm. CSA assumes that cuckoo is a point $\mathbf{x} = (x_0, \ldots, x_{n-1})$ on the n-dimension solution space. Of course, modeling the natural life of these creatures is impossible, so it is simplified due to the following points

- the size of population is constant in all iterations,
- each cuckoo can lay only one egg per iteration,
- cuckoo is identified with egg,
- host may detect unfamiliar egg with some propability, and in that case when egg is found, the cuckoo has to look for new place in the solution space.

At the beginning of algorithm, the population is created. Each cuckoo is placed on the solution space in random way. In each iteration, each cuckoo performs two actions. The first is a flight to another place which is modeled using Levy's equation (called *Levy's flight*) defined as

$$L(\mathbf{x}, \zeta, \eta) = \sqrt{\frac{\eta}{2\pi}} \frac{e^{-2\eta(\mathbf{x}-\zeta)}}{\sqrt{(\mathbf{x}-\zeta)^3}}. \tag{2}$$

In the above formula $\zeta, \eta \in \langle 0, 1 \rangle$ are coefficient. Using this formula, the cuckoo's movement is done using the following equation

$$\mathbf{x} = \mathbf{x} \pm L(\mathbf{x}, \zeta, \eta). \tag{3}$$

The second step is host decision because after the cuckoo flight, her egg is tossed to the nest. In fact, cuckoos can mask the tossed egg to make it look like other ones – the probability

of detection is minimized. Modeling this phenomenon involves defining a threshold value Ξ, with respect to which the egg will be detected. For random probability $\xi \in \langle 0, 1\rangle$ the following condition is checked

$$H(\mathbf{x}) = \begin{cases} \xi > \Xi & \text{the egg remains in the nest} \\ \xi \leq \Xi & \text{the egg is removed from the nest} \end{cases}. \tag{4}$$

In the case when the egg is removed from the nest, a new position is choosen for this cuckoo (in random way) – this action guarantees a constant size of the population. The algorithm is executed to satisfy a stop condition, which is most often the number of iteration in the algorithm. At the end, the best cuckoo is returned, which has the smallest or largest value of fitness function in the entire population (depending on the optimization problem being considered).

Data: number of iterations t, number of cuckoos k, fintess function $f(\cdot)$, the solution space, threshold value Ξ
Result: The best cuckoos
Start;
Generate a population of k cuckoos in random way in solution space;
$i := 0$;
while $i \leq t$ **do**
  $k = 0$;
  while $j \leq k$ **do**
    Change the position of j-th cuckoo using Eq. (2);
    Choose $\xi \in \langle 0, 1\rangle$ in random way;
    if $\xi > \Xi$ **then**
      Generate new position for j-th cuckoo;
    end
  end
end
Return cuckoo with the best fitness value $f(\cdot)$;
Stop;

Algorithm 1: Cuckoo Search Algorithm.

IV. Polar Bear Algorithm

Another heuristic algorithm is inspired by polar bear's behavior during hunting [9]. Polar bear must find the prey and if he doesn't find anything, he must use the ice floak as a way to move a certain distance. He uses it to drift to another location, where possibly he will find the prey (especially seals).

The basis of this algorithm are similar to the CSA, we have a population of k bears, where each of them can be represented as point $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$ in some solution space. At the beginning, an initial population is created with one difference – only 75% of individuals are generated, the remaining 25% depend on Arctic conditions and reproduction.

Each bear moves only if the new position is better than the current one in relation to the fitness function $f(\cdot)$. It can be defined as the following equation

$$(\mathbf{x}_j^t)^{(i)} = (\mathbf{x}_j^{t-1})^{(i)} + \text{sign}\,(\omega)\,\alpha + \gamma, \tag{5}$$

where α is a random number in $(0, 1\rangle$, ω is the distance between two spatial coordinates and γ is a random value in the range of $\langle 0, \omega\rangle$. The distance is defined using simple Euclidean metric between two points and described as

$$d\left((\mathbf{x})^{(i)}, (\mathbf{x})^{(j)}\right) = \sqrt{\sum_{k=0}^{n-1}\left((x_k)^{(i)} - (x_k)^{(j)}\right)^2}. \tag{6}$$

Polar bears do not hunt only on the surface, but also in the water. In a situation where the bear flows on the ice floe, he closely observes the surrounding water. In the case of noticing the future victim, the bear dives into the water and attack. There are occasions when a bear very suddenly throws himself on the victim, if there is a chance that he will run away. In the modeling of this phenomenon, the polar bear's area of view should be considered as

$$r = 4a\cos(\phi_0)\sin(\phi_0), \tag{7}$$

where $a \in \langle 0, 0.3\rangle$ is a visible distance, $\phi_0 \in (0, \frac{\pi}{2}\rangle$ is the angle of approaching to the victim. This equation is used to describe the local movement of these individuals, where each spatial coordinate is modified using

$$\begin{cases} x_0' = x_0 \pm r\cos(\phi_1) \\ x_1' = x_1 \pm [r\sin(\phi_1) + r\cos(\phi_2)] \\ x_2' = x_2 \pm [r\sin(\phi_1) + r\sin(\phi_2) + r\cos(\phi_3)] \\ \ldots \\ x_{n-2}' = x_{n-2} \pm \left[\sum_{k=1}^{n-2} r\sin(\phi_k) + r\cos(\phi_{n-1})\right] \\ x_{n-1}' = x_{n-1} \pm \left[\sum_{k=1}^{n-2} r\sin(\phi_k) + r\sin(\phi_{n-1})\right] \end{cases}, \tag{8}$$

where $\phi_1, \phi_2, \ldots \phi_{n-1} \in \langle 0, 2\pi\ )$.

Difficult living conditions on arctic surfaces may be detrimental to the individuals living there, for this purpose a random value $\kappa \in \langle 0, 1\rangle$ is introduced, which allows to introduction the conditions of freezing or dying from hunger as the following rule

$$\begin{cases} \text{Death} & \text{if } \kappa < 0,25 \\ \text{Reproduction} & \text{if } \kappa > 0,75 \end{cases}, \tag{9}$$

where, if the first condition is met, the weakest individual dies (on the condition that the population size is greater than half). In the second case, the two individuals in a given iteration (one is the best one in thole population relative to the fitness function and when the size of population is smaller than assumed) reproduce as

$$(\mathbf{x}_j^t)^{(new)} = \frac{(\mathbf{x}_j^t)^{(best)} + (\mathbf{x}_j^t)^{(i)}}{2}, \tag{10}$$

Table I: Test functions used in optimization.

Name	Function formula	Domain	Point **x**	Minimum
Ackley	$-20\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)-\exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right)$	$\langle -32.8, 32.8\rangle$	$(0,\ldots,0)$	0
Booth	$(x_1+2x_2-7)^2+(2x_1+x_2-5)^2$	$\langle -10, 10\rangle$	$(1,3)$	0
Easom	$-\cos(x_1)\cos(x_2)\exp\left(-(x_1-\pi)^2-(x_2-\pi)\right)$	$\langle -10, 10\rangle$	(π,π)	-1
McCormick	$\sin(x_1+x_2)+(x_1-x_2)^2-1.5x_1+2.5x_2+1$	$x_1\in\langle -1.5, 4\rangle$, $x_2\in\langle -3, 4\rangle$	$(-0.54719, -1.54719)$	-1.9133
Sphere	$\sum_{i=1}^{n}x_i^2$	$\langle -10, 10\rangle$	$(0,\ldots,0)$	0
Trid	$\sum_{i=1}^{n}(x_i^2-1)^2-\sum_{i=2}^{n}x_ix_{i-1}$	$\langle -10, 10\rangle$	$(i(n+1-i))$	$\frac{-n(n+4)(n-1)}{6}$
Zakharov	$\sum_{i=1}^{n}x_i^2+\left(\sum_{i=1}^{n}0.5ix_i\right)^2+\left(\sum_{i=1}^{n}0.5ix_i\right)^4$	$\langle -5, 10\rangle$	$(0,\ldots,0)$	0

where $(\mathbf{x}^t)^{(best)}$ and $(\mathbf{x}^t)^{(i)}$ are two individuals selected from the best in whole population.

Data: number of iterations t, number of polar bears k, fintess function $f(\cdot)$, the solution space
Result: The best polar bear
Start;
Generate a population of $0.75\cdot k$ bears in random way in a given solution space;
$i := 0$;
while $i \leq t$ **do**
  $k = 0$;
  while $j \leq k$ **do**
    Calculate new position using Eq. (5);
    if *new position is better* **then**
      Change the position of j-th bear;
    end
    Calculated a view distance using Eq. (7);
    Move bear closer to the victim using Eq. (8);
    Choose $\kappa\langle 0, 1\rangle$ in a random way;
    if $\kappa > 0.75$ *and the population size is correct* **then**
      Remove the weakest individual from the population;
    end
    if $\kappa < 0.25$ *and the population size is correct* **then**
      Take the two best individuals for the reproduction process according to Eq. (10);
    end
  end
end
Return polar bear with the best fitness value $f(\cdot)$;
Stop;

Algorithm 2: Polar Bear Algorithm.

V. Manipulation of the initial population

The idea is to place individuals at similar distances, which has a chance to guarantee a search of a larger area. The reason is that the placement of individuals in the original algorithms is based on random choice. This can lead to the situation that all individuals will be located in a similar area, omitting the rest.

Each individual is positioned according to the solution space, i.e. $\langle L_i, R_i\rangle$. Having k individuals, the area can be divided into m parts as

$$\bigcup_{j=0}^{m-2}\left\langle L_i + j\cdot\frac{R_j}{m}, L_i+(j+1)\cdot\frac{R_j}{m}\right\rangle. \tag{11}$$

In each subset, $\frac{k}{m}$ individuals are generated in random way. The boundary values of subsets are repeated in neighboring elements, but it does not extend the initial range.

VI. Experiments

As part of checking the validity of the proposed solution, the test function described in Tab. I were used (with dimension equal to 10). Tests for described algorithms with space manipulation and without were conducted for two different size of population – 100 and 1000 during 300 iterations. Note that the individuals will be placed at random, so to authenticate the results, each algorithm will be made 10 times, and the result averaged. Obtained solutions are presented in Tab. III and II.

According to the tables, the average results are in almost every case more precise, which allows us to stalk that it is worth manipulating the solution space and dividing it into smaller subsets in which individuals will move in heuristic algorithms.

VII. Conclusions

In this paper, we proposed an even distribution of individuals in the initial populations in order to increase the search of the full area. The tests were carried out using two algorithms inspired by nature – Cuckoo Search Algorithm and Polar Bear Algorithm. The obtained results indicate a much faster finding of the extreme, and compared to the same number of iterations – the solutions are more accurate. This allows to say that heuristics can work more effectively when their randomness is minismized.

Table II: Obtained average results for 10 individuals.

Function name	CSA	CSA with modification	PBA	PBA with modification
Ackley	0.02972880281961	0.0177714446875131	0.0897574672427762	0.071868300285129
Booth	-0.0158455184268977	-0.0089278921386823	0.063578710734648	0.0355476169081161
Easom	-0.905359405654184	-0.987498435744782	-0.992768202439308	-0.982191178613431
McCormick	-1.97623352281811	-1.94638200721307	-1.9460962896008	-1.86267427586381
Sphere	-0.0536565001372511	0.0149251524428488	-0.0643595185430532	0.0481632181667552
Trid	-209.889724268061	-210.042710510819	-210.021059083669	-210.065610355961
Zakharov	0.00851286580251198	-0.0125859739783155	-0.0245003661254888	0.0547493551181393

Table III: Obtained average results for 100 individuals.

Function name	CSA	CSA with modification	PBA	PBA with modification
Ackley	0.0234407469855811	0.011506596170644	-0.0205448465051804	0.0162699407465658
Booth	-0.0176004870876672	-0.0171003493475348	-0.0215751469119336	0.00856991690897929
Easom	-0.979836908462381	-0.996544835418547	-1.01681687238944	-0.989967549610368
McCormick	-1.90125860997073	-1.92803074770287	-1.92798707448323	-1.99966222023081
Sphere	0.0134152536808584	0.00213265699349087	0.0206986147541081	-0.0172196125691848
Trid	-210.005388625923	-209.982567193072	-210.007645864532	-209.994710194309
Zakharov	-0.0181133843577064	-0.0170223669058917	0.000409970479276949	0.0166918954551055

REFERENCES

[1] R. Damaševičius, "Optimization of svm parameters for recognition of regulatory dna sequences," *Top*, vol. 18, no. 2, pp. 339–353, 2010.

[2] D. Połap, "Human-machine interaction in intelligent technologies using the augmented reality," *Information Technology And Control*, vol. 47, no. 4, pp. 691–703, 2018.

[3] Y. Gao and Y.-J. Liu, "Adaptive fuzzy optimal control using direct heuristic dynamic programming for chaotic discrete-time system," *Journal of Vibration and Control*, vol. 22, no. 2, pp. 595–603, 2016.

[4] S. Lange, S. Gebert, T. Zinner, P. Tran-Gia, D. Hock, M. Jarschel, and M. Hoffmann, "Heuristic approaches to the controller placement problem in large scale sdn networks," *IEEE Transactions on Network and Service Management*, vol. 12, no. 1, pp. 4–17, 2015.

[5] Z. Pooranian, M. Shojafar, J. H. Abawajy, and A. Abraham, "An efficient meta-heuristic algorithm for grid computing," *Journal of Combinatorial Optimization*, vol. 30, no. 3, pp. 413–434, 2015.

[6] C. Jagtenberg, S. Bhulai, and R. van der Mei, "An efficient heuristic for real-time ambulance redeployment," *Operations Research for Health Care*, vol. 4, pp. 27–35, 2015.

[7] G. Graditi, M. L. Di Silvestre, R. Gallea, and E. R. Sanseverino, "Heuristic-based shiftable loads optimal management in smart micro-grids," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 271–280, 2015.

[8] X.-S. Yang and S. Deb, "Cuckoo search via lévy flights," in *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*. IEEE, 2009, pp. 210–214.

[9] D. Połap *et al.*, "Polar bear optimization algorithm: Meta-heuristic with fast population movement and dynamic birth and death mechanism," *Symmetry*, vol. 9, no. 10, p. 203, 2017.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Automatic Detection of Contraindications of Medicines in Package Leaflet

Jonas Žalinkevičius
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
jonas.zalinkevicius@hotmail.com

Rita Butkienė
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
rita.butkiene@ktu.lt

Abstract— **Before physicians prescribe medicines, they must take into consideration the patient's diseases and medicines they use. This is done to avoid complications that may occur. All information about possible contraindications is written in the medicine package leaflet. A system that can automatically detect contraindication mention in the Lithuanian text of leaflet applying natural language parsing is presented. This system gives a possibility to shorten the time needed for medicines prescription decision making. Results of the experiment showed that the created system successfully detected 56 per cent contraindications.**

Keywords— medicine contraindications, drug – drug interactions, shallow parsing, morphological analysis, noun phrase detection

I. Introduction

When patient is diagnosed with a new disease, additionally physician asks patient about his allergies, previous health problems, chronic deceases, what medications and food supplements he is using. After taking gathered information into consideration and evaluation of possible contraindications with prescribed medication physician assigns treatment and, if needed, changes previous assignments. Almost all information about contraindications can be found in medicine package leaflet. According to Lithuania's medicines registration procedure [1], every package must have leaflet written in Lithuanian. Information in the leaflet must be divided in six sections [2], although the text in section can be written in not structural manner. So, if physician needs to find possible contraindications, he must read all text in second section (Table 1) or search for information on the Internet. Usually health care information consists of unstructured data and that leads to inaccurate search results that contains hundreds of links to not relevant documents. And user must read through results in order to find relevant information.

Automatic information extraction tools can extract biomedical data, save it in a structural way, and minimize information search problem. However, automatic text analysis and information extraction from unstructured text in the medical domain is a challenging task [3]. The aim of this paper is to present a system that gives physicians the possibility of a faster and more accurate way of finding contraindications using automated contraindication detection in the medicine package leaflet.

A system that automates the extraction of contraindications from leaflet text is described is in Section 3. Using this system all leaflets of medicines registered in Lithuania were analyzed. Results of this analysis (contraindications extracted) are used in a commercial medications information system that is used by Lithuanian physicians for prescription of medications. Evaluation of obtained results is presented in Section 4.

II. Related Work

In Lithuania, it is established that each medicine registered in Lithuania must contain a package leaflet describing therapeutic indications, possible contraindications, safety precautions, and usage information in the Lithuanian language. In order to be sure that the patient not suffer from possible contraindication, the physician should read through all leaflet text before prescribing the medicine. Usually, the analysis of leaflet is time-consuming, so physicians tend to skip it and rely on the knowledge and experience they have gained.

There are lots of systems developed for analysis and information extraction from the biomedical text in the English language. But there is no solution for detection of contraindication (i.e. contraindication with disease or contraindication with the pharmacological group) mentions in Lithuanian written text. We have analyzed articles that describe similar problems when analyzing biomedical text. For example, a tool *Semantator* [4] was created for converting biomedical text to linked data. It used ontology-based information extraction using biomedical ontology terms hosted in *BioPortal* and ontology editor *Protégé* for text preprocessing. A semantic annotation and inference platform SENTIENT-MD [3] creates a dependency graph as the first step for dependency parsing which is one of the tasks of semantic annotation of medical knowledge in natural language text. Markus Bundschus [5] used probabilistic graphical models (Conditional Random Fields) to identify semantic relations.

Although all these authors work on texts written in English, we found that common rules and approaches could be applied to Lithuanian texts as well. In order to extract information from text, preprocessing is needed using natural language processing: text segmentation, morphological analysis should be performed and then syntactic parse tree or the dependency graph [6]. [7] should be formed. For semantic relations detection, existing ontologies or knowledge bases should be used.

III. System Description

In this section, a system for the detection of contraindication mentions in the medicine leaflet text written

in Lithuanian is presented. The system implements a text analysis pipeline of four analysis stages: extraction of contraindication text block, morphological analysis, noun phrase detection and annotation.

Additionally, all annotated phrases are checked is it in the database of noun phrases to be ignored or not. This database is manually filled and helps to obtain more precise results. The overall pipeline for detection of contraindication mentions is shown in fig. 1.

Below each stage of text analysis is discussed in more detail.

A. Extraction of contraindication text blocks

In Lithuania, when describing the medicine, a producer must follow a certain template of the package leaflet [2]. This template splits the description of leaflet into 6 sections listed in Table 1

TABLE I. MEDICINE PACKAGE LEAFLET SECTIONS

No	Section
1	What X is and what it is used for
2	What you need to know before you <take> <use> X
3	How to <take> <use> X
4	Possible side effects
5	How to store X
6	Contents of the pack and other information

Information which patient should be aware of before he or she takes the medicine is presented in section number two. Example of this section is shown in fig. 2 with highlighted contraindications phrases. So, the first task of our system is to find this section and extract its text for further analysis.

B. Morphological analysis

A morphological analysis forms a background for information extraction about contraindications. In this stage, a given text is split into lexical units (e.g. sentences, lexemes) and analyzed morphologically. For this task, a web service provided by system "*http://semantika.lt*" [8] is used. The web service returns morphological features for each given lexeme: part of speech, gender, number and so on.

C. Noun phrase detection

Phrases that express a specific contraindication usually are noun phrases, for example, *heart attack*, *type one diabetes*, *pancreatitis*, and so on. Therefore, we chose a phrase structure grammar method because it better fits for noun phrase detection than dependency grammar as it was suggested by Axel Halvoet in his monography [9]. Phrase structure rules are used to split natural language written sentence into its constituent parts: lexical and phrasal categories [9], [10], [11]. For the noun phrase detection in the medicine's leaflet, three phrase structure rules ware specified (see Table 2).

TABLE II. NOUN PHRASE STRUCTURE RULES

No	Rule
1	A lexeme is a part of noun phrase if it is a noun in genitive case and folows another noun in genitive case or adjective or numeral or participle.
2	A lexeme is a part of noun phrase if it is an attributive adjective in the same case, number and gender as base noun and folows noun in genitive case or adjective or numeral or participle.
3	A lexeme is a part of noun phrase if it is an attributive numeral in the same case, number and gender as the base noun and follows noun in genitive case, or adjective, or numeral, or participle.

An algorithm implemented for the noun phrase detection checks every lexeme in the sentence for the satisfaction of conditions of at least one rule presents in Table 2. If the condition is satisfied a lexeme is included in the noun phrase. The workflow of analysis of the noun phrase *Lėtinis reumatinis perikarditas* (Chronic rheumatic pericarditis) is shown in Table 3.

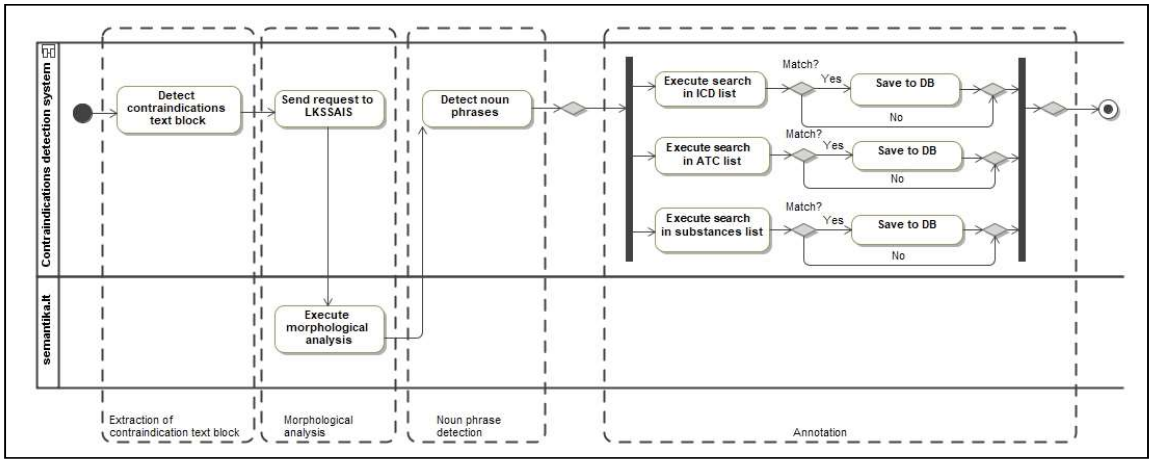
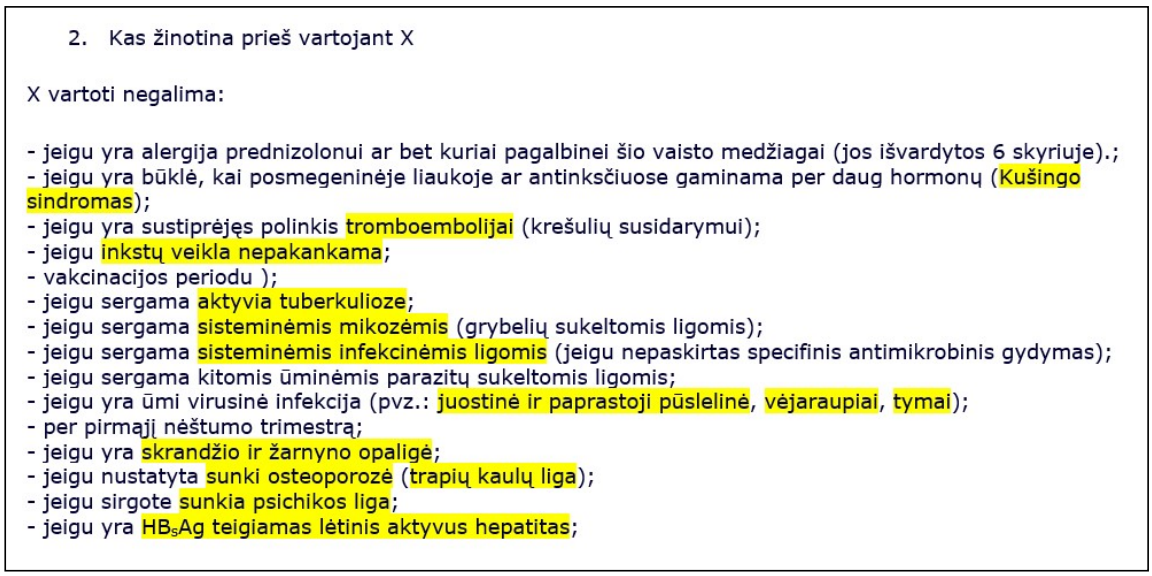


Fig. 1. Contraindications lookup process activity



2. Kas žinotina prieš vartojant X

X vartoti negalima:

- jeigu yra alergija prednizolonui ar bet kuriai pagalbinei šio vaisto medžiagai (jos išvardytos 6 skyriuje).;
- jeigu yra būklė, kai posmegeninėje liaukoje ar antinksčiuose gaminama per daug hormonų (Kušingo sindromas);
- jeigu yra sustiprėjęs polinkis tromboembolijai (krešulių susidarymui);
- jeigu inkstų veikla nepakankama;
- vakcinacijos periodu);
- jeigu sergama aktyvia tuberkulioze;
- jeigu sergama sisteminėmis mikozėmis (grybelių sukeltomis ligomis);
- jeigu sergama sisteminėmis infekcinėmis ligomis (jeigu nepaskirtas specifinis antimikrobinis gydymas);
- jeigu sergama kitomis ūminėmis parazitų sukeltomis ligomis;
- jeigu yra ūmi virusinė infekcija (pvz.: juostinė ir paprastoji pūslelinė, vėjaraupiai, tymai);
- per pirmąjį nėštumo trimestrą;
- jeigu yra skrandžio ir žarnyno opaligė;
- jeigu nustatyta sunki osteoporozė (trapių kaulų liga);
- jeigu sirgote sunkia psichikos liga;
- jeigu yra HB_sAg teigiamas lėtinis aktyvus hepatitas;

Fig. 2. Example of "What you need to know before use of X" section in medicine package leaflet

TABLE III. EXAMPLE OF NOUN PHRASE DETECTION WORKFLOW

Step	Action	Rule satisfaction
1	The first lexeme *Lėtinis* (Chronic) is an adjective in the nominative case, singular and of masculine gender	No rule condition is satisfied fully, but according to rule No. 2 the lexeme is a good candidate for the noun phrase.
2	The second word *reumatinis* is an adjective in the nominative case, singular and of masculine gender and folows the adjective *Lėtinis*	No rule condition is satisfied fully, but according to rule No. 2 the lexeme is a good candidate for the noun phrase.
3	The third word *perikarditas* is a noun in the nominative case, singular and of masculine gender It follows the adjectives *lėtinis* and *reumatinis* which are in the same case, number and gender .	The condition of rule No. 2 is satisfied. The noun is a base noun for the first two adjectives. They are attributive adjectives of the noun. So, the condition of rule No. 2 is satisfied as well. The analysis of the third lexeme completes construction of the noun phrase.

When construction of the noun phrase is complete the form of the head noun in the phrase is changed to its canonical form (lemma). This is done because the name of item registered in the International Classification of Diseases (ICD) [12], Anatomical Therapeutic Chemical Classification System (ATC) [13] or lists of active substances are in the canonical form, therefore, normalization is required to ensure correct comparison of values in the next stage of analysis.

D. Annotation

All noun phrases identified in the previous stage are reviewed and checked for contraindication. If a contraindication is identified, the phrase is annotated. For annotation three databases are used: ICD, ATC and the lists of active substances. The algorithm compares the noun phrase and name of the item from the database. If the noun phrase matches the name in ICD the phrase is tagged as contraindication with the disease. If phrase matches ATC item name, it is tagged as contraindication with a pharmaceutical chemical group, and if the phrase matches the name of the active substance, it is tagged as contraindication with an active substance.

It is worthy to mention that before comparison of the noun phrases all identified phrases are checked against phrases in the database of noun phrases to be ignored. In the text of medicine package leaflet, a lot of words (i.e. illness, hand and so on) that are irrelevant (do not express a contraindication) but are used in ICD, ATC and active substances lists could be found. The database of noun phrases to be ignored was filled manually with the help of a professional pharmacist.

IV. EXPERIMENT

The aim of the experiment is to evaluate the created system and check if a tool can achieve its target - to give physicians the possibility of a faster and more accurate way of finding contraindications. The experiment was done by manually annotating contraindications mentions in the package leaflet text block and comparing results with the system's results. This was done by professional pharmacist who works in JSC Skaitos kompiuterių servisas.

A. Plan

The experiment was organized as follow. From medicines database ten randomly selected leaflets were analyzed using the system created. The results of the analysis were automatically gathered into the table, which example is presented in Table 4 In the first column the code of item automatically found in the text of leaflet by the system is indicated. The second column represents the database (ATC, ICD or active substances) where the item is registered. The third column was used for the evaluation of annotation correctness.

TABLE IV. AUTOMATICALLY DETECTED CONTRAINDICATIONS RESULTS EVALUATION FOR SINGLE LEAFLET

Code	Domain	Is detection correct
J01CR	ATC	False
J05AE	ATC	True
I09.2	ICD	True

The same randomly selected leaflets were analyzed and annotated manually, and the table of the same structure was filled in with manual annotation results. Manually found contraindications were not interpreted or changed to synonyms. For example, *heart attack* and *myocardial infarction* is the same disease. But ICD contains only one name of this disease - *myocardial infarction*. The created system is not able to recognize the heart attack as a synonym of *myocardial infarction*.

Additionally, the active substances mentioned in the leaflet, were translated into the Latin language (nominative and genitive grammatical cases). This was done because the database of active substances, that was provided, has three versions of translation: Lithuanian, Latin in nominative case and Latin in genitive case.

B. Results

The results of the evaluation are presented in Table 5. The precision, recall and F-Score metrics have been calculated for each leaflet analyzed. Additionally, the ratio between the number of correctly detected contraindications and overall automatically detected contraindications was calculated as well. This metric allows to evaluate how accurate the results are and to use them in further calculations.

Results showed that the system developed is able to correctly detect 56% of relevant contraindications. The average number of links detected automatically is 1482.8 while manually detected links is 197.9. The number of links detected automatically in one leaflet is average four times higher, than detected manually. The average number of erroneous links to ICD is of 72%, to ATC - 90%, and to the list of active substances - 61%.

Calculations show that the system is able to achieve 0.25(±0.23) precision, 0.56(±0.32) recall, and 0.31(±0.19) F-score value. To give a better perspective where the system's failures were and possible reasons for that, Pearson correlation coefficient calculations between various indicators were done (Table 6). The biggest impact on F-Score had incorrectly detected links to ICD, a coefficient was -0.89. The reason why precision was so low is that of the high ratio between automatically and manually detected links.

TABLE V. EXPERIMENT RESULTS

ID	Auto. detected links	Auto. correctly detected links	Man. detected links	Precision	Recall	F-Score	Ratio of links amounts	Err. links to ICD	Err. links to ATC	Err. links to active substances
13092	1906	346	385	0,18	0,90	0,30	4,95	82%	100%	65%
13571	1899	367	444	0,19	0,83	0,31	4,28	81%	100%	58%
859	87	67	162	0,77	0,41	0,54	0,54	17%	100%	100%
1300	400	28	146	0,07	0,19	0,10	2,74	98%	100%	24%
10958	464	14	71	0,03	0,20	0,05	6,54	100%	25%	21%
1872	283	66	68	0,23	0,97	0,38	4,16	77%	100%	43%
5363	473	237	291	0,50	0,81	0,62	1,63	46%	88%	49%
13273	158	51	72	0,32	0,71	0,44	2,19	45%	100%	100%
10744	1199	150	175	0,13	0,29	0,18	6,85	87%	100%	100%
16551	1090	120	204	0,11	0,25	0,15	5,34	90%	87%	51%
Median	468,5	93,5	168,5	0,185	0,56	0,305	4,22	82%	100%	55%
Q1	312,25	54,75	90,5	0,115	0,26	0,158	2,328	54%	91%	45%
Q3	1171,75	215,25	269,25	0,298	0,825	0,425	5,243	89%	100%	91%
Avg	795,9	144,6	201,8	0,25	0,56	0,31	3,92	72%	90%	61%
Std dev	686,52	129,45	132,27	0,23	0,32	0,19	2,10	27%	23%	30%
Min	87	14	68	0,03	0,19	0,05	0,54	17%	25%	21%
Max	1906	367	444	0,77	0,97	0,62	6,85	100%	100%	100%

TABLE VI. CORRELATION OF ESTIMATES AND INDICATORS

		Estimates		
		Precision	***Recall***	***F-Score***
Indicator	Incorectly detected links to ICD list amount	-0,9655	-0,3114	-0,8939
	Incorectly detected links to ATC list amount	0,3292	0,4184	0,4382
	Incorectly detected links to active substances list amount	0,5229	0,1244	0,4523
	Automatically and manually detected contraindications ratio	-0,8119	-0,2583	-0,7682

C. Conclusions of experiment

The experiment shows that the system automatically successfully detected more than half of the relevant contraindication links (56%). But 75% of links were erroneous and the system lacks precision. Reason for that is a high number of incorrect links to ICD (r=-0.9655), this indicator has most negative impact on the precision and F-Score results. This might be because of commonly used phrases that are not contraindications but used in the ICD list. For example, the word *allergy* does not imply that this is a contraindication and must be ignored. Another reason for low estimates results is, the number of detected contraindications phrases. Calculations shows, that the higher is the difference between automatically and manually detected contraindications phrases, the lower are precision and F-Score results. The reason for that is, high number of noun phrases that are irrelevant to contraindications noun phrases, for example: pill, driving.

Additionally, considering why F-Score is so low (0.31) the assumption that this is because of low precision (0.25) can be done. To raise this indicator the list of phrases to be ignored (common word and phrases) must be used. The most frequent reasons for the incorrect detection of contraindications are:

- the context of phrase in the sentence is not taken into account;
- Conjunctions are not taken into account and two or more noun phrases (i.e. "…kidney and liver diseases…") are not identified;
- Brackets that are used to specify contraindication are not taken into account ("…liver tumor (malignant or benign)…").

To avoid errors caused by those reasons, users of "*https://gydytojams.vaistai.lt*" IS will be able to mark contraindication as erroneous and if pharmacist approves that it will be removed from the database.

V. CONCLUSIONS

In this paper the system which automatically detects contraindications and links them to existing "Skaitos kompiuterių servisas" databases have been introduced. System analyses text of medications leaflets, it extracts noun phrases and links them to corresponding items in ATC, ICD and active substances list. The system presented was used for the extraction of contraindications from leaflets of all medications registered in Lithuania. Extracted data was used in the pilot project for extending a functionality of system "*https://gydytojams.vaistai.lt*". The additional function supports physicians in search of possible contraindications that are relevant to patient medical records. Moreover, physicians have the possibility to give feedback about erroneous contraindications presented. In such a way they help in expanding the list of phrases to be ignored and eliminating incorrect contraindication links.

The experiment shows that approximately 56% of contraindications are found but only every fourth is correct. Several changes of the algorithm still remain for future work. First, before the noun phrase is looked up in databases, a context must be identified. This would reduce the number of incorrect links. Second, to detect phrases that show to medication analyzed and to ignore them.

ACKNOWLEDGMENT

Data for this system was provided by JSC Skaitos kompiuterių servisas

REFERENCES

[1] VVKT prie LR SAM, "Įsakymas 2015 m. liepos 3 d. Nr.(1.72E)1A-755 Dėl paraiškų registruoti vaistinį preparatą, perregistruoti vaistinį preparatą, pakeisti registracijos pažymėjimo sąlygas, teisės į vaistinio preparato registraciją perleidimo, nereglamentiniam pakuotės ir (ar," 03 07 2016. [Online]. Available: https://www.e-tar.lt/portal/lt/legalAct/d6b588f0215b11e5b336e9064144f02a/fnBoVLTlBQ. [Accessed 12 11 2016].

[2] European Medicines Agency, "European Medicines Agency," 02 2019. [Online]. Available: https://www.ema.europa.eu/documents/template-form/qrd-product-information-annotated-template-english-version-10_en.pdf.

[3] S. Sahay, E. Agichtein, B. Li, E. V. Garcia and A. Ram, "Semantic Annotation and Inference for Medical Knowledge Discovery," 2007. [Online]. Available: http://www.cc.gatech.edu/faculty/ashwin/papers/er-07-16.pdf. [Accessed 16 10 2016].

[4] C. Tao, D. Song, D. Sharma and C. G. Chute, "Semantator: Semantic annotator for converting biomedical text to linked data.," Journal of Biomedical Informatics, vol. 46, no. 5, pp. 882-893. 12p., Oct2016.

[5] M. Bundschus, M. Dejori, M. Stetter, V. Tresp and H.-P. Kriegel, "Extraction of semantic biomedical relations from text using conditional random fields.," BMC Bioinformatics, vol. 9, pp. 1-14, 2008.

[6] Y. Zhang, H.-Y. Wu, J. Xu, J. Wang, S. Ergin, L. Li and H. Xu, "Leveraging syntactic and semantic graph kernels to extract pharmacokinetic drug drug interactions from biomedical literature.," BMC Systems Biology, vol. 107, pp. 323-334 12p., 8/26/2016.

[7] R. Frank, Phrase Structure Composition and Syntactic Dependencies, vol. 38, Cambridge, Mass: The MIT Press, 2002, pp. 2-27.

[8] Kaunas University of Technology and Vytautas Magnus University, "Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema," [Online]. Available: http://semantika.lt/. [Accessed 04 02 2019].

[9] A. Holvoet, Bendrosios sintaksės pagrindai, Vilnius: Vilniaus Universitetas, Asociacija „Academia Salensis", 2009.

[10] D. Jurafsky and J. H. Martin, "Formal Grammars of English," in Speech and Language Processing (2Nd Edition), JAV, Prentice-Hall, Inc., 2009, pp. 396-408.

[11] D. Šveikauskienė, "Lietuvių kalbos sintaksinė analizė," Lietuvių kalba, vol. 7, 2013.

[12] Valstybinė ligonių kasa, "TLK-10-AM / ACHI / ACS elektroninis vadovas," [Online]. Available: http://ebook.vlk.lt/e.vadovas/index.jsp. [Accessed 05 02 2019].

[13] Norwegian Institute of Public Health, "WHOCC - Structure and principles," [Online]. Available: https://www.whocc.no/atc/structure_and_principles/. [Accessed 05 02 2019].

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Mathematical Model Library for Recombinant *E.coli* Cultivation Process

Mantas Butkus
Kaunas University of Technology, Department of Automation, Kaunas, Lithuania
mantas.butkus@ktu.edu

Vytautas Galvanauskas
Kaunas University of Technology, Department of Automation, Kaunas, Lithuania
vytautas.galvanauskas@ktu.lt

***Abstract*— Biotechnological processes are among the most complicated control objects that require deep knowledge about the process. These systems have nonlinear relationships between process variables and properties that vary over time. Usually such processes are hard to model and require exceptional knowledge and experience in this field. In this review article studies conducted within the last five years in the biotechnology field, that used various model types (mechanistic models, neural networks, fuzzy models) to model cultivation processes were analyzed. Recommendations on what type of models should be used taking into account available process knowledge and experimental data were provided. Mechanistic models are best suited if there is a lack in experience in this field, advanced models like neural networks, fuzzy logics or hybrid models should be used if there is enough experimental data and process knowledge since these models tend to model the process more precisely and take in to account parameters or phenomena that cannot be described by mechanistic models.**

Keywords—biotechnological processes, neural networks, fuzzy logics, cell growth modeling.

I. Introduction

Biotechnological processes are among the most complicated control objects that are characterized by all the properties complicating control: nonlinear relationships between the process variables, dynamic properties of such processes significantly change with time, the processes lack in reliable sensors for state monitoring [13]. Therefore, development of effective control systems is a relevant bioengineering task. Most of the control systems these days rely on mathematical models that are well known but not always describe the process well or simplify the process. *E. coli* is mostly used in biotechnology, since it is well known and researched [13]. However there are no clear recommendations what kind of models should be used in different cases. In order to enrich the understanding of biotechnological modeling and selecting the best suited model the authors compiled a review on the methods used to model *E. coli* cultivation.

The aim of this article is to present various kinetic models for recombinant cultivation processes and recommendations on what kind of models to use depending on the process and gathered data. In Section **II** the process how studies were selected and analyzed is presented. In Section **III**, an explanation how, biotechnological processes are modeled and various models that have been used in the selected researches are presented. Section **IV** provides recommendations on which models should be used depending on the process knowledge and availability of experimental data.

II. Methods

This review was conducted using Google Scholar database. Google Scholar is an open access scholarly search engine that consists of full-text journal articles, books, and other scholarly documents. Even though this database has been criticized by many scholars because of its shortcomings on bibliometric purposes [15, 16], it is still one of the mainly used databases because of its broader coverage. Relevant articles were filtered out and processed according to the following rules and criteria:

- „Modeling“ is mentioned in the topic of the article.
- The article was published after 2014.
- Biotechnological cultivation processes are only analyzed.
- E.coli cultivation processes are preferred.
- Article is an open access resource.
- Article is within the first 30 pages of Google Scholar search.

In Figure 1 the structure of the selection of articles is described.

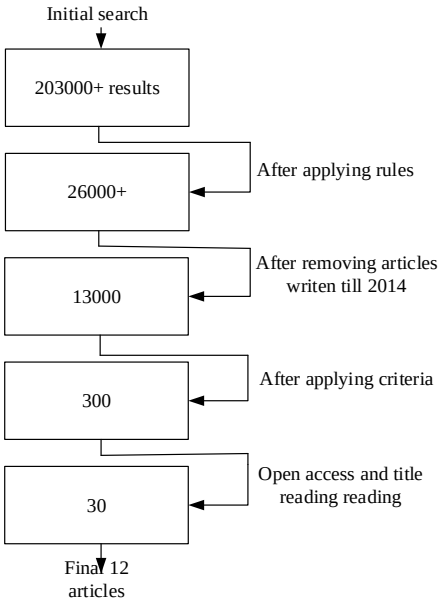


Fig. 1. Flow diagram of literature search

After selection of the relevant articles twelve studies [1-12] were selected and analyzed to determine what kind of models are used to model recombinant *E.coli* cultivation processes within the last five years.

III. BIOTECHNOLOGICAL PROCESS MODELLING

In order to model biotechnological processes, mass and energy balance equations for the modeled process should be created [13]. The balance equations are created in accordance with the mass conservation law. This means that the mass change in the bioreactor occurs due to:

- chemical reactions that occur in the bioreactor thus creating new products;
- quantity of material supplied by external material flows;
- the amount of culture medium containing the material in question is removed from the bioreactor.

The equation for mass balance of materials is described by:

$$\frac{d(C_1V)}{dt} = q_1C_2V + C_{in1}F_{in} - C_1F_{out}, \quad (1)$$

where, C_1 is the concentration of the material in the reactor, V is the volume of the medium. The amount of material in the medium will be equal to the product of these two variables. q_1 is the specific reaction rate relative to the concentration of C_2 material, in other words, this value indicates the amount of material C_1 formed per unit of mass C_2 per one-time unit. F_{in} and F_{out} are the input and output flows. The change in volume of the medium can only occur due to the flows into and out of the reactor. It can be described by the following differential equation:

$$\frac{dV}{dt} = F_{in} - F_{out}F_{out} \quad (2)$$

After the transformations of the equations (1)-(2) one gets final differential equation for the mass balance:

$$\frac{dC_1}{dt} = q_1C_2 + \frac{F_{in}}{V}(C_{in1} - C_1) \quad (3)$$

The change in concentration is not directly dependent on the outflow flow, and after taking a small sample, the concentration of the substance C1 will not change drastically, However, the outflow determines the volumetric variation of the medium, while the volume is already included in the equation.

The specific reaction rates in the previously discussed mass balance equations can be modeled by different types of models. The authors will further cover the mechanistic models of these reaction rates. The main growth indicator for microorganisms is the growth rate. For example, a new *E. coli* cell, using substrate, is generated in about 40 minutes if the temperature is 37 degrees Celsius, and some other types of bacteria divide even faster [13]. Naturally, the question is how to measure the number of cells that are formed. It is possible to estimate their number, but as the cells grow and divide, it is decided that the best way to characterize the number of cells is to determine their total mass, i.e. to calculate biomass amount. Growing biomass creates new cells that utilize nutrients and release vital products. Therefore, it is common to express these specific rates for biomass. In the 1930s, Monod described the growth of biomass at the specific rate of biomass growth, which is expressed by [13]:

$$\mu = \frac{1}{xV}\frac{d(xV)}{dt} = \frac{1}{X}\frac{dX}{dt}, \quad (4)$$

where X is the biomass amount, μ is defined as the relative increase in biomass per unit time. This quantity is not constant during the process and depends on various parameters:

- physiological state of microorganism culture,
- biomass concentration in the medium,
- concentration of substrates,
- pH of medium,
- temperature,
- pressure, etc.

The equation (4) can be used to determine the experimental biomass measurement data, but the modelling of the biomass balance equation is usually a function of certain variables. Below, the most often used kinetic models are presented.

A. Monod kinetics

Monod kinetics is the most commonly used μ relationship in biotechnological process modelling. The specific reaction rate depends on the concentration of the main substrate and is described by the formula:

$$\mu = \mu_{max}\frac{s}{K_s+s} \quad (5)$$

where μ is the specific growth rate of the microorganisms, μ_{max} is the maximum specific growth rate of the microorganisms, s is the concentration of the limiting substrate for growth, K_s is the "half-velocity constant" — the value of s when $\mu/\mu_{max} = 0.5$ μ_{max} and K_s are empirical coefficients to the Monod equation. They will differ between species and based on the ambient environmental conditions [1]. This kinetic model is usually used if the kinetics of the process is not well known. In a study conducted by Papic *et al.* [2] Monod kinetics was used since the relationship between the produced dsRNA and biomass are unknown. The results showed a 37% increase in the process productivity. Similary the Monod kinetics was used by S. Limoes [3] when modelling recombinant cellulase cultivation. In [7] several Monod kinetic models were used to model a multi-substrate environment. In all presented studies the model was sufficient and fit the experimental data.

B. Moser kinetics

Another well-known Monod kinetic modification is the model proposed by Hermann Moser [4]. Moser added another variable n, which integrates the microorganism mutation.

$$\mu = \mu_{max}\frac{s^n}{K_s+s^n} \quad (8)$$

When n is 1, this equation becomes the Monod equation. In the analysed studies [5, 6] Moser model was used to study the kinetic behaviour of the culture since the microorganism

was not well known. Results showed, that the Moser model is inferior compared with other classical kinetic models.

C. Powell kinetics

The original Monod equation was modified by Powell, introducing the terms of maintenance rate m which takes into account some of the limitations of Monod model. The Powell kinetic model is described by the equation:

$$\mu = (\mu_{max} + m)\frac{s}{K_s+s} - m \quad (10)$$

All of the described models are mostly used where no additional data from the process is gathered and are considered as "classical" models that should be used if the processes are not well known and there is not much experience gathered.

D. Blackbox and hybrid kinetics

Hybrid modelling techniques have emerged as an alternative to classical modelling techniques. Recently, these models are particularly widely used in the field of biotechnological process optimization [10,14]. Hybrid models include mechanistic models, artificial neural networks, fuzzy systems, and expert knowledge-based models into a single system, based on principled process management rules and new information. Mechanistic models are based on the application of fundamental principles and the use of certain simplistic assumptions to model phenomena in the process. Using engineering correlations, one can create different types of empirical models that describe well the nonlinear process properties. Using artificial neural networks, it is possible to successfully model functional relationships when there is a lot of measurement to identify a data model, and fundamental functional relationships between individual modeled state variables are not completely clear. In hybrid models, different parts of biotechnological processes are modelled in different ways. The main goal of modeling is to improve both process management and quality. Therefore, the aim is to model each process parameter as best as possible. Because process parameters are described in a variety of relationships, one way to model nonlinear relationships is to use artificial neural networks. An artificial neural network can be understood as a set of certain nonlinear mathematical relationships such as hyperbolic tangents, logarithmic or sigmoidal functions.

Another method, that is widely used, is the ensemble method [8]. It consists on building an ensemble of alternative models that comply with experimental observations. In particular, models with different complexity are generated and compared with respect to their ability to reproduce key features of the data. To overcome data scarcity and inaccuracies (noise), sampling-based approaches have become popular to yield surrogates for missing knowledge in parameter values [8]. In one of the studies [9] the researchers used random forest and neural networks for biomass and recombinant protein modeling in Escherichia coli fed - batch fermentations. The applicability of two machine learning methods, random forest and neural networks, for the prediction of cell dry mass and recombinant protein based on online available process parameters and two - dimensional multi - wavelength fluorescence spectroscopy was investigated. The researched models solely based on routinely measured process variables gave a satisfying prediction accuracy of about ± 4% for the cell dry mass, while additional spectroscopic information allows for an estimation of the protein concentration within ±12% [9]. These studies showed that hybrid models are capable of modeling complex biotechnological systems. According to [10] hybrid models have the following advantages over classical models:

- potentially fewer experiments required for process development and optimization;
- allow to study impact of certain variables without the execution of experiments, e.g., for the initial biomass concentration;
- may provide good extra- and interpolation properties.

E. Fuzzy logic models

An important feature of fuzzy logic is that it is possible to divide information into vague areas using non-specific sets [12]. In contrast to the classical set theory, where, according to a defined feature, the element is strictly assigned to one of the sets, the non-expressive set provides an opportunity to define a gradual transition from one set to another using membership functions. A model of fuzzy sets usually associates input and output variables by compiling if-rules such as:

> ***IF*** *the substrate concentration is low*
> ***AND*** *the specific rate of biomass growth is medium*
> ***AND*** *the concentration of dissolved oxygen is low*
> ***THEN*** *the speed of product production is medium.*

These kind of models can also be used to model the cell specific growth rate or can be used for model identification. In a study conducted by Ilkova [11] fuzzy logics were used to develop a structural and parametric identification of an *E. coli* fed-batch laboratory process. In this study the authors presented an approach for multicriteria decision making – InterCriteria Analysis to mathematical modelling of a fermentation process. It is based on the apparatus of index matrices and intuitionistic fuzzy sets. The approach for multicriteria analysis makes it possible to compare certain criteria or estimated by them objects. Basic relationships between different criteria in fed batch fermentation – biomass, substrate, oxygen and carbon dioxide were explored. This allowed to create an adequate model that was able to predict the experimental data.

In a study conducted by Liu [12] fuzzy stochastic Petri nets for modeling biotechnological systems with uncertain kinetic parameters were analyzed. In this research the authors applied fuzzy stochastic Petri nets by combining the strength of stochastic Petri nets to model stochastic systems with the strength of fuzzy sets to deal with uncertain information, taking into account the fact that in biological systems some kinetic parameters may be uncertain due to incomplete, vague or missing kinetic data, or naturally vary, e.g., between different individuals, experimental conditions, etc.. An application of fuzzy stochastic Petri nets was demonstrated. In summary, their approach is useful to integrate qualitative experimental findings into a quantitative model and to explore the system under study from the quantitative point of view. Fuzzy stochastic Petri nets provide a good means to consider parameter uncertainties in a model and to efficiently analyze how uncertain parameters affect the outputs of a model.

IV. Conclusions

After the analysis, the following recommendations can be taken into account when modeling biotechnological processes. It can be concluded, that the best suited model depends on the experience of the researcher and available measurement data:

1. If there is little experience and lack of knowledge about the process, then mechanistic models should be used to model the process and its dynamics. Monod kinetics are usually used to model biotechnological process biomass growth.
2. If there is sufficient experimental data, hybrid models that implement machine learning methods like neural networks and classical mechanistic models to model the researched process can be used, since these models consider processes parameters or dynamics that are not described or left out in mechanistic models. This type of models requires large sets of experimental data.
3. If there are experts, that have very high process knowledge, fuzzy models can be also used, since they consider atypical process behavior. By assessing the verbal knowledge of the experts complicated systems can be modeled and researched.

These recommendations can be used while deciding what kind of methods to use creating a biotechnological process model.

Acknowledgment

This research was funded by the European Regional Development Fund according to the supported activity "Research Projects Implemented by World-class Researcher Groups" under Measure No. 01.2.2-LMT-K-718.

References

[1] C. P. Jr. Grady, L. J. Harlow and R. R. Riesing, "Effects of the Growth Rate and Influent Substrate Concentration on Effluent Quality from Chemostats Containing Bacteria in Pure and Mixed Culture", Biotechnol. Bioeng., vol. 14, no. 3, pp. 391–410, 1972.

[2] L. Papić, J. Rivas, S. Toledo and J. Romero, "Double-stranded RNA production and the kinetics of recombinant *Escherichia coli* HT115 in fed-batch culture", Biotechnol. Rep., e00292, 2018

[3] S. Limoes, S. F. Rahman, S. Setyahadi and M. Gozan, "Kinetic study of *Escherichia coli* BPPTCC-EgRK2 to produce recombinant cellulase for ethanol production from oil palm empty fruit bunch", IOP Conf. Ser. Earth Environ. Sci., vol. 141, no. 1, 2018.

[4] Y. Pomerleau and M. Perrier, "Estimation of multiple specific growth rates in bioprocesses", AlChE J., vol. 36, no. 2, pp. 207–215, 1990.

[5] F. Ardestani, F. Rezvani and G. D. Najafpour, "Fermentative lactic acid production by lactobacilli: Moser and gompertz kinetic models", Journal of Food Biosciences and Technology, vol. 7, no. 2, pp. 67–74, 2017.

[6] J. C. Leyva-Díaz, L. M. Poyatos, P. Barghini, S. Gorrasi and M. Fenice, "Kinetic modeling of Shewanella baltica KB30 growth on different substrates through respirometry", Microb. Cell Fact., vol. 16, no. 1, 189, 2017.

[7] M. E. Poccia, A. J. Beccaria, R. G. Dondo, "Modeling the microbial growth of two *Escherichia coli* strains in a multi-substrate environment", Braz. J. Chem. Eng., vol. 31, no. 2, pp. 347–354, 2014.

[8] E. Vasilakou, D. Machado, A. Theorell, I. Rocha, K. Nöh, et al., "Current state and challenges for dynamic metabolic modeling", Curr. Opin. Microbiol., vol. 33, pp. 97–104, 2016.

[9] M. Melcher, T. Scharl, B. Spangl, M. Luchner, M. Cserjan, at el., "The potential of random forest and neural networks for biomass and recombinant protein modeling in *Escherichia coli* fed-batch fermentations", Biotechnol. J., vol. 10, no. 11, pp. 1770–1782, 2015.

[10] M. von Stosch, S. Davy, K. Francois, V. Galvanauskas, J. Hamelink, et al., "Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry", Biotechnol. J., vol. 9, no. 6, pp. 719–726, 2014.

[11] T. S. Ilkova and M. M. Petrov, "Intercriteria analysis for identification of *Escherichia coli* fed-batch mathematical model", J. Int. Sci. Publ.: Mater., Meth. Technol, vol. 9, pp. 598–608. (2015).

[12] F. Liu, M. Heiner and M. Yang, "Fuzzy stochastic petri nets for modeling biological systems with uncertain kinetic parameters", PLoS One, vol. 11, no. 2, e0149674, 2016.

[13] V. Galvanauskas and D. Levišauskas, "Biotechnologinių procesų modeliavimas, optimizavimas ir valdymas", Vilniaus pedagoginio universiteto leidykla, ISBN 978-9955-20-261-5, pp. 1–112, 2008.

[14] V. Galvanauskas, R. Simutis and A. Lübbert, "Hybrid process models for process optimisation, monitoring and control", Bioprocess Biosyst. Eng., vol. 26, no. 6, pp. 393–400, 2004.

[15] P. Jacsó, "Google Scholar duped and deduped–the aura of "robometrics"." *Online Information Review* 35.1pp. 154-160, 2011.

[16] I.F. Aguillo, "Is Google Scholar useful for bibliometrics? A webometric analysis." *Scientometrics* 91.2, pp. 343-351, 2012.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Lattice Based Merkle

Maksim Iavich
IT *dept. School of Technology*
Caucasus University
Tbilisi, Georgia
m.iavich@scsa.ge

Avtandil Gagnidze
Faculty of Business Management
Int. Black Sea University
Tbilisi, Georgia
gagnidzeavto@yahoo.com

Giorgi Iashvili
IT *dept. School of Technology*
Caucasus University
Tbilisi, Georgia
g.iashvili@scsa.ge

Sergiy Gnatyuk
IT Security dept. National Aviation University
Kyiv, Ukraine
sergio.gnatyuk@gmail.com

Vira Vialkova
dept.of Cyber Security
Taras Shevchenko National University
Kyiv, Ukraine
veravialkova@gmail.com

***Abstract*— Scientists are actively working on the creation of quantum computers. Quantum computers can easily solve the problem of factoring the large numbers. As the result of it quantum computers are able to break the crypto system RSA, which is used in many products. Hash based digital signatures are the alternative to RSA. These systems use cryptographic hash function. The security of these systems depends on the resistance to collisions of the hash functions that they use. The paper analyzes hash based digital signature schemes. It is shown, that hash and one way functions must be used many times during the implementation of the hash based digital signature schemes. Great attention must be paid on the security and the efficiency of these functions. Hash functions are considered to be resistant to quantum computer attacks, but the Grover algorithm allows us to achieve quadratic acceleration in the search algorithms. It means that hash functions must be complicated to be secure against quantum computers attacks. Scientists are working on determination of the cost of attacks on SHA2 and SHA3 families of hash functions. It is recommended to use lattice based constructions for one way and hash functions. Lattice based crypto systems are one of the alternatives to RSA. These crypto systems have very reliable security evidence, based on "worst-case hardness", and are resistant to attacks of quantum computers. The security of lattice based crypto system is based on the complexity of lattice problems, the main one of them is the shortest vector (SVP) problem.**

It is proposed to use the lattice-based hash function instead of the standard one, and to use lattice based one-way function as a one-way function in hash-based digital signature scheme. It is analyzed the possibility of using the family of one-way functions, suggested by Ajtai. In this paper, it is proposed to use the one-way functions offered by Ajtai and it can be considered as the initial idea. It is worth to consider the idea of using optimized one-way lattice based functions. As the result we get the secure hybrid of lattice based and hash based crypto systems, that can be used in post-quantum epoch.

Keywords*— *lattice, lattice-based crypto system, hash-based crypto system, Merkle crypto system

I. Introduction

Digital signature is a requisite of the electronic document, which is obtained by the cryptographic transformation and gives the possibility to check the true value of the information from the moment of digital signature formation. Digital signatures are very important in real life. The world leading scientists and experts are actively working on creating quantum computers. Recently it is published an article claiming that the corporations Google and NASA and Universities Space Research Association (USRA) has signed an agreement on cooperation with the producer of D-Wave quantum processors.

D-Wave 2X is the latest quantum processor, which contains 2048 physical qubits. In this model of quantum computer 1152 qubits are used to perform calculations. Each additional qubit also enlarges twice the search space so increases the speed of the calculations.

Quantum computer will be able to destroy most of all or completely all the traditional widely used cryptosystems, concretely, systems based on the integer factorization task (e.g. RSA). Some cryptosystems, such as RSA, with four thousand-bit key are considered to be safe against the classical computers attacks, but they are powerless against the quantum computer attacks. The security of digital signatures is based on the complexity of discrete algorithm solution and the large integers factorization problem. Quantum computers will easily overcome this problem and it will cause the breaking of digital signatures, implying the absolute failure.

II. Lattice based crypto systems

Lattice based crypto systems are one of the alternatives to RSA. These crypto systems have very reliable security evidence, based on "worst-case hardness", and are resistant to attacks of quantum computers. The security of lattice based crypto system is based on the complexity of lattice problems, the main one of which is the shortest vector (SVP) problem [1-3].

Lattice based one-way functions. Ajtai offered a family of one-way functions with the security based on the worst cases of approximate SVP with accuracy nt, where t is an integer[4]. Later Goldreich showed that this function is resistant to collisions, and it gives us the opportunity to use it as a hash function [5]. A lot of work is done to reduce the size of the constant and in recent works the constant is already equal to 1.

The function has parameters n, m, a and b, which are integers. The security of the function depends on the choice of *n*. In the case of hashing *m* must be greater than *nlog a/ log b*.
Matrix K from $Z^{n\times m}{}_a$ is chosen as a key. One-way function f works as follows:

$f(x) = Kx \bmod a$. The function transforms mlog b into nlog a bit. As we can see, all the arithmetic can be performed very effectively without using the precision of integers commonly used in cryptographic functions.

Hash-based crypto systems. Hash based digital signatures are also the alternative to RSA. These systems use cryptographic hash function. The security of these systems depends on the resistance to collisions of the hash functions, that they use [6,7].
One-time signatures. Lamport–Diffie one-time signature scheme.

Lamport–Diffie one-time signature scheme was offered [8].
For the signature key *X*, *2n* random lines of size *n* are generated.

$$X= (x_{n-1}[0], x_{n-1}[1], \ldots, x_0[0], x_0[1]) \in \{0,1\}^{n,2n} \quad (1)$$

Verification key $Y= (y_{n-1}[0], y_{n-1}[1], \ldots, y_0[0], y_0[1]) \in \{0,1\}^{n,2n}$

It is calculated as follows:

$$y_i[j] = f(x_i[j]),\ 0<=i<=n-1,\ j=0,1 \quad (2)$$

f – is one way function:

$$f: \{0,1\}^n \rightarrow \{0,1\}^n; \quad (3)$$

As we see, for generating *Y* the one-way function *f* is used *2n* times.

Signature of the message. To sign the message m, we hash:

$$h(m)=hash = (hash_{n-1}, \ldots, hash_0) \quad (4)$$

h- is a cryptographic hash function:

$$h: \{0,1\}^* \rightarrow \{0,1\}^n \quad (5)$$

The signature is calculated as follows:

$$sig= (x_{n-1}[hash_{n-1}], \ldots, x_0[hash_0]) \in \{0,1\}^{n,n} \quad (6)$$

The size of the signature is $\boldsymbol{n^2}$, one-way function f is not used.

Message verification. To verify the signature *sig,* the message is hashed

$$hash = (hash_{n-1}, \ldots, hash_0) \quad (7)$$

After that the following equality is verified:

$$(f(sig_{n-1}), \ldots, f(sig_0)) = (y_{n-1}[hash_{n-1}], \ldots, y_0[hash_0]) \quad (8)$$

If the equation is true, then the signature is correct.
For verification the one-way function *f* is used *n* times.

Winternitz one-time signature scheme. In the Lamport scheme key generation and signature generation are efficient, but the signature size is equal to n^2.
Winternitz one-time signature scheme is used to reduce the size [9]. In this scheme several bits of the hashed message are simultaneously signed by one line of the key.

The Winternitz parameter is the number of bits of the hashed message that will be signed simultaneously. It is chosen as *w>=2*.

After that we calculate:

$$p_1=n/w \text{ and } p_2= (\log_2 p_1 +1+w)/w,\ p= p_1+ p_2 \quad (9)$$

The signature keys are generated randomly:

$$X= (x_{p-1}[0], \ldots, x_0) \in \{0,1\}^{n,p} \quad (10)$$

The verification key is computed as:

$$Y= (y_{p-1}[0], \ldots, y_0) \in \{0,1\}^{n,p}, \text{ where}$$
$$y_i=f^{2^{w}-1}(x_i),\ 0<=i<=p-1 \quad (11)$$

Signature of the message. The lengths of the signature and the verification key are equal to ***np*** bits, one-way function *f* is used $p(2^w-1)$ times.

To be signed the message is hashed: *hash=h(m).* The minimum number of zeros is added to the hash, so that the hash would be a multiple of *w*. Afterwards it is divided into p_1 parts of size *w*.

$$\text{hash}=k_{p-1},\ldots, k_{p-p1} \quad (12)$$

The checksum:

$$c=\sum_{i=p-p1}^{p-1}(2^w-k_i) \quad (13)$$

As $c<= p_1 2^w$, the length of its binary representation is less than $\log_2 p_1 2^w+1$

The minimum number of zeros is added the binary representation, so that it would be a multiple of w, and it is divided into p_2 parts of the length *w*.

$$c=k_{p2-1},\ldots, k_0 \quad (14)$$

the message signature is calculated as follows:

$$sig=(f^{k_{p-1}}(x_{p-1}), \ldots, f^{k_0}(x_0)) \quad (15)$$

in the worst case f is used $\boldsymbol{p(2^w-1)}$ times. The size of the signature is equal to *pn*.

Signature Verification. To verify the signature $sig = (sig_{n-1}, \ldots, sig_0)$ bit string $k_{p-1}, \ldots, k_0$ are calculated.

Then the following equality is verified:

$$(f^{(2^w-1-k_{p-1})})(sig_{n-1}), \ldots, (f^{(2^w-1-k_0)})(sig_0) = y_{n-1}, \ldots y_0 \quad (16)$$

In the worst case function f must be used to verify the signature $p(2^w-1)$ times.

Comparison of Lamport and Winternitz one-time signature schemes

	Lamport	Winternitz
Use *f* to generate keys	*2n*	$p(2^w-1)$
Use *f* to calculate the signature	Is not used	$p(2^w-1)$
Use *f* to generate verify the signature	*n*	$p(2^w-1)$

Fig. 1. *Comparison of signature schemes.*

Merkle crypto-system. One time signatures are not convenient in use, because to sign each message a unique key is needed. The Merkle signature scheme allows to sign multiple messages with the same key. This system uses one-time signature and a binary tree a public key as a root.

Key generation. The size of the tree must be $H>=2$ and using one public key $2H$ documents can be signed. Signature and verification keys are generated; $Xi, Yi, 0<=i<=2H$. X_i- is the signature key, Y_i- is the verification key. Signature keys are hashed using the hash function $h:\{0,1\}^* \rightarrow \{0,1\}^n$ in order to get the leaves of the tree.

The concatenation of two previous nodes is hashed in order to get the parent node.

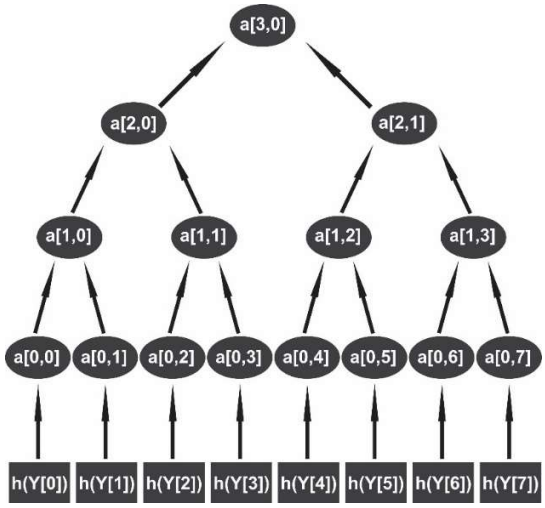


Fig. 2. Merkle tree with $H=3$.

$a[i,j]$ are the nodes of the tree;

$$a[1,0]=h(a[0,0] \,||\, a[0,1]) \quad (17)$$

The root of the tree is the public key of the signature - *pub,* 2^H pairs of signature keys must be generated in order to calculate the public *k*, and the hash function h is used $2^{H+1}-1$ times.

Message signature. A message of any size can be signed being transformed to size of *n* by means of hashing $h(m) = hash$,
An arbitrary one-time key X_{any} is used, and the signature is a concatenation of one-time signature, one-time verification key, index of a key and all fraternal nodes according to the selected arbitrary key with the index "any".

$$Signature= (sig||any|| Y_{any}||auth_0,\ldots,auth_{H-1}) \quad (18)$$

Signature verification. The one-time signature is checked using the selected verification key, if the verification is true, all the $a[i, j]$ are calculated using "auth", index "any" and Y_{any}. The signature is verified, if the root of the tree matches the public key.
The hash function in Merkle is used $2^{H+1}-1$ times, one-way function f is used $3p(2^w-1)$ times in the case of Winternitz, and $3n$ times in the case of Lamport. Hash functions are considered resistant to quantum computer attacks, but the Grover algorithm allows us to achieve quadratic acceleration in the search algorithms. It means that hash functions must be complicated to be secure against quantum computers attacks. Studies are conducted to determine the cost of attacks on SHA2 and SHA3 families of hash functions [10].

CONCLUSIONS

We propose to use the lattice-based hash function and a lattice based one-way function in hash-based digital signature schemes.

The family of one-way functions, suggested by Ajtai, can be used. As the key of hash functions, the matrix K from $Z^{n\times m}{}_a$ is selected, it transforms mlog b into $nlog$ a bits and $h(x)$ is calculated as $Kx \bmod a$.

The matrix K from $Z^{m\times m}{}_b$, is selected as the key of an one-way function,. It transforms $mlog\ b$ bits into $mlog\ b$ bits and $f(x)$ is computed as $Kx \bmod a$.

One-way functions offered by Ajtai are proposed in the paper and it can be considered as the initial idea. It is worth considering the idea of using optimized one-way lattice based functions.

ACKNOWLEDGEMENT

The work was conducted as a part of joint project of Shota Rustaveli National Science Foundation of Georgia and Science & Technology Center in Ukraine [№ STCU-2016-08]

effectively without using the precision of integers commonly used in cryptographic functions.

REFERENCES

[1]. Güneysu T., Lyubashevsky V., Pöppelmann T. (2012) Practical Lattice-Based Cryptography: A signature scheme for embedded systems. *Lecture notes in computer Sci.*, 7428: 530-547, Springer.

[2]. Akinyele, J.A., Garman, C., Miers, I. et al. (2013) Charm: a framework for rapidly prototyping cryptosystems. *Journal of cryptographic engineering*, 3. Springer: 111-128

[3]. Gagnidze A., Iavich M., Iashvili G., (2017) Analysis of post quantum cryptography use in practice. *Bulletin of the Georgian National Academy of Sciences*, 2, 12: 29-36

[4]. Ajtai, M.: Generating hard instances of lattice problems. In Complexity of computations and proofs, volume 13 of Quad. Mat., pages 1–32. Dept. Math., Seconda Univ. Napoli, Caserta (2004). Preliminary version in STOC 1996. 8. Babai, L.: On Lovász lattice reduction and the nearest lattice point problem. Combinatorica, 6:1–13 (1986).

[5]. Goldreich, O., Goldwasser, S., and Halevi, S.: Collision-free hashing from lattice problems. Technical Report TR96-056, Electronic Colloquium on Computational Complexity (ECCC) (1996).

[6]. Bernstein D.J., Buchmann J., Dahmen E., (2009) *Book: Introduction to post-quantum cryptography*, Springer.

[7]. Gagnidze A, Iavich M., Iashvili G., (2016) Some aspects of post-quantum cryptosystems. *Eurasian journal of business and management*, 5, 1: 16-20

[8]. Lamport, L.: Constructing digital signatures from a one way function. Technical Report SRI-CSL-98, SRI International Computer Science Laboratory, 1979.

[9]. Merkle, R.C.: A certified digital signature. Advances in Cryptology - CRYPTO ’89 Proceedings, LNCS 435, pages 218–238, Springer, 1989.

[10]. Amy M., Di Matteo O., Gheorghiu V., Mosca M., Parent A., Schanck J. (2017) Estimating the cost of generic quantum pre-image attacks on SHA-2 and SHA-3. *Lecture notes in computer science*, 10532: 10-31, Springer.

INFORMACINËS TECHNOLOGIJOS • IVUS 2019 • ISSN 2029-249X • eISSN 2029-4824

Context Based Number Normalization using Skip-Chain Conditional Random Fields

Linas Balčiūnas
Vytautas Magnus University
Kaunas University of Technology
Kaunas, Lithuania
linasb20@gmail.com

***Abstract*—Verbalizing numeric text tokens is a required task for various speech related applications, including automatic speech recognition and text-to-speech synthesis. In morphologically rich languages, such conversion involves predicting implicit morphological properties of a corresponding numeral. In this paper, we propose first-order skip-chain Conditional Random Field (CRF) models and various prepossessing techniques to leverage different contextual information. We show that our best skip-chain CRF models achieve over 80% accuracy on the set of 2000 Lithuanian sentences.**

I. INTRODUCTION

Number normalization is a task of replacing numeric tokens in a sentence by numerals (word tokens) using an appropriate inflected form of a numeral. Number normalization usually involves disambiguation as the same numeric token needs to be mapped into different word forms depending on the context (e.g. '5 vaikai eina' ↦ '*Penki* vaikai eina' (five children are going) vs. '5 vaikų nėra' ↦ '*Penkių* vaikų nėra' (five children are missing)). Although number normalization can be considered as a part of a broader task of text normalization, formulating it as separate task might be beneficial, since the process of number normalization can be quite complex depending on morphological features of a language. In this paper, we describe the process of building and evaluating a number normalization system for Lithuanian. However, some techniques and models are language-independent and might be applied for other languages. In Lithuanian, for example, number '5' depending on sentence context may represent any of 63 different words. Predicting such relationship directly is rather difficult and it would require huge data-set to properly learn Numeral grammar. Simpler approach is predicting Part Of Speech (POS) tag and then generating numerals accordingly. POS tag contains all necessary morphological information to use language-specific grammar based numbers-to-words system [1]. This way possible result classes are shared across all numbers and predicting POS tag can be formulated as sequence labeling, rather than sequence-to-sequence task, because of the one-to-one relationship.

II. RELATED WORK

As far as we know, there are no published works or publicly available applications performing number normalization based on the sentence context for Lithuanian. Although, many languages deal with similar morphological disambiguation problems. Russian and Lithuanian numbers share many morphological properties, including types (cardinal, ordinal), genders and cases. There are existing research and systems for Russian language on general text normalization, hand-written language-general grammar [2], Recurrent Neural Network (RNN) [3], and number normalization [4].

III. PREPROCESSING

A. Data

A small text corpus for training and evaluating text normalization models was collected and manually annotated. It consists of 1955 sentences containing 3143 numeric tokens. Sentences were inspected by linguists who suggested a numeral word form as an approriate replacement for every numeric token. In some ambiguous cases a few reasonable alternatives were proposed. Some ambiguities were related to the use of the pronominal numeral forms (e.g. '15 savaitė' (15th week) ↦ '*penkiolikta* savaitė' (non-pronominal form) or '*penkioliktoji* savaitė' (pronominal form)). Other ambiguities were related to numeral case (e.g. '2019 vasarį' (2019 February) ↦ '*du tūkstančiai devynioliktųjų* vasarį' or '*du tūkstančiai devynioliktaisiais* vasarį'). All suffixes that represented a 'normalisation hint' were eliminated from the data set (e.g. '2019-aisiais' was replaced by '2019'). This had an effect of making training subset of the corpus more interesting for the training algorithm, increased the complexity of the normalization task, and reduced the normalization accuracy estimates on the test subset of the corpus.

Sentences of the corpus were pre-processed by the Hidden Markov Models (HMM) based POS tagger [5]. Every text token was labelled with the so called 'detailed' (or composite) morphological label that contained the following information:

- Lemma
- Part of speech (Noun, Verb, Adjective,...)
- Case (Nominative, Genitive,...)
- Gender (Feminine, Masculine)
- Number (Singular, Plural)

Since important prediction decisions are based on tagger provided POS tags and lemmas, to ensure optimal performance, morphological annotations were hand-corrected in training-testing data-set. When using morphological analysis

data, it is beneficial to divide POS tags into sub-labels to build more abstract grammar rules and filter out redundant information.

B. Number grammar

Similar more saturated Natural Language Processing (NLP) sequence labeling tasks, POS tagging and Named Entity Recognition (NER), does not require hand-written language-specific grammar rules to achieve state-of-the-art [6] performance. Long-Short Term Memory neural networks coupled with Conditional Random Fields (LSTM-CRF) based data-driven sequence labeling approaches proves to be insufficient to achieve desired number normalization quality, considering training-data availability limitations. To efficiently leverage small data-set, morphosyntactic knowledge should be exploited for crafting language and task specific grammar rules.

All rules are constructed as conditional functions without any prior weighting. Here are different techniques used to approximate and generalize number relationship with sentence context:

- Lemma classification. (Replacing certain word lemmas with dedicated class name, for example, month names with '%Month')
- Number classification. (See Table I)
- Verb classification. (Based on syntactic features of controlling case of other POS)
- Syntactic linking (See Section V. Long-Distance Dependencies)

TABLE I
NUMBER CLASSIFICATION EXAMPLE

Num.	Roman	Int	Digit count	Req.Gen.*	Req.Sing.*
21	-	+	2	-	+
113.5	-	-	3	+	-
IV	+	+	1	-	-

*Requires Genitive and Requires singular signifies that countable noun of certain number must be of Genitive case or Singular

IV. MODELS

In this paper, mainly variations of Conditional Random Fields (CRF) are explored, since CRF got better baseline performance than its neural version (LSTM-CRF) and appears to be more suitable for particular data-set and grammar rule-set.

A. Sub-Label models

To create single sequence tagging model for this task, we would need 79 (different combinations of sub-labels shown in Table II) detailed morphological labels corresponding to the output classes. With the currently available corpus this would cause significant data scarcity problems. There are no training examples for a considerable number of classes and many others are barely represented. Good way to address this data scarcity problem is creating three independent CRF models for Case, Type and Gender prediction and to combine these predictions at a later stage. This is preferable since there are no direct dependency relationship between these morphological categories and operating with sub-labels allows creating more abstract rule-set. Although, it is worth noting that sub-label dependencies have been proven useful for NLP sequence labeling using CRF [7] in combination with composite labels. Additionally exploiting composite label dependencies might be beneficial for number normalization as well, and it is worth exploring in future research.

TABLE II
MORPHOLOGICAL SUB-LABELS

Case	Type/Number	Gender
Nominative	Cardinal	Feminine
Genitive	Ordinal singular	Masculine
Dative	Ordinal plural	Not applicable
Accusative	Ordinal definitive singular	
Instrumental	Ordinal definitive plural	
Locative	Cardinal multiple	
Not applicable	Month*	

*This class is designed for number that could be substituted with month name, for example '2019-02-03' and '2019-February-03'.

B. Skip-Chain CRF

Linear-chain structure is usually used for sequence labeling CRF, since additionally modeling non-linear relationships requires complicated inference algorithms and prior specification of such dependencies [8], [9]. For number normalization, simplified version of Skip-chain Conditional Random Field (S-CRF) can be used, as shown in gender prediction model comparison in Figure 1 and Figure 2 (for readability reasons, we only show English glossary sentence example of Lithuanian model). Both graphs are representations of Viterbi algorithm decoding (same structure is used for encoding). Circles correspond to nodes and arrows to transitions. Weight of node or transition is calculated as sum of its conditional feature-set weights (unigrams for node and bigrams for transition). 'f' and 'm' are notations for 'feminine' and 'masculine' genders, while '0' represents a class for non-number tokens, which are not to be changed by normalization task. Blue path is the correct path selected by Viterbi algorithm. In Linear-chain CRF (L-CRF) most bigram features are useless, since they connect with non-number tokens (in Figure 1 none of transition weights are significant). This means we effectively have zeroth-order CRF. Transitions that are actually important are between numbers. To implement such dependencies we make two sequences - full (original) and skip (numbers only). We build unigram features only for number tokens, but from full sequence. This way our unigram features exactly match those of linear-chain model. Next, bigram features are built from skip sequence. For encoding and decoding we use skip sequence as well, since we do not build any feature functions for non-number tokens. Skip-chain models, as described above, have unaltered unigram and improved bigram function sets (for number tokens), while being significantly faster (see graph simplification shown in Figure 1 and Figure 2). Our

implementation uses modified version of CRFSharp toolkit [10].

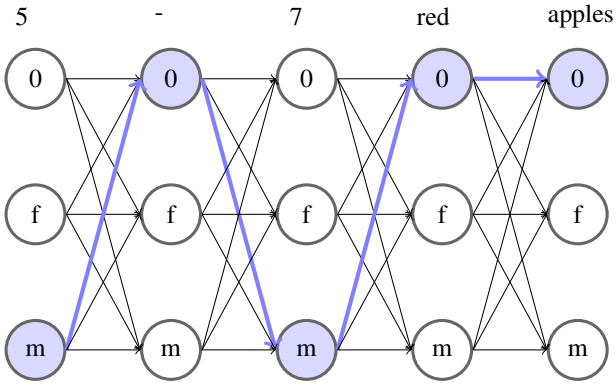


Fig. 1. First-Order Linear-Chain CRF

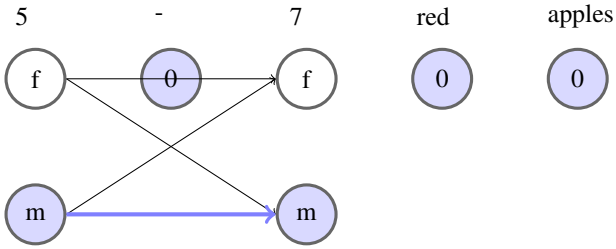


Fig. 2. First-Order Skip-Chain CRF

V. Long-Distance Dependencies

Although skip-chain structure quite reliably models some important long-distance relationships, it is not able to capture distant dependencies between number and non-number tokens (e.g. in '3 didžiųjų mobiliojo ryšio operatorių' $\mapsto$ '*trijų* mobiliojo ryšio operatorių' (three major mobile network operators) numbers '3' case is determined by words 'operatorių' (operators) case). CRF is generally unable to leverage such features, and requires either hybridization such as LSTM-CRF, or additional pre-processing. We propose identifying position-distance independent relationships using an ad-hoc set of linkage rules and formulating perceived syntactic links as conditional functions of CRF. In case of Lithuanian, we discern three directly related parts of speech (Noun, Verb, Preposition) in numeral normalization task. For each, we use different set of linkage rules, to identify related tokens to every number in sentence, effectively performing partial syntactic analysis. To link prepositions and verbs to numbers, our rules solely rely on morphological labels provided by the POS tagger. For nouns, the task of linking could be more precisely formulated as an identification of a noun which represents an object or quantity being counted by some number in the sentence. This is extremely important, since countable noun has crucial morphological information. For example, with successful identification we no longer need to predict the gender of a numeral, since it is directly determined by the noun.

A. Countable noun identification

We determine the most likely countable noun in a two-step process. First, for a given numeric token d we select all potentially countable noun tokens $\{n_i\}$ according to the ad-hoc set of linkage rules for nouns. We can not make an educated choice among selected nouns on the basis of available morphosyntactic annotation, since noun morphology does not have the property of 'countability'. To discriminate among potential countable nouns, semantic analysis is needed. We need to rate the set of selected nouns $\{n_i\}$ according to some 'countability' measure ξ that is dependent on the numeric token d being normalized and select the noun n_{best} with the highest $\xi(d, n_i)$ rating $n_{best} = \arg\max \xi(d, n_i)$

Suppose that we have vector embeddings $v(n), v(n) \in R_D$ for every noun n, that were obtained by an algorithm such as 'word2vec' [11]. Suppose that we also designed a mapping ϕ that maps every numeric token d into a vector $\phi(d), \phi(d) \in R_D$ such that $\phi(d)$ is representative embedding of the set of nouns that are frequently counted by the numeric token d. If both assumptions are correct, we can rate the set of potential countable nouns by estimating cosine similarity between each selected noun and the corresponding representative vectors, i.e.

$$\xi(d, n_i) = \text{cosine-similarity}(\phi(d), v(n_i)) \qquad (1)$$

We have tested a few different approaches to design the above mentioned mapping $\phi(d)$. We sought large unannotated text corpus for number and noun adjacent co-occurrences and made noun frequency lists per every numeric token that was found (around 350 thousand co-occurrences). Information present in a frequency list can be aggregated into a single vector by estimating the weighted average of noun embeddings making up that list. Thus a representative (or central) embedding vector can be obtained per every numeric token. Although this tabular mapping from numeric tokens into representative vectors can be used in (1), it has serious limitations. The table contains many unreliable vectors for rare numbers, because of lack of co-occurrences in unannotated corpus. To circumvent limitations of this tabular mapping we used Neural Network (NN) approach to build continuous co-occurrence model. We built two different neural networks: one with a single input (corresponding to the mathematical value of the numeric token) and one with 7 inputs, corresponding to the decomposition of the numeric token into sub-parts (thousands, hundreds,...) and including number features similar to Table I. NN had 200 output units.

Evaluation of these models are shown in Table III. The baseline performance is obtained by the simple rule "take the first potentially countable noun to the right of a numeric token". Accuracy is measured using whole CRF training data, extracting situations where choice between two or more nouns (2.41 avg.) is needed.

TABLE III
COUNTABLE NOUN LINKING

Method	Accuracy
Select first	68.77
1-input NN	84.11
7-input deep NN	87.40

VI. EVALUATION

We evaluate models with 5-fold cross-validation (except for countable noun identification in Section V, since training and testing data-sets were obtained from different sources). Accuracy of different models are shown in Table IV. Combined accuracy estimates accuracy of all three models. The combined answer is considered to be correct if all three sub-labels are correct.

It is worth noting, that our model is focused on grammatically correct, as 'spoken' number normalization. This might not be desirable for systems like text-to-speech synthesis, hence more standardized approach can be chosen. For Lithuanian language, numeral definiteness property could be removed from prediction model, since it is not strictly constrained by grammar. This would increase language correctness and improve Type prediction models and combined accuracy as shown last line of Table IV (best performing model without definiteness property).

Accuracies above represent the lower bound accuracies of the real-world number normalization performance. Firstly, in certain situations some sub-label prediction mistakes might be irrelevant for numeral generation. For example, both '5, Cardinal, Genitive, Feminine' and '5, Cardinal, Genitive, Masculine' will generate same word representation 'penkių'. Secondly, real-world sentences often contain suffixes (e.g. 'Kovo 11-ąją' $\mapsto$ 'Kovo *vienuoliktąją*' (March 11th)) that either offer an unambiguous hint that solves the number normalization problem or at least provides most of the needed morphological information, which can be used to correct prediction mistakes.

TABLE IV
EVALUATION

	Case	Type	Gender	Combined
L-CRF	77.19	89.00	94.79	67.52
S-CRF	78.89	89.82	95.17	69.43
S-CRF+class.*	83.81	93.64	95.17	76.49
S-CRF+class.+syn.**	86.05	94.01	98.51	80.91
without Definiteness	86.05	96.82	98.51	83.08

*classification, see Section III. Preprocessing
**syntactic analysis, see Section V. Long-distance dependencies

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we describe number normalization disambiguation model, which is needed to develop context dependant number-to-words system. Sequence-labeling approach allows us to normalize countable abbreviations and symbols (next to number) effortlessly, since countable noun morphological form can be extracted from predicted label (e.g. 'nuo 5%' (from 5%) $\mapsto$ 'nuo *penkių procentų*'). Our implementation based on this model is publicly available [12] and in future will be integrated into full Lithuanian text normalization system.

Number normalization errors are often directly dependant on morphological analysis mistakes and we are currently working on improving both vocabulary-grammar and disambiguation sides of Lithuanian POS tagging to consequentially increase number normalization accuracy.

Currently we use 'word2vec' [11] algorithm trained on relatively small text corpus to produce word embeddings. Although, various improvements have been made in encoding text semantic information to vectors [13], [14] and using more advanced method and larger corpus would likely improve our model performance.

Our achieved number normalization accuracy could be further improved by expanding annotated training data, since considerable amount of errors are direct result of data scarcity. Although, our approach generally lacks in semantic and syntactic language understanding, so performing full syntactic sentence analysis in preprocessing stage would be highly beneficial.

ACKNOWLEDGMENT

This research was supported by the project "Semantika 2" (No. 02.3.1-CPVA-V-527-01-0002). Special gratitude goes to our colleagues Lina Majauskaitė and Dovilė Stukaitė who helped us in collecting and annotating text corpus.

REFERENCES

[1] Virginijus Dadurkevičius. *dadurka/number-to-words-lt*. URL: https://github.com/dadurka/number-to-words-lt.

[2] Ke Wu, Kyle Gorman, and Richard Sproat. *Minimally Supervised Written-to-Spoken Text Normalization*. 2016. arXiv: 1609.06649.

[3] Richard Sproat and Navdeep Jaitly. *RNN Approaches to Text Normalization: A Challenge*. 2016. arXiv: 1611.00068.

[4] Kyle Gorman and Richard Sproat. "Minimally Supervised Number Normalization". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 507–519. URL: https://www.transacl.org/ojs/index.php/tacl/article/view/897/213.

[5] URL: http://donelaitis.vdu.lt/main_helper.php?id=4&nr=7_2.

[6] Alan Akbik, Duncan Blythe, and Roland Vollgraf. "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1638–1649. URL: http://aclweb.org/anthology/C18-1139.

[7] Miikka Silfverberg et al. "Part-of-Speech Tagging using Conditional Random Fields: Exploiting Sub-Label Dependencies for Improved Accuracy". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 259–264. DOI: 10.3115/v1/P14-2043. URL: http://aclweb.org/anthology/P14-2043.

[8] Michel Galley. "A Skip-chain Conditional Random Field for Ranking Meeting Utterances by Importance". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 364–372. ISBN: 1-932432-73-6. URL: http://dl.acm.org/citation.cfm?id=1610075.1610126.

[9] Jingchen Liu, Minlie Huang, and Xiaoyan Zhu. "Recognizing Biomedical Named Entities Using Skip-chain Conditional Random Fields". In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. BioNLP '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 10–18. ISBN: 978-1-932432-73-2. URL: http://dl.acm.org/citation.cfm?id=1869961.1869963.

[10] Zhongkai Fu. *zhongkaifu/CRFSharp*. URL: https://github.com/zhongkaifu/CRFSharp.

[11] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781 (2013). URL: http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

[12] URL: http://prn509.vdu.lt:9080/.

[13] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation". In: *In EMNLP*. 2014.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Improvement and Digitalization of Business Processes in Small-Medium Enterprises

Matas Dumčius
Department of Information Systems, Faculty of Informatics
Kaunas University of Technology
Studentų 50, Kaunas, Lithuania
matas.dumcius@ktu.edu

Tomas Skersys
Department of Information Systems, Faculty of Informatics
Kaunas University of Technology
Studentų 50-313a, Kaunas, Lithuania
tomas.skersys@ktu.lt

***Abstract*— As small-medium enterprises (SMEs) produce huge amounts of added value in the market, it is important to ensure that their business processes are optimized. Business Process Management (BPM) methodology is popular in large enterprises, however, there is lack of information on how it can be adapted by SMEs considering the financial and informational constraints they face. In this research, a set of qualitative and quantitative analysis methods for improving the quality of business process models are presented and applied to optimize the order management business process of a small Lithuanian optical retail business company. Further, a business process management system (BPMS) is presented to digitalize and support the execution of redesigned process. A brief discussion of applying process quality analysis methods within the selected business domain is presented. It is expected that our research will make BPM initiatives more feasible for businesses of similar type.**

***Keywords*— business process improvement, business process management, business process digitalization, business process management system, small-medium enterprise, BPMS, BPMN, BPM, quality management.**

I. Introduction

Small-medium enterprises are defined as companies with staff count of up to 250 employees or turnover lover than 50 million € or balance sheet total less than 43 million € by the European Commission. Such businesses produce vast majority of added value in the economics of European Union. As stated in the annual report on European SMEs 2017/2018 these kind of businesses accounted for 47% of the increase in the value added and 52% of the cumulative increase in employment in non-financial sector since the global financial crisis [1]. Most of the SMEs are working in the state of very limited resources, accessibility to information and experience in process management and automation. But these fields are of the highest importance when considering that the main goal of business is related to creating value through activity while utilizing resources in the most efficient way. Optimization and documentation of SMEs processes are important as efficient and effective process delivery is a key to a long-term success, business competitiveness, growth and viability [2],[3]. One of the most common ways of improving business processes running in one or across many enterprises is by implementing a Business Process Management (BPM) methodology.

BPM is a widely spread methodology, which concentrates on business processes and aims to improve their quality, efficiency, compliance, customer integration, employee engagement and agility [4], [2]. Arguably, the ultimate goal of BPM is a digitalized enterprise running on optimal business processes throughout its whole life cycle. Even though traditionally most of the successful BPM initiatives are associated with very large organizations, it has no restrictions on being implemented within the working environment of SMEs whatsoever. However, considering the financial and informational environment in which SMEs are working, the effectiveness of this methodology might be unpredictable, and the results might be even detrimental in case of straightforward BPM application in a small-medium sized organization. In order to get positive outcome of BPM initiative in SME, every step in BPM lifecycle should be analyzed in the context of such business.

The main task of this is to evaluate different types of qualitative process evaluation and optimization methods of small optical retail business order management process. Then test the redesigned processes by using business process simulation tool in order to get the best process optimization result. Such evaluation and example of process optimization might help to apply BPM methodology in SMEs process redesign phase easier for other businesses of the same type.

II. Basic Definitions and Related Work

A. Definition of Business Process Management

Business Process Management does not have one specific definition. Generally, it is a methodology which aims at improving organization performance by concentrating on business processes. BPM combines the business management, quality control and information technologies traditions. The origin and composition of BPM can be seen in Fig. 1.

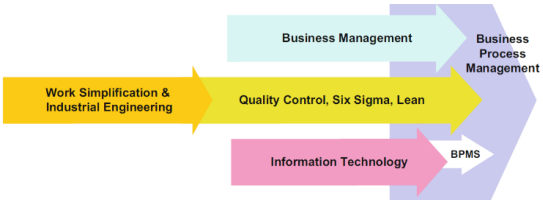


Fig. 1. Approaches of business process change [5]

BPM initiatives are carried out by following a predefined set of stages – a BPM lifecycle (Fig. 2). At first – a business problem is defined and a target process is selected. Then the current state of the process is documented (usually in a form of process model). When a process model is present qualitative and quantitative analysis are carried out so that any improvement could be measured. Then the process is redesigned according to the initially stated issues and a new

process model is created. This process model serves a basis for the next stage – process implementation. The changes are realized and the process is moved to the new state so that performance goals could be achieved. Usually management changes and process automation via development of information systems are implemented. Once changes are completed, process execution data is collected, analyzed, new problems or goals are defined and respective actions are taken.

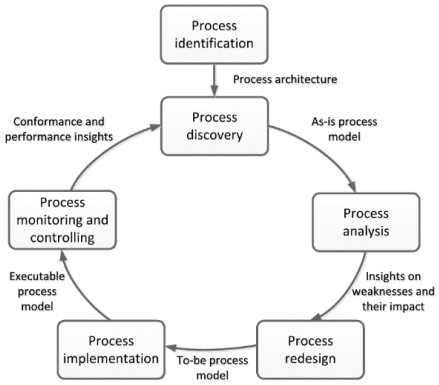


Fig. 2. BPM lifecycle [6]

B. Quality of the process

Quality is one of the most important parts through the whole BPM lifecycle. In this case we are talking not about the results of the process – quality of deliverables such as products or services – but about execution properties of the process. Quality models are defined as easily understandable and changeable models with little to no design errors. A modeled process reveals only activities and decisions, but no information can be retrieved directly about its quality metrics [12]. For this reason, there are numerous qualitative and quantitative process analysis methods which help to find bottlenecks of the process, sources of arising problems, compare different process models and come up with an optimized process model. But there is small amount of information about application of these methods in the scope of small-medium sized business.

C. Process model optmization mehods

Qualitative process analysis mostly helps to identify redundant or weak parts of the process. In the next section we will summarize some methods of qualitative process analysis.

Value-Added Analysis – the main goal of such analysis is to remove non-value adding tasks from the process model. The process is decomposed to the simplest tasks requiring one action of one process participant. Then those tasks are assigned one of three categories:

- Value-adding – task that contributes to the final product or service,
- Business value-adding – task that is necessary for business to be running,
- Non-Value adding – all remaining tasks.

By removing the non-value adding tasks we should be utilizing resources in a more efficient way.

Root Cause Analysis – this type of analysis is mostly conducted in manufacturing companies to find out the reason behind various incidents or defects [13]. Though this method might be adapted for thorough problem identification and analysis in business process model with an intention of optimizing it [14]. First step of such analysis is defining perspectives in which stated problems will be explored. 6M model (machine, method, material, man, measurement, milieu), some perspectives from Six Sigma methodology [15] or any other model might be applied. Then each stated problem is analyzed in every perspective while searching for the root cause. This type of analysis does not require a full process model, but the results of it might help redesign ongoing process.

Impact assessment and issue documentation is type of analysis that typically follows up root cause analysis. As mentioned before, root cause analysis defines problems and their cause. But it does not point out their impact to the whole process, so there is no formal way to prioritize them. One suggestion is to create issue register, in which each problem would have an impact assessment (qualitative or quantitative) such as impact on time, finances or any other metric [6]. If metrics are defined, we can conduct Pareto analysis. In practice Pareto analysis makes an assumption that 20% of problems make 80% of impact [16]. Of course, if there are only few problems stated in the issue register, this type of analysis is unnecessary.

All mentioned types of qualitative analysis help process analytics to identify existing problems and redesign process models. But these methods do not specify any information on how process model redesign could be evaluated, or process models could be compared. In such cases quantitative analysis must be performed to get the required data.

Qualitative analysis may be conducted in three different categories: analysis of process model metrics which are derived from software engineering, theoretical process execution analysis and process simulation.

Software quality quantitative metrics can be applied in business process model evaluation because of huge similarity between processes and software – they both process data, have a structure and are based on a static model [12], [17]. Calculating such metrics as model coupling, cohesion, complexity, modularity or size could give good indications about the quality of a process model. Process model metrics are valuable for process analysts, but process stakeholders are usually interested in execution metrics such as execution price, duration, quality of results and model flexibility. All these parameters can be calculated by performing Flow analysis or applying Queuing theory [6]. These methods can give valuable data, but they are hardly applicable in real world process models. Flow analysis can be conducted only when process models are simple, have only exclusive or parallel gateways and Queuing theory can only give data for one activity. Also Queuing theory calculations are complicated even for simple situations. For those reasons the most practical solution is process model simulation. The process model simulation can be done by using various tools such as *IBM WebSphere Business Modeler*, *ITP Commerce Process Modeler for Visio*, *ProSim* or an open-source solution – *BIMP* simulator [20]. Process simulation software instantiates huge amount of hypothetic process instances and records properties of each execution. Only process execution data, such as probabilities of various decisions and duration of activities are required as input parameters. This type of qualitative analysis allows process analysts to easily compare different process

models in different execution environments. Although process analysis is one step in BPM lifecycle and without process digitalization it creates small part of the BPM added value for the business.

Application of BPM is popular among large enterprises, but small-medium sized businesses differ in terms of available resources, process relation, work ethics and the speed of decision making [7]. Thus the application of BPM in SMEs must be investigated.

D. Case studies of BPM application in SMEs

A case study was conducted in three different SMEs in Belgium by C. Bauwens and T. Van Dorpe. They state that the maturity level of an SME must be assessed to understand where the organization is with its BPM development [8]. Hammer's Process and Enterprise Maturity Model [9] and McCormack's Business Process Orientation Maturity Model [10] are used to access SMEs under research. Authors conclude that SMEs have rather low maturity level, they are pointing to weak spots of the small businesses such as lack of documentation and limited inner efforts to process improvements. As the recommendations of improvement goals are provided, no details of how to improve process models is provided.

Another case study conducted within Australian Small Business by I. Dallas and M. T. Wynn goes through steps of BPM initiative and gives implications on what could be improved in the methodology and observations which could help other SMEs [2]. Though process evaluation and optimization get very little attention as the analyzed business was under establishment. Only some advantages of Business Process Management Systems (BPMS) such as automatic work allocation are introduced as main benefits for SMEs. The implementation of BPMS systems in small businesses is also widely discussed in the work of Veldhuizen R., Ravesteijn P. and Versendaal J. They distinguish main differences of SMEs and large enterprises and suggest an adapted BPMS implementation model [11].

As we can see from conducted case studies, despite that BPM is a process improvement methodology, there is almost no research in the area of a process evaluation and optimization for SMEs. Considering that it is one of the early stages of BPM lifecycle and BPMS creation stages, effective process optimization might save both time and financial resources of such companies.

III. PROCESS OPTIMIZATION IN A SMALL OPTICAL RETAIL BUSINESS

A. Optical retail business order management process

The case organization is an optical retail branch business, which have optical-shops located across Lithuania and is in the market since 1997. It fits all small business parameters defined in section 1. Only process of order management of prescription glasses production will be optimized in the scope of this research. Order management is quite complicated in this type of business. It always must adapt to dynamic market, new technologies and products introduced in the field and finally to always changing systems of suppliers. For this reason, rigid, hard-coded or universal off-the-shelf solutions are usually not suitable or are too expensive to deploy and support an ongoing process. The current (as-is) order management process does not have a supporting information system.

In order to manage the complexity of the process, it was segmented into three sub-sequential sub-processes (Fig. 1.) based on a three level SCOR model [18] – order initiation, order production and order completion (Fig. 1). Each of this subprocess was modeled separately by using BPMN 2.0 modeling language and Camunda Modeler tool [19].

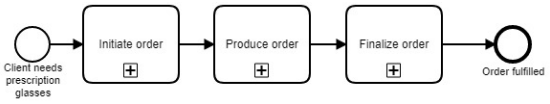


Fig. 3. Top level order management process model

Order initiation subprocess in target business is not documented or formalized, its specific order of execution is defined by optics sales assistant at the order initiation time for each instance separately. Input of this subprocess is client needs and output – a filled order. Order initiation subprocess is explicitly presented in Appendix A, which we think is enough to show the overall complexity of the underlying business logic of the analyzed business domain. Order production subprocess consists of order manufacturing internally or externally. Tasks related to order data sending to manufacturers, ordering lenses, sending order to production sites error management and quality control procedures. Order completion summarizes notifying client, handing finished order to a client, receiving final payment and generating invoices if clients ask. In the scope of order management in optical retail business there were 12 process models created in total.

As this process is executed the target business is facing following problems:

- The states of the orders are not tracked.
- Order fulfillment date often passes due date.
- Late order data retrieval and inaccuracies in it.
- There is no responsible person for each order.
- Delays in notifying clients.
- Sometimes not all required documentation is filled by the sales assistants.
- Long duration of changes implementation.
- Close to no control in order management by managers and other business authorities.

B. Process optimization

Creating an order management information system based on an unoptimized process would be inefficient. Considering the size of the optical retail company process optimization must be done at minimum cost. For these reasons qualitative and quantitative analysis will be performed and there will be a brief discussion about applicability of each method in SME under research. As mentioned in [2], the qualitative analysis in a small business company might be rejected or seen as redundant. Though without it, it is close to impossible to conduct process optimization.

1) Value-Added Analysis

As mention in section 2, the main task of this analysis is elimination of non-value-adding tasks. The presented order management process was broken down to a task list of 71 basic task. Each of the task was then given an assignee and a

category whether it was value-adding, business value-adding or non-value-adding task. The analysis was conducted with a supervision of the company CEO. 11 non-value-adding tasks were identified – most of them related to manual data entry tasks, filling of different forms, work related to a not unified process throughout the company. A plan for each of this task was made – either it was to be automated or eliminated. There were 31 business value-adding tasks which were also revised and if possible planned to automate by introducing business process management system.

2) Root Cause Analysis

During this analysis only problems stated in presentation of current order management process were evaluated using 6M perspective model presented in section 2. CEO of the company was involved in all stages of analysis. From the obtained results we can see that most of the problems in order management were arising from the way the process is executed (method), technical (machine) and human (man) factors. For readability analysis of each problem was depicted with cause-effect (Ishikawa) diagrams. Considering each problem, process model was revised, and a solution was suggested. After the analysis process was remodeled, a business process management system and knowledge system were introduced.

Impact assessment and issue register was not created because as defined in section 2, in the context of SMEs with relatively small amount of problems this method is excessive.

C. To-be model, simultaion results.

The main reason behind qualitative analysis was to find out how effective qualitative analysis is in terms of process execution properties such as execution price, duration and resources utilization. Qualitative analysis was conducted by simulating process models. Process model metrics and theoretical models were not applied because of lack of information and technologies that could be used in SME. As-is and to-be process models were simulated using opensource process model simulator BIMP [20]. Two-year orders historical data was used as input parameters for this analysis. By making the process unified and removing non-value adding tasks we have made the process 12.9% shorter in duration. The remodeled process included more tasks concerning quality control of produced prescription glasses but overall still was 4% shorter in duration and most importantly it reduced resource utilization by 15.9%. On the price point, average execution cost increased by marginal 1.32€, this was probably a result of so called devil's quadrangle [21] – by improving process quality and speed, we have increased its execution price. The optimization results on process flexibility were not tested. Though considering relatively small rate of process initiations in optical retail business flexibility might be more linked to the ability of changing process model than adapting to increased amounts of process instances.

D. Process model digitalization and automation by introducing Business Process Management System

A process quality analysis alone can have little to no impact on the execution of an as-is process in a company. Especially, when a process is remodeled with a supporting information system in mind. In case of analyzed small optical retail business order management process, a business process management system is presented which supports to-be process model. Prototype of this system was developed on an open-source Camunda BPMS platform in order to get highest amount of added-value from BPM initiative and process optimization. Two order management sub-process were completely digitalized. The main advantages of such system identified by the business stakeholders after the presentation of the solution are as follow:

- *Process execution based on an executable process model.* Process model-based execution ensures that all required documentation is present during order management and stored after the order is fulfilled.
- *Business rules automation.* DMN decision tables are integrated in the executable process model and are supported by the platform. Special offers, discounts for product groups are automatically applied during process execution. Most importantly data defining business rules be easily changed by an authorized user with no specific experience in information technologies. Because of this, the implemented solution is considered flexible.
- *Tasks allocation and required data presentation*. Business process management system (Camunda Tasklist component) allocates tasks and provides only relevant task data for the sales assistant at proper time. No excessive data is provided, nor it is required to look up for any data during order management process.
- *Automation of manual tasks*. Most of the manual tasks such as filling order contract, finding order or client data and sending notifications to clients were automated and are performed by the BPMS engine.

Other advantages of BPMS system such as automatic task list creation for each employee by task priorities, process execution data monitoring or the ability to implement changes to process model with little effort and minimal costs are expected to be identified by the business in a long-term testing of the created system. Advantage of information system being built on executable process model is considered an advantage for the developers or administrators of BPMS as it does not create direct value for the target business.

E. Discussion

After conducting qualitative analysis on an order management process of a small optical retail business we have obtained good results for further process reengineering. From the owners and managers of target business it was expected that this type of analysis will not be useful in the scope of small business. Although it pointed out the redundant tasks and weak spots of the process and it was easier to improve the process model. For a fluent qualitative analysis there were not enough information on how to perform it, though a lot of definitions of different methods can be found. It is very unlikely that such analysis could be performed in a SME business without external consultants. Value-Added analysis showed unnecessary tasks, but its results alone lacked information on what parts could be improved. We suggest that this analysis method should always be used with some problem-oriented method. In such method, like Root-Cause analysis, smaller number of perspectives than in 6M model is not advised as it may be difficult to point out which parts of the process require improvement. Finally, the qualitative analysis took more effort to complete than expected. This should be considered and its advantages clarified for the stakeholders as it can be rejected by businesses as not

necessary part of process optimization or automation in an early stage of initiative.

Quantitative analysis, as it stands for, gives specific, comparable results that are understandable for all process stakeholders. For this reason, it is much easier to conduct such analysis in the scope of SME than qualitative analysis. As expected, the qualitative analysis gave strong backup to the results of previously conducted analysis. Although process model simulation requires basic process modeling knowledge. Added the usability of the used simulation tool, it is same as in the case of qualitative analysis – it is unlikely that this type of analysis can be conducted inside SME business with no external help. To achieve this a more stable simulation tool must be developed and more information on how to use it must be provided for the user.

IV. Conclusions and further research

As we can see from an overview of a business process optimization methods and specific optimization case in a small optical retail shop, these methods can be applied in order to optimize small optical retail business order management process. Although for these methods to be applicable widely in small businesses there must be broader amount of information available on this topic. For example, a shared knowledge base with examples for process optimization should be accessible. Further aim of this initiative is to perform a long term BPMS system usability research in a small optical retail business. Employees attitude and effectiveness of BPMS must be investigated in such environment despite initial advantages recognized by the stakeholders of the process like process model based execution, business rules automation, tasks allocation and automation.

Further research on the topic of process optimization could be pointed to improving process simulation and modeling tools by making calculations of process model metrics such as cohesion in real time thus making the evaluation of those models faster and easier.

References

[1] EC. *Annual Report of European SMEs 2017-2018*. 2018. ISBN 9789279968228.

[2] DALLAS, Ian and WYNN, Moe Thandar. Business Process Management in Small Business: A Case Study. *Information Systems for Small and Medium-sized Enterprises* [online]. 2014. No. 2, p. 67–96. DOI 10.1007/978-3-642-38244-4. Retrieved http://link.springer.com/10.1007/978-3-642-38244-4

[3] FREUND, Jakob. *Camunda BPM Compared to Alternatives* [online]. 2015. Retrieved https://network.camunda.org/whitepaper/8

[4] LA ROSA, Marcello. Interview with Michael Rosemann on "The Role of Business Process Management in Modern Organizations." *Business & Information Systems Engineering* [online]. 2016. Vol. 58, no. 1, p. 89–91. DOI 10.1007/s12599-015-0419-8. Retrieved http://link.springer.com/10.1007/s12599-015-0419-8

[5] HARMON, Paul. *The Scope and Evolution of Business Process Management*. 2014. ISBN 9783642451003.

[6] DUMAS, Marlon, LA ROSA, Marcello, MENDLING, Jan and REIJERS, Hajo A. *Fundamentals of Business Process Management*. Heidelberg, New York, Dodrecht, London : Springer, 2013. ISBN 978-3-642-33142-8.

[7] BERNAERT, Maxime, POELS, Geert, SNOECK, Monique and DE BACKER, Manu. Information Systems for Small and Medium-sized Enterprises. *Information Systems for Small and Medium-sized Enterprises* [online]. 2014. No. 2, p. 67–96. DOI 10.1007/978-3-642-38244-4. Retrieved http://link.springer.com/10.1007/978-3-642-38244-4

[8] BAUWENS, Cedric and VAN DORPE, Thomas. *Business Process Management in SMEs*. Universiteit Gent, 2018.

[9] POWER, Brad. Michael Hammer's Process and Enterprise Maturity Model. *Business Process Trends*. 2007. No. July, p. 1–4.

[10] MCCORMACK, K. and JOHNSON, W. Business Process Orientation: Gaining the E-Business Competitive Advantage. . 2001.

[11] VELDHUIZEN, R., VAN RAVESTEIJN, P. and VERSENDAAL, J. BPMS implementations in SMEs: Exploring the creation of a situational method. *25th Bled eConference - eDependability: Reliable and Trustworthy eStructures, eProcesses, eOperations and eServices for the Future, Proceedings*. 2012. Vol. 1949, p. 84–98.

[12] HEINRICH, Robert and PAECH, Barbara. Defining the Quality of Business Processes. *Modellierung 2010 P-161* [online]. 2010. P. 113—148. Retrieved http://subs.emis.de/LNI/Proceedings/Proceedings161/P-161.pdf?origin=publicationDetail#page=134

[13] ROONEY, J J and VAN DEN HEUVEL, L N. Root Cause Analysis for Beginners. *Quality Progress*. 2004. No. July, p. 45–53.

[14] EDITORS, Series, BERNUS, Peter and SHAW, Michael J. *Handbook on Business Process Management 1* [online]. 2010. ISBN 978-3-642-00415-5. Retrieved http://link.springer.com/10.1007/978-3-642-00416-2

[15] DELSANTER, Judith. Six sigma. *Managing Service Quality: An International Journal*. 1992. Vol. 2, no. 4, p. 203–206. DOI 10.1108/09604529210029353.

[16] WILKINSON, Leland. Revising the Pareto Chart. *The American Statistician*. 2006. Vol. 60, no. 4, p. 332–334. DOI http://dx.doi.org/10.1198/000313006X152243.

[17] VANDERFEESTEN, Irene, REIJERS, Hajo A. and VAN DER AALST, Wil M P. Evaluating workflow process designs using cohesion and coupling metrics. *Computers in Industry*. 2008. Vol. 59, no. 5, p. 420–437. DOI 10.1016/j.compind.2007.12.007.

[18] HUAN, Samuel H., SHEORAN, Sunil K. and WANG, Ge. A review and analysis of supply chain operations reference (SCOR) model. *Supply Chain Management: An International Journal* [online]. 2004. Vol. 9, no. 1, p. 23–29. DOI 10.1108/13598540410517557. Retrieved http://www.emeraldinsight.com/doi/10.1108/13598540410517557

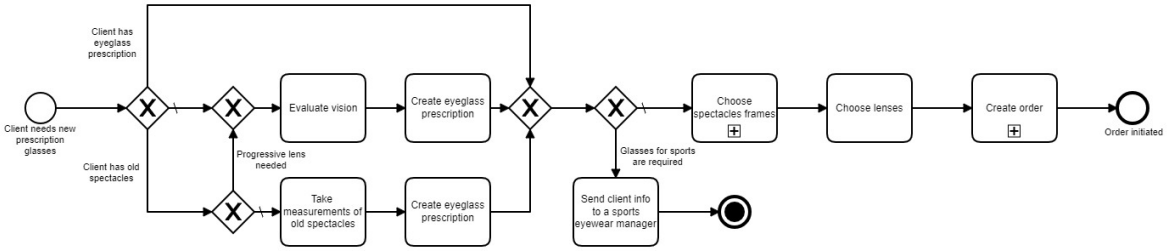
[19] CAMUNDA. Modeler. [online]. [Accessed 2019.02.7]. Retrieved https://camunda.com/products/modeler/

[20] BIMP. BIMP - The Business Process Simulator. [online]. [Accessed 2018.06.13]. Retrieved http://bimp.cs.ut.ee/

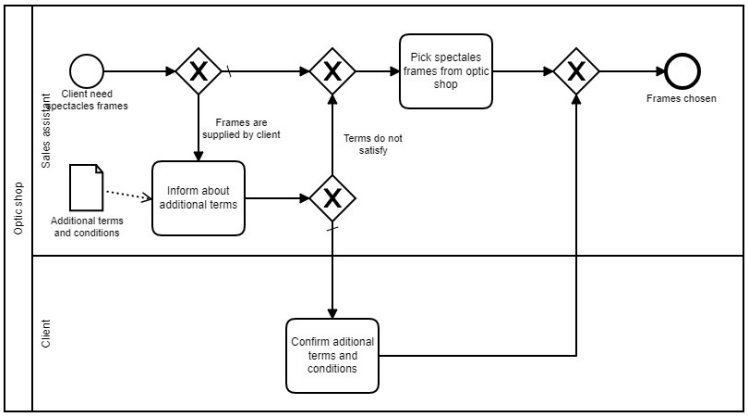
[21] REIJERS, H. A. and LIMAN MANSAR, S. Best practices in business process redesign: An overview and qualitative evaluation of successful redesign heuristics. *Omega*. 2005. Vol. 33, no. 4, p. 283–306. DOI 10.1016/j.omega.2004.04.012.

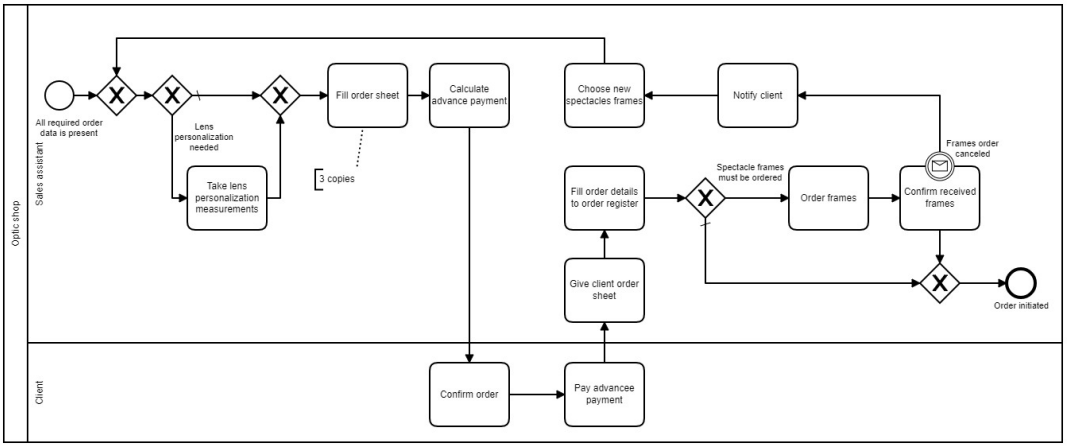
APPENDIX A. ORDER ITINIATION SUB-PROCESS

The following diagram presents the first sub-process of Order management business process, which is Order initiation.



The next two diagrams represent two sub-processes of Order initiation, namely, Choose spectacles frames and Create order.





INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Validation of VARK questionnaire using gaze tracking data

Simonas Baltulionis, Vilius Turenko, Mindaugas Vasiljevas, Robertas Damaševičius
Department of Software Engineering
Kaunas University of Technology, Kaunas, Lithuania
robertas.damasevicius@ktu.lt

Tatjana Sidekerskienė
Department of Applied Mathematics, Kaunas University of Technology,
Kaunas Lithuania

***Abstract*—We use gaze data (fixation time on Areas of Interest, AoIs) collected while reading educational materials to validate the VARK (Visual Auditory Reading Kinaesthetic) questionnaire. We analyse the dependencies between four types of AoIs (Title, Text, Graph, Formula) and the VARK scores for sensory modalities using correlation and linear regression analysis. Our results show significant correlations for Formula – Reading, Text – Visual, and Title – Kinaesthetic dependencies. The results of research can be used for objective evaluation of learning style of subjects using gaze tracking technology.**

Keywords—VARK; learning styles; gaze tracking; multimedia

I. Introduction

Learning styles were defined to justify individual preferences and differences in learning and understanding [1]. Notable models of learning style include Kolb's experiential learning, which introduces *accommodators, convergers, divergers* and *assimilators* [2]; Mumford's model, which has *activists, reflectors, theorists,* and *pragmatists* [3]; Barbe et al. model, which considers *auditory, visualising,* and *kinesthetic* modalities [4], and Index of Learning Styles (ILS), which considers, *active/reflective, sensing/intuitive, visual/verbal, and sequential/global* learning [5]. Learning styles can be employed for user modelling, developing effective pedagogical guidelines, personalization of learning scenarios and materials, and increasing interactivity of presentation in multimedia-based e-learning systems. The usefulness of the learning styles were proven in various and diverse fields of education such as computer programming [6] and nursing [7]. Different tools have been used to evaluate learning styles such as Visual Auditory Reading Kinaesthetic (VARK) [8], Visual Auditory Kinaesthetic (VAK) [9] and Learning Style Questionnaire (LSQ) [10]. However, as the use of questionnaires as a research tool is prone to subjectiveness and difficulty of interpretation, and have been criticized for weak empirical evidence, no correlation with learning outcomes [11] and the lack of independent research on the model [12].

The objective evaluation methods were suggested to use electroencephalogram (EEG) [13, 14, 15] and electrocardiogram (ECG) signals acquired from the learners [16]. Here we analyse the use of gaze tracking data recorded while learners read learning materials to evaluate their learning styles. The idea in itself is not new as gaze tracking has been used previously in this context [17, 18, 19] while aiming to detect correlations between assimilation of different types of information and different parameters like learning style. We specifically focus on the validation of the VARK model, which proposed four types of learners: visual, auditory learning, textual and kinaesthetic. Our novelty is that we focus on the validity of the VARK questionnaire in itself and aim to confirm the VARK scores by gaze related characteristics of subjects without analysing the differences in learning style and efficiency.

II. Method

A. VARK

VARK defines Visual, Aural, Read/write, and Kinaesthetic sensory modalities that are employed in the learning process. Visual (V) modality prefers the presentation of information using maps, diagrams, charts, graphs, and symbolic elements such as arrows and boxes. Aural / Auditory (A) prefers any information that can be heard and discussed. Read/write (R) modality prefers words (text). Kinaesthetic (K) modality prefers anything that is real, i.e., examples, personal experiences, or practice. Some individuals do not have a preferred modality and could be defined as Multimodal (MM).

B. Data collected by gaze tracking

During gaze tracking we collect the number and location of fixations, which are gaze points that are directed towards a certain part of an image, which is labelled as Area of Interest (AoI). Fixations are indications of visual attention. The eye movements between fixations are known as saccades. However, we do not use the saccade data in this study.

Following Yu [20], we introduce three types of AoI: Text (T1), Graph (G) and Formula (F). Title, a fourth type of AoI, we used, is also the Text, but it is used a separate element (T2), which provides the concise summary of the content. Note that due to the selected type of learning materials, which is static and does not include any interaction, the A modality does not have a preferred representation type.

C. Research hypotheses

We assume that subjects have their own preferred sensory modalities, which makes them unconsciously to pay more attention to a corresponding type of information. Based on this assumption, we formulate the following research hypotheses:

H1: V subjects prefer the G information.

H2: A subjects do not have a preferred type of information.

H3: R subjects prefer the T information.

H4: K subjects prefer the F information.

D. Testing of hypotheses

For testing of hypotheses we use the Pearson correlation:

$$r = \frac{1}{n-1}\sum \frac{(x_i - \overline{X})(y_i - \overline{Y})}{s_x s_y} \qquad (1)$$

here x_i, y_i are the data values for which the dependency is tested, $\overline{X}, \overline{Y}$ are means, s_x, s_y are standard deviations. The value of $r > 0$ indicates a positive relationship of X and Y, and $r < 0$ indicates a negative relationship.

The significance of the correlation value is calculated using the critical values of t-statistics as follows:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad (2)$$

here n is the size of a sample. Given a small sample of $n = 5$ in our case, the statistically significant ($p < 0.05$) correlation value must be at least $|r| > 0.86$.

We also construct the linear regression models between the dependent variables (T1, T2, G, F) and the independent variables (V, A, R, K). Linear regression is defined as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (3)$$

here Y_i is the value of dependent variable, X_i is the value of the independent variable for the i-th sample, β_0 is the free coefficient, β_1 is the slope, and ε_i is the random error. The sign of slope coefficient defines the direction of dependency (positive or negative), and the absolute value shows the strength of dependency.

The reliability of the linear regression model is evaluated using the significance of the coefficients (all must have $p < 0.05$), and the coefficient of determination r^2:

$$r^2 = \frac{explained\ variation}{total\ variation} = \frac{SSR}{SST} \quad 0 \le r^2 \le 1 \quad (4)$$

III. Experimental Setting and Results

A. Experimental setting

Five participants (one female, four male) were recruited for this study, ages between 23 and 45 with an average of 29.8 years (SD = 8.66). All participants had a Bachelor's degree in engineering and normal or corrected-to-normal vision. Participants were familiar with computers and had previous experience in using the internet. For each subject 7 slides that consisted of title, text, graph and formula were shown. Each slide was shown for 30 seconds interval, and the session took approximately 4 minutes. Subjects were instructed that they should try to memorize as much information as possible because at the end of the slide show a test will be taken. which consists of questions related to all different types of mathematical objects. For instance, to answer which formula, graph or text matches a given statement.

The Tobii 4C eye tracker was used to record eye movements of participants. The eye tracker uses infrared corneal reflection to measure point of gaze with data rates of 90 Hz. A 24 inch screen was used to show the slides. The eye tracker using instructions was mounted just below the visible screen area. The operating distance between the eye tracker and subjects' eyes was in the range 70-75 cm. For each subject the eye tracker was re-calibrated using a 5-point calibration to achieve most accurate results. Gaze monitoring system was used to measure the number and duration of fixations in the Areas of Interest (AOIs). The system consists of components listed below (see Fig. 1):

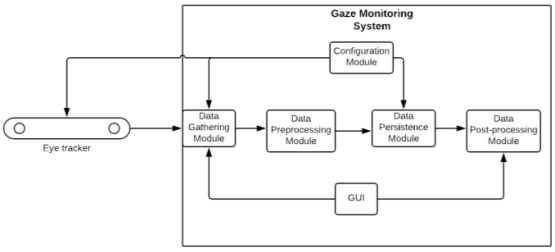


Fig. 1. Architecture of a system

- The Data Gathering Module reads the raw gaze data from the eye tracker device via USB.
- The Data Preprocessing Module filters noise, calculates additional metrics and characteristics like saccades.
- The Data Persistence Module saves the acquired gaze data to CSV, XML or database.
- The Data Post-processing Module maps persisted gaze data to AOIs and calculates additional data features such as the total and average number and duration of fixations.
- The Configuration Module configures how data is gathered and persisted in the system.

The stimulus was the educational materials from the "Mathematics 1" course delivered to the 1st year Bachelor students at Kaunas University of Technology. The topic of the educational materials was the integral calculus. Structurally arranged as a set of PowerPoint (Microsoft, USA) slides, each slide representing a learning unit had four components: Title, Text, Formula and Graph (see Fig. 2).

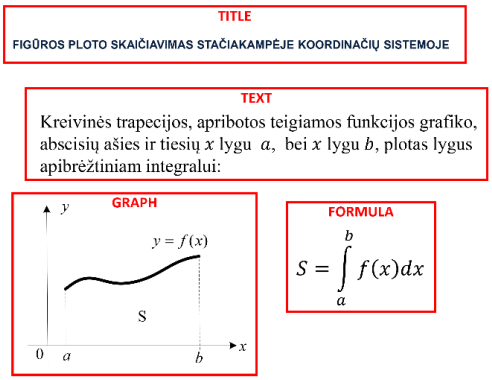


Fig. 2. Areas of Interest (AoI) in learning material.

This study examined visual attention and the reading behaviour of the subjects. Each participant took the VARK Questionnaire for the assessment of learning styles. Then the participants we asked to complete a calibration session followed by launching the learning material slides in full screen mode. Following that, participants were asked to read the slides presented at the computer screen. During the experiment, the eye tracker measured the learner's eye movements such as eye fixations and fixation durations. After completing the reading component, a knowledge assessment test was administered to participants on screen. The results of

the knowledge evaluation test were not used in this study, as the aim was to motivate participates to read attentively rather than evaluating their knowledge gained on the subject.

B. Results

The results of gaze time spent on each AoI are summarized in Fig. 3: most time (~35%) was spent on text, while least time (~6%) on the title of the learning material.

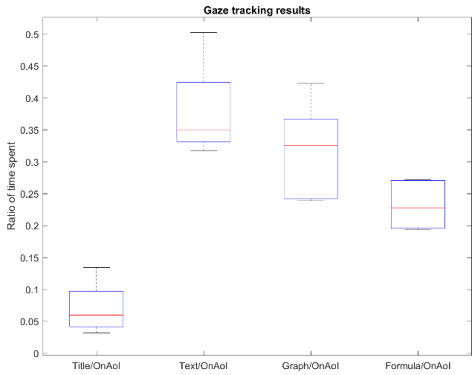


Fig. 3. Example of a figure caption. (*figure caption*)

The summary of the VARK scores are presented in Figure 4. On average, the highest score was assigned to Visual type (9.2), while the lowest score was assigned to Aural type (4.2).

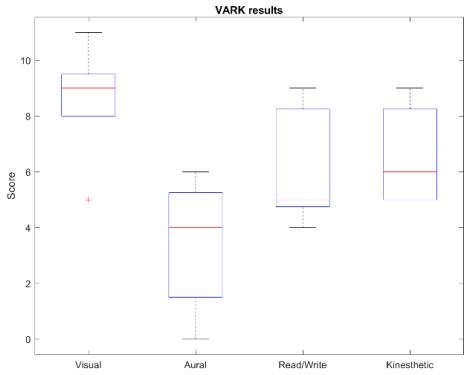


Fig. 4. Results of VARK questionnaire scores

We performed the correlation analysis on the ratio of time spent on the Title (T2), Text (T1), Graph (G) and Formula (F) AoIs vs the Visual (V), Aural (A), Read/Write (R) and Kinesthetic (K) scores from the VARK questionnaire. The results are presented in Fig. 5. We found significant correlations for Title ↔ Kinaesthetic ($r = 0.96$), Text ↔ Visual ($r = -0.94$), and Formula ↔ Read/Write ($r = 0.93$). We did not find any significant correlations for the A modality thus confirming the H2 hypothesis. We could not confirm the H1 hypothesis, however the results show that V subjects strongly do not prefer T information. We also could not confirm the H3 hypothesis, but we found that R subjects prefer F information. We also could not confirm the H3 hypothesis, but the results show that K subjects prefer T information.

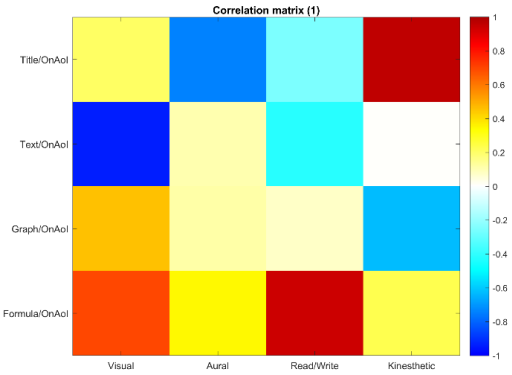


Fig. 5. Correlation matrix of the relative fixation times in Title, Text, Graph and Formula AoIs vs the Visual, Aural, Read/Write and Kinesthetic scores

We also explored more different types of relationship and analysed the dependencies between the grouped dependent variables (T1+T2, T1+G, T1+F, T2+G, T2+F, G+F) and independent variables (V, A, R and K). The results presented in Fig. 6. The significant correlations were found only for Title + Formula ↔ Kinaesthetic ($r = 0.86$), and Text + Graph ↔ Kinaesthetic ($r = -0.86$).

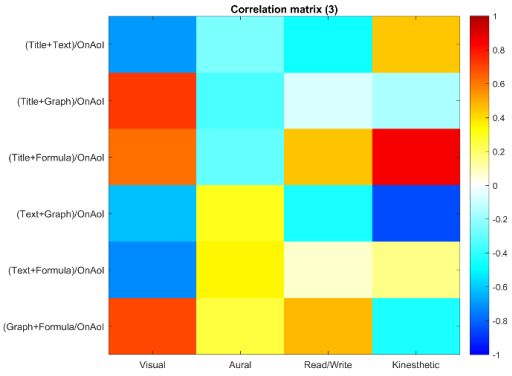


Fig. 6. Correlation matrix of the relative fixation times in Title+Text, Text+Graph, Title+Formula, Text+Graph, Text+Formula and Graph+Formula AoIs vs the Visual, Aural, Read/Write and Kinesthetic scores

Four linear regression models were constructed for each of the V, A, R and K modality scores as dependent variables and the Title (T2), Text (T1), Graph (G) and Formula (F) AoIs as independent variables (see a summary presented in Fig. 7). All models are reliable ($p < 0.001$ for all coefficients and $r^2 > 0.99$ for all models). When considering the value of slope coefficient, the V modality is mostly influenced by Title (39.5, positively) and Graph (28.8, positively), the A modality is mostly influenced by Title (-74.8, negatively), the R modality is mostly influenced by Formula (94.7, positively) and Title (-65.1, negatively), and the K modality is mostly influenced by Title (71.1, positively).

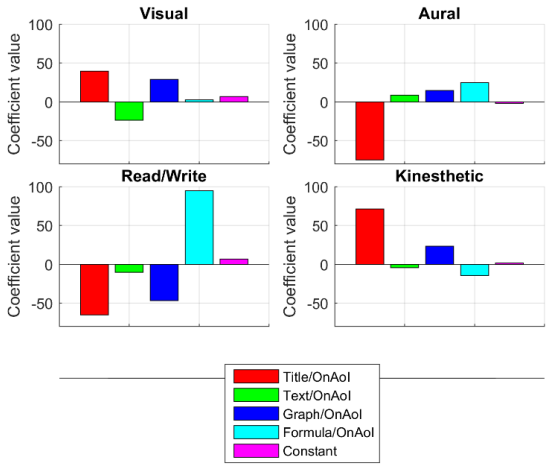


Fig. 7. Summary of V, A, R and K linear regression models

We also constructed the inverse linear regression models were constructed for the Title (T2), Text (T1), Graph (G) and Formula (F) AoIs as independent variables and the V, A, R and K modality scores as dependent variables (see a summary presented in Fig. 8). In this case, only one model for Title was reliable ($p < 0.001$ for all coefficients and $r^2 > 0.99$). When considering the value of the slope coefficient, the time spent on Title AoI is mostly influenced by the K modality (0.017, positively), which agrees with the corresponding linear regression model for the K modality presented in Fig. 9.

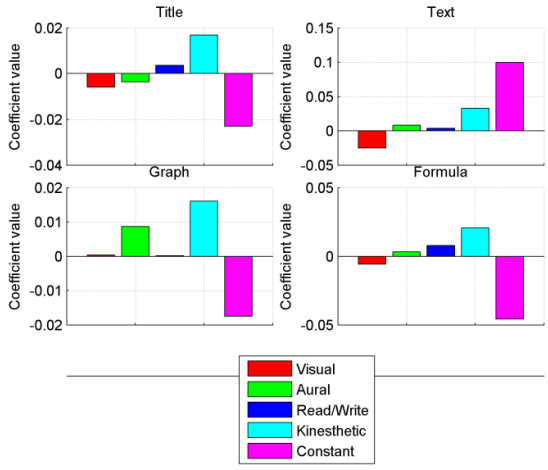


Fig. 8. Summary of Title (T2), Text (T1), Graph (G) and Formula (F) linear regression models

Finally, we evaluate how much of variance in the data for the sensory modalities is explained by the variance in the AoI (Fig. 9) and vice versa (Fig. 10). We can see that the V modality is most influenced by the Formula (+51%) and Text (-24%) AoIs. The A modality is most influenced by the Text (+35%) and Formula (+31%) AoIs. The R modality is most influenced by the Title (+35%) and Graph (-32%) AoIs. The K modality is most influenced by the Title (+29%) and Formula (+27%) AoIs.

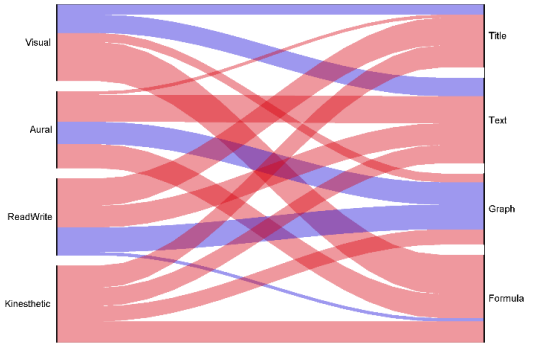


Fig. 9. Variance in sensory modalities explained by the type of AoI (red – positive influnce, blue – negative influence)

The attention on the Title AoI is most influenced by the V (+38%) and K (+30%) modalities. The attention on the Text AoI is most influenced by the V (-52%) and K (+30%) modalities. The attention on the Graph AoI is most influenced by the K (+57%) and Title (-29%) modalities. The attention on the Formula AoI is most influenced by the V (+75%) and K (+20%) modalities.

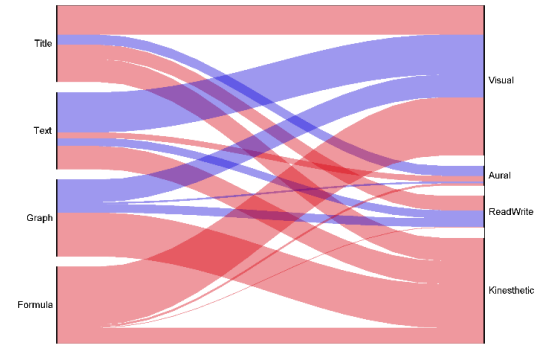


Fig. 10. Variance in the type of AoI explained by sensory modalities (red – positive influnce, blue – negative influence)

C. *Evaluation*

Our findings are in line with Al-Wabil et al. [21], who analysed Index of Learning Styles (ILS) using gaze tracking, also found that verbal learners pay attention to textual content more than multimedia, and visual learners scan the text and direct more attention to multimedia elements than textual content. Hoffler et al. [22] analysed the Object-Spatial Imagery and Verbal Questionnaire (OSIVQ) and found significant correlations between dwell time and the object and spatial visualizers, while no correlation was found for verbalizers. Our results confirm common knowledge, such as Visual subjects do not like Text but do like Graphs, however also provide interesting insights such as Kinaesthetic subjects liking Titles, which represent a condensed ('tangible') form of information, and Visual subjects liking Formulas, which although are a form of mathematical notation, yet share many similarities to the visual representation of information.

D. Threats to validity

A small sample of subjects and biased selection of participants (all subjects have a strong background in computer science) may render the results of our study as less reliable. Furthermore, the factors of stress, emotion and gender have not been accounted for in this study, although our previous research has demonstrated their significant influence on gaze characteristics [23, 24, 25]. Also note that the types of the AoIs analysed can not be separated strictly: in some cases text and graphs also contained elements of mathematical notations such as the names of variables.

IV. Conclusions

Our results demonstrate significant positive correlation between the attention on the Title Area of Interest (AoI) and the Kinaesthetic sensory modality ($r = 0.96$), significant negative correlation between the Text AoI and Visual modality ($r = -0.94$), and significant positive correlation between the Formula AoI and the Read/Write modality ($r = 0.93$). The linear regression models show the importance of Titles for the Visual, Aural and Kinaesthetic modalities and the importance of Formula for the Read/Write modality. The inverse linear regression model shows the significant attention of the Visual modality to Titles. The latter is confirmed by the variance analysis, which shows that Visual subjects prefer Formulas and dislike Text, Aural subjects like Text and Formulas, Read/Write subjects like Titles and dislike Graphs, and Kinaesthetic subjects like Titles and Formulas.

Our results show that there is a possibility for the VARK questionnaire to be another valid tool to analyse cognitive types of subjects. On the other hand, the gaze tracking data could possibly provide valuable objective information and insights on the cognitive preference of subjects that might possibly supplement the results of the subjective questionnaire. Future work will focus on collecting a larger dataset of gaze tracking data and extending the experiment to a more diverse set of AoIs.

References

[1] Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R.A. (2008). Learning styles: concepts and evidence. Psychological Science in the Public Interest. 9 (3): 105–119.

[2] Kolb, D.A. (2015). Experiential learning: experience as the source of learning and development (2nd ed.). Pearson Education.

[3] Mumford, A. (1997). Putting learning styles to work. Action learning at work. Aldershot, Hampshire, Brookfield, VT: Gower. 121–135.

[4] Barbe, W.B., Swassing, R.H., & Milone, M.N. (1979). Teaching through modality strengths: concepts practices. Zaner-Bloser.

[5] Felder, R.M., & Silverman, L.K. (1988). Learning and Teaching Styles in Engineering Education. Engr. Education, 78(7), 674-681.

[6] Abbott, M. R. B., & Shaw, P. (2018). Multiple modalities for APA instruction: Addressing diverse learning styles. Teaching and Learning in Nursing, 13(1), 63-65. doi:10.1016/j.teln.2017.08.004

[7] Diaz, F.S., Rubilar, T.P., Figueroa, C.C., & Silva, R.M. (2018). An adaptive E-learning platform with VARK learning styles to support the learning of object orientation. 2nd IEEE World Engineering Education Conference, EDUNINE 2018, 1-6. doi:10.1109/EDUNINE.2018.8450990

[8] Prithishkumar, I. J., & Michael, S. A. (2014). Understanding your student: Using the VARK model. Journal of Postgraduate Medicine, 60(2), 183-186. doi:10.4103/0022-3859.132337

[9] Apipah, S., Kartono, & Isnarto (2018). An analysis of mathematical connection ability based on student learning style on visualization auditory kinesthetic (VAK) learning model with self-assessment. Journal of Physics: Conference Series, 983(1). doi:10.1088/1742-6596/983/1/012138

[10] Kappe, F. R., Boekholt, L., den Rooyen, C., & Van der Flier, H. (2009). A predictive validity study of the learning style questionnaire (LSQ) using multiple, specific learning criteria. Learning and Individual Differences, 19(4), 464-467. doi:10.1016/j.lindif.2009.04.001

[11] Husmann, P. R., & O'Loughlin, V. D. (2019). Another nail in the coffin for learning styles? disparities among undergraduate anatomy students' study strategies, class performance, and reported VARK learning styles. Anatomical Sciences Education, 12(1), 6-19. doi:10.1002/ase.1777

[12] Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). Learning styles and pedagogy in post-16 learning: a systematic and critical review. London, England: Learning & Skills Research Centre.

[13] Thepsatitporn, S., & Pichitpornchai, C. (2016). Visual event-related potential studies supporting the validity of VARK learning styles' visual and read/write learners. Advances in Physiology Education, 40(2), 206-212. doi:10.1152/advan.00081.2015

[14] Jawed, S., Amin, H. U., Malik, A. S., & Faye, I. (2018). Differentiating between visual and non-visual learners using EEG power spectrum entropy. International Conference on Intelligent and Advanced System, ICIAS 2018, doi:10.1109/ICIAS.2018.8540571

[15] Alhasan, K., Chen, L., & Chen, F. (2018). Mining learning styles for personalised elearning. 2018 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovations, SmartWorld/UIC/ATC/ScalCom/CBDCom/IoP/SCI 2018, 1175-1180. doi:10.1109/SmartWorld.2018.00204

[16] Granero-Molina, J., Fernández-Sola, C., López-Domene, E., Hernández-Padilla, J. M., Romão Preto, L. S., & Castro-Sánchez, A. M. (2015). Effects of web-based electrocardiography simulation on strategies and learning styles. Revista Da Escola De Enfermagem, 49(4), 645-651. doi:10.1590/S0080-623420150000400016

[17] Mehigan, T. J., Barry, M., Kehoe, A., & Pitt, I. (2011). Using eye tracking technology to identify visual and verbal learners. IEEE International Conference on Multimedia and Expo, 1-6. doi:10.1109/ICME.2011.6012036

[18] Koć-Januchta, M., Höffler, T., Thoma, G., Prechtl, H., & Leutner, D. (2017). Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures – an eye-tracking study. Computers in Human Behavior, 68, 170-179. doi:10.1016/j.chb.2016.11.028

[19] Winoto, P., Tang, T. Y., Huang, Z., & Chen, P. (2017). "Thinking in pictures?" performance of chinese children with autism on math learning through eye-tracking technology. In: Zaphiris P., Ioannou A. (eds) Learning and Collaboration Technologies. Technology in Education. LCT 2017. Lecture Notes in Computer Science, vol 10296. Springer, Cham, 215-226. doi:10.1007/978-3-319-58515-4_17

[20] Yu, X. (2015). Literature Review of Applying Visual Method to Understand Mathematics. MATEC Web of Conferences, 22, 1063. doi:10.1051/matecconf/20152201063

[21] Al-Wabil, A., ElGibreen, H., George, R. P., & Al-Dosary, B. (2010). Exploring the validity of learning styles as personalization parameters in elearning environments: An eyetracking study. ICCTD 2010 - 2010 2nd International Conference on Computer Technology and Development, 174-178. doi:10.1109/ICCTD.2010.5646127

[22] Höffler, T. N., Koć-Januchta, M., & Leutner, D. (2017). More evidence for three types of cognitive style: Validating the object-spatial imagery and verbal questionnaire using eye tracking when learning with texts and pictures. Applied Cognitive Psychology, 31(1), 109-115. doi:10.1002/acp.3300

[23] Vasiljevas, M., Gedminas, T., Ševčenko, A., Jančiukas, M., Blažauskas, T., & Damaševičius, R. (2016). Modelling eye fatigue in gaze spelling task. 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing, ICCP 2016, 95-102. doi:10.1109/ICCP.2016.7737129

[24] Raudonis, V., Maskeliūnas, R., Stankevičius, K., & Damaševičius, R. (2017). Gender, age, colour, position and stress: How they influence attention at workplace? Computational Science and Its Applications – ICCSA 2017. Lecture Notes in Computer Science, 10408. Springer, Cham, 248-264. doi:10.1007/978-3-319-62404-4_19

[25] Liaudanskaitė, G., Saulytė, G., Jakutavičius, J., Vaičiukynaitė, E., Zailskaitė-Jakštė, L., & Damaševičius, R. (2019). Analysis of affective and gender factors in image comprehension of visual advertisement. Artificial Intelligence and Algorithms in Intelligent Systems. CSOC2018. Advances in Intelligent Systems and Computing, vol 764. Springer, Cham, 1-11. doi:10.1007/978-3-319-91189-2_1

Analysing program source code reading skills with eye tracking technology

Vilius Turenko, Simonas Baltulionis, Mindaugas Vasiljevas, Robertas Damaševičius
Department of Software Engineering
Kaunas University of Technology, Kaunas, Lithuania
robertas.damasevicius@ktu.lt

***Abstract*—Many areas of software engineering require good program code reading skills. We analyse the process of program reading using gaze tracking technology. We performed a study with six subjects, who performed four code reading tasks. The errors the embedded into program sources code and the lines of code with the areas were analysed as Areas of Interest (AoI). We formulated a research hypothesis and tested it using a one-way analysis of variance (ANOVA) test. The results of the study confirmed our research hypothesis that the number of fixations on AoI is larger than the number of fixations on other areas.**

Keywords—program comprehension, code reading; eye tracking, gaze tracking, human-centered computing.

I. Introduction

Program code reading skills are important in many areas of software engineering, especially, in adopting good code writing practices and techniques, understanding how programs work, identifying cases of poor programming style and bad design, and delivering effective software maintenance. Examples include program tracing and searching for bugs, code smells and design anti-patterns [1]. As automatic methods for finding bugs and poor coding practices are still not very effective [2], source code reading and analysis by human experts remain as relevant as ever. Program comprehension is a crucial part of computer science education, providing an important part of an understanding of complexity of information technology (IT) systems [3]. The interest on applying gaze tracking in the context of multimedia supported learning is on the rise [4]. Gaze data had been successfully applied to analyze changes in cognitive load during assimilation of learning materials and are starting to be incorporated into adaptive e-Learning systems [5]. However, currently there are no effective strategies in evaluating code reading skills and assessing program comprehension. Recently, eye tracking and was proposed as a viable research instrument for evaluating source code reading [6]. The outcomes of gaze tracking studies are especially relevant in the context of Evidence-based Software Engineering (EBSE) in order to provide detailed insights regarding different practices in software engineering [7].

Eye movements are directly related to cognitive and information processing processes, and through these processes, visual information is used to stimulate the brain and to understand the given task. There are two assumptions related to cognitive processes and fixations: 1) if a person is seeing an object (such as a word), he/she tries to understand it; 2) a person fixates his/her gaze on an object until he/she understands it. A fixation is an aggregation of gaze points based on a specified area and time span. An Area of Interest (AoI) is a part of a visual stimulus that is of special importance Other important characteristics are a scan path, which is a series of fixations that indicate the path and tendency of eye movements, and a heat map, which identifies the focus of visual attention [8]. For example, Uwano et al. [9] studied graduate students conducting code reviews and discovered that their gaze patterns followed a common scanpath, first reading code top to bottom, and then rereading a few parts in more depth. Chandrika et al. [10] confirmed the positive relationship of eye tracking traits over source code lines and comments for code comprehension. Melo et al. [11] analysed how programmers debug code with embedded pre-processor directives. Jbara and Feitelson [12] analysed how code repeatability impacts the number of fixations in a predefined area of interest (AOI), and the total fixation time. Beelder and Plesis [13] analysed how the number and durations of fixations are influenced by syntax highlighting. Yennigall et al. [14] also used fixation counts and duration to analyse how programming novices understood program code.

In this paper, we describe the results of gaze tracking study on evaluating and analysing the code reading skills of software programmers, specifically focusing on the ability to find errors in program code.

II. Methodology

A. Program reading tasks

The study consisted of 4 tasks:

a. In Task 1, the aim was to read the program source code with the aim of finding the result returned (printed) (Fig. 1).

b. In Task 2, the aim was to identify the purpose of the algorithm and discover the hidden error associated with the incompatibility of the variable types (Fig. 2).

c. In Task 3, the aim was to find three syntactic errors related to the incorrect use of variable names, types and basic methods (Fig. 3).

d. In Task 4, the aim was to determine whether the algorithm would perform the specified function, and to find a hidden semantical error (Fig. 4).

```
1
2 /*Duota klasė stačiakampis, kurioje duoti ilgio ir plot
3 io parametrai,
4 *klasėje užklogi operatoriai daugybos ir dalybos vei
5 ksmams
6 *
7 */
8 public class Stačiakampis
9 {   //Kintamieji
10   public int ilgis { get; set; }
11   public int plotis { get; set; }
12
13   //Konstruktorius
14   public Stačiakampis(int ilgis, int plotis)
15   {
16       this.plotis= ilgis;
17       this.ilgis  =plotis ;
18   }
19
20   //Ilgių daugyba
21   public static int operator *(Stačiakampis x, Sta
22 čiakampis y)
23   {
24       return x.ilgis / y.ilgis;
25   }
26   //Plotų dalyba
27
28   public static int operator /(Stačiakampis x, Sta
29   čiakampis y)
30   {
31       return x.plotis * y.plotis;
32   }
33
34}
35 class Program
36{
37   static void Main(string[] args)
38   {
39       Stačiakampis s1 = new Stačiakampis(1, 5);
40       Stačiakampis s2 = new Stačiakampis(5,10);
41       Console.WriteLine("{0} {1}", s1 / s2, s2 * s1
42   }; }}
43
            ATS 5 0
```

Fig. 1. Program source code with Area of Interest (AoI) highlighted for Task 1: calculate output of a program

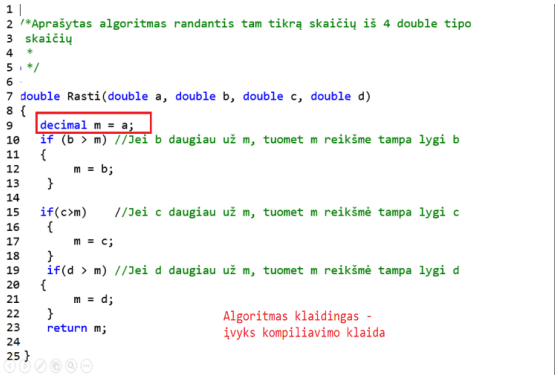


Fig. 2. Program source code with Area of Interest (AoI) highlighted for Task 2: find syntactic error

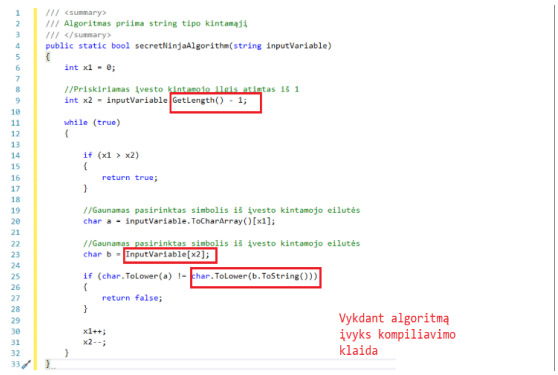


Fig. 3. Program source code with Area of Interest (AoI) highlighted for Task 3: find multiple syntactic errors

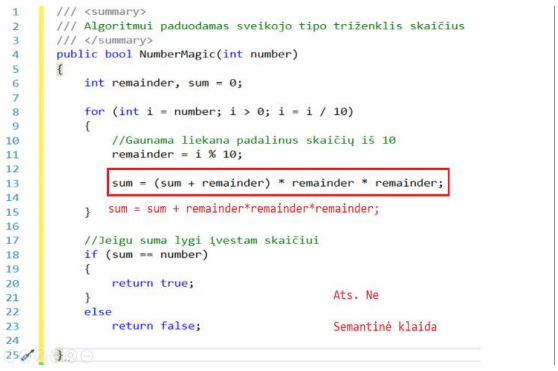


Fig. 4. Program source code with Area of Interest (AoI) highlighted for Task 4: find semantic error

B. *Data collected by gaze tracking*

During gaze tracking we collect the number and location of fixations, which are gaze points that are directed towards a certain part of an image, which is labelled as Area of Interest (AoI). Fixations are indications of visual attention. Here we analyze the distribution of the number of fixations between and out of AoIs. The eye movements between fixations are known as saccades. However, we do not use the saccade data in this study. A scan path is a directed path created by saccades between eye fixations.

C. *Research hypotheses*

We assume that subjects are thinking about the object of interest when they are looking directly at it. Based on this assumption, we formulate the following research hypothesis:

H1: The number of fixations on Areas of Interest is larger than the number of fixations on other areas.

D. *Testing of hypotheses*

For testing of hypotheses we employ a statistical one-way analysis of variance (ANOVA) test. The test, which is a standard statistical method, confirms or rejects the equality of the averages of two or more samples by examining the variances of samples. ANOVA compares the variance between the data samples to variance within each particular sample. If the between-sample variance is much larger than the within-sample variance variation, the average values of different samples can not be equal.

III. Experimental Setting and Results

A. *Experimental settings*

Six participants (1 female and 5 male) were recruited for this study, ages between 20 and 25 with an average of 22.8 years. All participants had normal or corrected-to-normal vision. Participants were familiar with computers and had previous experience in using the internet and all of them were studying or working in programming sphere. An informed consent was obtained from subjects before the study.

All subjects were given the same laptop Dell which had an additional monitor used for experiment and the Tobii Eye Tracker 4C eye-tracking device used to record eye movements and gaze fixations. The eye tracker uses infrared corneal reflection to measure point of gaze with data rates of 90 Hz. A 24 inch screen was used to show the slides which consisted of programming source code. The eye tracker using instructions was mounted just below the visible screen area. The operating distance between the eye tracker and subjects' eyes was between 70-75 cm. Efforts were made to ensure good lighting and a device calibrated before the test. For each subject the eye tracker was re-calibrated using an integrated 5-point calibration to achieve most accurate results.

Before the start of the experiment, the subjects were asked to fill in the Google Form questionnaire before the start of the study on their demographical characteristics (gender, education, age, level of programming skills). All responses were anonymized. After filling personal characteristics subjects had a chance to read some common information about tasks that they will face in this experiment, this way subjects were informed about some important rules, for example, no additional libraries or other extensions were used, also that some tasks were bug free and some had hidden bugs, the idea was to stimulate the subjects to be focused by not telling what tasks had bugs and what were bug free. After introducing tasks in common, the presentation with the slides containing the source code of tasks was opened, the observation session started at the start of each task and the session was stopped after the task was completed, each task had a separate observation session. 3 and 4 tasks had some brief information about given algorithms, for example, definition of palindrome and Armstrong's number and examples of each case. To complete each task, 90 seconds were given. After the completion of each task, the participants were asked to provide the answers in a Google Form on what is the result of program execution (Task1), what is the purpose

of an algorithm (Task 2), and is the program correct (Task 3, Task 4).

B. *Experimental system*

Gaze monitoring system was used to measure the number and duration of fixations in the Areas of Interest (AOIs). The system consists of components listed below (see Fig. 5).

- The Data Gathering Module reads the raw gaze data from the eye tracker device via USB.
- The Data Preprocessing Module filters noise, calculates additional metrics and characteristics like saccades.
- The Data Persistence Module saves the acquired gaze data to CSV, XML or database.
- The Data Post-processing Module maps persisted gaze data to AOIs and calculates additional data features such as the total and average number and duration of fixations.
- The Configuration Module configures how data is gathered and persisted in the system.

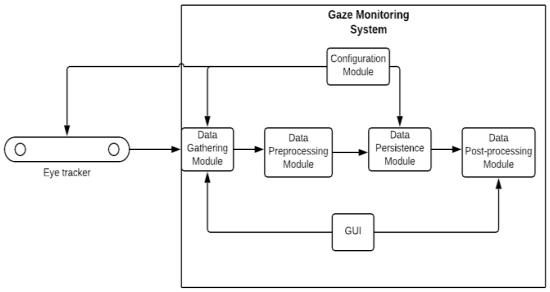


Fig. 5. Architecture of a system

System offers four types of data stream which are used to gather fixations and saccades directly from gaze tracking device.

- Unfiltered gaze
- Lightly filtered gaze
- Sensitive fixation
- Slow fixation

For this experiment, sensitive fixation type was chosen because of its accuracy and unnecessary noise reduction.

In addition, the system is running in the background and it has no effect on the stimulus, thus the subject's attention is concentrated only to source code.

Besides types of data stream, before starting gaze tracking session, user has an option to choose to record his screen, but for now it is only a prototype version, which needs to be improved for better accuracy, also session can have additional information about subject, for example name, age and other description, if it is not necessary, user can select anonymous session. In the near future, system will offer an option to choose screen resolution manually, which will allow to select concrete zones of interest.

C. *Results*

The results of participants (number of fixations) are summarized according to tasks and subjects in Fig. 6.

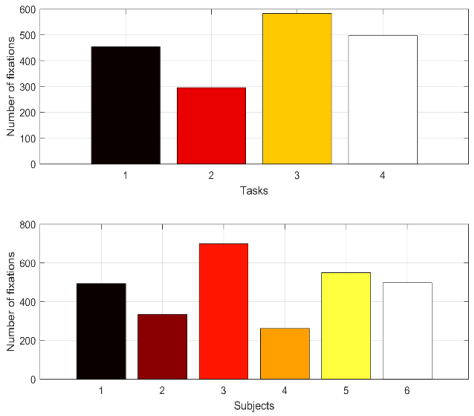


Fig. 6. Summary of the number of fixations according to subjects and tasks

An example of the gaze path generated from gaze tracking data is presented in Fig. 7. The gaze path shows how and in what sequence the subject has read the code. Note the order of reading is clearly not linear.

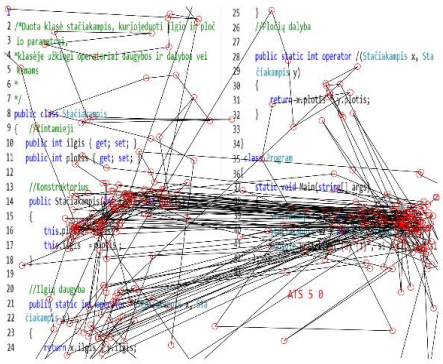


Fig. 7. Example of a gaze path (Task 1, Subject 1)

An example of the heatmap generated from gaze tracking data is presented in Fig. 8. Note that most of attention was focused on and around the Area of Interest centred on code line 42 (see also Fig. 1).

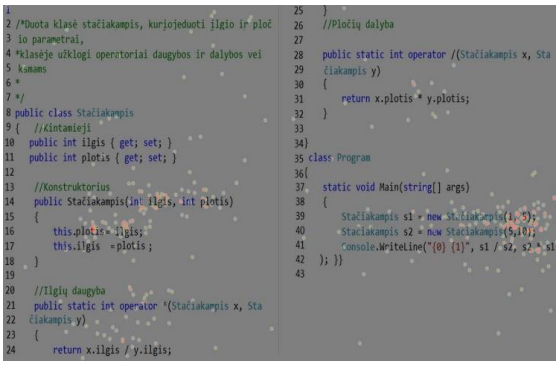


Fig. 8. Example of a gaze fixation heatmap (Task 1, Subject 1)

In Fig. 9, the average gaze fixation numbers for AoI and Non-AoI areas is presented. We can see that for all tasks, the number of fixations on AoIs was larger, although the difference was not statistically significant for Task 2 (also see the results of statistical testing using ANOVA in Table I).

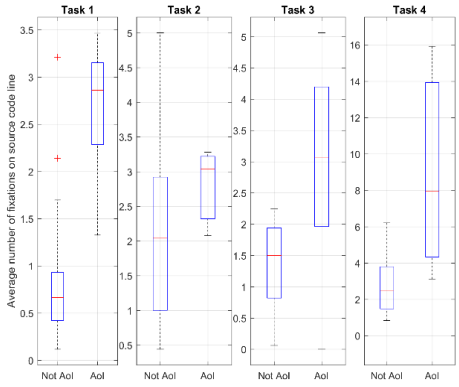


Fig. 9. Average number of fixations on AoI vs non-AoI source code lines

The results of statistical testing using ANOVA are presented in Table I. We found statistically significant differences in the number of fixations on the Areas of Interest (AoI) vs non-AoI for Tasks 1, 3 and 4. However, we did not find such differences for Task 2.

TABLE I. RESULTS OF STATISTICAL TESTING

Task	Results of ANOVA	
	F-value	*p-value*[a]
1	37.79	0 (***)
2	0.66	0.4245
3	14.73	0.0006 (***)
4	15.58	0.0006 (***)

[a] *** - statistically significant

D. Limitations and threats to validity

The study is based on the assumption that humans think about objects when look at them, however we cannot be sure that is assumption is correct. Our eye-tracking experiment only explores the processing of cognitive response to visual stimulus without considering the quality of responses. Moreover, due to a small sample of subjects and gender bias (all participants were male) we could not analyse the gender and affective differences, which have been noted as significant in other gaze tracking studies [15]. To minimize threats to validity, the participants did not know about the hypothesis formulated for the research. They only knew that they would be in helping us to understand how program code is read and understood.

In three tasks of four performed we were able to confirm our research hypothesis. In one, task the hypothesis could not be confirmed. We think that we reason was in poor design of the task, which we hope to improve in our further research.

IV. CONCLUSION

We have presented a study aimed at comprehending how programmers read and debug program code. Our results indicate that gaze tracking can be used successfully to follow and assess the cognitive behaviour of programmers as they correctly identify the errors embedded into the source code. The number of gaze fixations is a significant parameter when assessing the level of attention attributed to a particular Area of Interest.

Future work will focus on the methodological improvement of our research study and collection of a larger dataset of data from more subjects.

REFERENCES

[1] Obaidellah, U., Al Haek, M., & Cheng, P. C. (2018). A survey on the usage of eye-tracking in computer programming. ACM Computing Surveys, 51(1) doi:10.1145/3145904

[2] Gupta, A., Suri, B., Kumar, V., Misra, S., Blažauskas, T., & Damaševičius, R. (2018). Software Code Smell Prediction Model Using Shannon, Rényi and Tsallis Entropies. Entropy, 20(5), 372. doi:10.3390/e20050372

[3] Damaševičius, R. (2009). On The Human, Organizational, and Technical Aspects of Software Development and Analysis. In Information Systems Development (pp. 11–19). Springer US. doi:10.1007/b137171_2

[4] Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. Computers and Education, 125, 413-428. doi:10.1016/j.compedu.2018.06.023

[5] Rosch, J. L., & Vogel-Walcutt, J. J. (2013). A review of eye-tracking applications as tools for training. Cognition, Technology and Work, 15(3), 313-327. doi:10.1007/s10111-012-0234-7

[6] Busjahn, T., Schulte, C., & Busjahn, A. (2011). Analysis of code reading to gain more insight in program comprehension. In Proceedings of the 11th Koli Calling International Conference on Computing Education Research - Koli Calling '11. ACM Press. doi:10.1145/2094131.2094133

[7] Sharafi, Z., Soh, Z., & Guéhéneuc, Y. (2015). A systematic literature review on the usage of eye-tracking in software engineering. Information and Software Technology, 67, 79-107. doi:10.1016/j.infsof.2015.06.008

[8] Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., & Ertl, T. (2017). Visualization of eye tracking data: A taxonomy and survey. Computer Graphics Forum, 36(8), 260-284. doi:10.1111/cgf.13079

[9] Uwano, H., Nakamura, M., Monden, A., & Matsumoto, K. (2006). Analyzing individual performance of source code review using reviewers' eye movement. In Proceedings of the 2006 symposium on Eye tracking research & applications - ETRA '06. ACM Press. doi:10.1145/1117309.1117357

[10] Chandrika, K. R., Amudha, J., & Sudarsan, S. D. (2017). Recognizing eye tracking traits for source code review. In 2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE. doi:10.1109/etfa.2017.8247637

[11] Melo, J., Narcizo, F. B., Hansen, D. W., Brabrand, C., & Wasowski, A. (2017). Variability through the Eyes of the Programmer. In 2017

IEEE/ACM 25th International Conference on Program Comprehension (ICPC). IEEE. https://doi.org/10.1109/icpc.2017.34

[12] Ahmad Jbara and Dror G. Feitelson. 2015. How programmers read regular code: a controlled experiment using eye tracking. In Proceedings of the 2015 IEEE 23rd International Conference on Program Comprehension (ICPC '15). IEEE Press, Piscataway, NJ, USA, 244-254.

[13] Beelders, T., & du Plessis, J.-P. (2016). The Influence of Syntax Highlighting on Scanning and Reading Behaviour for Source Code. In Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists on - SAICSIT '16. ACM Press. https://doi.org/10.1145/2987491.2987536

[14] Yenigalla, L., Sinha, V., Sharif, B., & Crosby, M. (2016). How Novices Read Source Code in Introductory Courses on Programming: An Eye-Tracking Experiment. In Lecture Notes in Computer Science (pp. 120–131). Springer International Publishing. https://doi.org/10.1007/978-3-319-39952-2_13

[15] Liaudanskaitė, G., Saulytė, G., Jakutavičius, J., Vaičiukynaitė, E., Zailskaitė-Jakštė, L., & Damaševičius, R. (2019). Analysis of affective and gender factors in image comprehension of visual advertisement. Artificial Intelligence and Algorithms in Intelligent Systems. CSOC2018. Advances in Intelligent Systems and Computing, vol 764. Springer, Cham, 1-11. doi:10.1007/978-3-319-91189-2_1

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

CWW Enhanced Fuzzy SWOT Evaluation for Risk Analysis and Decision Making under Uncertainty

Žygimantas Meškauskas
Department of Computer Sciences
Kaunas University of Technology
Kaunas, Lithuania
zygimantas.meskauskas@ktu.lt

***Abstract*—The SWOT analysis is a worldwide method used to assist in the decision making in industrial and business management, banking, planning of military operations, science, as an obligatory tool on the governmental level as well as in the public or private life. Up until now all data had to be collected from the experts and the decision makers in numerical form and the results are also presented numerically. In this paper, we aim to enrich the SWOT analysis with Computing with Words paradigm for expert knowledge extraction in a verbal form. This way of conveying information allows the experts to express their opinion with uncertainties. Moreover, enriched SWOT analysis results are extremely useful for the risk analysis and decision making.**

Keywords—SWOT analysis, computing with words, fuzzy, risk analysis, decision making, uncertainty

I. Introduction

Many different tools and activities are used in various fields of activity for the extraction of the experts' knowledge. If some information about specific area is needed, it is not mandatory to have a deep knowledge in that area. This is the case where field experts take a major role and the method itself is only needed to save an extracted information in a structured form. Generally, data extraction and structuring process can be defined as:

Data → Information → Knowledge → Wisdom.

Data extraction is always performed in a certain form of a dialogue. Experts from different fields often use different terminology to describe same objects from different perspectives. The biggest challenge is to have a successful conversation with an expert so that the opinion would be expressed adequately. For this purpose, a widely used SWOT analysis method enriched with computing with words paradigm was used for verbal knowledge expression and uncertainties evaluation. The results of such analysis can also be expressed in linguistic form and provide information for the risk management and decision making.

Chapter 2 contains a related work section, chapter 3 describes CWW enhanced SWOT analysis methodology, chapter 4 describes risk management and decision making, in the chapter 5 experimental simulation is presented and chapter 6 concludes everything with remarks.

II. Related works

SWOT analysis enhanced Computing with Words paradigm methodology is described in [10]. This article mainly focuses on the use of analysis under uncertainties for experts' knowledge extraction and the use of analysis results in a risk management and decision making. The idea that risk is not simply a loss, multiplied by the probability, but there are also positive risk options that are described in [4]. Risk management part in this work is based on composed risk formula in [7], that links risk analysis inputs and SWOT analysis outputs.

III. CWW enhanced SWOT analysis

It is known that SWOT stands for strengths (ST) , weaknesses (WK), opportunities (OP) and threats (TH) surrounding any idea, plan, or project to be investigated and / or implemented. Opportunities and threats are usually defined as external issues of the project and signify possible positive and negative achievements when the project will be realized. At the same time strengths and weaknesses mean internal issues that enable and impede the achievement of the main goals and the development the projects. A quantitative interaction between OPs, THs, STs and WKs usually is expressed by a numerical SWOT matrix which shows the influence of STs and WKs on strengths and threats [10].

This article aims to find ways on how to use verbal qualitative evaluation in the process of delivering descriptions of data necessary for SWOT analysis. Attempting to perform necessary SWOT computations and deliver the obtained SWOT analysis results in a verbal form OPs, THs, STs and WKs were characterized by means of using words. It indicates that CWW (Computing with Words) methodology enriches SWOT methodology and creates a possibility for SWOT users and decision makers to communicate using words of common language. We propose and investigate new possibilities applied to enrich SWOT analysis mechanism using elements of artificial intelligence, and, especially, the computing with

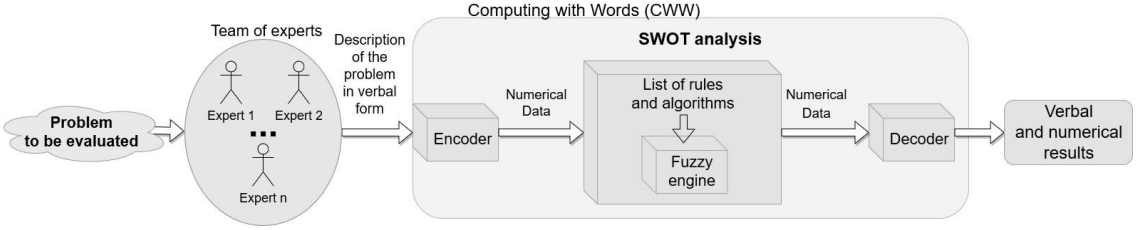


Fig. 1. Functional structure of SWOT+CWW methodology.

words paradigm. This approach is novel due to the originality of the encoding of input words that describe the investigated situation in a new functional organization of the SWOT engines, and the originality of the method used for decoding and aggregation of numerical outputs into a verbal form. Main idea of CWW enhanced SWOT analysis is to take verbal descriptions as input, convert that data into numbers for internal computation using fuzzy logic engine and translate the result to the user into a verbal form as shown in Fig. 1.

It is not necessary to have the knowledge on a specific domain – this is the role of experts. A certain number of experts can describe the situation and all the dynamics of the domain. The main focus is to collect required expert information for analysis and data storing. The most convenient way to describe a situation for any human being is to express it verbally instead of using numbers but some level of uncertainty arises from those words. Computational systems are based on a numerical data, so data encoding, and decoding is needed. In the line with CWW paradigm, all inputs and outputs to the user (expert) are in a verbal form. All the internal SWOT analysis computations using CWW paradigm are performing by black box principle. When an expert characterizes the information and dynamics for the domain, all this information is used in a data processing by translated list of rules and algorithms. Rules and algorithms are determined by expert's described dynamics of the field and used to translate between numerical and verbal data using fuzzification and defuzzification with fixed membership functions (displayed in Fig. 2).

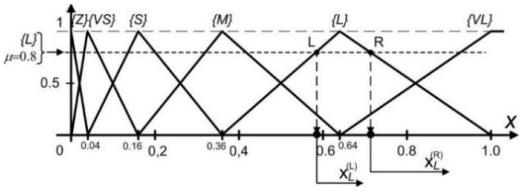


Fig. 2. Fixed CWW fuzzy membership functions.

Fuzzy logic engine calculates a numerical value of a given verbal term and a value of uncertainty by assigned membership function. Number of different verbal terms describes input words as possibilities. But according to the so called "Miller's law" [6] (The Magical Number Seven, Plus or Minus Two), a human can differentiate approximately up to seven different verbal evaluations. This CWW enhanced SWOT analysis verbal data input and output dictionary are selected based on this law. It used six different terms:

- "Zero" ({Z}),
- "Very small" ({VS}),
- "Small" ({S}),
- "Medium" ({M}),
- "Large" ({L}),
- "Very large" ({VL}).

Each verbal term from selected dictionary has its triangular form. Peak of each triangle on the X axis represents numerical value for verbal term in case of an absolute certainty. Left and right shoulders of the triangle represent uncertainty. In an example (Fig. 2), an expert expressed an opinion as "Large" with a degree of certainty (μ) as 0.8. Left shoulder of term "Large" ($X_L^{(L)}$) is the pessimistic value of uncertainty and the right shoulder ($X_L^{(R)}$) is optimistic.

When all data needed for SWOT analysis is submitted in that form, aggregated opportunity OP_Σ and TH_Σ values are calculated. Due to the data being translated in two ways (pessimistic and optimistic), there is a possibility for multiple perspectives of the results that can serve as a possible input data for risk analysis methodology.

IV. Risk analysis and decision making

Risk is the level of uncertainty of action (results). Most of methodologies interpret that risk directly depends on threats. In our approach we reference to Hillson [4] and state that risk is symbiosis of opportunities and threats. To implement this idea, we have associated risk components with SWOT analysis.

A. Risk analysis

In the context of a risk analysis, opportunities and threats can be associated with SWOT analysis components of opportunities and threats components, with efforts and hesitancies also making an impact. Efforts can be expressed as investments in a risk analysis process and hesitancies are the level of uncertainty. In our approach, risk can be described as:

$$\boldsymbol{R} = \mathbf{R}(\boldsymbol{EFF}\uparrow; \boldsymbol{OP}\downarrow; \boldsymbol{TH}\uparrow; \boldsymbol{HES}\uparrow) \quad (1)$$

The concept of risk combines:

- Activity (EFF/efforts/input/ ...);
- Potentially positive results (OP/ achievements/attainments/ ...);
- Potentially negative results (TH/ losses/defeats, …);
- Uncertainties (HES/hesitations/instabilities/options/probabilities/ ...).

OP and TH components of risk are strictly related to SWOT analysis outcomes ($\boldsymbol{OP}_\Sigma$ and $\boldsymbol{TH}_\Sigma$). Risk can be evaluated by combining it with an expert evaluation about required efforts (EFF) and (if needed) uncertainties (HES) evaluation. Risk evaluation can be estimated, and actions taken if necessary. Furthermore, verbal advices or visual representation of the results can be done.

B. Decision making

A decision is a commitment to a proposition, or a plan of an action based on the information and values associated with the possible outcomes. The process operates in a flexible timeframe that is free from the immediacy of evidence acquisition and the real time demands of action itself. Thus, it involves deliberation, planning, and strategizing [8]. The study of decision making is a multidisciplinary field. It occurs in psychology, statistics, economics, finance, engineering (e.g., quality control), political science, philosophy, medicine, ethics, and jurisprudence. There are many conflicting criterions that need to be evaluated in making decisions in our daily or professional lives.

Research on a multi-criteria decision support developed two main groups of methods, i.e., American and European schools. Methods of the American school of decision support are based on a functional approach, more precisely the utility or value function. Researchers from the European school

emphasize the fact that many methods do not consider the variability and uncertainty of expert judgments. However, the most common solution to this problem is to use granular mathematics, e.g., fuzzy sets theory or interval arithmetic [5].

V. Experimental Simulation

Generally, a lot of SWOT analysis tools were created, but they lack verbal operations. For this reason, a prototypical SWOT enhanced CWW analysis tool was created and used to test the effectiveness of the described methodology. Pilot testing was made on "Construction of a new hotel complex in a particular area" example from [11]. The example itself has already been analyzed in article and all SWOT analysis data is accessible for the use and the comparison of the results.

A. Data input

SWOT enhanced CWW tool data input is processed by one component at a time. There are two groups of identical data input:

1. opportunities and threats;
2. strengths and weaknesses.

User must enter a title and a short acronym of every SWOT analysis component (row number is generated automatically if not specified). When user submits OP or TH information, a degree of importance (impact) and value of truth (membership value) evaluations needs to be specified. Estimation itself is entered in a verbal form. The input of the opportunity is shown in Fig. 3

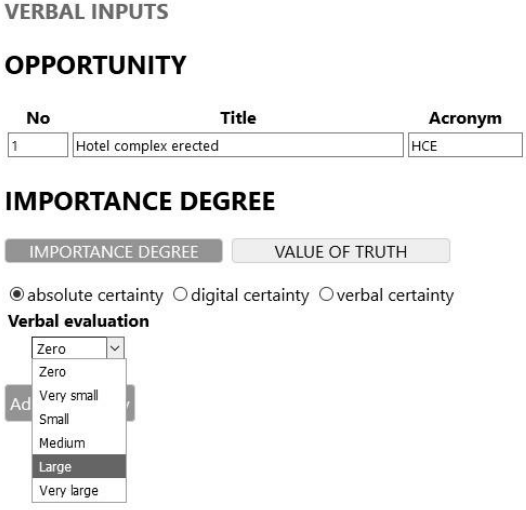


Fig. 3. Opportunity input.

Second step in data input procedure is ST and WK information as well as the data of influences. Information about strength or weakness is entered analogous to opportunities and threats. Procedure of the influence input is as follows: user chooses ST or WK component from existing list and then specifies influenced component (OP or TH). Value of influence is entered in a verbal form. There are three ways to express certainty about the given evaluation:

1. Absolute certainty – used, when there is no doubt about given estimate;
2. Digital certainty – used, when there is some uncertainty which can be evaluated;
3. Verbal certainty – possibility to express both evaluation and a level of certainty about that evaluation in verbal form.

Strength input is shown in Fig. 4.

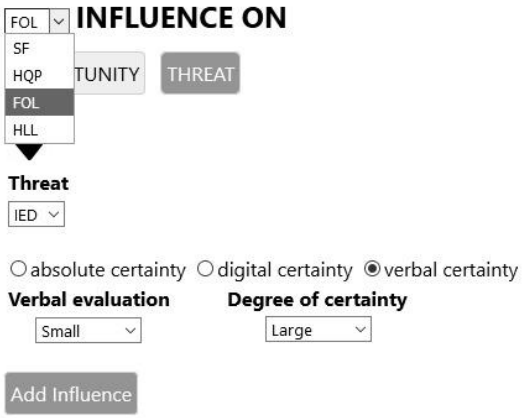


Fig. 4. Strength influence on threat.

B. Testing situation

Pilot testing was done using example from [11]. List of opportunities is shown in Fig. 5.

No.	Acronym	Name
1	HCE	Hotel complex erected
2	MID	Modern infrastructure developed
3	HPO	High profit obtained

Fig. 5. List of opportunities.

List of threats is shown in Fig. 6.

No.	Acronym	Name
1	IED	Increased erosion of dunes
2	IPE	Increased pollution of environment

Fig. 6. List of threats.

List of strengths is shown in Fig. 7.

No.	Acronym	Name
1	SF	Significant financing
2	HQP	High quality of personnel
3	FOL	Flexibility of law
4	HLL	High level of lobbying

Fig. 7. List of strengths.

List of weaknesses is shown in Fig. 8.

No.	Acronym	Name
1	LOI	Lack of infrastructure
2	HLS	High level of storms
3	IPC	increasing protests of local community

Fig. 8. List of weaknesses.

All SWOT analysis components and evaluations are presented in table. SWOT evaluation table is shown in **Error! Reference source not found.**.

	c	ρ	SF	HQP	FOL	HLL	LOI	HLS	IPC
HCE	L	S	L				M		
MID	L	S		S	M	M			M
HPO	M	M	L				M	M	
IED	L	S			S	L	S		M
IPE	L	S	M			M		M	

Fig. 9. Evaluation table.

"Degrees of importance" (c), "Values of truth" (ρ) and influences are shown in verbal form (S – small, M- medium, L- large). Some of the words (Z - zero, VS - very small and VL - very large) did not occur in our model.

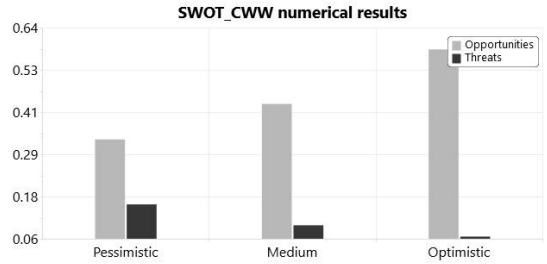
C. Experimental results

The final evaluation of summarized opportunities $OP_{\sum}$ as well as threats $TH_{\sum}$ is performed according to formulas (1) and (2):

$$OP_{\sum} = \sum_{o=1}^{O}\{c_o(\rho_o + \sum_{s=1}^{S} ST_{os} + \sum_{w=1}^{W} WK_{ow})\} \quad (1)$$

$$TH_{\sum} = \sum_{t=1}^{T}\{c_t(\rho_t + \sum_{s=1}^{S} ST_{ts} + \sum_{w=1}^{W} WK_{tw})\} \quad (2)$$

SWOT analysis results are shown in Fig. 10.



Verbal evaluation

Pessimistic	Opportunities: **VS:0.4, S:0.6**	Threats: **VS:0.67, S:0.33**
Medium	Opportunities: **VS:0.13, S:0.87**	Threats: **VS:0.91, S:0.09**
Optimistic	Opportunities: **S:0.83, M:0.17**	Threats: **Z:0.1, VS:0.9**

Fig. 10. Numerical and verbal results.

By given SWOT analysis evaluations, results are calculated and presented in three ways:

1. Optimistic – the best possible result of an overall Opportunities and Threats evaluation (Best opportunities size);
2. Pessimistic – the worst possible result of an overall Opportunities and Threats evaluation (Worst threats size);
3. Medium – the average result of overall Opportunities and Threats evaluation (Realistic view);

The tool shows numerical results in a graphical form and verbal results are shown at the bottom as the value and the certainty. Looking at the pessimistic perspective of this model, the resulting opportunities are estimated as very small (VS) with 0.4 certainty and as small (S) with 0.6 certainty. Meanwhile in the optimistic perspective common opportunities are estimated as small (S) with 0.83 certainty and as medium (M) with 0.17 certainty. These results reflect the hotel complex building in Palanga Lithuania) example from the article [11].

VI. Concluding Remarks

This paper suggests the use of verbal descriptions for SWOT analysis data input. A new prototypical software tool based on Hillson's ideology and methodology about enriching SWOT analysis with CWW paradigm was created. Successful experiment simulation based on a created tool was made and simulation results were presented. Those results can serve as expert information for risk management and decision making.

Further research objective is to create a network of tools for more complex situation analysis with more than one SWOT analysis possibility. The main idea of SWOT enhanced CWW network is to use one SWOT analysis results as influence on another connected SWOT analysis results.

Acknowledgment

I wish to thank prof. Raimundas Jasinevičius for his methodological assistance and guidelines and prof. Egidijus Kazanavičius for creating an environment for the research.

References

[1] L. A. Zadeh, „Towards Human Level Machine Intelligence - Is It Achievable? The Need for Paradigm Shift," *IEEE Computational Intelligence Magazine*, t. 3, nr. 3, pp. 11-22, September 2008.

[2] S. K. Pal, R. Banerjee, S. Dutta ir S. S. Sarma, „An Insight Into The Z-number Approach To CWW," *Fundamenta Informaticae*, t. 124, nr. 1-2, pp. 197-229, 2013.

[3] M. J. Kochenderfer, Decision Making Under Uncertainty: Theory and Application, London: The MIT Press, 2018.

[4] D. Hillson, Effective Opportunity Management for Projects: Exploiting Positive Risk, New York: Marcel Dekker, Inc., 2004, p. 316.

[5] S. Faizi, T. Rashid, W. Sałabun, S. Zafar and J. Wątróbski, "Decision Making with Uncertainty Using Hesitant Fuzzy Sets," *International Journal of Fuzzy Systems*, vol. 20, no. 1, pp. 93-103, January 2018.

[6] G. A. Miller, „The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, t. 63, nr. 2, pp. 81-97, 1956.

[7] Balžekienė, Aistė; Gaulė, Eglė; Jasinevičius, Raimundas; Kazanavičius, Egidijus; Petrauskas, Vytautas, „Risk Evaluation: The Paradigm and Tools," įtraukta *Information and Software Technologies: 21st International Conference, ICIST 2015*, Druskininkai, Lithuania, 2015.

[8] M. N. Shadlen ir R. Kiani, „Decision Making as a Window on Cognition," *Neuron*, t. 80, nr. 3, pp. 791-806, 30 October 2013.

[9] "Artificial Intelligence: How knowledge is created, transferred, and used," Elsevier, 2018.

[10] Petrauskas, Vytautas; Jasinevičius, Raimundas; Kazanavičius, Egidijus; Meškauskas, Žygimantas;, „CWW elements to enrich SWOT analysis," *Journal of Intelligent and Fuzzy Systems*, t. 34, nr. 1, pp. 307-320, January 2018.

[11] R. Jasinevičius and V. Petrauskas, "Dynamic SWOT Analysis as a Tool for Environmentalists," *Environmental Research, Engineering & Management*, vol. 43, no. 1, pp. 14-20, 2008.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Modelling Principles for Blockchain-based Implementation of Business or Scientific Processes

Mantas Jurgelaitis
Information Systems Department
Kaunas University of Technology, Informatics Faculty
Kaunas, Lithuania
mantas.jurgelaitis@ktu.lt

Vaidotas Drungilas
Information Systems Department
Kaunas University of Technology, Informatics Faculty
Kaunas, Lithuania
vaidotas.drungilas@ktu.lt

Lina Čeponienė
Information Systems Department
Kaunas University of Technology, Informatics Faculty
Kaunas, Lithuania
lina.ceponiene@ktu.lt

Rita Butkienė
Information Systems Department
Kaunas University of Technology, Informatics Faculty
Kaunas, Lithuania
rita.butkiene@ktu.lt

Evaldas Vaičiukynas
Information Systems Department
Kaunas University of Technology, Informatics Faculty
Kaunas, Lithuania
evaldas.vaiciukynas@ktu.lt

***Abstract*—Blockchain technology and smart contract development currently lacks clarity in its implementation. The complicated architecture of blockchain is an obstacle that developers face during design and implementation of blockchain-based systems. In this paper we propose a method based on Model Driven Architecture, which could be used for defining and specifying blockchain structure and behavior. Such approach could be used as one of the ways for describing blockchain-based systems in a more general language in order to facilitate blockchain development process.**

Keywords—blockchain, smart contract, MDA, UML

I. INTRODUCTION

Blockchain is a decentralized, distributed database that is shared and replicated across all the parties participating in the network. It takes a form of a public ledger for the transactions contained in the chain [1]. The blockchain is made up of blocks, where each block contains a hash of the previous block, thus linking to it [2]. The main advantages provided by the blockchain are a peer-to-peer exchange (i.e. decentralization) and robust public record keeping with the possibility to eliminate the intermediary between parties [3].

Blockchain enables development of distributed software architecture where networks of untrusted participants can establish agreements on shared states for decentralized and transactional data in a secure way. Blockchain ensures trust among parties in decentralized systems without the need of centralized supervisor in charge of verifying the correctness of the records in the ledger [4]. Since blockchain digital currencies combine features of money with those of a payment system, central banks started to investigate the technology. From the industry side, over one hundred corporations have joined blockchain working groups or consortia and the number of patents filled increased to more than three thousand in 2017 [5].

Blockchain itself cannot handle large amount of information, because the main purpose is to store simple transactional logs. As a result, scalability of blockchain is a concern for the developers [4]. Even though the technology has been around for a while, there are only a few academic blockchain-related studies. Most of the available non-scientific papers about blockchain technology are whitepapers, documentations or technical documents; some of which may be incomplete or are still being updated. A practical approach to blockchain development also demands considerable investment [6]. Developers need to have a clear understanding of the capabilities and limitations of blockchain technology. Additionally, they need to acquire the necessary skills to implement the technology and to figure out how the new trust architecture would affect the application.

Currently, the blockchain technology is most broadly applied for cryptocurrencies, but other areas are in early adoption as well. Solutions for automating the insurance process [7], supply chain management [8] even some proposals for scientific reproducibility or decentralization of scientific data exist [9] [10]. But various authors describe the same concepts differently and only a few define their development process [8] [11] [12]. The capability for replicating blockchain technology solutions is low, as it is highly dependent on the authors' environment. No formal methods exist, and common elements are difficult to find or to associate. This is particularly noticeable with the more complex elements that support more complicated scenarios [13]. The common process for development of blockchain-based system could ease the applicability in other areas than the financial sector and bridge the knowledge gap among stakeholders of blockchain technology.

Modelling is a common tool for facilitating communication, visualizing the development process and even automating the implementation of some development artefacts. The strategies of Model Driven Development [14] are often applicable in the area of software engineering. Model Driven Architecture (MDA) is a methodology, which encompasses a set of guidelines for specifying models during software development process [15]. Each MDA-based specification has three levels: A Computation Independent Model (CIM), a Platform-Independent Model (PIM), and one or more Platform-Specific Models (PSM). Unified Modelling Language (UML) [16] is commonly used for MDA model development. MDA models can represent system at different levels of abstraction, from various viewpoints, from enterprise architectures to technology implementations [17]. Model-Driven Architecture principles could be adapted to blockchain and smart contract domain together with guidelines for developing blockchain structure and smart contract behavior.

In this paper we present our idea of blockchain technology development method, based on MDA principles. The proposed method could be used for different processes for the definition of common elements and the identification of potential application areas. The method could provide a more structured approach for the development of blockchain elements, which could potentially shorten the time of development. In the future we plan on developing the method further.

The rest of the paper is organized as follows. The second section discusses the background of blockchain technology and its evolution. The third section presents current areas of blockchain application and research on improving blockchain-based systems development process. The fourth section describes our proposition on applying principles of model-driven architecture for blockchain system development. The last section summarizes concludes our main insights and outlines future work.

II. BACKGROUND

This chapter presents the basic concepts and categories of blockchain technology. In this way, we intend to define the domain in order to outline a common understanding of these technologies and attain a more comprehensive communication in this area. Although the terms "blockchain" and "distributed ledger" are used interchangeably, the blockchain is a type of distributed ledger technology and blockchain mostly differs from distributed ledger in the way the data is stored [5]. For this reason, we refer to these technologies using the terms "blockchain" and "blockchain technologies" as they cover both concepts.

A. Blockchain technology

Blockchain is a distributed database that is shared across participants [18]. Participants can independently verify information because copies of records are available in the blockchain. If a node fails, the remaining ones can continue to operate. Verification process does not depend on a centralized authority. Information is kept in a digital ledger. The transactions in the blockchain are recorded near real time. Once transactions are included in the ledger it is nearly impossible to delete or rollback the changes. Each block is timestamped, and each block has a pointer referring to the data stored in the previous block in the chain.

The data in blocks are hash sealed. The participants can interact with the blockchain only by using a generated address so that the identity of the user is not revealed. Any transaction refers to some previous transactions. Once the current transaction is recorded into the blockchain, the state of referred transactions change. That way transactions can be tracked and verified once needed.

Participants in the network authenticate and approve transactions before inclusion to the blockchain. Few different methods for reaching consensus exist. Consensus algorithms in blockchain are used to maintain data consistency in a distributed network. Usually, the basis of such algorithms is that the majority of network participants needs to approve the correctness of transaction. This way the need for a third party is avoided. In a traditional centralized transaction system, each transaction needs to be validated by a central trusted agency (e.g., the central bank).

The blockchain technologies are divided into three generations, based on the complexity of the components. The first generation of blockchain was all about cryptocurrency and its exchange possibilities. The 2.0 generation focuses heavily on the use of smart contracts built using scripting language of the blockchain. The third generation of blockchain supports decentralized applications based on blockchain technologies in other previously unsupported areas like government, health, science and culture.

1) Blockchain 1.0: Bitcoin and cryptocurrency

Blockchain technology began with Bitcoin, and many developers around the world still consider that the main blockchain example. Blockchain technology relies on a shared public ledger that entire cryptocurrency networks share and depend on. Traditional currency system participants rely on the bank to authenticate the integrity of a ledger, but blockchain relies on peer-to-peer network transfer thus eliminating the need of the third parties. Today, the first generation of blockchain technologies is mainly defined by cryptocurrencies like Bitcoin [19]. Bitcoin has proven to be an effective decentralized digital currency. The relative simplicity of Bitcoin and its inability to handle contracts limits its ability to serve a wider range of use cases.

2) Blockchain 2.0: Ethereum and smart contracts

Bitcoin introduced a very basic scripting language, that allowed some form of contractual complexity. The extension of this scripting language to handle more complex data manipulations within blockchain came to be defined as a second-generation blockchain technology. The second generation is mainly represented by Ethereum. Ethereum proposed a structure in which blockchain technology could be used to facilitate the management of digital assets. Ethereum offers new functionality through so-called smart contracts, which can manage agreements between parties on the blockchain [20]. A smart contract can manage itself, events can be triggered without the need of any party input. Second-generation blockchains can leverage distributed network for computing power, this way smart contracts can execute complicated logic. In such cases, parties do not need to pay a 'trusted' third party and could leave agreements to execute autonomously.

Unfortunately, these blockchain technologies are known to struggle with scaling difficulties. Additionally, neither Ethereum nor Bitcoin is fundamentally integrable with other decentralized currencies or platforms; meaning that in most cases users wishing to transfer value from one platform into another must do so through via exchange services [21].

3) Blockchain 3.0

The new generation blockchains come into existence with a focus to address the issues in both Blockchain 1.0 and 2.0 via different protocols, techniques and frameworks. High scalability, interoperability, adaptability, sustainability, privacy as well as instantaneous transactions are features that should separate Blockchain 3.0 from its previous iterations [19] [22]. The third generation of blockchain is at the time being developed and there are no specific blockchain solutions which define this generation. A candidate for flagship blockchain example in this category should address present flaws of existing solutions.

B. Smart Contracts

An additional implementation of more advanced data manipulation mechanism for blockchain enabled application layer development in the form of smart contracts. Basically, a smart contract is a deployed program that can be executed on

the blockchain network following the principle of trigger causing an appropriate reaction [23] [24]. Smart contracts can express triggers, conditions, and even cover entire business processes [11]. A contract can be viewed as a simple class, or it can contain complex structures, functions, modifiers, events for the implementation of various level of logic [25]. Usually, blockchains have a built-in scripting language, which is used to execute additional business logic triggered by a transaction. Recent generations of blockchains (e.g. Ethereum and Hyperledger) use integrated programming language executable by a virtual machine [2].

The consumer deals directly with the transactions on the blockchain, a smart contract holds value which is released at the time certain conditions are met, this way the contracts have lower transactional costs unlike traditional contracts [12]. Smart contracts could theoretically cover entire software applications, but most smart contracts currently are like traditional contracts for creating legally binding agreements between certain parties. Other areas of applications for smart contract hold entertainment value (e.g. CryptoKitties [26]), unlike aforementioned contracts, these contracts are most likely developed by people with interest in Solidity (contract-oriented programming language for writing smart contracts on Ethereum blockchain) [11]. Smart contract code generation would potentially simplify the smart contract development process, raise the abstraction level and increase potential usage in various domains.

Blockchain technologies are rapidly evolving and the area of their application broadens. Our research focusses on analyzing the applicability of these technologies in various areas and possibilities of improving the development process of blockchain-based systems.

III. RELATED WORK

In this section we overview applications of blockchain technologies in various fields and discuss difficulties of application of blockchain technology to diverse domains. The research on the applicability of modelling techniques to blockchain-based system development is also overviewed.

One of the blockchain application areas is tracking the provenance of assets. Solutions like supply chain management are one of the business issues that could benefit from automation with smart contracts. Evaluation of provenance is generally difficult not only because of the number of goods that are handled in complex supply chains but also because of the amount of information for tracking product location, physical characteristics. As a solution to this problem, simple data models of the ontology were described. These models were later used to develop smart contract implementation [8].

Another way to utilize smart contract on blockchain is insurance-related contracts. The extended solution in a form of framework exists to help developers deploy more secure and less costly contracts. Smart contracts can automate the processes of insurance operations such as client registration, policy assignment, premium payouts, submission claims, and processing of refunds without or with minimal involvement of third parties [7].

As well as business processes, there are many suggestions to improve scientific processes using blockchain technologies [9] [10]. Proposals for implementing blockchain technologies could be grouped into three groups: the first that uses blockchain as a storage unit or as a token indicating possession and the second that proposes using blockchain to work on scientific computations giving monetary rewards or free of charge. There are suggestions for blockchain to be used as a platform to store medical data [27]. Storing medical data in blockchain makes medical data more accessible for medical staff and general public alike. Also, there are propositions to use blockchain technology as a proof of intellectual work [10]. This could enable scientists or members of the general public to store ideas in the blockchain with a timestamp of the proposed idea. The second category is run by business and scientific organizations alike [28] [29]. The third category uses blockchain technologies for both storing data and performing computational tasks. An example of the third category could be blockchain used for federated learning and also analyzed the latency aspects of their proposed architecture [30].

Blockchain use cases continue to grow in scope and complexity, that is why the need for common guidelines becomes apparent [31]. For blockchain to be accepted as a technology in other industries besides financial, it is essential that stakeholders have a common understanding of the technology and possibilities of blockchain [32]. A few proposals for standardization of software engineering of blockchain technologies have emerged during the last year [11] [33] [34] [35]. The author of [11] proposes a development method which includes smart contract development approach based on MDA. In [33] a general proposal is presented for extending existing modelling notations to include specific blockchain concepts or integrations. An approach for the modelling blockchain business networks via layer-based modeling and ontology design is presented in [34]. Authors describe abstraction layer which can be used to describe blockchain and develop a Blockchain Business Network Ontology which depicts common terms for blockchain networks. A model-driven approach for generation of smart contract code of is described in [35]. Authors develop BPMN process model for collaborative business process and use it for generation of smart contract code. In the area of information systems, the application of blockchain technology and cryptocurrencies is still quite limited [2] [5]. In order to adapt blockchain technology to specific needs, the main attention should be paid to the development and implementation methodologies of such technologies.

The lack of a formal and unified methodology complicates the application of blockchain technology. Only a small portion of authors describe their development process, unfortunately these are often specialized for specific use cases. A universal method for development could ease the design and implementation process of blockchain technology-based systems [36]. Introduction of some standard way for defining and specifying blockchain structure and behavior could facilitate the blockchain development process [37]. There are proposals suggesting that blockchain components could be modeled using BPMN and UML [33], although researchers are not proposing to apply MDA to the whole development process. A similar proposal of using modelling to define smart contracts have been proposed in [38]. An approach to Ethereum smart contract development was also proposed [11]. The analyzed proposals for modelling of blockchain and smart contracts are in the early stages and have not yet been extensively validated or tested.

Based on the analysis of blockchain application areas and proposed techniques for its implementation, introduction of a

common methodology for blockchain-based system development could facilitate the development process and broaden the scope of its applications.

IV. Applying the Principles of Model-Driven Architecture for Blockchain-Based Systems Development

To facilitate the development of a blockchain-based system, we propose a methodology based on MDA principles. We believe, that such methodology could help to describe blockchain-based systems concepts and behavior in a more general language [37]. Furthermore, some development actions could be accelerated by automation. This methodology is seen as covering five system development stages (Figure 1). For each stage, a certain type of resource required and a certain set of outcomes is identified. UML and its extension in a form of UML profile for blockchain specified using domain specific language (DSL) [39] are proposed as a language for modelling different aspects of the system. Below the methodology is explained in more detail.

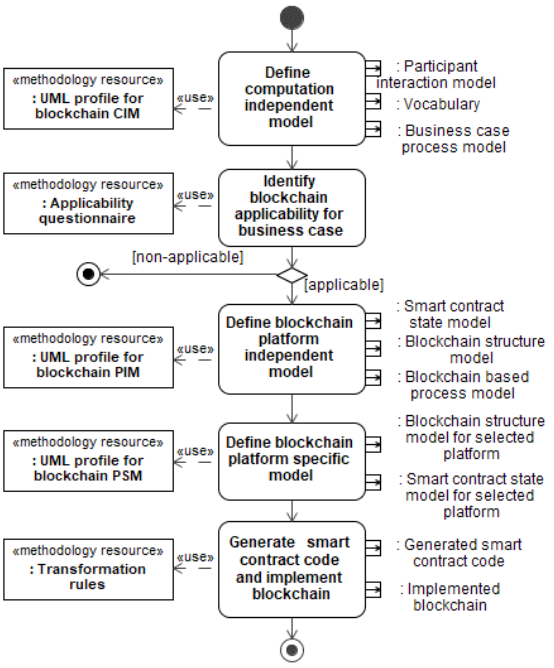


Figure 1. The proposed process for blockchain-based business process implementation

In the first development stage Computation Independent Model (CIM) should be specified. The purpose of this model is to show how blockchain could be adapted for specific processes' restructuring, reorganization, and integration. In MDA, a Computation Independent Model (CIM) is often referred to as a business or domain model. It presents the context of the system under development and what the system is expected to do but hides all information technology related specifications to remain independent. A model should be created using the provided UML profile for blockchain CIM. It would consist of a participation interaction model, vocabulary and a business case process model.

Due to the fact that blockchain technology is not applicable in every case, solutions helping to determine the suitability of blockchain exists [40]. These questions/frameworks help to assess the advantages and limitations of blockchain technology and the applicability of the technology for general purposes. Following this example, the second step of the method would be to identify whether blockchain is applicable in the specific case. Using the defined CIM and applicability questionnaire the specific case would be assessed, and the outcome would be the conclusion whether to proceed with the development of blockchain.

If blockchain technology is applicable, the development of the blockchain would continue with the design of the Platform Independent Model (PIM). Traditionally PIM represents the design of the system without the details about its implementation. Considering the method is tailored for the blockchain, the defined models would include details about implementation, but would not be based on any specific blockchain technology (e.g. Hyperledger or Ethereum). The previously described CIM models would be used as an input for the development of the PIM. Using the blockchain PIM, a smart contract state model, blockchain structure model and blockchain-based process model can be defined. These defined models would help to select a specific platform for further development because different platforms differ in terms of chain architecture, transaction structure, number of participants, smart contract capabilities, consensus algorithms and so on.

Afterwards the development of blockchain Platform Specific Model (PSM) would take place. MDA suggests automating the production of a PSM from previously defined models. It requires to define transformation rules which specify how models are transformed based on parameters defined by developers [15] [17]. The PSM of a system is defined and tailored for a specific platform. In our proposed method a PSM would be developed for a specific blockchain implementation. The previously defined PIM model would be used to define blockchain structure and smart contract model for the selected platform. Specialized model for a particular blockchain solution could be provided in a form of DSL.

Finally, in the last stage applying transformation rules for the specific blockchain PSM model the code for smart contract and blockchain implementation of that particular platform would be generated. The generated artefacts could be used to start building specific blockchain technology implementation.

The proposed methodology would not only help to define the blockchain artefacts, but also help to identify whether it is reasonable to adopt the technology. It could also offer guidelines for selecting appropriate blockchain platform. The solution would provide the possibility to facilitate and at least partially automate the blockchain technology-based system development process by providing a more standardized way of describing such systems, expanding the potential uses of blockchain for different business goals by restructuring current business processes.

V. Conclusion

The blockchain technologies are currently most broadly applied in the financial sector, and the application in other areas is still quite limited. The design process of blockchain technology-based systems is quite difficult, because no formal or formalized design, development methodologies exist. A proposed application of MDA principles in the process of

development of blockchain technology-based systems should help to determine whether it is possible to model blockchain structure and smart contract logic and whether the business logic could be conveyed in the smart contracts. The methodology could be used for modelling blockchain and smart contracts of business processes and thus relocating these to the blockchain. In addition, an extensive analysis of business processes is still required for determining how the relocation of the business logic to the blockchain could affect the current processes.

Going forward, it is essential to thoroughly examine the possibilities of adopting MDA principles for blockchain technology-based system development process. It is important to analyze blockchain implementations and find common elements. The results should help to determine how to model blockchain structure and smart contract logic, and how business logic can be conveyed by the smart contracts.

REFERENCES

[1] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," 2018. [Online]. Available: https://bitcoin.org/bitcoin.pdf .

[2] F. Glaser, "Pervasive Decentralisation of Digital Infrastructures: A Framework for Blockchain enabled System and Use Case Analysis," in Proceedings of the 50th Hawaii International Conference on System Sciences , 2017.

[3] M. Swan, Blockchain Blueprint for a New Economy, O'Reilly Media, 2015.

[4] S. Raval, Decentralized Applications: Harnessing Bitcoin's Blockchain Technology, O'Reilly Media, 2016.

[5] P. Tasca and C. J. Tessone, "Taxonomy of Blockchain Technologies. Principles of Identication and Classication," 2018. [Online]. Available: https://dx.doi.org/10.2139/ssrn.2977811.

[6] D. W. Cearley, B. Burke, S. Searle and M. J. Walker, "Top 10 Strategic Technology Trends for 2018," Garnter, 2017.

[7] M. Raikwar, S. Mazumdar, S. Ruj, S. S. Gupta, A. Chattopadhyay and K.-Y. Lam, "A Blockchain Framework for Insurance Processes," in 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2018.

[8] H. M. Kim and M. Laskowski, "Toward an ontology-driven blockchain design for supply-chain provenance," Intelligent Systems in Accounting, Finance and Management, 28 March 2018.

[9] b8d5ad9d974a44e7e2882f986467f4d3, "Towards Open Science: The Case for a Decentralized Autonomous Academic Endorsement System," 12 4 2016. [Online]. Available: 10.5281/zenodo.60054.

[10] M. Sharples and J. Domingue, "The Blockchain and Kudos: A Distributed System for Educational Record, Reputation and Reward," in Learning: Proceedings of 11th European Conference on Technology Enhanced Learning (EC-TEL 2015), Lyon, 2016.

[11] K. Boogaard, A Model-Driven Approach to Smart Contract Development, 2018.

[12] V. Buterin, "A Next-Generation Smart Contract and Decentralized Application," 2014. [Online]. Available: http://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf.

[13] D. Furlonger and R. Valdes, "Practical Blockchain: A Gartner Trend Insight Report," Gartner, 2017.

[14] O. Pastor, S. España, J. I. Panach and N. Aquino, "Model-Driven Development," Informatik Spektrum , 2008.

[15] O. Pastor and J. C. Molina, Model-Driven Architecture in Practice, Springer, 2007.

[16] O. M. Group, "UML 2.5 Specification," 01 03 2015. [Online]. Available: http://www.omg.org/spec/UML/2.5/PDF.

[17] Object Management Group, "Model Driven Architecture (MDA) MDA Guide rev. 2.0," 18 June 2014. [Online]. Available: https://www.omg.org/cgi-bin/doc?ormsc/14-06-01.

[18] Deloitte, "Blockchain @ Media | A new Game Changer for the Media Industry?," 2017 . [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/tr/Documents/technology-media-telecommunications/deloitte-PoV-blockchain-media.pdf.

[19] B. Smith, "What are the three generations of blockchain, and how are they similar to the web?," 2018. [Online]. Available: https://www.coininsider.com/three-generations-of-blockchain/.

[20] G. Wood, "Ethereum: a Secure Decentralised Generalised Transaction Ledger," 2014. [Online]. Available: https://gavwood.com/paper.pdf.

[21] B. Smith, "The blockchain can succeed like the web – here's how," Coin Insider, 18 December 2018. [Online]. Available: https://www.coininsider.com/the-blockchain-can-succeed-like-the-web-heres-how/.

[22] "What is Blockchain Technology?," 2018. [Online]. Available: https://blockgeeks.com/guides/what-is-blockchain-technology/.

[23] M. Vincenzo, Business Innovation Through Blockchain: The B3 Perspective, 2017, p. 101–124.

[24] S. Omohundro, "Cryptocurrencies, smart contracts, and artificial intelligence," AI Matters, vol. 1, no. 2, pp. 19-21, 2014.

[25] N. Prusty, Building Blockchain Projects, Packt Publishing, 2017.

[26] U. W. Chohan, The Leisures of Blockchains: Exploratory Analysis, SSRN, 2017.

[27] A. Azaria, A. Ekblaw, T. Vieira and A. Lippman, "MedRec: Using Blockchain for Medical Data Access and Permission Management," in International Conference on Open and Big Data, Vienna, 2016.

[28] J. Zawistowski, P. Janiuk, A. Regulski and A. Skrzypczak, "The Golem Project," 2016. [Online]. Available: https://golem.network/crowdfunding/Golemwhitepaper.pdf.

[29] G. Fedak, W. Bendella and E. Alves, "Blockchain-Based Decentralized Cloud Computing," 2018. [Online]. Available: https://iex.ec/wp-content/uploads/pdf/iExec-WPv3.0-English.pdf.

[30] H. Kim, J. Park, M. Bennis and S.-L. Kim, "On-Device Federated Learning via Blockchain and its Latency Analysis," CoRR, 2018.

[31] E. Piscini, D. Dalal, D. Mapgaonkar and P. Santhana, "Blockchain to blockchains," in Tech Trends 2018: The symphonic enterprise, 2017.

[32] J. d. Kruijff and H. Weigand, "Understanding the Blockchain Using Enterprise Ontology," 2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-59536-8_3.

[33] H. Rocha and S. Ducasse, "Preliminary Steps Towards Modeling Blockchain Oriented Software," in 1st International Workshop on Emerging Trends in Software Engineering for Blockchain, 2018.

[34] S. Seebacher and M. Maleshkova, "A Model-driven Approach for the Description of Blockchain Business Networks," in Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

[35] X. Xu, I. Weber and M. Staples, "Model-Driven Engineering for Blockchain Applications," in Architecture for Blockchain Applications , Springer, 2019, pp. 149-172.

[36] C.-F. Liao, S.-W. Bao, C.-J. Cheng and a. K. Chen, "On Design Issues and Architectural Styles for Blockchain-driven IoT Services," in IEEE International Conference on Consumer Electronics , Taiwan, 2017.

[37] A. B. Tran, X. Xu, I. Weber, M. Staples and P. Rimba, "Regerator: a Registry Generator for Blockchain," in CaiSE2017: 29th International Conference on Advanced, Essen, Germany, 2017.

[38] M. Marchesi, L. Marchesi and R. Tonelli, "An Agile Software Engineering Method to Design Blockchain Applications," in Software Engineering Conference Russia, Moscow, 2018.

[39] A. V. Deursen, E. . Visser and J. . Warmer, "Model-Driven Software Evolution: A Research Agenda," , 2007. [Online]. Available: http://swerl.tudelft.nl/twiki/pub/eelcovisser/modeldrivensoftwareevolutionaresearchagenda/dvw07.pdf. [Accessed 30 1 2019].

[40] K. Wüst and A. Gervais, "Do you need a Blockchain?," 2017.

Generation of high order primitive matrix elements with elements of abelian multiplicative groups with different power for post-quantum key exchange protocol

Richard Megrelishvili
Math *dept. Tbilisi State University*
Tbilisi, Georgia
richard.megrelishvili@tsu.ge

Melkisadeg Jinjikhadze
Math dept. Akaki Tsereteli State University
Kutaisi, Georgia
mjinji@yahoo.com

Avtandil Gagnidze
Faculty of management.
Bank of Georgia University
Tbilisi, Georgia
gagnidzeavto@yahoo.com

Maksim Iavich
School of technology. Caucasus University
Tbilisi, Georgia
m.iavich@scsa.ge

Giorgi Iashvili
School of technology. Caucasus University
Tbilisi, Georgia
g.iashvili@scsa.ge

***Abstract*— Active work is performed to create quantum computers. Quantum computers can break existing public key cryptography. So they can break Diffie-Hellman key exchange protocol. Matrix algorithms of key exchange can be considered as the alternative of Diffie-Hellman key exchange protocol.**

The improved method of key-exchange protocol is offered in the article. The method deals with the original matrix one-way function and the generalized method of processing the corresponding high order matrix multiplicative finite commutative group.

The general method of the insertion-enlarging method of building the primitive elements of the field is derived with elements of the matrix groups with different power.

The article describes the results that give us the prospect of generating the high order multiplicative Abelian matrix groups and of creating key-exchange protocol resistant to quantum computers attacks by means of this groups.

Keywords*— *Matrix One-way Function, Abelian Finite Field, Asymmetric Cryptography, High order finite matrix Field, Primitive Matrix Element, quantum computers, post-quantum cryptography.

I. Introduction

Scientists and experts are actively working on the creation of quantum computers. GOOGLE Corporation, NASA the association USRA (Universities Space Research Association and D-Wave teamed-up to develop quantum processors.
Quantum computers can break existing public-key crypto systems. Quantum computer solves the discrete logarithm problem both for finite fields and elliptic curves. Being able to efficiently calculate discrete logarithms it can break Diffie-Hellman key exchange protocol.
Quantum computer also solves the factorization problem, so it can easily break RSA cryptosystem.
Public-key cryptography is used in different products on different platforms and in various fields. Many commercial products use public-key cryptography, the number of which is actively growing. Public-key cryptography is also widely used in operating systems from Microsoft, Apple, Sun, and Novell. It is used in secure phones, Ethernet, network cards, smart cards, and it is widely used in cryptographic hardware. Public-key technology is used in protected Internet communications, such as S / MIME, SSL and S / WAN. It is used in government, banks, most corporations, different laboratories and educational organizations. Breaking existing public-key crypto-systems will cause complete chaos [1,2].
Public-key crypto systems resistant to quantum attacks are developed. But nowadays successful attacks are recorded on these systems [3,4].

II. One-way matrix function

One of the modifications of Diffie-Hellman's well-known method of cryptographic key exchange is the matrix algorithms of the exchange, the basis for which is the high order cyclic multiplicative matrix groups in the GF (2) field.
Suppose that the P matrix is a primitive element of a cyclic matrix group. While (P) is a multiplicative group formed by this matrix, with the power $2^n - 1$, where n is the size of the square matrix.
The matrix algorithm for general key development is the following:

- The sender sends to the receiving party via the open channel $u_1 = vP_1$ vector, where $P_1 \in \langle P \rangle$ is the secret matrix selected by the sender, and $v \in V_n$ is commonly known (V_n – is vector space on $GF(2)$ field);
- The receiving party chooses $P_2 \in \langle P \rangle$ to send a secret matrix and send to the sender $u_2 = vP_2$ vector;
- Sender calculates $k_1 = u_2 P_1$ vector;
- Receiver calculates $k_2 = u_1 P_2\, v$, where k_1 and k_2 – are secret keys.

It is evident, $k_1 = k_2 = k$, because of $k = vP_1P_2 = vP_2P_1$, while $\langle P \rangle$ is commutative group. Let $v = (v_1, v_2, v_3, \cdots, v_n) \in V_n$ and $u = (u_1, u_2, u_3, \cdots, u_n) \in V_n$ are non-secret vectors from above mentioned algorithm and

$$P_1 = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \in \langle P \rangle$$

is the secret matrix. Then according to algorithm:

$$vP_1 = \begin{pmatrix} v_1a_{11} + v_2a_{21} + \cdots + v_na_{n1} \\ v_1a_{12} + v_2a_{22} + \cdots + v_na_{n2} \\ \vdots \\ v_1a_{n1} + v_2a_{n2} + \cdots + v_3a_{n3} \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad (1)$$

The number of variables in the system of linear equations is the square of the number of equations. Generally, solutions in such cases are not uniquely defined and the system has infinitely many solutions. However, because we deal with a special types of matrices, the solution is defined uniquely. In addition, it is obvious that the solution of the system is very time consuming and is practically impossible in real time if the size of the matrix is large enough.

All of the above makes it necessary to generate high order Abelian multiplication matrix group, whose primitive element will be a high order quadratic matrix.

III. FINITE MATRIX GROUPS

Let's consider $(1+\alpha)^j$, where $j = 0, 1, 2, \cdots$, and α represents the root of primitive polynomial in the $GF(2^n)$ field with the module $p(x)$.

$$\begin{aligned} (1+\alpha)^0 &= 1 && 1 \\ (1+\alpha)^1 &= 1+\alpha && 11 \\ (1+\alpha)^2 &= 1+\alpha^2 && 101 \\ (1+\alpha)^3 &= 1+\alpha+\alpha^2+\alpha^3 && 1111 \\ (1+\alpha)^4 &= 1+\alpha^4 && 10001 \\ (1+\alpha)^5 &= 1+\alpha+\alpha^4+\alpha^5 && 110011 \end{aligned}$$

The polynomial coefficients generated by the above structure are known as the Serpinsky triangle. Serpinsky's structure contains a number of sub-structures that can be used as a generator (generating matrix) for multiplication groups, i.e. primitive elements. Such is, for example,

$$P_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad P_5 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix},$$

$$P_7 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \quad (2)$$

And many others. Their natural powers form the Abelian multiplicative cyclic group.

It's easy to see that natural powers of matrices P_3 , P_5 , P_7

P_3^k , P_5^k , P_7^k , $k = 1,2, \dots, 2^k - 1$ (3)

Form the Abelian multiplicative cyclic group:

$$P_3^1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, P_3^2 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \ P_3^3 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, P_3^4 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$
$$P_3^5 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, P_3^6 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \ P_3^7 = P_3^0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (4)$$

Therefore we derived the insertion-enlargement method of second order enlargement of the basic structure of P_3 [5].

Let's keep the structure of the matrix P_3 and enlarge them by elements of the set (4) as following [5]:

$$P_{3^2}(i,j)=\begin{pmatrix} P_3^i & P_3^j & P_3^j \\ P_3^j & 0 & 0 \\ P_3^j & P_3^j & 0 \end{pmatrix}, \text{ where } i,j=0..6. \qquad (5)$$

P_3 matrix is called basic structure and let's call P_3^i and P_3^j matrices the first and the second enlarged matrices and $P_{3^2}(i,j)$ matrix let's call the second order $(\boldsymbol{i},\boldsymbol{j})$ enlargement of P_3.

The set P_3^k, $k=1,2,\dots,2^3-1$ is called the primary group for $F(P_{3^2}(i,i+1))$ group.

We have proved the following statement:

Any second order $(i,i+1)$ enlargement of P_3 $P_{3^2}(i,i+1)$, $i=0..5$, is a primitive element and forms a finite Abelian multiplicative group $F(P_{3^2}(i,i+1))$, with power $2^{3^2}-1$.

For example, the matrix $P_{3^2}(0,1)$ is primitive and the matrix $[P_{3^2}(0,1)]^{2^{2\cdot 3^1}+2^{3^1}+1}$ is diagonal matrix:

$$[P_{3^2}(0,1)]^{2^{2\cdot 3^1}+2^{3^1}+1}=\begin{pmatrix} P_3^3 & 0 & 0 \\ 0 & P_3^3 & 0 \\ 0 & 0 & P_3^3 \end{pmatrix} \qquad (6)$$

All powers $\left([P_{3^2}(0,1)]^{2^{2\cdot 3^1}+2^{3^1}+1}\right)^i, i=1,2,\dots$ of the diagonal matrix are also diagonal and, because of the set $F(P_{3^2}(0,1))$ is a finite group, when $i=2^{3^1}-1$, we see:

$$\left([P_{3^2}(0,1)]^{2^{2\cdot 3^2}+2^{3^2}+1}\right)^i=\begin{pmatrix} (P_3^3)^i & 0 & 0 \\ 0 & (P_3^3)^i & 0 \\ 0 & 0 & (P_3^3)^i \end{pmatrix}$$

$$\begin{pmatrix} (P_3^3)^i & 0 & 0 \\ 0 & (P_3^3)^i & 0 \\ 0 & 0 & (P_3^3)^i \end{pmatrix}=\begin{pmatrix} P_3^{3i \bmod i} & 0 & 0 \\ 0 & P_3^{3i \bmod i} & 0 \\ 0 & 0 & P_3^{3i \bmod i} \end{pmatrix} \qquad (7)$$

As we perform the matrix operations with respect to the module of the primary group, the matrix (7) is the identity matrix. That means, that the set $F(P_{3^2}(0,1))$ is a finite group.

Note that we can consider any element of (4) as the basic structure. There exist the enlargement of this element using P_3^0 and P_3^1 matrices, that is primitive.

For example, following enlargements are primitive:

$$\begin{pmatrix} P_3^1 & P_3^1 & 0 \\ 0 & 0 & P_3^1 \\ P_3^0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} P_3^1 & 0 & P_3^0 \\ P_3^1 & P_3^1 & P_3^1 \\ 0 & P_3^1 & P_3^1 \end{pmatrix}, \begin{pmatrix} 0 & P_3^1 & 0 \\ 0 & P_3^1 & P_3^1 \\ P_3^0 & 0 & P_3^1 \end{pmatrix} \qquad (8)$$

Consider higher order sub-structures of the Serpinsky triangle:

$$P_5=\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

It is easy to check that P_5 is a primitive element. It means that P_5^k, $k=1,2,\dots,2^5-1$ is a finite Abelian multiplicative group.

Consider basic structure P_3 and enlarge it by P_5^0 and P_5^1 matrices.

There exist such enlargements of P_3 matrix by P_5^0 and P_5^1 matrices, that are primitive elements. For example:

$$\begin{pmatrix} P_5^1 & P_5^1 & P_5^1 \\ P_5^1 & 0 & 0 \\ P_5^0 & P_5^1 & 0 \end{pmatrix}, \begin{pmatrix} P_5^1 & P_5^1 & P_5^1 \\ P_5^1 & 0 & 0 \\ P_5^1 & P_5^0 & 0 \end{pmatrix} \qquad (9)$$

matrices are primitive elemenents. Therefore the set $F\left(P_{3\times 5^1}(P_5^0,P_5^1)\right)$ is a finite Abelian multiplicative group.

It is easy to check that

$$[P_{3\times 5^1}(P_5^0,P_5^1)]^{2^{2\cdot 5^1}+2^{5^1}+1}=\begin{pmatrix} P_5^2 & 0 & 0 \\ 0 & P_5^2 & 0 \\ 0 & 0 & P_5^2 \end{pmatrix} \qquad (10)$$

is a diagonal matrix. With the analogy of (7) we see that $\left([P_{3\times 5^1}(P_5^0,P_5^1)]^{2^{2\cdot 5^1}+2^{5^1}+1}\right)^i, i=1,2,\dots$ matrices are diagonal, and when $i=2^{5^1}-1$, we see

$$\begin{pmatrix} P_5^{2i \bmod i} & 0 & 0 \\ 0 & P_5^{2i \bmod i} & 0 \\ 0 & 0 & P_5^{2i \bmod i} \end{pmatrix} \qquad (11)$$

the identity matrix. That means, the set $F\left(P_{3\times 5^1}(P_5^0,P_5^1)\right)$ a finite Abelian multiplicative group with power of $2^{3\times 5^1}-1$.

Consider now enlargements of order k=2 P_5^i, $i = 1,2,\dots,2^5-1$ of the primitive element P_5 using elements $P_{5^k}(P_5^0, P_5^1), k = 2$. There exist such enlargements, that form a primitive matrix. For example:

$$P_{5^k}(P_5^0, P_5^1) = \begin{pmatrix} P_5^1 & P_5^1 & P_5^1 & P_5^1 & P_5^0 \\ P_5^1 & 0 & 0 & 0 & 0 \\ P_5^1 & P_5^1 & 0 & 0 & 0 \\ P_5^1 & 0 & P_5^1 & 0 & 0 \\ P_5^1 & P_5^1 & P_5^1 & P_5^1 & 0 \end{pmatrix}, k = 2 \quad (12)$$

Using the software we developed it could be seen that the matrix $\left(P_{5^k}(P_5^0, P_5^1)\right)^i$, where $i = 2^{4\cdot 5^{k-1}} + 2^{3\cdot 5^{k-1}} + 2^{2\cdot 5^{k-1}} + 2^{5^{k-1}} + 1,\ k = 2$ is a diagonal matrix:

$$\left(P_{5^k}(P_5^0, P_5^1)\right)^{2^{4\cdot 5^{k-1}}+2^{3\cdot 5^{k-1}}+2^{2\cdot 5^{k-1}}+2^{5^{k-1}}+1} = \begin{pmatrix} P_5^4 & 0 & 0 & 0 & 0 \\ 0 & P_5^4 & 0 & 0 & 0 \\ 0 & 0 & P_5^4 & 0 & 0 \\ 0 & 0 & 0 & P_5^4 & 0 \\ 0 & 0 & 0 & 0 & P_5^4 \end{pmatrix}$$

All powers $\left(P_{5^k}(P_5^0, P_5^1)\right)^{i\times j}$, are diagonal matrices, with elements P_5^i, $i = 1,2,\dots,2^5-1$ from primary group on diagonal. When $j = 2^{3^{k-1}} - 1$, we have

$$\begin{pmatrix} P_5^{4j \bmod j} & 0 & 0 & 0 & 0 \\ 0 & P_5^{4j \bmod j} & 0 & 0 & 0 \\ 0 & 0 & P_5^{4j \bmod j} & 0 & 0 \\ 0 & 0 & 0 & P_5^{4j \bmod j} & 0 \\ 0 & 0 & 0 & 0 & P_5^{4j \bmod j} \end{pmatrix} = \begin{pmatrix} P_5^0 & 0 & 0 & 0 & 0 \\ 0 & P_5^0 & 0 & 0 & 0 \\ 0 & 0 & P_5^0 & 0 & 0 \\ 0 & 0 & 0 & P_5^0 & 0 \\ 0 & 0 & 0 & 0 & P_5^0 \end{pmatrix}$$

the identity matrix.

CONCLUSIONS

The results described above give us the prospect of generating the high order multiplicative Abelian matrix groups and of the creating key-exchange protocol resistant to quantum computers attacks.

ACKNOWLEDGMENT

The Work Was Conducted as a Part of Research Grant of Joint Project of Shota Rustaveli National Science Foundation and Science & Technology Center in Ukraine [№ STCU-2016-08]

REFERENCES

[1]. Gagnidze A.G., Iavich M.P., Iashvili G.U., Analysis of Post Quantum Cryptography use in Practice, Bulletin of the Georgian National Academy of Sciences, vol. 11, no. 2, 2017, p.29-36

[2]. Song F. (2014) A Note on Quantum Security for Post-Quantum Cryptography. In: Mosca M. (eds) Post-Quantum Cryptography. PQCrypto 2014. Lecture Notes in Computer Science, vol 8772. Springer, Cham

[3]. Antonio Acín, Nicolas Brunner, Nicolas Gisin, Serge Massar, Stefano Pironio, and Valerio Scarani Phys. Rev. Lett. 98, 230501 – Published 4 June 2007

[4]. 1. Dinh H., Moore C., Russell A. (2011) McEliece and Niederreiter Cryptosystems That Resist Quantum Fourier Sampling Attacks. In: Rogaway P. (eds) Advances in Cryptology – CRYPTO 2011. CRYPTO 2011. Lecture Notes in Computer Science, vol 6841. Springer, Berlin, Heidelberg

[5]. 5. R. Megrelishvili, M. Jinjikhadze, M. Iavich, A. Gagnidze, G. Iashvili, Post-quantum Key Exchange Protocol Using High Dimensional Matrix; CEUR Workshop Proceedings (http://ceur-ws.org/Vol-2145/), Vol-2145 2019

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Public-Key hybrid cryptosystem based on Blowfish and RSA

Elza Jintcharadze
Faculty of Informatics and Control Systems
Georgian Technical University
Tbilisi, Georgia
elza.jincharadze@gmail.com

Maksim Iavich
Cyber Security Department
Caucasus University,
Tbilisi, Georgia
m.iavich@scsa.ge

***Abstract*— Nowadays data security is one of the important issues, especially for increasing transactions via the internet. This paper presents a hybrid cryptosystem using RSA (Asymmetric) and Blowfish (Symmetric) algorithm. Hybrid encryption is a combination of symmetric and asymmetric encryption methods. Symmetric algorithms are mostly used for encryption of messages than asymmetric.**

The objective of this research is to evaluate the performance of RSA, Blowfish cryptography algorithms and RSA&Blowfish hybrid cryptography algorithm. The performance of the implemented encryption algorithms is evaluated by means of encryption and decryption time and memory usage. To make comparison experiments, for those algorithms is created program implementation. The programming language Java is used for implementing the encryption algorithms.

Keywords— Symmetric cryptography, Asymmetric cryptography, Data encryption, Ciphertext, Decryption, Hybrid cryptosystem.

I. INTRODUCTION

Nowadays strength of the cryptosystem can not be totally ensured. Main goal of all cryptography algorithms is to offer best security, but due to fact that technology is rapidly developing proposed security systems becoming less resistant to every known or new attacks.

Both symmetric and asymmetric key algorithms have their advantages and disadvantages. Symmetric key algorithms are faster that asymmetric algorithms. Main requirement is that secret key must be shared in a secured way. Asymmetric systems provide secure transmission of keys, but this process needs much more time. To improve this problem is used the hybrid algorithm, which means using different type of cryptosystems together [22].

II. RSA

RSA is founded in 1977 is a public key cryptosystem. RSA is an asymmetric cryptographic algorithm named after its founders Rivest, Shamir &Adelman [5]. In general, RSA cryptosystem is used to provide privacy and ensure authenticity of digital data. Nowadays RSA is implemented in many commercial systems. RSA is used to ensure privacy and authenticity for web servers and browsers, to provide security for web Email and remote login sessions for credit-card payment systems. RSA is frequently used in applications where security of digital data is important.

RSA generates two keys: public key for encryption and private key to decrypt message. RSA algorithm can be divided into three steps: first step is to generate key which can be used as key to encrypt and decrypt data; Second step is encryption, where plaintext is converted into cipher text; and third step is decryption, where encrypted text is converted in to plain text at other side. RSA is based on factoring problem of finding product of two large prime numbers. Key size is 1024 to 4096 bits [5].

The negative side of RSA algorithm is low speed of encryption. Because encryption and decryption process with RSA algorithm needs more time than other algorithms. As other symmetric encryption systems, RSA uses two different keys: A public and a private one. Both keys work corresponding to each other, which means that a message encrypted with one of them can only be decrypted by its counterpart. The latter is usually available to the public because private key cannot be calculated from the public key.

III. BLOWFISH

Blowfish is one of symmetric key algorithm with 64-bit block cipher and it was developed by Bruce Schneider [1]. Blowfish is a block cipher, the encryption process and the decryption, Blowfish divides a message into blocks of equal size in length, i.e. 64 bits. Nowadays blowfish provides good security level and there is no any successful crypto attack against it. By encryption time Blowfish is faster than DES, but the weak point for this algorithm is it weak key.

IV. DESCRIPTION OF HYBRID CRYPTOSYSTEM

Hybrid encryption is a method of encryption that combines two or more encryption systems. It integrates a combination of asymmetric and symmetric encryption to take a benefit from the strengths of each form of encryption. These strengths of algorithm are defined as speed and security of this algorithm. Hybrid encryption is considered a highly secure type of encryption as long as the public and private keys are fully secure [4].

A hybrid encryption scheme is one that combines the convenience of an asymmetric encryption scheme with the effectiveness of a symmetric encryption scheme. There are various advantages for combination of encryption methods. One is that users have the ability to communicate through hybrid encryption. Usually during encryption process asymmetric algorithm is slowing down the encryption process [15]. But hybrid cryptosystem is using symmetric

encryption synchronously so both forms of encryption (symmetric and asymmetric) are improved. The result of hybrid encryption process has additional security level with overall improved system performance.

Symmetric and asymmetric cryptography algorithms have their own advantages and disadvantages. In general, symmetric ciphers are considerably faster than asymmetric ciphers, but require all parties to somehow share a secret key. Also, we have to take into consideration that asymmetric algorithms allow public key arrangements and key exchange systems, but this slowdowns encryption process speed [4]. A hybrid cryptosystem is using multiple ciphers of different types together, each to its best advantage. One common method of hybrid cryptosystem is to generate a random secret key for a symmetric cipher, and then encrypt this key via an asymmetric cipher using the recipient's public key. After this step the plaintext is encrypted using the symmetric cipher and the secret key. After encryption process the encrypted secret key and the encrypted message will be then sent to the receiver.

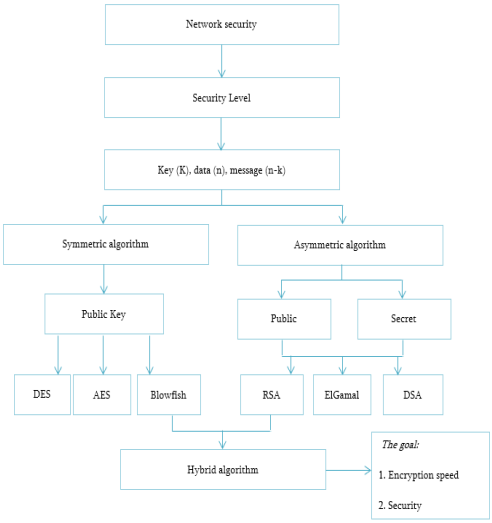


Fig. 1. Main idea of proposed hybrid cryptosystem

The main goal of hybrid cryptosystem is to generate random key for symmetric system and after this encrypt this key for asymmetric system. So, we will get secret key which will be used for encryption plaintext. During Hybrid encryption process data are transferred using unique session keys along with symmetrical encryption. Public key encryption process is implemented for random symmetric key encryption. After receiver gets encrypted message, public key encryption method is used to decrypt the symmetric key. After recovering of the symmetric key, then it is used to decrypt the message.

V. PROPOSED WORK – HYBRID CRYPTOSYSTEM WITH COMBINATION OF BLOWFISH AND RSA

To create strong encryption algorithm there is proposed combination of two encryption algorithms -Blowfish and RSA. There was done experiments on proposed algorithms by terms of their encryption speed, used memory and system requirements. The programming language Java is used for implementing the encryption algorithms. To make more exact calculations was used console work with Java NetBeans IDE.

In general encryption time is connected to algorithm architecture. Table 1 shows encryption and decryption results on Blowfish algorithms. Size of used key is 16 bits.

TABLE I. STATISTICAL RESULTS OF BLOWFISH ENCRYPTION AND DECRYPTION PROCESS

Plain text size (KB)	Plaintext size (Bytes)	Blowfish Encryption Time (Nanoseconds)	Blowfish Decryption time (Nanoseconds)	Blowfish Encrypted File size (KB)	Blowfish used RAM (Bytes)
32	32710	10753053	1984528	59241	9762104
64	65420	12169867	2743007	119493	10696784
128	130840	12567266	5602025	236670	12556416
256	261680	18200673	9356337	475738	16252696
512	523360	23987822	16802548	954280	23511600
1024	1048460	35550482	26062972	1915678	15407800
2048	2096920	43489299	40463494	3804367	28875368
4096	4193840	62097598	56950097	7552059	55642240

The same experiment was done on RSA system, where was used different size of plaintext. Table 2 shows used encryption time in nanoseconds.

TABLE II. STATISTICAL RESULTS OF RSA ENCRYPTION AND DECRYPTION PROCESS

Plaintext size (KB)	Plaintext size (Bytes)	RSA Encryption time (nanoseconds)	RSA Decrypted file size (KB)	RSA Decryption Time (Nanoseconds)	RSA Used RAM (Bytes)
32	32710	1536637771	118780	55542452	5611360
64	65420	3208498484	237689	121344997	4677800
128	130840	6149709140	474654	284935252	62035768
256	261680	10574937240	946614	671696785	72146728
512	523360	20368096461	1896331	1991097468	117161952
1024	1048460	41504791208	3795983	6934459468	238824584
2048	2096920	89946149790	7586016	27974097086	371242008
4096	4193840	181620236481	15179673	121238321204	572478144

Proposed hybrid cryptosystem works as following at first system reads plaintext and generates secret key with RSA and public keys are generated automatically. Next step is to generate Blowfish symmetric key which will be encrypted with RSA system. This provides high security for key, because usage of RSA algorithm decreases decryption probability of public key. So, when we share public key, will be shared also RSA secret key. After these steps, plaintext is encrypted using Blowfish, because as other symmetric algorithms Blowfish is fast. Decryption process is reverse process of above described encryption.

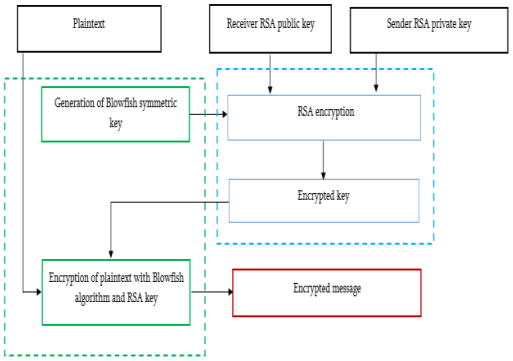


Fig. 2. RSA + Blowfish - the proposed hybrid system architecture

There was created program, implementation for this hybrid cryptosystem on Java programming. Table 3 shows program execution results on different size plaintext.

TABLE III. BLOWFISH + RSA HYBRID SYSTEM ENCRYPTION TIME

Plaintext size (KB)	Plaintext size (Bytes)	RSA+Blowfish Encryption time (nanoseconds)	RSA + Blowfish Encrypted File size (KB)	RSA + Blowfish Decryption time (nanoseconds)	RSA + Blowfish Used RAM (Bytes)
32	32710	9047797	59355	1881211	9498968
64	65420	12203366	118428	2189046	22598000
128	130840	13555651	237417	5057937	26353056
256	261680	14240434	477370	9345405	27380576
512	523360	29886045	951418	18116046	29011368
1024	1048460	40855251	1898922	25666278	30336472
2048	2096920	43979084	3813804	44415486	43218240
4096	4193840	63542269	7624638	54848853	56610432

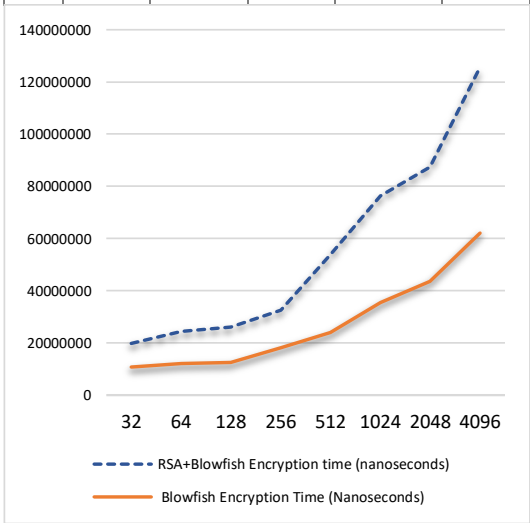


Fig. 3. Comparison of Blowfish and RSA + Blowfish cryptosystems encryption time

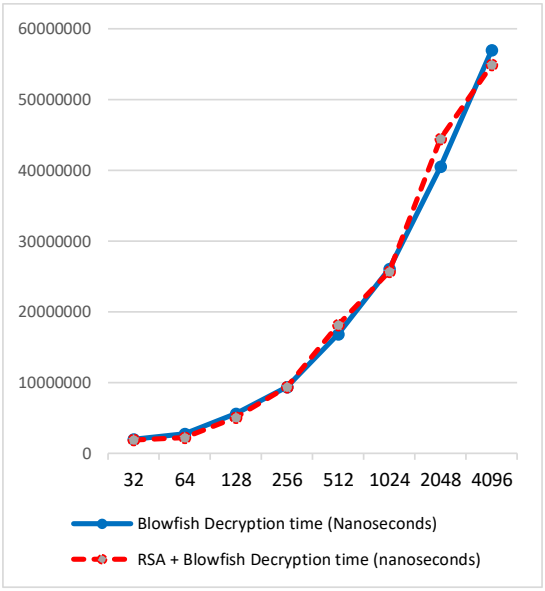


Fig. 4. Comparison of Blowfish and RSA + Blowfish cryptosystems decryption time

VI. CONCLUSION AND SCOPE OF FUTURE WORK

This paper provides description and comparative analyses of new hybrid cryptosystem model. New hybrid model combines Blowfish (symmetric) and RSA (Asymmetric) cryptosystems. Paper shows program implementation and experimental research results with java programing language. Described algorithms and hybrid model are evaluated by terms of encryption speed, memory usage, encrypted file size and ensured security level. Taking into account the time and consumption of the technical resources, Blowfish is the best one else than the other reviewed.

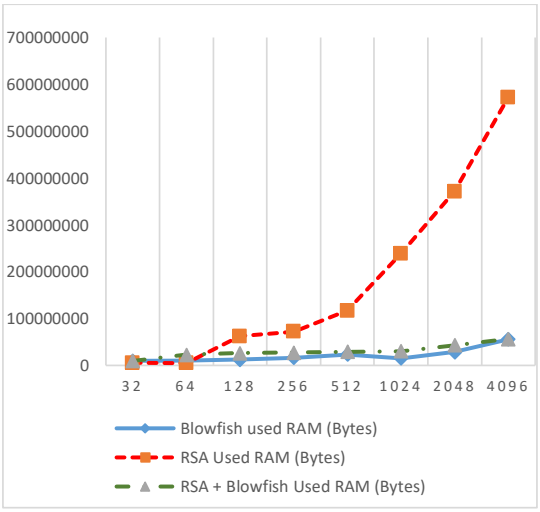


Fig. 5. Used memory comparison chart - Blowfish, RSA and Blowfish+RSA

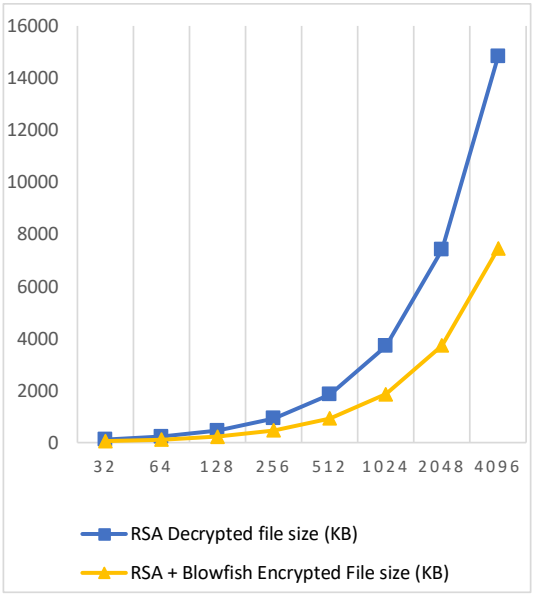


Fig. 6. Encrypted file size comparison chart - RSA and Blowfish+RSA

As a conducted experimental result shows provided new hybrid model is significantly faster and secure, because it takes all advantages and strength of symmetric and asymmetric systems. The experiment showed the following results:

- If Blowfish, RSA and Blowfish + RSA hybrid algorithms are compared according to the memory used, the highest technical resources require RSA algorithm, and Blowfish is slightly behind the Blowfish + RSA hybrid scheme.
- Considering the option of encryption Blowfish keeps its initial first position and is the fastest of these systems. However, the Blowfish + RSA hybrid algorithm is far below and significantly faster than RSA. And RSA takes the longest time to encrypt and is very slow.
- Observation of the decryption time parameters has shown that the Blowfish + RSA hybrid algorithm and the blowfish algorithm are almost equally fast with the decryption process and are faster than the RSA algorithm.
- As an overview of the encrypted file size setting, the lowest memory needs Blowfish system, the following is Blowfish + RSA, and the RSA algorithm increases the size of an encrypted file with the highest rate.

For future is possible to review another hybrid model of symmetric and asymmetric algorithms. It is possible to conduct a series of entropy research of the different cryptographic algorithms and above-presented hybrid model. This will allow us to identify the sustainability of each algorithm against different types of attack, including the frequency analysis of encrypted text.

References

[1] B. Schneier, “Description o f a New Variable-Length Key,64-Bit Block Cipher (Blowfish)”, Fast Software Encryption, Cambridge Security Workshop proceedings (December 1993), Springer-Verlag, 1994, pp. 191-204

[2] "The Digital Millennium Copyright Act of 1998" (PDF). United States Copyright Office. Retrieved 26 March 2015.

[3] Cramer, Ronald; Shoup, Victor (2004). "Design and Analysis of Practical Public-Key Encryption Schemes Secure against Adaptive Chosen Ciphertext Attack"

[4] Hofheinz, Dennis; Kiltz, Eike (2007). "Secure Hybrid Encryption from Weakened Key Encapsulation". Advances in Cryptology - CRYPTO 2007

[5] Johhanes A. Buhman, Introduction to Cryptography, Second Edition, 2000

[6] Alfred J. Menezes, Paul C. van Oorschot, Scott A. Vanston, Handbook of Applied Cryptography, Massachusetts Institute of Technology, June 1996

[7] Ilya KIZHVATOV, Physical Security of Cryptographic Algorithm Implementations, L’UNIVERSITÉ DU LUXEMBOURG, 2009

[8] Simson Garfinkel, Alan Schwartz, Gene Spafford, Practical UNIX and Internet Security, 3rd Edition Securing Solaris, Mac OS X, Linux & Free BSD

[9] The official Advanced Encryption Standard". Computer Security Resource Center. National Institute of Standards and Technology. Retrieved 26 March 2015.

[10] Баричев С. В. Криптография без секретов. – М.: Наука, 1998.

[11] Шнайер Б. Прикладная криптография. Протоколы, алгоритмы и исходные тексты на языке С, 2-е изд . – М.: Вильямс, 2003.

[12] "Quantum cryptography: An emerging technology in network security". - Sharbaf, M.S. IEEE International Conference on Technologies for Homeland Security. 2011

[13] Adleman, Leonard M.; Rothemund, Paul W.K.; Roweis, Sam; Winfree, Erik (June 10–12, 1996). On Applying Molecular Computation To The Data Encryption Standard. Proceedings of the Second Annual Meeting on DNA Based Computers. Princeton University.

[14] Cramer, Ronald; Shoup, Victor (2004). "Design and Analysis of Practical Public-Key Encryption Schemes Secure against Adaptive Chosen Ciphertext Attack"

[15] Hofheinz, Dennis; Kiltz, Eike (2007). "Secure Hybrid Encryption from Weakened Key Encapsulation"

[16] Taher ElGamal (1985). «A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms

[17] Ященко В. В. Введение в криптографию. СПб.: Питер, 2001.

[18] Hamdan O. Alanazi, B. B. Zaidan, A. A. Zaidan, Hamid A. Jalab, M. Shabbir and Y. Al-Nabhani, “New Comparative Study Between DES, 3DES and AES within Nine factors”, Journal of Computing, Volume, 2, Issue 3, March 2010, pp. 152-157.

[19] Dr. Prerna Mahajan and Abhishek Sachdeva, “ A study of Encryption Algorithms AES, DES and RSA for Security”, Global Journal of Computer Science and Technology Network, Web & Security, Volume 13 Issue 15 Version 1.0 Year 2013, pp. 15-22.

[20] Deepak Kumar Dakate and Pawan Dubey, “Performance comparison of Symmetric Data Encryption Techniques”, International Journal of Advanced Research in Computer Engineering and Technology, Volume 3, No. 8, August 2012, pp. 163-166.

[21] Sumitra, “Comparative Analysis of AES and DES security Algorithms”, International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013, pp. 1-5.

[22] “Hybrid Encryption Model of AES and ElGamal Cryptosystems for Flight Control Systems”, Maksim Iavich ; Sergiy Gnatyuk ; Elza Jintcharadze ; Yuliia Polishchuk ; Roman Odarchenko, Oct. 2018

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Evaluation of the impact on energy consumption of MQTT protocol over TLS

Edgaras Baranauskas
Department of Computer Sciences
Kaunas University of Technology
Kaunas, Lithuania
edgaras.baranauskas@ktu.edu

Jevgenijus Toldinas
Department of Computer Sciences
Kaunas University of Technology
Kaunas, Lithuania
eugenijus.toldinas@ktu.lt

Borisas Lozinskis
Department of Computer Sciences
Kaunas University of Technology
Kaunas, Lithuania
borisas.lozinskis@ktu.lt

***Abstract*— Message Queuing Telemetry Transport (MQTT) protocol is widely used in device-to-device communications. While MQTT has three quality of service (QoS) levels, it does not integrate security mechanisms. Transport Layer Security (TLS) is the standard protocol on top of the Transmission Control Protocol (TCP) to secure data in communications. In this paper, we evaluate the impact on energy consumption of MQTT protocol using its QoS levels over TLS.**

Keywords—IoT, MQTT, TLS, battery energy consumption

I. INTRODUCTION

According to Gartner prediction spending on Internet of Things (IoT) endpoint security solutions worldwide will reach 631 millions of dollars in 2021 [1]. The term Internet of Things define smart objects that are interconnected using various network interfaces and protocols such as Constrained Application Protocol (CoAP), Message Queuing Telemetry Transport (MQTT), MQTT-SN (for sensor networks), Extensible Messaging and Presence Protocol (XMPP), Web Application Messaging Protocol (WAMP) and many others. In machine-to-machine (M2M) application layer protocols most popular is MQTT, well-known cloud platforms, such as Amazon AWS, Microsoft Azure, and IBM Watson expose their services through MQTT [2]. MQTT has a low memory footprint, low power consumption, and better distribution of information to recipients [3]. Because of that, MQTT protocol is widely used in device-to-device (D2D) communications, where one of the major issue is to ensure the security of devices and D2D communications [4]. MQTT has three quality of service (QoS) levels and does not integrate security mechanisms. Transport Layer Security (TLS) is the standard protocol on top of the Transmission Control Protocol (TCP) to secure data in communications. OASIS [5] explicitly recommends the utilization of TLS as security decision at transport layer. In this paper, we evaluate the impact on energy consumption of MQTT protocol using its QoS levels over TLS.

II. RELATED WORK

In [4] authors declare the user's responsibility to address security issues for MQTT, MQTT-SN protocols and suggests enabling security for them by envisaging SSL/TLS, but due to IoT heterogeneity it is cumbersome to manage certificates and keys. Thus, authors [4] propose attribute based encryption for secure MQTT that augments security feature for the existing MQTT protocol and its variants.

Use of Datagram Transport Layer Security (DTLS) for securing data communications over User Datagram Protocol (UDP) adds at least 33 bytes to the original packet header, and while IoT devices run on batteries, efficient secure communication scheme is needed [6].

A novel security mechanism introduced for MQTT environments is based on AugPAKE via a secure side channel, where authentication and authorization tokens are transported in the same field [7] of the topic name. In [8] the most known application layer protocols are compared: CoAP, MQTT, XMPP, HTTP, AMQP and WebSocket, All the protocols mentioned above use TCP as transport layer (CoAP uses UDP) and TLS/SSL as security layer (CoAP uses DTLS). In terms of Message Oriented Approach (MOA), MQTT stands out [8]. Requirements for authentication, authorization, data integrity, and confidentiality do not included in the MQTT specification. Authors [9] argue that the lack of security requirements in the MQTT protocol standard is related to:

- MQTT focuses only on message dispatching.
- Reducing the overhead that is related to security features is used to keep the protocol as light as possible.
- Historical implementations of MQTT were based on private networks.
- Significantly different security functionalities required while MQTT is used from IoT devices to Facebook messenger mobile application.

The authors [9] are inclined to believe that a good mid-term solution to large-scale MQTT security problems could be represented by implementation of TLS. Current open-source MQTT implementations compared in table I.

TABLE I. OPEN-SOURCE MQTT IMPLEMENTATION

MQTT implementation	**MQTT implementation property**		
	Definition	***Security***	***QoS***
Mosquitto [10]	Most commonly used implementation	SSL/TLS support	QoS0, QoS1, QoS2
eMQTTC [11]	Asynchronous Erlang MQTT Client Requires Erlang R17+	TCP/SSL Socket Support	QoS0, QoS1, QoS2
Apollo [12]	Is a faster, more reliable, easier to maintain messaging broker built from the foundations of the original ActiveMQ	SSL/TLS Support	QoS0, QoS1, QoS2
Artemis [13]	Implementation arising from ActiveMQ	SSL support	QoS0, QoS1, QoS2

In [14] proposed potential methodologies to extend the Common Architectures and Network services found in the IEEE 1451 Family of Standard into applications that utilize MQTT. The authors installed the Mosquitto MQTT client

onto ESP-32s, MQTT broker onto Raspberry Pi 3 and experimentally conclude that MQTT is an effective communication protocol when it comes to small-scale systems, security is a major area for future investigation. MQTT has its downsides in security but is being greatly adapted in the world of IoT today and the hope is extend that adaptation to the IEEE 1451 Family of standards [14].

MQTT is a simple protocol designed for devices with low processing power and it tries to minimize the processing needed to exchange messages, which means that serious security problems arise such as lack of: authentication, authorization, confidentiality and integrity [15].

The security challenges of the IoT industry with focus on standardized communication protocols explored and implementation details for the security levels mandated by the Constrained Application Protocol provided in [16]. MQTT implementations also offer out of the box the security certificates mode that could be achieved in the Java Paho library or as part of the Mosquitto framework. In fact, the MQTT broker also offers the possibility to maintain a list of revoked certificates that can be used to disable rogue endpoints [16].

The most critical issues with the aim of guiding future research directions on the IoT security panorama highlighted [17]. According to the author conclusion, the most vulnerable level of the IoT system model is the perception layer due to the physical exposure of IoT devices, to their constrained resources and to their technological heterogeneity. Thus, it is crucial, in the next future, to start working on the critical issues of this level implementing lightweight security solutions that can adapt to the heterogeneous environments with resource-constrained devices.

Smart city solutions have to be energy-efficient, cost-efficient, reliable, secure, to do that IoT devices should operate in a self-sufficient way without compromising QoS in order to enhance the performance with uninterrupted network operations. Therefore, the energy efficiency and life span of IoT devices are key to next generation smart city solutions [18]. With the increase in IoT applications for smart cities, energy-efficient solutions are also evolving for low-power devices. Energy-efficient solutions such as Lightweight Protocols, Scheduling Optimization, and Predictive Models for Energy Consumption, Cloud-Based Approach, Low-Power Transceivers, and Cognitive Management Framework can reduce energy consumption or optimize resource utilization. Possible future directions for energy management in smart cities are [18]:

- Energy-efficient mechanisms for software-defined IoT solutions, which can provide scalable and context-aware data and services.
- Directional energy transmission from dedicated energy sources for wireless power transfer.
- Energy efficiency and complexity of security protocols are crucial aspects for their practical implementation in IoT; thus, it is important to investigate robust security protocols for energy constraint IoT devices.
- Fog computing can lead to energy saving for most of the IoT applications; therefore, it is important to study energy consumption of fog devices for IoT applications.

In [19] authors evaluate MQTT (QoS0) vs HTTPS, send performance, battery energy consumption and conclude that while HTTPS is slightly more efficient in terms of establishing connection, MQTT is much more efficient during transmission.

III. MQTT QUALITY OF SERVICE LEVELS

MQTT provides three levels of QoS [20]:

At most once (Fig. 1) - sometimes called "fire and forget". The message is delivered at most once, or it is not delivered at all.

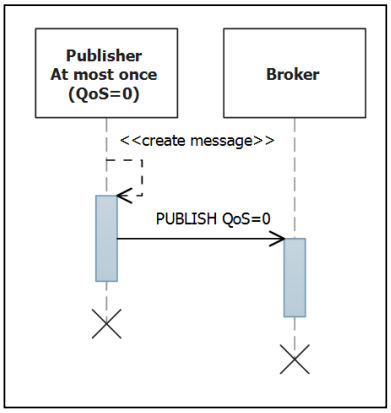


Fig. 1. MQTT QoS level “At most once"

At least once (Fig. 2), it is the default mode of transfer. The message is always delivered at least once. If the sender does not receive an acknowledgment, the message is sent again with the DUP flag set until an acknowledgment is received.

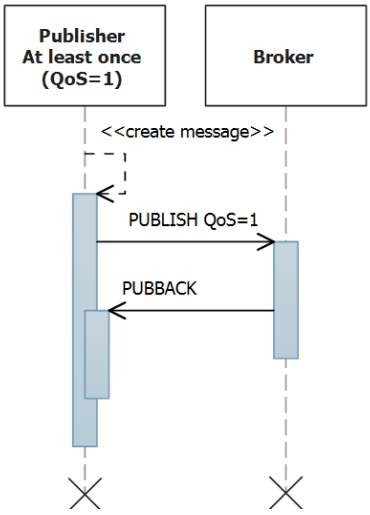


Fig. 2. MQTT QoS level “At least once"

Exactly once (Fig. 3), the message is always delivered exactly once. The message must be stored locally at the sender and the receiver until it is processed. **Exactly once** is the safest, but slowest mode of transfer.

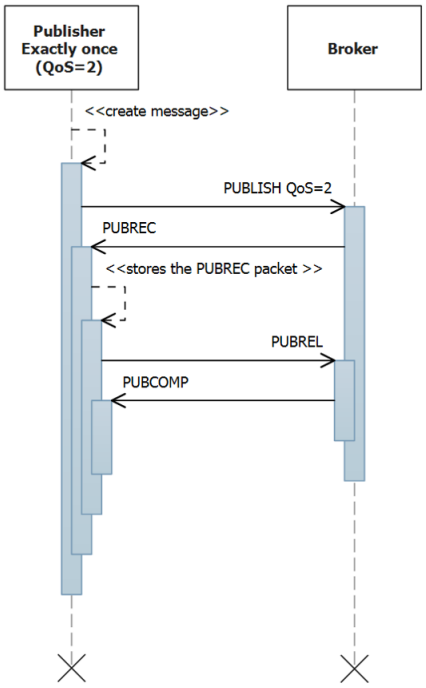


Fig. 3. MQTT QoS level "Exactly once "

IV. EVALUATION FRAMEWORK AND EXPERIMENTAL SETUP

A general framework for evaluation of the impact on energy consumption of MQTT protocol over TLS is shown in Fig. 4. T-diagram is linking together three evaluation domains: security, reliability and energy consumption.

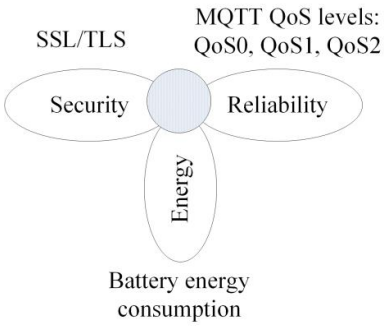


Fig. 4. General framework for evaluation of the impact on energy consumption of MQTT protocol over TLS

The framework also outlines a context of the selected domains: for security domain it is SSL/TLS , the MQTT QoS levels ensure reliability, and battery energy consumption measurement for energy domain. Our experiments are performed using (see Fig. 5):

- **Access point** – Wi-Fi router TP-Link.
- **Broker** – Raspberry Pi2 with Broadcom BCM2837 Arm7 Quad Core CPU, clock frequency 900MHz, 1GB RAM, 802.11b/g/n Wi-Fi communication protocols.
- **Subscriber/Publisher** – IoT Module ESP32 with Tensilica L106, 32-bit, RISC CPU, clock frequency 160 MHz, 802.11b/g/n Wi-Fi communication protocols.
- **Measuring instrument** – digital multimeter MASTECH MS8050.
- **Power supply** for ESP32 – llithium battery LS903052, 3.7V, 1200mAh.

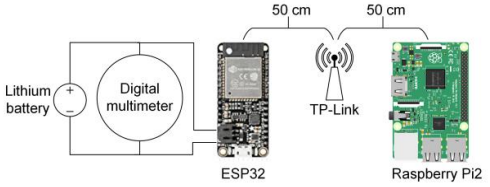


Fig. 5. Experimental setup

The ESP32 module integrates ESP8266EX is and is recommended for tests or for further development. For our evaluation, we use the Mosquitto MQTT broker that configured to use TLS. We create a simple scenario to establish encrypted connection between broker and client similarly as encrypted connection between web server and web client. To create certificates we use OpenSSL v1.1.1a software for Windows [21]. In our case, we create Certification authority (CA) in a computer with Windows OS. Certificate creation and installation in the Mosquitto MQTT broker (in our case Raspberry Pi2) and in the subscriber/publisher (in our case ESP32) is shown in Fig.6.

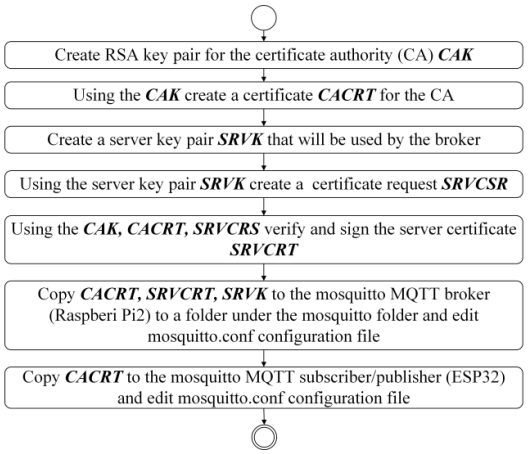


Fig. 6. Certificate creation and installation in the mosquitto MQTT broker and subscriber/publisher

V. EXPERIMENTAL RESULTS

The results of measurements are presented in Figs. 7-9. Fig. 7 shows the battery voltage level for MQTT "At most once" over TLS,

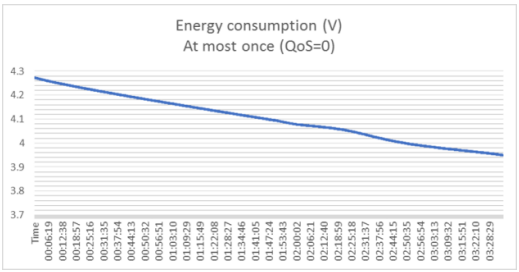


Fig. 7. Battery voltage level for MQTT "At most once" over TLS

Fig. 8 shows the battery voltage level for MQTT "At least once" over TLS.

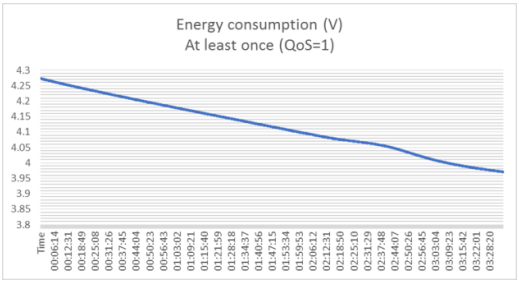


Fig. 8. Battery voltage level for MQTT "At least once" over TLS

Fig. 9 shows the battery voltage level for MQTT "Exactly once" over TLS.

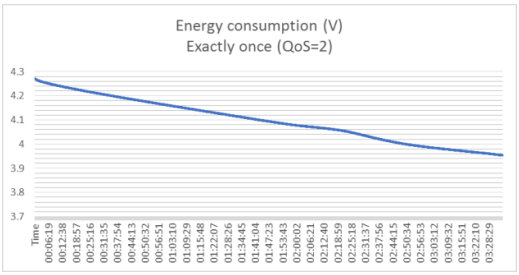


Fig. 9. Battery voltage level for MQTT "Exactly once" over TLS

The results of measurements are summarized in table II. Based on these results we can evaluate the difference in energy consumption of three MQTT protocol QoS levels over TLS. Less energy consumes "At least once (QoS1) over TLS – voltage drop 0.3026V. Most energy consumes "At most once (QoS0) over TLS – voltage drop 0.3228V and "Exactly once (QoS2)" over TLS consumes more energy than QoS1 and less than QoS0 - voltage drop 0.3176V.

TABLE II. EVALUATION OF THE IMPACT ON ENERGY CONSUMPTION OF MQTT PROTOCOL OVER TLS

MQTT QoS Level	**Energy consumption**	
	Voltage drop (V)	***Consumed time (hh:mm:ss)***
MQTT "At most once (QoS0)" over TLS	0.3228	03:34:45
MQTT "At least once (QoS1)" over TLS	0.3026	03:34:36
MQTT "Exactly once (QoS2)" over TLS	0.3176	03:34:45

VI. CONCLUSION

The energy consumption of MQTT protocol with various QoS over TLS levels is highly different. The main results of this paper are as follows:

1) The real time measured values for energy consumption securing MQTT over TLS are achieved with various QoS levels.

2) The results of energy consumption measurements when performing secure communication using MQTT protocol over TLS can be used to reliably predict energy consumption of three QoS levels:

- QoS "At least once (QoS1)" over TLS consumes less energy than the others two QoS levels (QoS0 over TLS and QoS2 over TLS),
- QoS "At most once (QoS0)" over TLS consumes more energy than the others two QoS levels (QoS1 over TLS and QoS2 over TLS),
- QoS "Exactly once (QoS2)" over TLS consumes 5 % more energy than QoS "At least once (QoS=1)" over TLS",
- QoS " At most once (QoS0)" over TLS consumes 6,7 % more energy than QoS "At least once (QoS1)" over TLS,
- QoS "Exactly once (QoS=2)" over TLS consumes 1,7 % less energy than QoS "At most once (QoS0) over TLS".

REFERENCES

[1] Gartner Says Worldwide IoT Security Spending Will Reach $1.5 Billion in 2018. STAMFORD, Conn., March 21, 2018. Gartner, Inc. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2018-03-21-gartner-says-worldwide-iot-security-spending-will-reach-1-point-5-billion-in-2018 [Accessed February 07, 2019]

[2] R. Giambona, A. E.C. Redondi, M. Cesana, "Demonstrating MQTT+: An Advanced Broker for Data Filtering, Processing and Aggregation", In 21st ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM '18), October 28-November 2, 2018, Montreal, QC, Canada. ACM, New York, NY, USA. [Online]. Available: https://doi.org/10.1145/3242102.3243317

[3] S. B. Kenitar, S. Marouane, A. Mounir, "Evaluation of the MQTT Protocol Latency over Different Gateways", SCA2018, October 2018, Tetuan, Morocco. [Online]. Available: https://doi.org/10.1145/3286606.3286864

[4] M. Singh, R. MA, S. VL, and B. P, "Secure MQTT for Internet of Things (IoT)".2015 Fifth International Conference on Communication Systems and Network Technologies. IEEE Computer Society, 2015, pp. 746-751. DOI 10.1109/CSNT.2015.16

[5] MQTT Version 3.1.1 Plus Errata 01. OASIS Standard Incorporating Approved Errata 01. [Online]. Available: http://docs.oasis-

open.org/mqtt/mqtt/v3.1.1/errata01/os/mqtt-v3.1.1-errata01-os-complete.html [Accessed February 08, 2019]

[6] M. H. Amaran, M. S. Rohmad, L. H. Adnan, N. N. Mohamed, H. Hashim, "Lightweight Security for MQTT-SN". International Journal of Engineering & Technology, 7 (4.11) (2018) pp. 223-226

[7] M. Calabretta, R. Pecori, L. Veltri, "A Token-based Protocol for Securing MQTT Communications". 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2018 . DOI:10.23919/SOFTCOM.2018.8555834

[8] A. L. Marra, F. Martinelli, P. Mori, A. Rizos, and A. Saracino, "Introducing Usage Control in MQTT". S. K. Katsikas et al. (Eds.): CyberICPS 2017/SECPRE 2017, LNCS 10683, Springer International Publishing AG 2018, pp. 35–43. https://doi.org/10.1007/978-3-319-72817-9_3

[9] G. Perrone, M. Vecchio, R. Pecori, and R. Giaffreda, "The Day After Mirai: A Survey on MQTT Security Solutions After the Largest Cyber-attack Carried Out through an Army of IoT Devices". In Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security (IoTBDS 2017), p.p. 246-253. DOI: 10.5220/0006287302460253

[10] Eclipse Mosquitto. [Online]. Available: http://mosquitto.org/ [Accessed February 08, 2019]

[11] Erlang MQTT Client. [Online]. Available: https://github.com/emqtt/emqttc [Accessed February 08, 2019]

[12] Apollo. [Online]. Available: http://activemq.apache.org/apollo/ [Accessed February 08, 2019]

[13] Apache ActiveMQ Artemis. [Online]. Available: http://activemq.apache.org/artemis/ [Accessed February 08, 2019]

[14] J. Velez, R.Trafford, M. Pierce, B. Thomson, E. Jastrzebski, and B. Lau, "IEEE 1451-1-6: Providing Common Network Services over MQTT". IEEE Sensors Applications Symposium (SAS). 2018. DOI: 10.1109/SAS.2018.8336750

[15] S. H. Ramos , M. T. Villalba, and R. Lacuesta, "MQTT Security: A Novel Fuzzing Approach". Wireless Communications and Mobile Computing. Volume 2018, Hindawi, 11 pages. https://doi.org/10.1155/2018/8261746

[16] S. Zamfir, T. Balan, I. Iliescu, and F. Sandu, "A Security Analysis on Standard IoT Protocols". 2016 International Conference on Applied and Theoretical Electricity (ICATE). DOI: 10.1109/ICATE.2016.7754665

[17] M. Frustaci, P. Pace , G. Aloi, and G. Fortino, "Evaluating Critical Security Issues of the IoT World: Present and Future Challenges". IEEE Internet of things journal, vol. 5, No. 4, august 2018, pp. 2483-2495. DOI: 10.1109/JIOT.2017.2767291

[18] W. Ejaz, M. Naeem, A. Shahid, A. Anpalagan, and M. Jo, "Efficient Energy Management for the Internet of Things in Smart Cities". IEEE Communications Magazine, January 2017, pp. 84-91. DOI: 10.1109/MCOM.2017.1600218CM

[19] J. L. Espinosa-Aranda, N. Vallez, C. Sanchez-Bueno, D. Aguado-Araujo, G. Bueno, O. Deniz, "Pulga, a tiny open-source MQTT broker for flexible and secure IoT deployments". 1st IEEE Workshop on Security and Privacy in the Cloud, Florence (Italy), September 30, 2015

[20] IBM, "Qualities of service provided by an MQTT client". [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSFKSJ_8.0.0/com.ibm.mq.dev.doc/q029090_.htm [Accessed February 8, 2019]

[21] 20-Nov-2018 OpenSSL 1.1.0j is now available, including bug and security fixes [Online]. Available: https://www.openssl.org/ [Accessed January 4, 2019]

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Autenticated key agreement protocol using Schnorr identication

Donatas Bartkus
Kaunas University of Technology
NFQ Technologies
Kaunas, Lithuania
donatas.bartkus@nfq.lt

Eligijus Sakalauskas
Kaunas University of Technology
Department of Applied Matematics
Kaunas, Lithuania
eligijus.sakalauskas@ktu.edu

***Abstract*— The optimized authenticated key agreement protocol is presented by joining Diffie-Hellman key agreement protocol and Schnorr identification. The optimization is achieved by reducing computations in the most consuming operation in Schnorr identification protocol, namely in the computation of whitness. In addition, the number of communications is reduced to 3 instead of 5, when these protocols operate separately. The security analysis to several attacks is performed. Results of numerical simulation show, that additional computation time required to realize protocol makes up only 6% of computations required to realize one step of Diffie-Hellman protocol. This is the benefit to user when he uses mobile communication device connecting to eBanking system.**

Keywords—Authenticated key agreement protocol, Diffie-Hellman key agreement protocol, Man-in-the-Middle attack.

I. INTRODUCTION

The one of major e.Banking vulnerability is well known Man-in-the-Middle - MIM attack. The problem is that Bank as an organization is protected by its public key certificate and there is no problem for User's browser to perform Bank authentication. But on the other end of communication point is an ordinary user having only either assigned list of passwords or code generator. So Users identification is considerable weak. Situation can facilitated when during the authentication protocol, User must confirm a code he receives in his mobile phone.

But nevertheless, the extra security measures are welcome in such a sensitive area as e-banking.

In this paper we propose an extra authentication measures which can be added to standard SSL/TLS protocol. The solution is based on joining together Diffie-Hellman key agreement protocol (DH KAP) [1] with Schnorr identification protocol [2]. Key agreement protocol with Schnorr identification we name as authenticated key agreement protocol (AKAP). Notice that DH KAP is included in SSL/TLS and so in https protocol.

We do not consider here the standard Bank authentication protocol using https but describe only AKAP itself.

We show, that by joining these protocols together the security of united protocol is considerable higher and the number of communications can be reduced to 3 instead of 5 by executing these protocols separately. The other benefit is that realizing this protocol there is no need to use complicated Public Key Infrastructure (PKI) for Users providing them with public key certificates. The registration of User's public key in Bank is only required. We assume that User can generate his private and public keys either by himself or can use his identity card with his open access public key.

Since Id protocol is restricted in time, the values of exponents used by client in exponentiation operations can be reduced so reducing the number of arithmetic operations executed on-line. For example, if time response from client to bank is restricted to 5-10 minutes, the exponents used by client can be restricted in the interval between 128 to 256 bits, so making impossible for adversary to find these exponents in this short time slot by trying to solve discrete logarithm problem (DLP) or arranging brute force attack [3]. This significantly saves the computation resources of client communication device.

Before executing AKAP, User must register his public key to the Bank. This procedure is described in the next section.

II. INITIATION AND USER REGISTRATION

Beforehand Bank possesses public parameters (PP) corresponding to ElGamal cryptosystem [4]. These parameters are large strong prime number p and generator g of corresponding group $\boldsymbol{Z}_p^*$ = {1,2,3, …, p-1}. Strong prime p is of order ~2^{2048} and is defined in the following way

$$p = 2 * q + 1, \tag{1}$$

where p and q are primes. So PP consist of a vector with two components

$$\text{PP} = (p, g). \tag{2}$$

Bank possesses also his private key PrK_B and corresponding public key PuK_B computed using PP following to ElGamal cryptosystem requirements. Bank chooses at random number z in the interval $2 < z < \text{p-2}$ which is his private key PrK_B

$$\text{PrK}_\text{B} = z. \tag{3}$$

Bank's public key is computed in the following way

$$\text{Puk}_\text{B} = b = g^z \bmod p. \tag{4}$$

For User registration Bank must supply User by public parameters PP = (p,q) defined in (1) and with his public key Puk_B in (4). We assume that *after registration phase* User *trusts* Bank and his parameters.

Then User, for example using authorized software, generates his private key PrK_U and corresponding public key PuK_U following the same procedure.

His PrK_U is found by choosing at random number z in the interval $2 < z < p$-2

$$PrK_U = x. \quad (5)$$

Then User computes his PuK_U

$$PuK_U = u = g^x \bmod. \quad (6)$$

PuK_U serves as User's identification and is handed to Bank together with his passport when he opens an account in the Bank.

Registration phase is completed when User register his PuK_U in the Bank. Bank trusts that User's PuK_U is *authentic*.

We do not consider here how to simplify registration procedure by providing user with authorized and certified script to compute its PrK_U and PuK_U.

III. AUTHENTICATED KEY AGREEMENT PROTOCOL (AKAP)

AKAP is constructed on the basis of DH KAP using Schnorr identification protocol since public parameters (PP) of both protocols are the same, namely PP consist of large strong prime number p and generator g defined above.

This protocol is based on the discrete logarithm and Diffie-Hellman problem complexity assumptions [5].

The essence of the joint protocol is that both User and Bank proof to each other that they know their private keys corresponding to their private keys.

After the registration phase, User can connect to the Bank using https protocol. As we noticed above the standard procedure of Bank authentication we omit from our consideration.

AKAP execution is performed by the following steps.

1. User generates short random number y in the interval between 128 to 256 bits and computes a whitness w

$$w = g^y \bmod p, \quad (7)$$

and sends w to the Bank.

2. Analogously Bank computes a challenge c by choosing at random number h in the interval $2 < h < p$-2

$$c = g^h \bmod p, \quad (8)$$

and sends c to the User.

3. User possessing its private key x and having y computes response r and choosing at random $2 < t < p$-2 computes value v

$$r = y + cx \bmod (p\text{-}1), \quad (9)$$

$$v = g^t \bmod p,$$

and sends r, v to the Bank.

Then Bank having trusted User's $PuK_U = u$ in (6) verifies if

$$g^r = wu^c. \quad (10)$$

If (10) is valid, then User proved to the Bank, that he possesses his PrK_U corresponding to his PuK_U. The correctness of verification is proved from the following identities

$$g^r = g^{y+cx} = g^y g^{cx} = w(g^x)^c = wu^x. \quad (11)$$

Taking into account that Bank associated User's PuKU with his identity during registration procedure, User proved to Bank its identity as well.

After that User and Bank computes their common secret key k referencing to the 1-3 steps of protocol. User computes

$$k_{UB} = c^t = g^{ht}. \quad (12)$$

Bank computes

$$k_{BU} = v^h = g^{rh}. \quad (13)$$

Since $g^{ht} = g^{th}$, then

$$k_{UB} = k_{UB} = k. \quad (14)$$

Presented protocol allows an additional Bank identification to the User. Since User has trusted Bank's $Puk_B = b$, then after the User's identification and common secret key agreement, the secure secret channel is created between the Bank and User. Then Bank can transfer confidential information to the User encrypted with agreed secret key k. Together with this encrypted information Bank possessing its private key z in (3) and having y computes its response

$$rr = h + wz \bmod (p\text{-}1), \quad (15)$$

encrypts it together with the other data d by forming message $m = rr \parallel d$, where $\parallel$ means a concatenation rr with d. Using symmetric encryption function Enc_k with agreed secret key k Bank obtaining ciphertext cc

$$cc = Enc_k(rr \parallel d), \quad (16)$$

and sends cc to the User.

User decrypts cc with decryption function Dec_k using the same secret key k

$$Dec_k(cc) = rr \parallel d. \quad (17)$$

User extracts rr and analogously to (10), (11) and having trusted Bank's $PuK_B = b$ in (4) verifies if

$$g^{rr} = wb^w. \quad (18)$$

Correctness of verification is based on the following identities

$$g^r = g^{y+cx} = g^y g^{cx} = w(g^x)^c = wu^c. \quad (11)$$

So this step of identification does not require an additional transaction between the User and the Bank.

According to this AKAP, the computation of additional exponent function is required to compute w in (7), but according to our recommendation, this computation is performed with the small exponent value y.

IV. Security analysis

Taking in mind that identification protocols are restricted in time, User can save his computation resources by reducing the exponent y in whitness computation in (7). Normally, to provide sufficient security requirements, in ElGamal cryptosystem and Diffie-Hellman protocol exponents are of the same order as prime number p, i.e. of order $\sim 2^{2048}$. Let us assume that during on line communication with https protocol, time slot for User's respond is 5-10 min.

We have showed that instead of choosing y exponent of the order $\sim 2^{2048}$ it sufficient to choose it of the order less than $\sim 2^{256}$.

For User impersonation adversary needs to find y solving discrete logarithm problem (DLP), say Pollard's rho algorithm [6]. The running time of this algorithm is approximately $O(\text{sqrt}(p))$. So it is infeasible to solve DLP for the adversary with such a big p within 5-10 min.

In the case if adversary tries to guess y using brute force attack, it fails since under the current security assumption this attack is ineffective if total scan area is of order 2^{112}.

For security analysis the following assumption are made:

1. The discrete logarithm problem (DLP) is infeasible for public parameters PP = (p,g) generated according to (1), (2).
2. The DLP is infeasible to solve (7) within a time interval of 1 hour, when exponent y has 128 bits length.

We consider the following attacks scenarios of active adversary [8]:

A1. Private key PrK compromization.

A2. Man-in-the-Middle attack (MiM) when adversary tries to impersonate User.

A3. Agreed secret key k compromising.

A1 is trivially impossible due to DLP assumption 1 and therefore neither User's PrK_U nor Bank's PrK_B can not compromised.

A2 is impossible since adversary can not compute right r in (9). Adversary do not know $PrK_U = x$ due to security assumption 1 and is not able to find y due to assumption 2. Determination of y after the session does not helps, since the other value of y will be chosen during the next session.

A3 fails since according to assumption 1 it is infeasible for the adversary. In general, to make protocol resistant to A3, assumption 1 can be replaced by weaker assumption, namely by Diffie-Hellman assumption [8].

V. Results of numerical simulation

To estimate performance of AKAP it is needed to estimate an extra exponentiation operation introduced in (7). In section above we have shoved that exponent y can be of order less than $\sim 2^{256}$. We have estimated the comparative time of computations whitness w in (7) with exponents y having 128, 256 and 2048 bit length. Computations were performed using MatLab functions which are not optimized to our problem. It is a good reason to see the differences in computation time.

The following public parameters were chosen.

p=90141004504084853150703550577638275328418853
89638000071063126866737278485607079315515440158083833442407246770520315798882215016022163341672020711624873317473466044729108359784053448327105936938668505952166343746951640603447840668358848939153746984291553038882551512200193627754713641437012051833782928210284365850008943093991693579078973450072587825105401934030706342561385045566865094072689220247264806572569465417199009917670013436170225222078342224452954089460540644981343071332069086085944663530055655166662964498526920472782512976690245697007033748148668453565177683876502832151118368967913438681327530651209597852328615469738889576413262248810180556687608 1

q=16539597963984575801338779528129055057767873281475214510502277081029721145735573765583992235867057951281107763085565117022611464882304358285839753706658299883658386706664670515184005380258011621033212413646074885452489270867538999416082225570190786690280499297651496335436443136462221629782207487432850416880163379102824991187343872949529571896386748800724430404100634647603949328976103978625177426176680349283185691444017001518955764983055076419858742491203285841663846121773735752989218899656654882833620340168875965971259173003355656745569339132201390248119828495128560713077858394774924572423818473868910364011111

The results of numerical simulation are performed in table below, where the mean values were taken executing 10 experiments in every case.

Bit length of exponent y	Average time of $w = g^y \bmod p$ computation in seconds
128	12.33 ± 1.14
256	22.52 ± 2.37
2048	210.09 ± 22.42

Fig. 1. Results of numerical simulation

As we see, using considerable small exponent of y of 128 bits length, computation time realizing AKAP increases insignificantly. It is about 6% of computation time needed to compute whitness w in (7) with exponent y having 2048 bits length.

VI. Conclussions

Authenticated key agreement protocol (AKAP) based on joining Diffie-Hellman key agreement protocol (KAP) and

Schnorr identification protocol (ID) is constructed. The main benefits of proposed solution are the following.

1. This composition is effective since public parameters of both protocols are the same.
2. The number of communications is reduced to 3, while executing these protocols separately it is equal to 5.
3. Introduction of Schnorr ID requires a little additional exponent function computations since the exponent values can be chosen small, i.e. bit length of additional exponent is about 20 than the convenient requirements.
4. Proposed AKAP is optimized in computation time resources. It is recommended to use a small exponent having 128 bits length in the first step of algorithm execution. Results of numerical simulation showed that expected computation time is larger only by 6% of single exponentiation operation used in Diffie-Hellman KAP.

REFERENCES

[1] Sivanagaswathi Kallam, "Diffie-Hellman: Key Exchange and Public Key Cryptosystems", Internet: http://cs.indstate.edu/~skallam/doc.pdf, Sep. 30, 2015 [Feb. 26, 2019].

[2] David Mandell Freeman, "Schnorr Identification and Signatures", Internet: http://web.stanford.edu/class/cs259c/lectures/schnorr.pdf, Oct. 20, 2011 [Feb. 26, 2019].

[3] Nolan Winkler, "The discrete log problem and eliptic curve cryptography", Internet: http://math.uchicago.edu/~may/REU2014/REUPapers/Winkler.pdf, [Feb. 26, 2019].

[4] Andreas V. Meier, "The ElGamal Cryptosystem", Internet: http://wwwmayr.in.tum.de/konferenzen/Jass05/courses/1/papers/meier_paper.pdf. Jun. 8, 2005 [Feb. 26, 2019].

[5] Feng Bao, Robert H. Deng, Huafei Zhu, "Variations of Diffie-Hellman Problem", Internet: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.3007&rep=rep1&type=pdf, [Feb. 26, 2019].

[6] Stephen McConnachie, "Pollard's rho-Algorithm, And Its Applications to Elliptic Curve Cryptography", Internet: https://ir.canterbury.ac.nz/bitstream/handle/10092/10097/mcconnachie_2007_report.pdf?sequence=1&isAllowed=y, [Feb. 26, 2019].

[7] Mahtab Mirmohseni, Panos Papadimitratos, "Active Adversaries from an Information-Theoretic Perspective: Data Modification Attacks", Internet: https://people.kth.se/~papadim/publications/fulltext/active-adversaries-information-theoretic-data-modification-attacks.pdf, 2014 [Feb. 26, 2019].

[8] Dan Boneh, "The Decision Diffe-Hellman Problem", Internet: http://crypto.stanford.edu/~dabo/papers/DDH.pdf, [Feb. 26, 2019].

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Comparative Analysis of Enterprise Architecture Frameworks

Jovita Bankauskaite
Department of Information Systems
Kaunas University of Technology
Kaunas, Lithuania
jovita.bankauskaite@ktu.lt

***Abstract*— Complex, critical systems require to apply model-based system engineering (MBSE) practices and use standardized methodologies and frameworks that help define the system in a commonly recognized way. The enterprise architecture framework helps determine how information, business, and technology work together. It brings more discipline to the organization by standardizing and consolidating processes to ensure better consistency. This has become a necessity for companies seeking to organize various architectural perspectives into a holistic and unified view. There are several frameworks that help companies implement architecture efficiently. This opens up the question of what set of criteria based on a system of systems principles can be used for comparative analysis of enterprise architecture framework in order to select the best one. This paper proposes five criteria that include weights and presents the results of a comparison of six enterprise architecture frameworks.**

Keywords*— *Comparative analysis, Enterprise architecture framework, Unified Architecture Framework

I. INTRODUCTION

In the age of innovation, people are surrounded by many systems designed to facilitate everyday life, accelerate processes, or even save human life. The growing complexity of the problems requires the problem-solving to be transferred to the systemic level. Nowadays, there are problems that need to be taken into account through a system thinking in order to address new challenges such as the internet of things, autonomous traffic management and so on.

Model-Based Systems Engineering (MBSE) is systems engineering methodology which emphasizes the application of rigorous visual modeling principles. Models are created to deal with complexity, they allow to understand an area of interest, encourage reuse and improve quality [1]. Complex real-life problems require to apply MBSE practices in the level where independently from one another evolving in time systems communicate to achieve a common goal. This is the level of system of systems (SoS). The Department of Defense (DoD) Defense Acquisition Guidebook defines the SoS as a "set or arrangement of systems that results when independent and useful systems are integrated into a larger system that delivers unique capabilities" [2]. SoS is a large complex system that needs to be defined accurately and consistently.

Any complex system can be viewed from several different angles, each of which can be depicted in various architectural perspectives. In order to organize these diverse architectural perspectives into a holistic and unified view, it is necessary to use an enterprise architecture framework that was originally designed by John Zachman. Fig. 1 provides the Zachman framework.

According to the [3] "the framework successfully combines people, data and technology to show a comprehensive view of the inter-relationships within an information technology organization". Framework helps to develop a complex, integrated, cohesive and comprehensive solution and can speed up the architecture development process.

The framework structure the architecture description into domains, layers or images, and suggests using views - diagrams and matrices - to document each concept. The structured description of architecture allows to make systemic decisions on all system components and make long-term decisions about new design requirements, sustainability and support. Organization architecture framework (EAF) defines principles and practices how to design an enterprise or system of systems architecture.

	Why	How	What	Who	Where	When
Contextual	Goal List	Process List	Material List	Organizational Unit & Role List	Geographical Locations List	Event List
Conceptual	Goal Relationship	Process Model	Entity Relationship Model	Organizational Unit & Role Rel. Model	Locations Model	Event Model
Logical	Rules Diagram	Process Diagram	Data Model Diagram	Role relationship Diagram	Locations Diagram	Event Diagram
Physical	Rules Specification	Process Function Specification	Data Entity Specification	Role Specification	Location Specification	Event Specification
Detailed	Rules Details	Process Details	Data Details	Role Details	Location details	Event Details

Fig. 1. The Zachman framework for Enterprise Architecture [4]

In this paper, we focus on the criteria for comparing enterprise architecture frameworks. The question is how to evaluate different enterprise architecture frameworks: what set of criteria based on a system of systems principles to use in order to select the best framework.

In this paper, we propose a new set of criteria including weights and criteria rating which can be used for comparative analysis of enterprise architecture frameworks.

The rest of this paper is structured as follows: in Section 2, the related works are analyzed; in Section 3, the overview of enterprise architecture is provided; in Section 4, the set of criteria with weight are provided to perform a comparison; in Section 5, the achieved results, conclusions, and future work directions are indicated.

II. RELATED WORKS

There is a large number of research papers on the comparison of enterprise architecture frameworks. Most of them are proposing are criteria how to compare EAF, other ones propose criteria on how to compare EAF in the specific area like SOA, EAF implementation etc.

A number of authors compare organization architecture frameworks to offer a more comprehensive and accurate EAFs comparison approach, these studies has been defined in [5], [6], [7], [8]. Paper [5] investigates the concept of architecture by examining six AF: ZF, 4+1 Views, FEAF, RM-ODP, TOGAF, DoDAF. Authors of this paper proposed to compare EAF by fundamental elements such as their goals, inputs and outcomes. The proposed criteria were evaluated on the basis of three estimates: „Y" – explicitly supports an element, „N" – does not support an element, „P" – partially supports or eludes to support an element. In [6] article is compared and contrasted four distinct approaches to the representation and management of models relating to enterprise complexity, ZF, ISO 15704, ISO/CEN 19439 and ISO/IEC 15288. Approaches compare has been performed by archetype dimension, prototype models, purposive dimension, life history, populating with artifacts, profile of change and managing change. Paper [7] provides a guidance in the selection of an EAF that meets the needed criteria. In this paper is performed a comparison of five frameworks: ZF, DoDAF, FEAF, Treasury Enterprise Architecture Framework (TEAF) and TOGAF. The authors of this paper proposed to compare EAFs by tree criteria: (i) views/perspectives - planner, owner, designer, builder, subcontractor, user; (ii) abstractions - what, how, where, who, when, why; (iii) the systems development life cycle – planning, analysis, design, implementation, maintenance. In [8] article author compares following four leading EAF: ZF, TOGAF, FEAF and GEAF. A comparative analysis is performed by ten criteria: taxonomy completeness, process completeness, reference-model guidance, practice guidance, maturity model, business focus, governance guidance, partitioning guidance, prescriptive catalog, vendor neutrality, information availability, time to value. Each EAF is ranked in each of ten criteria and the rating from 1 (very poor) to 4 (very good) is assigned.

Comparative analysis of enterprise architecture frameworks has been performed in other studies to compare EAF in various areas than the framework definition, these study has been defined in [9], [10]. In paper [9], a comparison of EAF has been designed based on the identifies parameters. These parameters identify the gaps in the maturity model. In [10] publication is proposed three major aspects to compare enterprise architecture framework implementation methods: (i) concepts – definition of EA, alignment between IT and business, the association and communication among artifacts; (ii) modeling - notation, syntax and semantics; (iii) process - activities and steps for enterprise architect and business analyzer in EA implementation.

In conclusion, all the analyzed papers and articles to compare enterprise architecture frameworks encounter several common issues: (i) unified architecture framework (UAF) is not included in the comparison, (ii) unsupported weight of the comparison criteria, (iii) it is difficult to interpret the results of EAF comparison, there is not provided the formal result of a comparison.

Overall, researches carried out in this area mainly provide a set of criteria for a general EAF comparison, regardless of the area where that EAF will be applied. I am proposing a more specific, easy to apply set of EAF comparison criteria, applicable to the majority of EAF. The proposed approach in combination with SoS principles will provide a set of comparison criteria with weights to compare EAF. This will help to make a more accurate comparison that is specifically based on the SoS domain and will provide the appropriate EAF selection that will be applied to complex system modeling.

III. Overview of Enterprise Architecture Frameworks

Below is provided a brief description of six enterprise architecture frameworks that are used in this study.

Department of Defense Architecture Framework (DoDAF). This framework is developded for the United States Department of Defense (DoD) that provides visualization infrastructure for specific stakeholders concerns through viewpoints organized by diffrent views. It helps to ensure that architectural artefacts are defined and characterized consistently according to the specific project or mission needs, in order to be "fit-to-purpose". DoDAF organization framework assists managers to make critical decisions more effectively by organizing information sharing across the Department, Joint Capability Areas (JCAs), Mission, Component, and Program boundaries [11]. DoDAF focuses on architectural data rather than architecture artifacts. The framework defines how to specify systems of system using the architectural terms within DoD [12]. The development and documentation of weapons and IT systems in USA must be conducted in accordance with the DoDAF guidelines.

NATO Architecture Framework (NAF). NAF is developed by the North Atlantic Threaty Organization (NATO) and derived from DoDAF Enterprise Architecture. The goal of NAF is to provide a standard for developing and describing architectures for both military and business use [13]. The NAF is designed to ensure that architectures developed adhering to it can be understood, compared, justified and related across many organizations, including NATO and other National Defense initiatives. NAF defines: methodology, viewpoints, stakeholder viewpoints and meta-model [13].

Ministry of Defense Architecture Framework (MODAF). This framework is developed for the British Ministry of Defense to support defense planning and change management activities. MODAF ensures accurate, comprehensive and consistent collection and presentation of information, helping to understand complex issues [14]. The main benefits of MODAF are the improvement of interoperability and implementation between Systems. The framework supports a variety of MOD processes, such as: capability management, acquisition and sustainment. MODAF architectures are designed as consistent, adjacent models that provide a comprehensive view of the enterprise. MODAF defines set of various relationships that can be used to integrate the architectural elements [14].

Unified Architecture Framework (UAF). UAF was initially created as UPDM 3.0, responding to the needs of UML / SysML and military communities to create a standardized and consistent enterprise architecture based on the U.S. Department of Defense Architecture Framework (DoDAF) and the UK Ministry of Defense Architecture Framework (MODAF) [15]. UAF consists of three main components: (i) framework – a collection of domains, model kinds, and

viewpoints, (ii) metamodel – a collection of types, tuples, and individuals used to construct views according to the specific viewpoints, (iii) profile – SysML based implementation of the metamodel to apply model-based systems engineering principles and best practices while building the views. UAF provides a set of rules to allow users to create a consistent enterprise architecture (as models) based on common enterprise and system concepts with rich semantics. These models then become the repositories from which various views can be extracted [16].

Federal Enterprise Architecture Framework (FEAF). This framework is developed for the U.S. federal government. It provides a common approach for the integration of strategic, business and technology management as part of organization design and performance improvement [17]. The government through organizations practice to define the enterprise architecture, used the EAF to assist the development of large, complex systems development processes. Architectural segments are created individually, according to the structural guidelines, each segment is considered to be its own enterprise within the Federal Enterprise.

The Open Group Architectural Framework (TOGAF). This framework is based on the Department of Defense's Technical Architecture Framework for Information Management [18]. TOGAF focuses on mission-critical business applications that use open systems building blocks. TOGAF provides and explains the rules, creating good principles for system architecture development. TOGAF includes three levels of principles: (i) support decision-making throughout the enterprise, (ii) provide guidance of IT resources; (iii) support architecture principles for development and implementation.

IV. Comparison of Enterprise Architecture Frameworks

Currently there is a wide selection of enterprise architecture frameworks. Comparison analysis is required to select the most appropriate framework. In order to more accurately compare the EAF, we suggest using the comparison criteria including ratings and weights. The criteria for the comparative analysis of the enterprise architecture frameworks are as follows:

- **Domain support (DS)** – level of domain support by EAF. The criterion identifies the universality of the framework.
- **Modeling languages openness (MLO)** – level of modeling languages openness. The criterion helps to evaluate whether the modeling language used by the EAF can be modified. Indicates whether the organization that manages the modeling language is open or private.
- **Information availability (IA)** – level of information availability of EAF. The criterion specifies how easily a user can find additional information, material, presentations that help to improve a user knowledge of certain EAF.
- **Tool support (TS)** - level of framework support by modeling tools. The criterion identifies the availability to use the framework in practice through a modeling tool.
- **Prevalence by researchers (PR)** - level of framework prevalence by the research's community. The criterion helps to evaluate whether the framework is being investigated or elaborated in scientific works.

In order to more accurately compare EAF, the set of criteria which are provided above should be ranked. TABLE I provides the rating definitions.

TABLE I. Criteria Ratings

Scale	Rating	Definitions
4	Very Good Full Support	Very good or fully criteria support.
3	Good Acceptable	More than weak criteria support.
2	Weak Less than Acceptable	Inadequate or very poor criteria support.
1	Very Poor Unacceptable	No criteria are met. Very poor criteria support.

To determine the importance of the criteria, it is proposed to assign a weight for each criterion. The total number of assigned weights should be 1. TABLE II provides the weighted rating of criteria.

TABLE II. Quantitively criteria and weights

Criteria	Weight	Justification
Tool support	0.3	The criterion refers to the practical application of the framework.
Domain support	0.3	The criterion refers to the application of the framework in different domains, which allows the company to define various areas using the same framework.
Modeling languages openness	0.2	The criterion refers to modifications to the modeling language.
Information availability	0.1	The criterion refers to the level of dissemination of the framework information.
Prevalence by researchers	0.1	The criterion refers to the level of framework popularity by researcher's community. It shows whether there are ongoing studies in this area.

When each criterion is ranked, it is necessary to calculate a weighted average that helps to show the best framework from the others. Below is provided the comparison formula (1).

$$\bar{X} = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i} \quad (1)$$

$\bar{X}$- weighted average
ω_i– weighted criteria
x_i– rate criteria

Table III provides the results of compared six enterprise architecture frameworks which are briefly introduced in section III. The comparison has been performed using the proposed set of criteria.

Table III. Comparison of enterprise architecture frameworks

EAF/ Criteria	DoDAF	NAF	MoDAF	UAF	FEAF	TOGAF
Domain support	1	1	1	3	1	2
Modeling Language openness	1	1	1	4	1	1
Information availability	3	2	3	1	1	4
Tool support	3	3	3	2	2	3
Prevalence by researchers	4	2	3	1	1	4
TOTAL	1.9	1.6	1.8	2.8	1.2	2.3

According to the results of the comparison, the best frameworks according to the criteria are listed below.

- Domain support – UAF
- Modeling Language Openness – UAF
- Information Availability – TOGAF
- Tool Support – DoDAF, NAF, MoDAF, TOGAF
- Prevalence by researchers - DoDAF, TOGAF

However, the Unified Architecture Framework according to the comparison result has been identified as best framework from other.

CONCLUSIONS

In this paper, we have analyzed the set of criteria which are used to perform a comparative analysis of enterprise architecture framework. The analysis disclosed that a wide variety of different sets of criteria are used which help to select the best framework. However, most of the criteria do not include weights that help determine the priorities of the criteria. The lack of criteria weights and ratings make it difficult to interpret the results of the comparison. Also, none of the proposed set of criteria is used to compare the newest framework – unified architecture framework. We have determined the need for criteria with weights and ratings.

In this paper, we propose a new set of criteria including weights and criteria rating which can be used to carry out an accurate and detailed comparative analysis of enterprise architecture frameworks. The set of criteria includes five criteria: domain support, modeling language openness, information availability, tool support and prevalence by researchers. For each criterion is assigned a weight indicating the importance and priority. Also, the criteria have a rating that determines the framework support under certain criteria, a rating of 1 (very poor) to 4 (very good).

Currently, this paper is oriented to the set of criteria which helps to evaluate different enterprise architecture frameworks in order to select the best framework. In the near future, we are planning to expand our research on enterprise architecture frameworks, especially on Unified Architecture Framework, to explore the possibility of performing an engineering analysis and behavioral modeling using a standard-based method.

REFERENCES

[1] J. Bankauskaite, A. Morkevicius, "An Approach: SysML-based Automated Completeness Evaluation of the System Requirements Specification," in *International Conference on Information Technologies (IVUS)*, Kaunas, Lithuania, 2018.

[2] "Defense Acquisition Guidebook," 26 February 2017. [Online]. Available: https://contractingacademy.gatech.edu/wp-content/uploads/2014/06/Defense-Acquisition-Guidebook-%e2%80%93-Feb.-26-2017.pdf.

[3] T. E. Consortium, "Enterprise Architecture Body of Knowledge," [Online]. Available: http://eabok.org/about.html. [Accessed 10 02 2019].

[4] J. A. Zachman, "The Concise Definition of The Zachman Framework by: John A. Zachman," 2008. [Online]. Available: https://www.zachman.com/about-the-zachman-framework. [Accessed 5 January 2019].

[5] A. Tang, J. Han and P. Chen, "A comparative analysis of architecture frameworks," in *1th Asia-Pacific Software Engineering Conference*, Busan, South Korea, 2004.

[6] R. Martin, E. Robertson, "A Comparison of Frameworks for Enterprise Architecture Modeling," *Conceptual Modeling - ER 2003*, vol. 2813, pp. 562-564, 2003.

[7] L. Urbaczewski, S. Mrdalj, "A comparison of enterprise architecture frameworks," no. Inform Syst. 7, 2006.

[8] R. Sessions, "Comparison of the Top Four Enterprise Architecture Methodologies," 2007.

[9] S. Pulparambil and Y. Baghdadi, "A Comparison Framework for SOA Maturity Models," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, Chengdu, 2015.

[10] B. D. Rouhani, M. N. Mahrin, F. Nikpay and P. Nikfard, "A Comparison Enterprise Architecture Implementation Methodologies," in *2013 International Conference on Informatics and Creative Multimedia*, Kuala Lumpur, 2013.

[11] Department of DefenseOffice of the Assistant Secretary of Defense (OASD) for Network Infrastructure and Integration, "The DoDAF Architecture Framework Version 2.02," 2010.

[12] Department of Defense, "DoD Architecture Framework Version 2.02, Manager's Guide," 2015.

[13] North Atlanic Treaty Organization, "NATO Architecture Framework, Version 4," 1 August 2018. [Online]. Available: https://www.nato.int/cps/en/natohq/topics_157575.htm. [Accessed 9 January 2019].

[14] Minestery of Defense, "Guidance MOD Architecture Framework," 12 December 2012. [Online]. Available: https://www.gov.uk/guidance/mod-architecture-framework. [Accessed 02 02 2019].

[15] OMG Group, "Unified Architecture Framework Profile (UAFP) version 1.0," 2017.

[16] OMG Group, "UAF specification at OMG," 2017.

[17] Executive office of the president of the united states, "FEA Consolidated Reference Model Document Version 2.3," October 2017. [Online]. Available: https://web.archive.org/web/20090202182509/http://www.whitehouse.gov/omb/assets/fea_docs/FEA_CRM_v23_Final_Oct_2007_Revised.pdf. [Accessed 14 January 2019].

[18] The Open Group, "The Open Group Architectural Framework," 2015. [Online]. Available: www.opengroup. org /architecture/togaf7-doc/arch/. [Accessed 02 02 2019].

INFORMACINËS TECHNOLOGIJOS • IVUS 2019 • ISSN 2029-249X • eISSN 2029-4824

Comparison of DevOps maturity models

Monika Gasparaitė, Saulius Ragaišis
Institute of Computer Science
Vilnius University
Vilnius, Lithuania
monika.gasparaite@mif.stud.vu.lt, saulius.ragaisis@mif.vu.lt

***Abstract*—Since software development using DevOps is still a new phenomenon, analysis of DevOps maturity models is insufficient. The purpose of this work is to investigate whether perception of DevOps is similar in different models. Models that have been investigated are as follows: BTopham, Samer I. Mohamed, and Focus Area models. In order to compare the models, assessments of maturity and capability assured by the models have been performed. The results have shown that the examined models are different in terms of their interpretation of DevOps.**

***Index Terms*—DevOps, Maturity Model, Process Assessment, Process Capability**

I. Introduction

Software process is a set of interrelated or interacting activities that are performed while creating a software product. Although software industry has improved significantly over the last few decades, many software companies face such problems as projects being behind schedule, exceeding the budget, customer dissatisfaction with product quality [1]. Eventually, it was acknowledged that most of the problems arise due to immature software process of the company [2]. Software process models were created to improve and assess software process. Software process model defines essential elements of the process, which can be used to assess the maturity of an organization or the capability of individual processes.

New software development methods are emerging over time. One of the rapidly growing phenomena is DevOps, the primary goal of which is to bridge the gap between development and operations. That can be achieved by combining the objectives of different disciplines and tools, forcing interdisciplinary professionals to communicate frequently.

Since software development using DevOps practices is a quite new concern, there is a lack of analysis related to DevOps maturity models. The purpose of this paper is to examine whether DevOps concept can be interpreted the same way in various DevOps maturity models.

This paper is organized as follows. Section II describes DevOps. Section III presents existing DevOps maturity models. Section IV defines comparison method which is used when assessing models in sections V-VI.

II. DevOps

The literature review has shown that definition of DevOps is ambiguous [3]. Even though DevOps has no formal definition, there are a few prevailing opinions about it. Some state that it can be defined as a job title which requires additional skills, while others strongly believe that DevOps is more like a movement with specific concepts covered. However, in simple terms, it can be defined as a combination of development and operations. Development is represented by software developers, while operations involve experts who maintain software in production environment such as database administrators and network specialists [4].

Today DevOps covers culture, collaboration, automation, lean practices, continuous improvement and delivery, user satisfaction. Culture can be interpreted as shared goals and values, responsibility and effortless communication. Collaboration is about adopting cross-functional teams. Automation involves building, testing, and deployment. Lean practices aim to eliminate waste. DevOps also encompasses practices related to monitoring and measurement, which result in continuous improvement. Another purpose of DevOps concerns releasing software and reacting on feedback faster [3].

III. DevOps maturity models

This section provides an overview of three considered models as follows: BTopham, Samer I. Mohamed, and Focus Area maturity models. From all DevOps models that can be found, only three DevOps models were selected for further investigation. The reason being is that they are more comprehensive in comparison with other models.

A. BTopham maturity model

Shani Inbar, Sayers Yaniv, Pearl Gil, Schitzer Eran, Shufer Ilan, Kogan Olga, Srinivasan Ravi present DevOps maturity model which consists of five maturity levels [5]. For the purpose of simplicity, this maturity model is hereinafter referred to as BTopham maturity model.

Authors define three dimensions - process, automation, and collaboration. It is essential to mention that term "dimension" is not used correctly in this context as it has a different meaning in traditional maturity models. In this case, dimension is an equivalent term to process area in CMMI [6]. In fact, other models give a different name for it. For example, ISO/IEC 15504 [7] uses term "process" while AgilityMOD [8] model names it aspect. In order to avoid misunderstandings, in this paper dimensions will be called aspects.

Even though BTopham maturity model is designed to cover the entire life-cycle of an application or a service for large enterprises, this model lacks clarity and specificity. This model

can be rather perceived as guidelines which state general ideas for software process improvement.

B. Samer I. Mohamed maturity model

Samer I. Mohamed maturity model presents an improved version of the previously examined BTopham maturity model [9]. This model has identical five maturity levels, but it defines slightly different aspects as follows: communication, automation, quality, and governance.

Communication refers to effective communication between teams. Automation specifies improvement of delivery speed, throughput, and repeatability. Quality is about delivering in more lean and faster manner while governance is responsible for controlling how those aspects work seamlessly together.

Although this model is an improved version of BTopham maturity model, disadvantages remain the same. Both models are quite abstract in order to assess maturity of real-life organization.

C. Focus Area maturity model

Rico de Feijter, Rob van Vliet, Erik Jagroep, Sietse Overbeek, Sjaak Brinkkemper present another DevOps maturity model which is the largest and the most comprehensive model of all models that have been found [3]. For the purpose of simplicity hereinafter it will be referred to as Focus Area model.

Focus Area model is a non-traditional model because of its focus area architecture. In this context, staged and continuous representations are traditional. Focus area architecture is different for two reasons. First of all, there can be unlimited quantity of maturity and capability levels. Secondly, each focus area (which is a synonym for process area in CMMI maturity model) has different evaluation intervals. For example, possible capability levels for communication focus area starts from A to E, while configuration management can be assessed from A to C. For this reason, meaning of C capability level is different in both contexts. Even though focus areas are named differently in traditional maturity models, original term in this paper will be used instead. This is because of focus area architecture which is quite specific.

This model describes sixteen focus areas which must be taken into consideration while trying to adopt or improve DevOps practices in company. All focus areas are logically grouped into three groups: 1) culture and communication, 2) product, process, and quality, 3) foundation.

IV. Comparison method

In order to determine how different DevOps maturity models perceive DevOps, comparison of maturity models has been performed. That comparison is accomplished by assessing maturity models in accordance with Focus Area model.

Model assessment is composed of several steps. Firstly, maturity assessment of selected model, for example, model X, is done. This assessment is performed as follows. First of all, it is assumed that the company meets the requirements of a certain maturity level of the selected model X. Based on those requirements, the assessment is carried out in accordance with Focus Area assessment method. It gives information about maturity level that can be achieved in accordance with Focus Area maturity model. The assessment is performed until all maturity levels of the selected model X are assessed.

After maturity assessment, the capability assessment is performed. This assessment is based on focus areas instead of maturity levels as in maturity assessment. It gives information about capability level for each focus area that can be assured by the highest maturity level of model X. If a certain level of capability is not achieved, the higher capability levels are no longer assessed. This assessment provides more information about the model being assessed.

Finally, reverse assessments are performed. Reverse assessments are analogous to previously mentioned assessments. The only difference is that it is based not on Focus Area maturity model but on the selected model X.

All assessments rely not only on model requirements that were presented but also on requirements that are established and well known in IT industry nowadays. For instance, version control system usage. If assessments were based only on the explicitly stated requirements, the results would be slightly worse. In fact, the purpose of this assessment is to get results that are as realistic as possible.

V. Comparison of BTopham and Focus Area maturity models

A. BTopham model maturity assessment

The assessment of BTopham model maturity allows to determine which level of maturity in accordance with Focus Area model each maturity level can assure.

Assessment begins with the first maturity level. BTopham model does not have any requirements for the first maturity level. For this reason, each organization is at this level by default. Focus Area maturity model defines maturity level 0, which is equivalent to maturity level 1 in BTopham model as there is no focus area assigned.

One of the Focus Area model requirements for maturity level 1 is “functional and non-functional requirements and incidents are gathered from and prioritized with internal stakeholders and customers”. However, BTopham maturity model does not provide any information about requirements and incidents. For this reason, maturity level 2 in BTopham maturity model cannot assure maturity level 1 in Focus Area model. Oddly, all other requirements for maturity level 1 are met.

The assessment of maturity levels 3 - 5 corresponds to assessment of maturity level 2 because previously mentioned requirement still cannot be satisfied.

In short, the results of BTopham model maturity assessment show that all maturity levels in BTopham maturity model can ensure only maturity level 0 in the Focus Area model.

B. BTopham model capability assessment

As mentioned before, capability assessment allows to gain additional knowledge about models. The assessment result is

provided in Table I. A dash indicates that the lowest capability level cannot be assured. Also, maximum capability levels are provided as they are different for each focus area.

TABLE I
THE RESULT OF BTOPHAM MODEL CAPABILITY ASSESSMENT

Focus area	Achieved capability level	Maximum capability level
Communication	E	E
Knowledge sharing	C	D
Trust and respect	A	C
Team organization	B	D
Release alignment	C	C
Release heartbeat	-	F
Branch and merge	A	D
Build automation	B	C
Development quality improvement	-	E
Test automation	-	E
Deployment automation	C	D
Release for production	-	D
Incident handling	-	D
Configuration management	A	C
Architecture alignment	A	B
Infrastructure	A	D

As shown in Table I, communication and release alignment focus areas assure the highest capability level. It means that perception of communication is similar in both models. It is worth to mention that requirement for capability level C is incorrect because it is composed of requirements for lower capability levels. Therefore, if lower capabilities (A and B) are achieved, the highest capability level C will be always achieved as well.

C. Focus Area model maturity assessment

The assessment of Focus Area model maturity allows to determine which level of maturity in accordance with BTopham maturity model each maturity level can assure.

As mentioned before, Focus Area model does not have any requirements for maturity level 0 because there are no focus areas assigned to this maturity level. For this reason, maturity level 0 in Focus Area maturity model is analogous to maturity level 1 in BTopham maturity model.

Maturity level 1 in Focus Area model cannot ensure maturity level 1 in BTopham maturity model because requirements "automation process is documented and partially automated", "regular sync meetings are held", "there is frequent communication between the teams" are not fulfilled.

The assessment of maturity levels 2-10 corresponds to previous assessment because Focus Area model does not have requirements related to documentation.

D. Focus Area model capability assessment

The result of Focus Area capability assessment is provided in Table II. The highest capability level that can be achieved is 5 as model provides exact five levels. It is worth to mention that in traditional models the lowest capability level is 0 which means that aspect is either not performed or partially performed. In this case, the lowest capability is 1.

TABLE II
THE RESULT OF FOCUS AREA MODEL CAPABILITY ASSESSMENT

Aspect	Achieved capability level
Process	2
Automation	1
Collaboration	1

E. Disadvantages of BTopham and Focus Area maturity models

The obtained results have shown that BTopham and Focus Area models are contrastive. First of all, a few requirements in BTopham maturity model are abstract, making it difficult to check whether the requirement is met, while the Focus Area model provides very specific requirements. For instance, Focus Area model requires manual code quality monitoring and examples of how it can be done are provided, while BTopham maturity model requires that the quality of the overall process should be measured but it is not clear which areas and how to measure.

What is more, requirements that can be checked, such as standardization or documentation of the process, are not required in Focus Area model. Assessments have shown that BTopham maturity model has some drawbacks. For instance, requirements for higher maturity levels can be satisfied even though requirements for lower maturity levels are not met. Meanwhile, the absolute majority of the requirements in Focus Area model are formulated so that higher level requirements cannot be met unless lower level requirements are satisfied.

Another disadvantage of BTopham maturity model is that this model does not define the necessary level implementation of the requirements for maturity level. Traditional models do not require complete implementation of the requirements. For example, ISO/IEC 15504 states that the process attribute is “fully achieved” if at least 86% requirements are met, while CMMI requires no fundamental shortcomings.

Unfortunately, Focus Area model has downsides as not proper requirements. For example, requirement "a software build is created manually" are always fulfilled because it cannot be done in a worse way. Furthermore, requirements for release alignment focus area are not correct because they overlap each other.

VI. COMPARISON OF SAMER I. MOHAMED AND FOCUS AREA MATURITY MODELS

A. Samer I. Mohamed model maturity assessment

The results of BTopham and Samer I. Mohamed maturity assessments are identical. The reason being is that both models do not define any requirements related to requirements and incidents gathering.

In short, Samer I. Mohamed maturity model can assure only maturity level 0 in accordance with Focus Area model.

B. Samer I. Mohamed model capability assessment

The result of Focus Area capability assessment is provided in Table III.

TABLE III
THE RESULT OF SAMER I. MOHAMED MODEL CAPABILITY ASSESSMENT

Focus area	Achieved capability level	Maximum capability level
Communication	E	E
Knowledge sharing	**A**	**D**
Trust and respect	A	C
Team organization	**A**	**D**
Release alignment	C	C
Release heartbeat	-	F
Branch and merge	A	D
Build automation	B	C
Development quality improvement	**A**	**E**
Test automation	-	E
Deployment automation	C	D
Release for production	-	D
Incident handling	-	D
Configuration management	A	C
Architecture alignment	A	B
Infrastructure	A	D

As shown in Table III, the assessment result is similar comparing with BTopham model. Almost all achieved capability levels are the same except for knowledge sharing, team organization, and development quality improvement focus areas.

C. Focus Area model maturity assessment

The assessment of Focus Area model maturity allows to determine which level of maturity in accordance with Samer I. Mohamed maturity model each level can assure.

Maturity level 0 in Focus Area model is analogous to maturity level 1 in Samer I. Mohamed maturity model because both maturity levels do not have any requirements.

Maturity level 1 in Focus area cannot ensure maturity level 2 in BTopham maturity model because requirements "automation process is documented but not yet executed as a standard", "defect tracking/management is done using proper tools" are not satisfied.

The assessment of maturity level 2 corresponds to previous assessment because Focus Area model does not have requirements related to documentation and defect tracking.

Even though requirements related to defect tracking/management are met in maturity level 3, requirement for documentation still cannot be fulfilled. For this reason, maturity level 3 in Focus Area model cannot assure maturity level 1 in Samer I. Mohamed model.

The assessment of maturity levels 4-10 is identical to assessment of maturity level 3. In short, maturity level 1 in Focus Area model assures maturity level 1 in Samer I. Mohamed maturity model.

D. Focus Area model capability assessment

The result of Focus Area capability assessment is provided in Table IV.

TABLE IV
THE RESULT OF FOCUS AREA MODEL CAPABILITY ASSESSMENT

Aspect	Achieved capability level
Communication	3
Automation	1
Governance	2
Quality	2

E. Disadvantages of Samer I. Mohamed and Focus Area maturity models

Assessments have shown that Samer I. Mohamed maturity model has the same drawbacks as BTopham maturity model.

CONCLUSIONS

The assessments have shown that considered models perceive DevOps in a different manner. This conclusion can be justified by the fact that the highest maturity level in BTopham and Samer I. Mohamed maturity models corresponds to the lowest maturity level in Focus Area maturity model and vice versa. Nevertheless, communication perception is similar. Assessments have exposed that all considered maturity models have weaknesses.

REFERENCES

[1] Stasys Peldžius, "Software process assessment using multiple process assessment models," Ph.D. dissertation, Vilnius University, 2014, (in Lithuanian).
[2] W. S. Humphrey, W. Sweet, R. Edwards, G. LaCroix, M. Owens, and H. Schulz, "A method for assessing the software engineering capability of contractors," 1987.
[3] R. de Feijter, R. van Vliet, E. Jagroep, S. Overbeek, and S. Brinkkemper, "Towards the adoption of devops in software product organizations: A maturity model approach," Department of Information and Computing Sciences Utrecht University, Utrecht, The Netherlands, Tech. Rep. UU-CS-2017-009, 2017.
[4] J. Roche, "Adopting DevOps Practices in Quality Assurance," *Commun. ACM*, vol. 56, no. 11, pp. 38–43, Nov. 2013.
[5] S. Inbar, S. Yaniv, P. Gil, S. Eran, S. Ilan, K. Olga, and S. Ravi, "DevOps and OpsDev: How Maturity Model Works," https://community.softwaregrp.com/t5/All-About-the-Apps/DevOps-and-OpsDev-How-Maturity-Model-Works/ba-p/306787, 2013, [Accessed: 2018-06-15].
[6] C. P. Team, "CMMI for Development, version 1.2," 2006.
[7] ISO/IEC 15504 – 1 Information technology – Process Assessment – Part 1: Concepts and vocabulary, "ISO/IEC," 2004.
[8] Ozcan Top, Ozden, "Software Agility Assessment Reference Model v3.0 (AgilityMOD)," 2014.
[9] S. I. Mohamed, "DevOps shifting software engineering strategy-value based perspective," *International Journal of Computer Engineering*, vol. 17, no. 2, pp. 51–57, 2015.

Accuracy of throwing distance perception in Virtual Reality

Karolis Butkus
Department of Information Systems
Kaunas University of Technology
Kaunas, Lithuania
e-mail: k.butkus@ktu.edu

Tautvydas Čeponis
Department of Multimedia Kaunas
University of Technology Kaunas,
Lithuania
e-mail: tautvydas.ceponis@ktu.edu

***Abstract* — This article investigates how people perceive distances in virtual reality (VR) and use that information to execute a representation of a real life throwing motion. In order to measure accuracy, this research proposes a throwing motion testing framework, which acquires metrics data from both the real and virtual environments. The results show, that the examinees tend to throw more accurately at longer distances and use excessive amounts of force.**

Keywords—virtual reality, perception, accuracy, throwing motion

I. INTRODUCTION

During the last decade virtual reality technology has significantly improved and is used in different technological spheres. The visual representation is becoming more realistic and looks more natural. Although technology is evolving, it is hard to replicate human senses. Therefore, this study tries to analyze how accurately people perceive virtual world distances when executing a throw.

This study presents a throwing motion testing framework to determine the differences between the virtual and real world's environment perception capabilities. It will discuss similar studies in the field related to perception and motion tracking, explain the testing framework and methodology, the experiment's process, discussion about the results and drawbacks of this study and the conclusion, possible future.

A similar project [1] to determine the perception of virtual reality was carried out in 2008 by researchers from Aachen, Germany. In their experiment, they asked 23 participants to estimate distances to virtual reality objects in three different environments. Results show that people tend to underestimate distances and that visual surroundings did not affect results considerably.

Another article checked people's ability to locate themselves in a virtual environment. Their task was to point at themselves in a VR platform using a pointer. The experiments results stated that participants most commonly locate themselves at the upper region of their face and that draws a conclusion that people in a virtual environment are more head-centered. [2]

A more recent study [3] was carried out by researchers from Iowa State University. The group examined prior attempts at improving distance perception in a Virtual environment (VE) and proposed a more thorough methodology to measure the results by isolating unaccounted variables in past studies. The experiment tested the participant's size and distance perception in a VE replica of a real world room with half of the examinees having seen the room prior to the experiment and half participating blindly. The first tested method for improved distance perception was visual replication of a real world environment, the second was walking interaction, which allowed participants to move around the virtual environment prior to testing. The results concluded that walking interaction significantly increased the accuracy of distance perception and size perception to a lesser degree. Furthermore, it was more effective than visual replication in both scenarios.

A similar study to research [3] was carried out at Clemson University in 2011 [4]. In this experiment, researchers investigated near-field egocentric distance estimations in an Immersive Virtual Environment and compared it to real world distances. The experiment examined two methods: verbal and reach measurements. Participants had to report distances verbally and then show it with their reach. Results show that both verbal and reach methods tend to underestimate distance and that with an increase in distance deviation also increased. Another interesting fact was that the verbal method was less accurate than the reach method.

The study [5] made by three researchers from the Dresden University of Technology attempted to find out what factors mostly affect people's estimations for distance in the virtual world. They arranged the factors in four groups: measurement methods, technical, compositional and human factors. The research concluded that people tend to underestimate distance and that to improve human distance recognition skills - a rich, detailed environment and powerful technical hardware must be ensured. Such as high quality graphics, carefully adjusted camera settings and virtual environment with a regularly structured ground texture.

As mentioned a few times in other researches people tend to underestimate distances in virtual reality and according to Steven M. LaValle, the cause for that could be different gaps between pupils [6]. If pupils in the real world are closer than in the virtual world, the virtual environment looks larger to the user and the other way round if the pupils are further apart in the real world.

II. THROWING MOTION TESTING FRAMEWORK

To determine the differences in perception between reality and a virtual environment, we focused on the different aspects of throwing kinematics in reality and VR. Three main characteristics are taken into consideration: throwing distance in reality, throwing distance in virtual reality and the initial velocity of the hand tracker in a throwing motion.

To measure the above mentioned features a throwing simulation framework was created. During the testing procedure participants throw a 10 gram ball to three different distances (2 meters, 3 meters, 4 meters) and a tracker attached to their hand transmits VR data which is recorded digitally, while real life distance is measured with a ruler. Each participant has three attempts at three distances with the virtual reality headset being used and another with it mounted on top of their head for tracking accuracy.

The testing system is developed using *Unity Engine* and *HTC Vive Pro* VR headset and tracker. The framework's visual environment is a replica of the room where the simulation was performed so it would not cause distractions to the participants. Distances at which the ball is thrown and standing position are marked in both the real and virtual environments (Fig. 1 and Fig. 2).

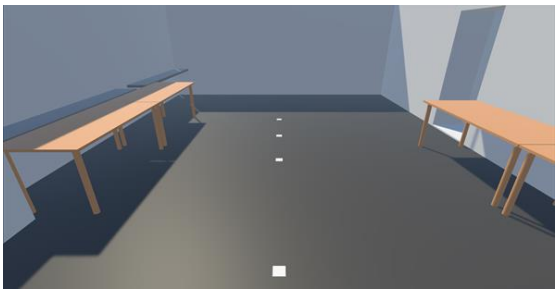


Fig. 1. *VR testing platform (user view)*

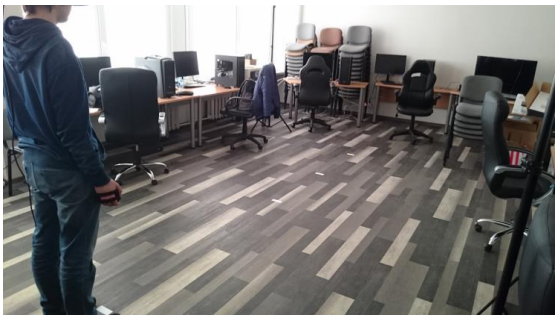


Fig. 2. *Real testing scene view*

In the experiment, *HTC Vive Pro* virtual reality headset and tracker are both connected to a personal computer with *Windows 10* operating system. The tracker's data collecting *Base stations 2.0* were placed at 5 meter distances from each other, at opposing corners of the room. The testing framework was built in *Unity 2018.3.5.f1* with an implemented *SteamVR* plugin. The original plugin's view for the ball throw scene was edited so that it would replicate the experiment room and the stock throw function was modified so that it didn't require any buttons to be pushed. The throw in the system is initiated when the tracker is swung and the velocity of the tracker starts to slow down after the constant increase in velocity at the start of the throw. The simulated environment replica consists of a 9 meter by 6 meter square room with an open top. The layout is positioned at the exact locations of real world objects.

To collect quite accurate motion data the tracker is attached to the palm of the participant and the ball is put on top of the device (Fig. 3). When the person executes a throw the tracker captures the initial velocity, and upon slowing down the system initiates a throw in virtual reality and sends the collected speed and data about the ball's collision with a ground surface to a text file. The real distance is measured with a ruler and all collected digital and non-digital data is saved in a spreadsheet.

Fig. 3. *Ball throw in the experiment*

III. EXPERIMENT

The main goal of the experiment is to determine how accurate is a human's perception at determining distances using a virtual throwing mechanism compared to a real world throw.

The experiment participants were six people: 4 males and 2 females. The participants age ranged from 19 to 25 years (mean age 22.3), all of them were healthy and didn't suffer from VR sickness. At the beginning of the test, the participants were given time to practice throwing in virtual reality and get used to it. Then the examinees did three consecutive throws at specified distances without a headset and then they had three attempts with the virtual reality device. This process was repeated three times at three different shooting distances. During the experiment, participants were not allowed to move from the starting position. The collected distance and velocity data was saved in a spreadsheet.

The experiment's results are presented in Table I where every user's average thrown distance is shown in a centimeters format. Results of shots with virtual reality equipment and without it are separated and the total average of each baseline distance is calculated.

From Table I it is easy to see that people throw the ball most accurately at a distance of 3 or 4 meters when using the VR headset, whereas at the 2 meter mark there is a 10 centimeters deviation. However, with unobstructed vision people throw the ball more accurately at the first and third distances and in this case there is about a 10 centimeters deviation from the second distance. This data shows that with an increase in distance people's throws tend to become more accurate, whereas near distances are more difficult to judge.

TABLE I. AVERAGE THROWN DISTANCE WITH VR AND WITHOUT IT

USER	Reality			VR		
	200 cm	*300 cm*	*400 cm*	*200 cm*	*300 cm*	*400 cm*
1	189.667	298.000	378.000	215.000	326.000	414.000
2	207.333	301.000	389.000	204.667	307.000	398.333
3	212.000	295.333	382.667	198.000	301.667	375.000
4	167.667	268.667	389.000	164.000	277.667	478.500
5	183.000	286.000	415.667	178.333	301.667	407.000
6	217.333	278.667	428.667	182.000	301.667	347.667
AVG[a]	196.167	287.944	397.167	190.333	302.611	403.417
SD[b]	17.565	11.458	18.464	17.222	14.090	40.216

[a.] AVG – Average

[b.] SD – Standard deviation

In addition, from the bar chart shown in Fig. 4, which represents the average miss distance from a mark (negative value if it is shorter than the baseline distance and positive if the average value is greater), it is noticeable that the experiment participants tend to underestimate distances and throw the ball at a shorter distance. Only two columns show a slight ball overthrow and both belong to results achieved in virtual reality .

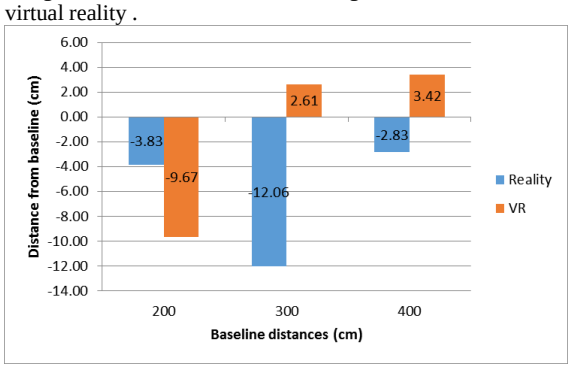


Fig. 5. *Average distance from baseline mark*

Table II shows every participant's standard deviation of three throws and average standard deviation which is about 17 centimeters. Therefore, it can be said that the experiment needs an increase in participants and throw attempts to make the experiment's data even more accurate.

TABLE II. THE STANDARD DEVIATION OF EACH PARTICIPANT THROWS

USER	Reality			VR		
	200 cm	*300 cm*	*400 cm*	*200 cm*	*300 cm*	*400 cm*
1	7.409	8.287	29.063	17.795	32.934	39.047
2	10.656	4.967	13.928	19.754	15.895	2.625
3	11.225	17.632	2.494	28.891	14.055	18.239
4	8.807	24.253	14.900	12.832	14.055	3.500
5	30.342	4.546	13.888	4.989	28.170	0.816
6	15.965	30.214	29.915	9.899	45.492	36.736

Data about the average initial velocity is presented in a clustered columns chart and a scatter graph (Fig. 5 and Fig. 6) where the baseline distances and different environments are separated. Besides average values, medians are given to make the data more accurate.

From Figures 5 and 6 it is noticeable that people tend to throw the ball with more power when they are in a virtual environment than when they are in the real world. This statement also is reaffirmed by the medians of all throws in real and virtual worlds.

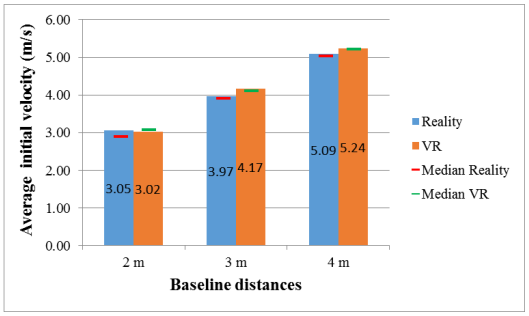


Fig. 4. *Average initial velocity and medians*

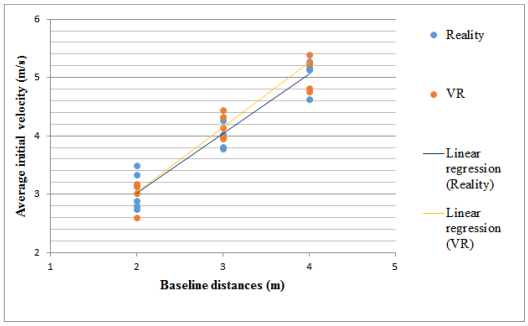


Fig. 6. *Average initial velocity linear regression*

IV. DISCUSSION

This study was conducted to find out how accurately people perceive the virtual environment and decide what amount of power is needed to throw the ball. To achieve this goal 6 participants took part in the experiment where they had to throw a ball at 3 distances with and without a VR headset.

After all tests, the collected data shows that people's accuracy with VR tends to increase with an increase in distance and that the average initial speed tends to be higher than pitching the ball without the headset. To explain the increase in velocity we could say that because people are more head-centered [2] in a virtual environment, they sense that distance is further than it actually is. Moreover, people are more likely to underthrow than overthrow the ball in real life and the increase in velocity when using VR allows their shots to be more precise. But when people are throwing close range shots the distances spread out and accuracy decreases.

These results show, that the described method can be used to calibrate hand strength in Virtual Environment fields, such as gaming [7], simulations [8], gesture recognition systems [9]. The motion force a person outputs in a fully

immersed virtual system has to be decreased by 3 – 5 % to assure that the user's perception of his virtual strength matches the real world results and compensates their depth perception in a VE.

To acquire more accurate estimations we cannot forget that all velocity data is collected by a wireless tracker and the real ball that was put on the tracker could interfere with results and that could be a reason why the standard deviation for a few participant's throws was so high.

In addition, to help the person better comprehend the depth of a virtual world during the experiment it could be allowed for the participants to walk around the room as shown in research [3] and not undertake the whole experiment from a standing position while only having to trust their vision.

Furthermore, it was brought to the examinees attention, that to get more accurate results the participants had to do a bigger backswing while performing the throwing motion to get a more consistent velocity and more suitable throw initialization timings.

V. Conclusion and Future Works

In this study, we concluded, that people perceives 2 – 4 meter distances nearly the same as in real life. Moreover, people tend to use 3 to 5 % more power when throwing a ball in virtual reality than in real life. However, the used methodology needs improvement (some throws standard deviation is as high as 45 centimeters) to eliminate unnecessary factors, such as inaccuracy of manual real world measurements and signal integrity loss from ball position relative to the sensor. Furthermore, a larger pool of participants is needed to achieve precise data averages and calculations. There is also the possibility to attach a separate sensor to the ball that is being thrown by the participants, thus eliminating the need for real world measurements by allowing us to compare the data between both throws directly. Although the research is not perfect it has considerable potential to be used as a calibration tool for various virtual reality fields which involve hand motion and arm strength.

References

[1] C. Armbrüster, M. Wolter, T. Kuhlen, W. Spijkers and B. Fimm, "Depth Perception in Virtual Reality: Distance Estimations in Peri- and Extrapersonal Space", *CyberPsychology & Behavior*, vol. 11, no. 1, pp. 9-15, 2008. Available: 10.1089/cpb.2007.9935

[2] A. H. V. D. Veer, A. J. T. Alsmith, M. R. Longo, H. Y. Wong, and B. J. Mohler, "Where am I in virtual reality?," *Plos One*, vol. 13, no. 10, 2018.

[3] J. W. Kelly, L. A. Cherep, B. Klesel, Z. D. Siegel, and S. George, "Comparison of Two Methods for Improving Distance Perception in Virtual Reality," *ACM Transactions on Applied Perception*, vol. 15, no. 2, pp. 1–11, Mar. 2018

[4] P. Napieralski et al., "Near-field distance perception in real and virtual environments using both verbal and action responses", *ACM Transactions on Applied Perception*, vol. 8, no. 3, pp. 1-19, 2011. Available: 10.1145/2010325.2010328.

[5] R. Renner, B. Velichkovsky and J. Helmert, "The perception of egocentric distances in virtual environments - A review", *ACM Computing Surveys*, vol. 46, no. 2, pp. 1-40, 2013. Available: 10.1145/2543581.2543590.

[6] S. LaValle, *Virtual Reality*. Cambridge University Press, 2016, pp. 153-154. Available: http://vr.cs.uiuc.edu/vrbooka4.pdf

[7] R. Buzys, R. Maskeliūnas, R. Damaševičius, T. Sidekerskienė, M. Woźniak and W. Wei, "Cloudification of Virtual Reality Gliding Simulation Game", *Information*, vol. 9, no. 12, p. 293, 2018. Available: 10.3390/info9120293.

[8] E. Danevičius, R. Maskeliūnas, R. Damaševičius, D. Połap and M. Woźniak, "A Soft Body Physics Simulator with Computational Offloading to the Cloud", *Information*, vol. 9, no. 12, p. 318, 2018. Available: 10.3390/info9120318.

[9] A. Vaitkevičius, M. Taroza, T. Blažauskas, R. Damaševičius, R. Maskeliūnas and M. Woźniak, "Recognition of American Sign Language Gestures ina Virtual Reality Using Leap Motion", *Applied Sciences*, vol. 9, no. 3, p. 445, 2019. Available: 10.3390/app9030445.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Cryptocurrencies short-term forecast: application of ARIMA, GARCH and SVR models

Dovilė Kuizinienė
Department of Applied Informatics
Vytautas Magnus University
Kaunas, Lithuania
dovile.kuiziniene@vdu.lt

Aušra Varoneckienė
Department of Applied Informatics
Vytautas Magnus University
Kaunas, Lithuania
ausra.varoneckiene@vdu.lt

Tomas Krilavičius
Department of Applied Informatics
Vytautas Magnus University
Kaunas, Lithuania
Baltic Institute of Advanced Technology
tomas.krilavicius@vdu.lt

***Abstract*— Cryptocurrency are difficult to forecast due to it's globality and availability to everyone and every time. There is no Friday or Holidays effect, seasonality, market news and other aspects, which influence the course direction. It is the phenomena of the market and it is useful to spread forecast methods research to find out the best fitting model for this phenomenon. In this paper is presented short-term forecast of five different cryptocurrencies (Bitcoin, BitcoinCash, Ethereum, Litecoin, Ripple). Forecast methods split in two groups: 1) real value (ARIMA and SVR models) 2) volatility (GARCH and SVR models). The model's suitability is evaluated by RMSE and MAE. The best results for real value forecast were achieved using ARIMA, for volatility forecast - SVR. In further research it would be useful to analyze methods variety of Artificial Neural Networks and others connected models' modifications.**

Keywords — Cryptocurrency, Bitcoin, forecast, ARIMA, GARCH, SVR.

I. INTRODUCTION

Around a ago decade cryptocurrency was presented as a new phenomenon on global financial markets. By providing an alternative money and investment opportunity, they function outside centralized financial institutions [1]. The basic idea of cryptocurrency sustains of electronic payment system based on cryptographic proof instead of trusted third party [2]. Distributed or decentralized cryptocurrency system allows its users transfers make faster, cheaper and secure, without any intermediate. Scientists are interested in this new cryptocurrency phenomena as well. Holub & Johnsonkuri (2018) analyzed 13.5 thousand research papers from 20 different databases (EBSCOhost, Elsevier, JSTOR, SSRN and other) on Bitcoin and other cryptocurrency topics. This fast-growing topic is across many disciplines: technical fields (29.9%), economics (24.9%), regulation (17.1%), finance (8.3%) and others [3]. Miau & Yang (2017) estimated that the number of literatures on Blockchain, Bitcoin and other cryptocurrencies topics is still increasing over the year in different disciplines [4]. Looking from financial perspective, cryptocurrency literature review can be split in to several main themes: factors valuation, analysis on returns, forecasting, market speculation, market efficiency. It is complicated to make forecast models for cryptocurrencies. Cryptocurrencies are traded all over the world, 24 hours a day, in more than 200 market places. Comparing the same cryptocurrency at the same time in different market places, the price may vary by a few or even several tens of dollars. The main aim of this paper is to compere different forecast models for short-run cryptocurrency: Bitcoin (BTC), Bitcoin-Cash (BCH), Ethereum (ECH), Litecoin (LTC), Ripple (XRP).

II. LITERATURE REVIEW

Forecast technics are being developed and applied in order to make better decisions. In this article we discuss time series forecast whose observations depends on the time. We offer to split Time series forecast models in three deferent groups, from each group shortly presented one model used in further research:

1. *Classic forecast methods* (moving average (MA), autoregressive models (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and others). These methods can be split in *data-driven forecasting methods* (averaging or smoothing), where there is no difference between a predictor and a target, and *sophisticated model-driven forecasting methods* (ARIMA) [5]. An ARIMA(p, d, q) is a model in which the time series has been differenced d times, and the resulting values are predicted from the previous p actual values and q previous errors [6]. The p, d, q values identifies ACF and PACF plots.
2. *Volatility forecast methods* (autoregressive conditional heteroskedasticity (ARCH), generalized autoregressive conditional heteroskedasticity (GARCH) and others). These methods are different from classic because it fits for non-stationary time series, where mean and variance changes over time. GARCH statistical model which includes not only the variation of error, but also the variance itself. It's ARMA variation of the model for the variance variable.
3. *Advanced forecast methods* (Support Vector Machines for Regression (SVR), Artificial Neural Networks (ANNs)). These artificial intelligence methods are easier to use, do not include assumption constraints such as linearity, normal distribution and a specific observation number and produce more accurate forecasts [7]. The main idea behind SVR is to minimize error in hyperplane, knowing error tolerance rate and by using support vectors. Forecast results depend on the proper selection of the hyper-parameters (kernel, degree of deviations (C) and support vector used (ε)) [8]. Thera are many different flavours of ANNs, and there are very interesting, but in this stage we investigate only very basic models, like in [9]. However we plan to investigate performance of ANNs in future.

III. Data and research metodology

We have chosen five cryptocurrencies with highest trading volume: Bitcoin (BTC), Bitcoin-Cash (BCH), Ethereum (ECH), Litecoin (LTC), Ripple (XRP), for period from 1st of June to 30th of September 2018. Period was selected due to availability download data from coindesk website [10]. The periods of cryptocurrency and dollar course is 5 minutes; hence we have 35136 different values for one cryptocurrency. During the period 68 records of data were missing (excluding Bitcoin), for 5 – 10 min. period. The missing values were replaced by the previous one. Table 1 and fig. 1 shows cryptocurrencies rate statistics of the period.

TABLE I. Cryptocurrency descriptive statistic

	BTC	BCH	ETH	LTC	XRP
Min.:	5785	413	170.3	47.67	0.25
1st Qu.:	6387	532.4	277.8	58.02	0.33
Median:	6594	699.6	421.8	76.55	0.45
Mean:	6792	691.4	384.5	75.94	0.4364
3st Qu.:	7222	810.9	473.6	85.08	0.51
Max.:	8479	1206.3	624.5	127.37	0.77

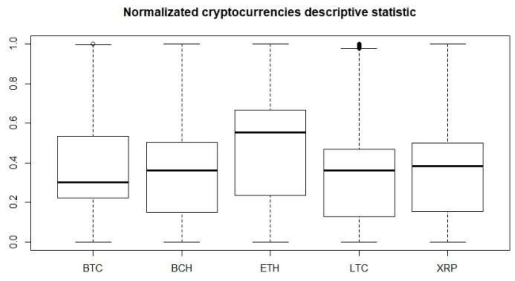


Fig. 1. Cryptocuriencies normalizated descriptive statistic

Forecast research was separated in two groups: 1) real value and 2) volatility. ARIMA, SVR methods and data normalization were used for the *first forecast group*. Data normalization is used to compere data results between different cryptocurrencies, as in:

$$x_{new} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x_{new} – new value, x_i – real value, x_{min} – minimum value of x sample, x_{max} – maximum value of x sample.

GARCH and SVR methods were used for the *second forecast group*. Variable of interest is counted, as in:

$$y_t = \frac{x_t - x_{t-1}}{x_{t-1}} \quad (2)$$

where y – interest, t – time, x – real value [11].

Research results are evaluated by RMSE and MAE coefficients, as in:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(a_t - f_t)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|a_t - f_t| \quad (4)$$

where f_t – forecast expected value, a_t – real value [6]. These measures of error were chosen due to possibility to compare results with other authors research. RMSE and MAE coefficients are calculated for test sample and 3 forecast samples: 1 step forward, 1 hour forward (12 steps) and 1 day forward (288 steps).

Limitations: ru

The basic version of model is analyzed. The chosen models' coefficients are used to all cryptocurrencies the same. Only in GARCH model ARFIMA hypothesis for high-frequency components coefficients vary for different cryptocurrencies.

IV. Results

A. Real value forecast (ARIMA and SVR)

ARIMA. Assumption of stationarity can be evaluated with visual inspection by autocorrelation function (ACF) and partial autocorrelation (PACF) plots [13] and by using Augmented Dickey-Fuller (ADF) test [6]. We rejected hypothesis of stationarity, because of trend in fig. 1 and ADF test results (p-value >0.5 (from all cryptocurrencies)). So, differencing order is equal 1. The parameters of p (stands of AR model) and q (stands of MA model) depends ACF and PACF plots. ACF and PACF plots are the same for ARIMA models: 110, 011 or 111. These models were checked by Akaike Information Criteria (AIC) [6] and the best results showed, that the best model for most cryptocurrencies – ARIMA(1,1,1). The ARIMA and SVR results are presented in table 3.

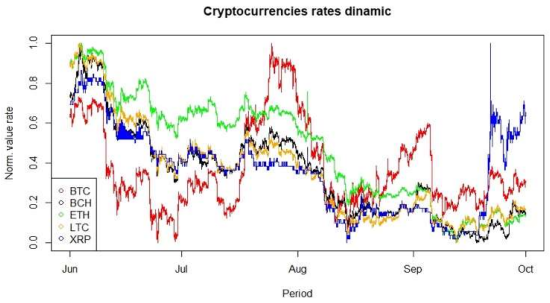


Fig. 2. Cryptocuriencies normalizated value rate dinamic

SVR. Only basic linear kernel was used due to high calculation requirements of other kernels (kernel = 2, C = 3 and ε=0.1). SVR model on Bitcoin normalized prices presented in fig. 3.

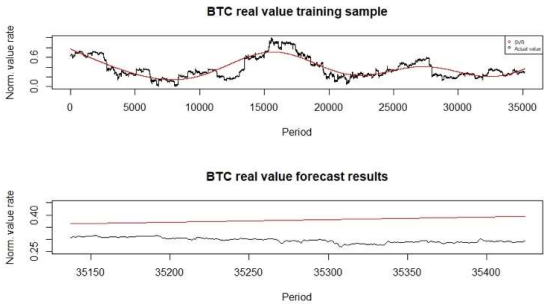


Fig. 3. Real value forecast BTC results

B. *Volatility foracast (GARCH and SVR)*

GARCH. In research is used GARCH (1,1) model, also known as basic GARCH. For each cryptocurrency we found out the best ARFIMA model by AIC and BIC criteria's results and Ljung-Box test confirmation (table 2). The ARFIMA model is compound GARCH method part for mean model calculations.

TABLE II. ARFIMA MODEL SELECTION FOR GARCH MODEL

	BTC	BCH	ETH	LTC	XRP
ARFIMA	(0,0,0)	(1,0,2)	(0,0,2)	(1,0,3)	(1,0,1)

SVR. Due to limited speed resources was used the same basic linear regression based SVR model (kernel = 2, C = 3 and ε=0.1). SVR model on Bitcoin volatility presented in fig. 3. The GARCH and SVR forecast results are presented in table 4.

TABLE III. ARIMA AND SVR MODELS RESULTS

*10^{-2}		TRAINING SAMPLE					FORECAST (1 step forward - 5 minutes)				
		BTC	BCH	ETH	LTC	XRP	BTC	BCH	ETH	LTC	XRP
RMSE	ARIMA	**0.438**	**0.220**	**0.168**	**0.249**	**0.747**	**0.135**	**0.268**	**0.110**	**0.020**	**0.019**
	SVR	11.585	5.492	4.78	4.674	7.335	5.951	6.002	2.760	3.974	9.570
MAE	ARIMA	**0.240**	**0.134**	**0.093**	**0.160**	**0.270**	**0.135**	**0.268**	**0.110**	**0.020**	**0.019**
	SVR	8.896	4.213	3.846	3.726	4.673	5.951	6.002	2.760	3.974	9.570
*10^{-2}		FORECAST (12 step forward – 1 hour)					FORECAST (288 step forward – 1 day)				
		BTC	BCH	ETH	LTC	XRP	BTC	BCH	ETH	LTC	XRP
RMSE	ARIMA	**0.796**	**0.690**	**0.269**	**0.490**	**2.586**	**1.217**	**0.832**	**0.538**	**0.926**	**3.255**
	SVR	5.385	5.685	2.651	3.629	7.559	8.421	7.126	3.673	5.487	14.52
MAE	ARIMA	**0.762**	**0.667**	**0.251**	**0.405**	**2.230**	**1.000**	**0.718**	**0.464**	**0.772**	**2.359**
	SVR	5.381	5.683	2.649	3.620	7.451	8.222	7.031	3.629	5.36	13.79

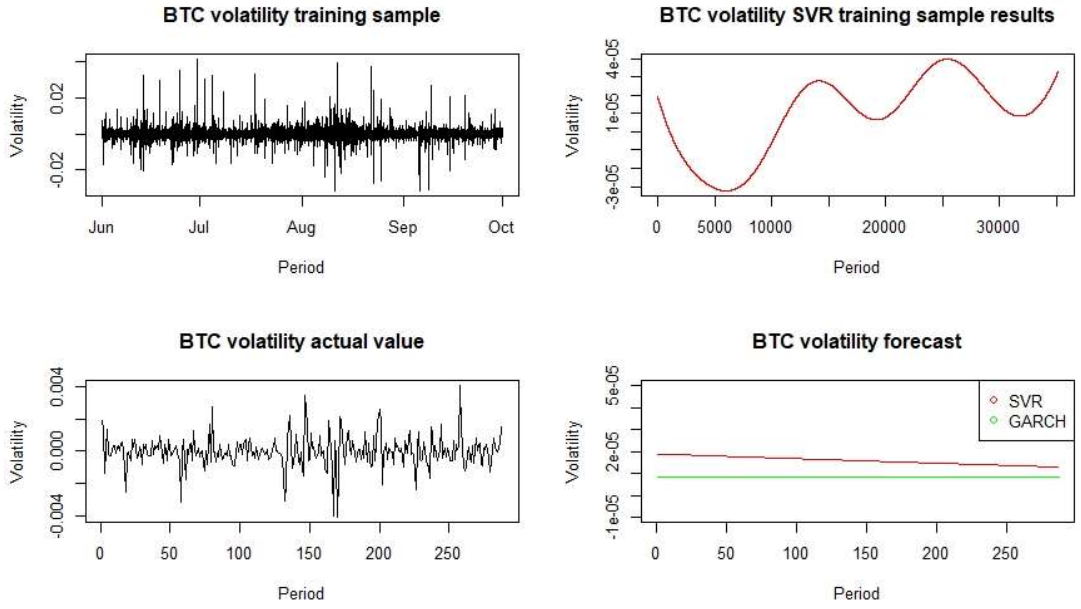


Fig. 4. Volatility forecast BTC results

TABLE IV. GARCH AND SVR MODELS RESULTS

*10^{-3}		TRAINING SAMPLE					FORECAST (1 step forward - 5 minutes)				
		BTC	BCH	ETH	LTC	XRP	BTC	BCH	ETH	LTC	XRP
RMSE	GARCH	**0.008**	**0.191**	**0.400**	**0.313**	**2.648**	1.876	4.680	0.750	0.266	**0.117**
	SVR	1.764	2.540	2.100	2.758	9.056	**1.866**	**4.277**	**0.525**	**0.257**	0.723
MAE	GARCH	**0.008**	**0.566**	**0.223**	**0.201**	**1.047**	1.876	4.680	0.750	0.266	0.117
	SVR	0.955	1.574	1.178	1.765	3.056	**1.866**	**4.277**	**0.525**	**0.257**	**0.723**
*10^{-3}		FORECAST (12 step forward – 1 hour)					FORECAST (288 step forward – 1 day)				
		BTC	BCH	ETH	LTC	XRP	BTC	BCH	ETH	LTC	XRP
RMSE	GARCH	0.912	2.759	0.990	1.741	9.785	**0.940**	2.755	**1.350**	**1.129**	**8.380**
	SVR	**0.909**	**2.701**	**0.965**	**1.720**	**9.601**	0.941	**2.753**	1.351	1.136	8.407
MAE	GARCH	0.673	2.183	0.857	1.299	**5.663**	**0.6318**	**2.114**	1.041	**0.871**	**3.778**
	SVR	**0.671**	**2.161**	**0.812**	**1.291**	6.011	0.6321	2.115	**1.038**	0.878	4.320

V. CONCLUSIONS

Globalization and market availability let us to trade in cryptocurrencies every day at any time. Cryptocurrencies is difficult to forecast due to complexity of the involved parties, low regulation and other reasons. The main aim of this research was to try different forecast models and find out one, most suitable for cryptocurrency. Five cryptocurrencies (BTC, BCH, ECH, LTC, XRP) were selected due to their market capitalization and spread availability to buy or sell in different market places. These cryptocurrencies were analyzed for period from the 1st of June to the 30th of September 2018.

The strategy was to start from the simple models. Analyses was performed on (1) real values (ARIMA and SVR) and (2) volatility (GARCH and SVR). Performance was evaluated using RMSE and MAE for test sample and 3 forecast samples: 1 step forward, 1 hour forward (12 steps) and 1 day forward (288 steps).

The main results are:

1) The best results in real value forecast were achieved using ARIMA for ETH.

2) The best results for volatility forecast were achieved using SVR for BTC.

3) If we compare this analysis with Peng and others (2017) research, this analysis GARCH (1,1) results are better due to the intensity of the data (5 minutes not 1 day) and test sample length (1 day not 1 months). But Peng and others analyzed 9 different GARCH methods and the best results received by combined the traditional GARCH model with Support Vector Regression (SVR-GARCH) [13].

In future research we are planning to investigate performance of SVR-Garch, and different ANN (especially Deep Learning, e.g. LSTM) models.

REFERENCES

[1] P. Ciaian, M. Rajcaniova and d. Kancs, "Virtual relationships: Short- and long-run evidence from BitCoin and altcoin markets," *Journal of International Financial Markets Institutions and Money,* vol. 52(C), p. 173–195, 2017.

[2] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf. [Accessed 13 2 2019].

[3] M. J. J. Holub, "Bitcoin research across disciplines," *Information Society,* Vols. 34:2, 1, pp. 14-126,, 2018.

[4] S. Y. J. Miau, "Bibliometrics-based evaluation of the Blockchain research trend: 2008 – March 2017.," *Technology Analysis & Strategic Management,* vol. 30:9, pp. 1029-1045, 2017.

[5] V. Katu and B. Deshpande, "Chapter 10 - Time Series Forecasting," in *Predictive Analytics and Data Mining*, Elsevier Inc. , 2015, pp. 305-327.

[6] J. Mount and N. Zumel, Exploring Data Science, Manning Publications, 2016.

[7] A. C.H. and E. Eğrioğlu, Advanced Time Series Forecasting Methods, ProQuest Ebook Central, 2012.

[8] B. Ojemakinde, "Support Vector Regression for Non-Stationary Time Series," Master's Thesis, University of Tennessee, 2006.

[9] N. Ahmed, A. Atiya, N. E. Gayar and H. El-Shishiny, "AN EMPIRICAL COMPARISON OF MACHINE LEARNING," *Econometric Reviews,* vol. 29, no. 0747-4938, p. 29(5–6):594–621, 2010.

[10] coindesk, 2018. [Online]. Available: https://www.coindesk.com/. [Accessed 30 09 2018].

[11] "ARCH/GARCH Models," The Pennsylvania State University, 2018. [Online]. Available: https://newonlinecourses.science.psu.edu/stat510/node/85/. [Accessed 05 12 2018].

[12] N. R., Fuqua School of Business Duke University, 2018. [Online]. Available: https://people.duke.edu/~rnau/411arim3.htm. [Accessed 05 12 2018].

[13] Y. Peng, P. H. M. Albuquerque, A. J. A. Padula and M. R. Montenegro, "The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression.," *Expert Systems with Applications,* vol. 97, p. 177–192, 2017.

INFORMACINËS TECHNOLOGIJOS • IVUS 2019 • ISSN 2029-249X • eISSN 2029-4824

Detection of Different Types of Vehicles from Aerial Imagery

Jonas Uus
Applied Informatics faculty
Vytautas Magnus University
Kaunas, Lithuania
Email: jonas.uus@bpti.lt

Tomas Krilavičius
Vytautas Magnus University
Kaunas, Lithuania
Baltic Institute of Advanced Technology
Vilnius, Lithuania
Email: tomas.krilavicius@bpti.lt

***Abstract*—Accurate detection of vehicles in large amounts of imagery is one of the harder objects' detection tasks as the image resolution can be as high as 16K or sometimes even higher. Difference in vehicles size and their position (direction, they face) is another challenge to overcome to achieve acceptable detection quality. The vehicles can also be partially obstructed, cut off or it may be hard to differentiate between object colour and its foreground. Small size of vehicles in high resolution images complicates the task of accurate detection even more. CNN is one of the most promising methods for image processing, hence, it was decided to use their implementation in YOLO V3. To deal with big high resolution images method for splitting/recombining images and augmenting them was developed. Proposed approach allowed to achieve 81.72% average precision of vehicles detection. Results show practical applicability of such approach for vehicles detection, yet to reach higher accuracy on tractor, off-road and van categories of the vehicles the count in different vehicle categories needs to be balanced, i.e. more examples of the mentioned vehicles are required.**

I. INTRODUCTION

Vehicles' detection from aerial photography is a very important and quite a difficult task, especially when it is performed in real time or high resolution aerial or satellite images are used for vehicle detection, such as 18000x18000 px. resolution images in COWC [1] dataset. As the drones are used in more and more sectors (according to "cbinsights", currently unmanned aerial vehicle (UAV) could be used in 38 different sectors [2]), for that reason the volume of video and photo material from drones is increasing, the need to create solution for making use of this unprecedented amount of data has become pronounced (at the moment of writing this paper YouTube returns more than 3.3 million results with "aerial footage" query). For human to annotate vehicles from videos or high resolution images it takes a lot of resources. Thus vehicle detection task needs to be automated.

In this paper we investigate applicability of Convolutional Neural Networks. Due to good performance [3], we use YOLO V3 (You only look once) [4] CNN as a tool to apply proposed splitting/merging images method.

Moreover, we split the image into fixed overlapping rectangular frames (*a sliding window method*).

Some results show that Yolo V2 performs quite well with aerial imagery only with applied modifications: "First making the net shallower to increase its output resolution. Second changing the net shape to more closer match the aspect ratio of the data." [5].

In another vehicles detection solution newer YOLO version was used [6]. Images were taken from 3 publicly available datasets: VEDAI, COWC and DOTA. The model had good test results for small objects, rotating objects, as well as compact and dense objects, with 76.7% mAP and 92% recall.

None of these solutions used splitting and remerging technique with images' overlapping. They used already presplitted images.

II. PROBLEM

As computing speed is increasing, and technology is advancing neural networks are being optimised, it had been decided to apply best image augmentation/splitting/remerging methods for vehicle detection. In the application of neural network the following set of problems becomes apparent:

1) Having a variety of different resolution images in dataset (HD, Full HD, 2K ...).
2) Uneven vehicles' sizes in a dataset which are influenced by different ground sample distances (GSD).
3) Uneven vehicles' count in categories by having more cars than other vehicle categories combined.
4) Almost all of the fully connected convolutional neural networks have a fixed-size first layer and all images should be resized to fit the first layer.
5) Vehicles can be partially obstructed (only part of the vehicle could be seen).
6) Hard to differentiate vehicles from foreground (for example, black car parked in a shadow).
7) Vehicles may be facing multiple directions depending on the camera flight direction and its rotation.
8) Available vehicle detection solutions are limited to detecting a small number of features.
9) After re-merging splitted images, the same vehicle may be detected multiple times.

Currently existing vehicles' detection solutions are subject to company trade secrets and companies do not openly discuss technical specifications and application results (for example web platform Supervisely [7]). That is why it is difficult to adapt or even sometimes impossible to add additional functionality, some solutions are based on older versions of

neural networks (for as long as they are functional) and they detect few vehicle categories. For example, one of the vehicle detection solutions [8] detects vehicle features based only on their size (either a small or a large vehicle). Also, currently available solutions which uses CNNs mostly work with fixed size input images or rescale them to fixed size as existing deep convolutional neural networks (CNNs) require a fixed-size (e.g. 224x224) input images [9]. As rescaling images is detrimental for small objects features within images, the images thus are split into smaller pieces, then after the process of individual detection of vehicles in each piece, every image is remerged into full sized image. For example, if a high resolution image such as 4K is rescaled to 608 by 608 pixels, then a rear glass of a car is about 20 by 10 px. after rescaling, and thus the window width becomes about 6 times smaller and its height about 3.5 times smaller and the size of a window decreases to about 6 by 6 px., as a result it becomes harder to differentiate between a van and a car and the probability of misidentification increases. In case of multiple detections in the overlapping image pieces, the NMS (Non-Maximum Suppression) [10] is used to remove duplicate detections as NMS retains only the overlapping bounding box with highest probability (if its area overlaps more than preset value).

The herein discussed practice of YOLO application encompasses the attempt to solve all of the above problems.

III. Dataset

MAFAT tournament [11] provided images which were used for training, validation and csv file with boxes and classes, but the csv file was created with classification task in mind and it was not used. The images were adapted for object detection task as the original dataset was initially created for classification task and not every object was annotated, or *false positives* [12] (also called a false detection, vehicle is annotated where there is none) were assigned. Every image was manually annotated and some of them were removed. Those images that were removed were not taken orthogonal to ground, they were taken at an angle. Only images with top-down view were kept. For image augmentation horizontal and vertical flipping and rotation at $45°$ intervals was used.

Following is the count of dataset images:

1) 1712 images were chosen as training images, about 80% of original training dataset images.
2) After splitting training images into 500x500 pixel pieces, images count rose to 9141.
3) 1986 images were chosen for validation, about 78% of original validation dataset images.
4) 12 227 vehicles were annotated manually by me in the training dataset, Fig. 1.
5) 10 914 vehicles were annotated manually by me in the validation dataset, Fig. 2.

The number of vehicles used in the training images is presented in Fig. 1 and the validation datasets are presented in Fig. 2.

The characteristics of dataset images:

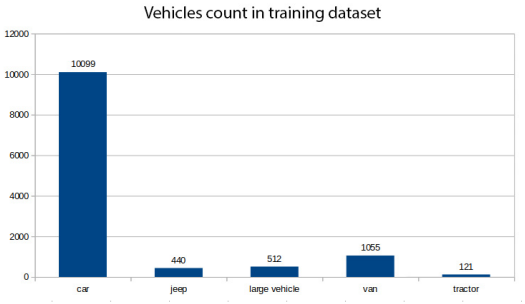


Fig. 1: Vehicles count in training dataset

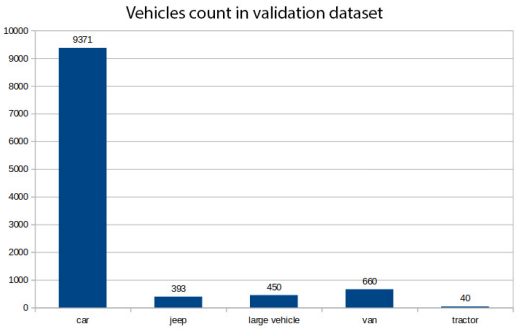


Fig. 2: Vehicles count in validation dataset

1) Images were taken from a variety of locations, some were taken in cities, others in rural areas.
2) Images were taken at a different time of a day.
3) Vehicles were lit from different sides.
4) The resolution of images were different from 900x600 px. to 4010x3668px.
5) Some parts of images were darkened out (for example one half of image was made completely black, while another half of image has picture).
6) GSD (Ground sample distance) of images varied between 5 and 15 cm.
7) Objects in images might have been obstructed by trees or cut off, only part of vehicle might have been seen (for example, a car parked in a garage, a car near the edge of the image).

Couple of images examples taken from dataset Fig 3.

See variation in image resolutions in table I.

The categories of vehicles that were being detected:

1) Car,
2) Off-road vehicle,
3) Large Vehicle,
4) Van,
5) Tractor.

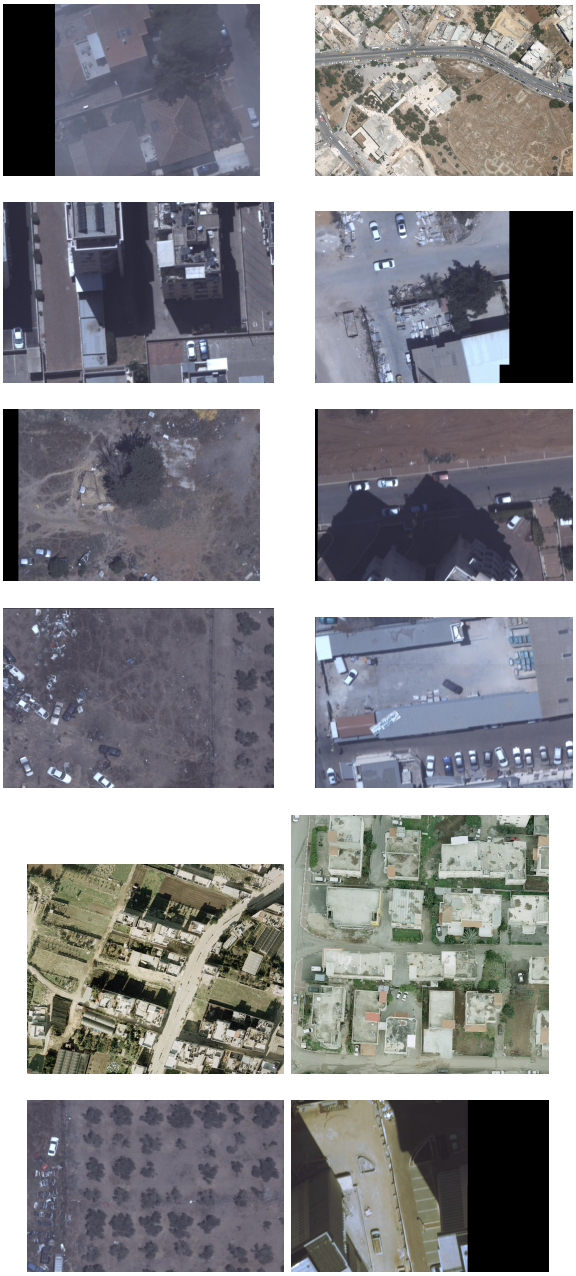


Fig. 3: Examples of images in dataset

TABLE I: Distribution of images with different resolution in dataset

Image resolution (px)	Validation dataset	In Training dataset
900 x 600	1975	1592
1057 x 800	2	3
1332 x 1283	1	0
2026 x 1649	6	37
4010 x 2668	2	40

The above dataset was considered sufficient for the evaluation of developed method.

IV. Proposed solution

The objective was to develop a method for identification of diverse vehicles.

Image resolution and sizes

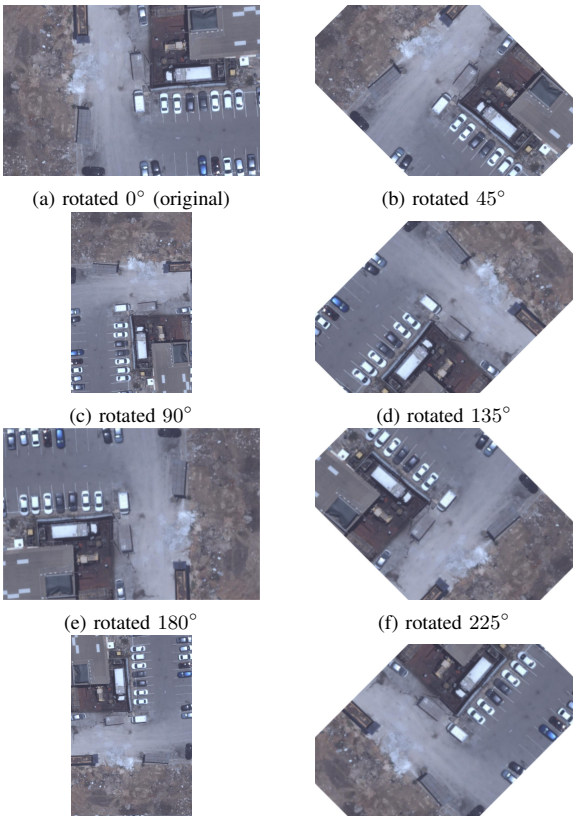
The use of CNNs is complicated due to the dataset having a variety of different resolution images (HD, Full HD, 2K ...) and uneven vehicles' sizes in a dataset, see Sect. III. The different sizes in the images are influenced by different ground sample distances (GSD) [13]. As almost all of the convolutional neural networks have a fixed-size first layer [9], all images are resized to that layer size, so if an image resolution is as high as 16K and it is being resized to, for example, 608x608 px. all of the small vehicle features will disappear from the subsequent image. For this reason we propose to split the image into fixed overlapping rectangular frames (*a sliding window method*). This produces double detection problem as vehicles may be detected on both images. To remove duplicates, NMS (Non-Maximum Suppression) is used [10]. If two or more bounding boxes overlap with same vehicle category, then the box with highest detection probability is kept, while the others are removed. Amount of overlapping is determined by finding largest possible vehicle size in the dataset. This ensures that if the vehicle was cut off on one of the images, it would be fully visible in another image.

Image obstruction

One more problem with vehicle detection in images is that the vehicles can be partially obstructed (only part of the vehicle could be seen) for example when car are half parked in garage, or when car are parked alongside tree ant tree branches obstruct car features, or when car is on the edge of image.

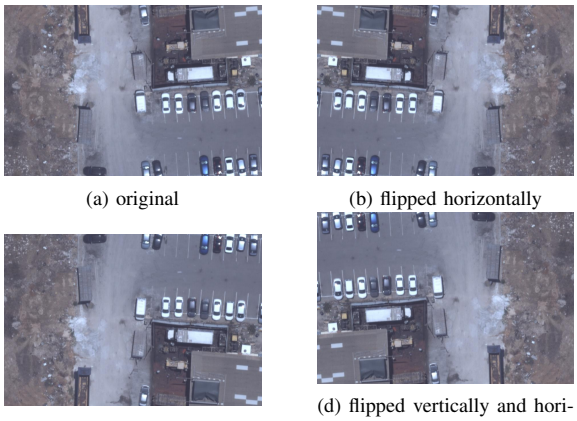
Orientation

As vehicles orientation in images are not constant they may be facing multiple directions depending on the camera flight direction and its rotation. To solve different vehicles orientation problem, the images are augmented with random rotation at $45°$ intervals Fig. 4.



(a) rotated 0° (original) (b) rotated 45° (c) rotated 90° (d) rotated 135° (e) rotated 180° (f) rotated 225° (g) rotated 270° (h) rotated 315°

Fig. 4: Images augmentation by rotating



(a) original (b) flipped horizontally (c) flipped vertically (d) flipped vertically and horizontally

Fig. 5: Images augmentation by flipping

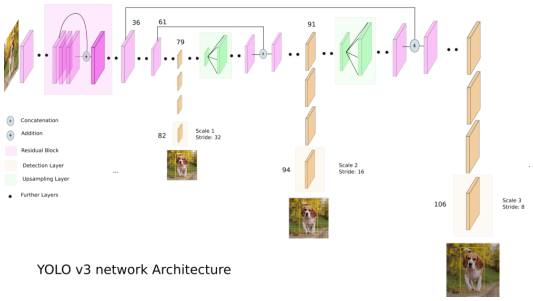


Fig. 6: YOLO V3 architecture [14]

Image augmentation

To increase images' count, images were augmented by rotating them at 45 degrees intervals. Additionally, dataset images were augmented by flipping them vertically, horizontally and by flipping both horizontally and vertically Fig. 5.

V. Experiments

A. Tools

For experiments, convolutional neural network YOLO V3 was used on Darknet framework. YOLO V3 architecture is presented in Fig. 6.

On original YOLO repository the problem was that while training, detection loss climbed to infinity, when any single parameter was changed, thus another forked repository [15] from github was used instead, as it does not have the same issue. For YOLO V3 to work with splitting/ merging workflow, original source code was modified. To know when the training had to be terminated, an average loss value was observed. It was observed that if any bigger change was to be carried out on neural network, such as adding new object category, the neural network should be trained from previous weights in which neural network had been more generic at detections. Training after changing parameters from scratch would be even better, but that would take longer. It was observed that YOLO detects new class better when previous best weights are not used.

Also, it is hard to differentiate an off-road from a car when looking from above, as the body shape of an off-road may differ only slightly (for example, be wider), thus off-road was annotated as a car. Jeep category is hard too, as the only difference between a car and a jeep is that a jeep has a rear spare tire attached or it has a truck bed (like a pickup).

B. Dataset

Vehicle categories like cars, jeeps, large vehicles, vans and tractors need to be detected from the aerial photographs and their position needs to be marked by drawing bounding box around each the object. At first, cars' class had been divided into hatchbacks and sedans, but during manual objects' annotation it was observed, that if a car is half obstructed and only its front part can be seen, it is impossible to tell whether it is a sedan or a hatchback as the only differentiating factor is the size of rear glass and only the trunk/ boot can be seen. For this reason, sedans and hatchbacks were merged into one vehicle class.

As the dataset contained mostly cars, YOLO learned that if unsure, it should ascribe an object to a car category, that way it could reach better mAP result in a long run than guessing rarer classes. This non-homogeneous dataset problem shows up, if dataset has different number of vehicles for given class in dataset. This non-homogeneous dataset problem could be solved by adding images in which rarer classes' vehicles are shown or by augmenting a larger number of rearer class images than images with other vehicles.

Cross-validation statistical method was used during YOLO training, the dataset was divided into images for training and validation. The neural network can not see any of the validation images during training, it can only see them when its performance is validated. This method is used to prevent overfitting. The following modifications were performed for the purposes for training and validation images in a dataset:

1) Images' slicing/ overlapping parameter values modification.
2) Fixing wrongly annotated vehicle data and their bounding boxes' locations in the datasets.
3) Changing vehicles' count of classes by adding, merging existing, then reannotating dataset.
4) Choosing images from dataset for training/ validation.
5) Experimenting with images' manipulations (vertical/ horizontal flipping, image rotation), this drastically improved dataset size. These manipulations were manually coded as YOLO, unlike Tensorflow, does not have these image manipuliations integrated.

The following modifications which were done on YOLO:

1) Change of YOLO layer resolution (mostly first layer, as all images are resized to the same resolution as the first layer size).
2) Experiments with different YOLO configurations and different layers' count.
3) Change of network parameters (such as anchors, recalculating certain layer size after vehicles' classes modifications, learning speed).
4) Adding a module to darknet for easier work with split images and for external communication with other programs.

C. Experiments results

To evaluate performance PASCAL VOC evaluation metrics were used and the results were compared using AP (average precision) [16]. This metric uses Jaccard index [17] for calculating IOU (intersection over union) to compare between ground truth and detection boxes.

After training the YOLO V3 neural network it managed to detect cars with 78.69% average precision (AP) Fig. 7, large vehicles with 44.85% average precision (AP) Fig. 8. Other vehicle categories such as jeeps, vans and tractors were detected but they were wrongly categorised. That was the reason vehicle category detection average precision was very low. To solve this problem, the dataset needs to have more unified count of vehicles in every category.

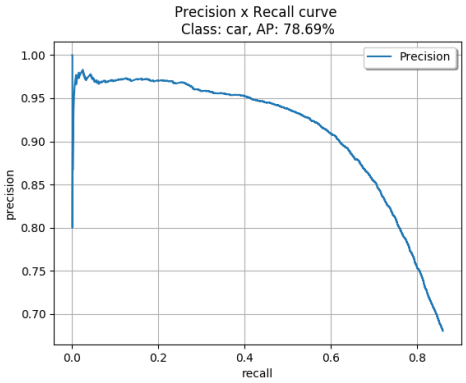


Fig. 7: Precision and recal curve for cars category

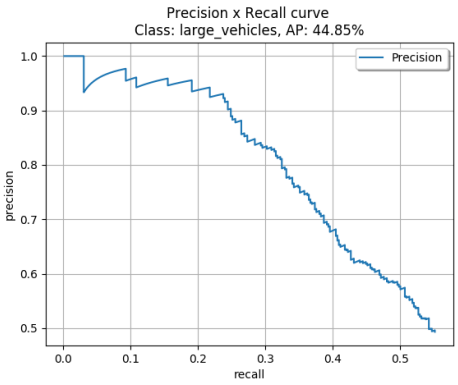


Fig. 8: Precision and recall curve for large vehicle category

The above figures show how precision and recall are corelated, for example, if we choose precision at 95%, then 45% of cars were detected in validation images at that level of precision. *F-Score* [18] at this precision level is equal to 0.61, if recall increases to 80% then the precision drops to 75%. *F-Score* at 75% is equal to 0.77. When all categories were merged into one and then results were validated again, average precision increased to 81.72% Fig. 9. This indicates that in order detection precision is increased, YOLO V3 needs to classify categories more accurately.

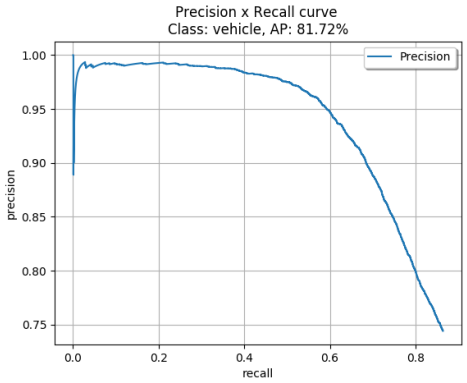


Fig. 9: Precision and recal graph when all vehicles are merged to one category

VI. Conclusions

This application could be used for statistics (to count how many vehicles are there in a given image), vehicles tracking, prediction of further vehicle movement direction and real-time vehicle detection from real time video feed. A vehicles' detection application was created so as users could easily configure it and make vehicles' detection task easier. The user only needs to input images and a couple of parameters to execute vehicles' detection with CNN.

Results:

1) Dataset was prepared for vehicles detection task by manually annotating all of the vehicles in dataset images.
2) Images' were augmented to increase dataset size.
3) Method for combining splitting and joining images and using convulutional neural network for vehicles detection was proposed.
4) Proposed method performance was tested by using YOLO V3 CNN

Conclusions:

1) When YOLO V3 is used together with proposed method is capable of detecting cars with 79% accuracy and large vehicles with 45% accuracy.
2) When proposed method is used, YOLO V3 CNN still has difficulty detecting characteristics of other vehicles, such as off-road, tractors and vans which makes the final detection result lower.
3) Proposed method helps to avoid losing vehicles and their features that would otherwise be lost by resizing high resolution images.
4) The dataset used for training and validation should have more unified count of vehicles categories (more photos with tractors, large vehicles and jeeps should be added to the dataset).

For future work R-CNN and SSD networks will be trained on Tensorflow framework as those networks are also widely used CNN's for object detection tasks and they will be tested using same proposed method. Also, as currently used images' dataset is relatively small, it needs to be increased from freely available datasets and photos taken from drones. As the dataset should have more unified count of vehicles categories, more photos with tractors, large vehicles and jeeps should be added to the dataset.

References

[1] Wesam A. Sakla Kofi Boakye T. Nathan Mundhenk, Goran Konjevod. A large contextual dataset for classification, detection and counting of cars with deep learning. *arXiv:1609.04453*, 2016.
[2] CBINSIGHTS. 38 ways drones will impact society: From fighting war to forecasting weather, uavs change everything. Accessed: 2019.02.22.
[3] Joseph Redmon and Ali Farhadi. Yolo: Real-time object detection. Accessed: 2019.02.22.
[4] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.
[5] Jennifer Carlet and Bernard Abayowa. Fast vehicle detection in aerial imagery. *CoRR*, abs/1709.08666, 2017.
[6] J. Lu, C. Ma, L. Li, X. Xing, Y. Zhang, Z. Wang, and J. Xu. A vehicle detection method for aerial image based on yolo. *Journal of Computer and Communications*, pages 98–107, 2018.
[7] Supervise. The leading platform for entire computer vision lifecycle. Accessed: 2019.02.22.
[8] Alexey. Object detection on satellite images. Accessed: 2019.02.22.
[9] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv:1406.4729v1*, 2014.
[10] Adrian Rosebrock. Non-maximum suppression for object detection in python. Accessed: 2019.02.22.
[11] yuvalsh. Mafat challenge - fine-grained classification of objects from aerial imagery. Accessed: 2019.02.22.
[12] Google. Classification: True vs. false and positive vs. negative. Accessed: 2019.02.22.
[13] Wikipedia contributors. Ground sample distance. Accessed: 2019.02.22.
[14] Ayoosh Kathuria. What's new in yolo v3? Accessed: 2019.02.22.
[15] Alexey. Yolo-v3 and yolo-v2 for windows and linux. Accessed: 2019.02.22.
[16] Jonathan Hui. map (mean average precision) for object detection. Accessed: 2019.02.22.
[17] Wikipedia. Jaccard index. Accessed: 2019.02.22.
[18] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. volume Vol. 4304, pages 1015–1021, 01 1970.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Content Analysis Methods for Estimating the Dynamics of Facebook Groups

Rasa Kasperienė
Faculty of humanities
Vytautas Magnus univercity
Kaunas, Lithuania
rasa.kasperiene@gmail.com

Tomas Krilavičius
Faculty of informatics
Vytautas Magnus univercity
Kaunas, Lithuania
tomas.krilavičius@vdu.lt

***Abstract*— The relationship between the content that is generated by the users of social networks and their dynamics has been analyzed by many scholars. However, due to favorable data policies, the majority studies have been carried out by analyzing Twitter data. In addition, such research on Facebook (FB) groups (esp. political) is usually qualitative. The present study analyses the dynamics as well as topic dynamics of radical right political groups on FB by employing a quantitative research methodology. The current paper draws on a large data set that is comprised of posts from FB groups. Overall, there are 79 728 posts which are made up of more than 2 million words and were generated within the timespan ranging from 2010 to 2018. The experimental set up compares the general dynamics and the dynamics of activity on four topics in two radical right FB groups (i.e., pro-Russian and other radical right) in Lithuania. The results show that the year 2014 was important for the radical right FB groups in Lithuania. Newly created pro-Russian FB groups started growing rapidly, whereas the posting activity in other radical right FB groups started to decrease. The topic word Lithuania is relevant for the whole activity time when it comes to all the radical right FB groups. Such topic words as Russia and land correlate with national and international political crisis.**

Keywords—Facebook, radical right, groups dynamics, timestamp.

I. INTRODUCTION

In recent years, the European Union has been witnessing the growth of radical political communities throughout Europe, including Lithuania. Many European countries are witnessing elections in which people vote for far-right and nationalist parties, even though they are at the opposite ends of a wide political spectrum. The migrant crisis accelerated a backlash against the recent political balance, but the wave of discontent also taps into long-standing fears about globalization and dilution of national identity. The increase in the percentage of radical wing voters substantially surpasses the percentage of immigration inflow [1].

The political radicals are more avid and enthusiastic to adopt new technology and have thus found the virtual space to be a uniquely useful place [2]. Through membership in groups, one can define and confirm his/her values and beliefs through incoming information or discussion. When members of such groups face uncertain situations, they can gain reassuring information about their problems and find security in companionship [3]. It is also important to highlight the fact that social media provides fertile ground for the dissemination of propaganda and disinformation as well as the manipulation of people's perceptions and beliefs [4]. Social networks can become a tool for manipulating the masses and fighting wars with little to no cost.

The present paper proposes a framework for carrying out research on posts from Facebook (further FB) groups as a means to reveal information dissemination and group behavior patterns in communication by information transmission dynamics in groups. In particular, the aim of this study is to analyze the establishment of radical right FB groups in relation to the political events of the time as well as the dynamics of the most prominent themes by using the data retrieved from FB groups and R toolset. This article investigates the launch of Lithuanian radical right FB groups in a wider political context. It is important to understand the dynamics and the reasons behind the activity of such groups. Another important issue is to pinpoint when the topics discussed in the aforementioned FB groups become relevant and no longer relevant. Finding the answers to these questions can provide a deeper insight into the social processes of radical right groups on FB.

Such social networks as FB and Twitter have become the most popular social networks in the world. In 2017, Twitter had more than 330 million active users, whereas FB had more than 2.13 billion monthly active users with a 14 per cent increase every year [5]. This giant flow of information has already shown to be useful for event detection [6], identifying public health issues [7], behavioral information propagation [1], community discovery [8], sentiment analysis [9], identification of communication roles [9], and recently as a means to aid political uprising [10] as well as a medium that can help to pinpoint and analyze the act of triggering an (upcoming) uprising [11].

II. DATA SET

FB groups are the place for small group communication and for people to share their common interests and express their opinion. Such groups allow people to come together around a common cause, issue or activity in order to mobilize, express their objectives, discuss issues, post photos and share related content [12]. All FB groups have a title and a group description that indicate the common cause of group activity. FB groups can be public or closed. In the first scenario, every FB user can access group content. In the latter, content can be accessed only with a permission given by the group administrator. To comply with the ethical aspects of doing research, the present study only reports on data that has been retrieved from public FB groups.

The data were downloaded by using the FB graph API [13]. The Graph API is created to get data into and out of the FB platform. This FB platform uses low-level HTTP-based API access that can only be obtained by a user who is registered as a FB developer. For API requests, it is necessary

to have the access token (app id) together with its app password and the access token.

FB API requests return the following group data [13]: post author id (from_id) as numeric string, post author name (from_name) as string, post text (message) as string, post creation date (created_time) as string, post type (type) as string, link in the post (link) as string, post id (id) as numeric string, daily entry (story) as bool, likes count (likes_count) as number, comment counts (comments) as number, shares count (shares_count) as number.

To analyze the posts of FB groups as a means of information dissemination together with the patterns of group behavior in terms of communication by information transmission dynamics in groups, the following subset of data was used: post text (message) as string, post creation date (created_time) as string, post id (id) as numeric string.

To handle the large dataset more efficiently, fingerprint of each FB group was created, and it contains the names and ids as well as the names of the dataset that come from the FB groups in focus.

The radical right groups on social network FB were identified through the Facebook search engine. The supporters of radical right diverge from other individuals though manifestation of nationalism, strong nation [14] and xenophobic ideology [15]. Nationalistic ideology relates to ethnocentrism and Euroscepticism. Xenophobic ideology relates to anti-immigration policy, hostility to ethnical minorities, and intolerance to sexual minorities. To identify radical FB groups by using the FB search engine, their most prominent characteristics were taken into consideration, and based on that, the following keyword list was compiled: *Lithuania, Lithuanians, land sale, European Union, NATO, refugees, refugee crisis, Muslims, Jewish restitution, Jew, Russian, Roma tabor, gay pride, gay mountaineering*. More than 20 most recent posts in each group that match the keywords were analyzed. After the analysis that aimed to pinpoint the FB groups which openly exhibit radical ideology, only 10 groups that proved to endorse radical ideology were chosen for a more in-depth analysis. The FB group selection criteria were the following: the presence of radical left ideological features on group titles, description and latest posts, the size of the group (more than 100 members), activity – the most recent post published at least 2 days prior to the analysis.

The data retrieved from FB groups were divided into two datasets, pro-Russian and other radical right groups. The analysis reveals that some radical right groups in addition to the nationalistic ideology manifest pro-Russian and pro-socialism ideology. Even though in some cases the titles and descriptions of the group's manifest nationalism and the idea of strong Lithuania, there was also support for Russian politics or a sense of nostalgia for the Soviet Union.

Each data set is comprised of five FB groups. As was previously indicated, to be able to handle such large amounts of data, the datasets were supplemented with additional records, i.e., the group and cluster ids. The dataset of pro-Russian FB groups consists of more than 70 150 posts. The second dataset, i.e. that of the other radical right groups, is comprised of 9 578 posts. The former dataset of groups has 13 940 members, whereas the latter has 6 126.

TABLE I.

Short data info	
Posts published period	4th of March 2010 – 1st of January 2018
Number of posts	79728
Download date	12th of February 2018
Max length of word	15 symbols
Min length of word	1 symbol

Lithuanian is a highly inflectional language, i.e. there are two grammatical genders for nouns and there are three genders for adjectives, numerals, participles, and pronouns. Every word must follow the gender and the number of the noun. All these features produce a substantial number of inflective forms of lemma. To avoid any loss of data, the lemmatization of the texts in FB posts was not used.

III. METHODS

To analyze the dynamics of the topics discussed in groups, the most frequent words were employed as features [16]–[18]. In addition, social networks post timestamp modelling was applied to analyze the behavior of online users [19], [20]**.** This paper proposes to study the posts from FB groups as a means of information dissemination and group behavior patterns in communication by information transmission dynamics in groups. The proposed approach is based on the following observation: the amount of information passed from one period to another in the social network may be quantified in different ways. For example, in the dataset of FB groups, the amount of information can be quantified by the time that passed from one post's appearance to other. The quantity of published group posts in a social network by looking at the time frame can show group behavior.

To grasp the information transmission when it comes to the group dynamics, the datasets of FB groups were expanded by adding *fingerprints* entries. Let a pro-Russian FB groups dataset be denoted by D_1 and another radical right group dataset be denoted by D_2. W represents time window (W = 6 months). Denote each Facebook post $\overline{as\ e_{ij}}$ where $i=1$ represents that a post belongs to D_1 and $i=2$ represents that post belongs to D_2; $j = 1; n_i$ where n_i is the number of posts in group D_i. Each post e_{ij} consists of p_{ij}, t_{ij}, g_{ij}. Each post p_{ij}, consists of a set of words $p_{ij} = (w_{ij1}, w_{ij2} \dots w_{ijk})$, where k is the number of words in p_{ij}.

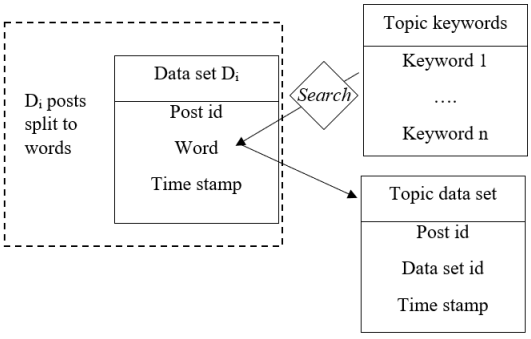


Fig. 1. Datasets of Facebook groups with expanded *fingerprints* entries

To compare the dynamics of the users in the two datasets, the transformed dates were stored from string to POSIXct objects. To transform the dates, *Lubridate* [21] package for R was used. In order to visualize the distribution of groups' activities through time, *ggplot2* [22]] package for

R was used. It helps to visualize the distribution of a single continuous variable by dividing the x-axis into bins and counting the number of observations in each bin. To make the text of the post tidy and the datasets lighter, *Tidytext* [23] and *Stringr* [24] packages for R were employed. By using these packages, English and Lithuanian stop words were removed. To estimate the dynamics of the topics in the collected posts, each entry (in form of sentences) was split into words. Once again, to keep track of data, every split word was supplemented with a post and dataset id, group name, and timestamp entries.

IV. Experiment

The preliminary analysis identified two types of radical right ideology in FB groups under investigation. The visualization in "Fig. 2" compares the dynamics of pro-Russian and other radical right groups' activity. It includes the posts (message) of both groups' members and post creation time (created_time). It also shows the peak activity periods that can be noticed in the datasets (within a time window of six months).

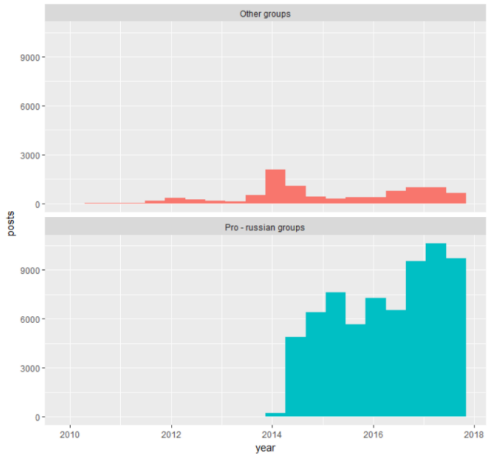


Fig. 2. The dynamics of radical right groups' activity on Facebook

The experiment shows that the activity of radical right FB groups starts in 2010, whereas pro-Russian groups emerge on FB four years later, in 2014. The pro-Russian groups that were created on the same year reached three times greater activity compared to other radical right groups on FB. From 2014 to 2017, the activity of pro-Russian groups has been increasingly growing. The activity has reached the maximum peak in 2017 with 23 413 posts per year "table 2". Until 2014, the radical right groups were witnessing the growth of posting activity every year, too. The year 2014 was important for the radical right FB groups as new ideology-following radical right groups started appearing and rapidly growing. After the appearance of pro-Russian groups on FB, the data spread in other radical right groups started decreasing, but the activity of pro-Russian groups on FB increased each year. This is evident because in 2015, the activity of pro-Russian groups on FB was 61 per cent greater than in previous year. Finally, in 2017, the posting activity in pro-Russian FB groups is 44 per cent greater than it was initially in 2014.

TABLE II.

The dynamics of radical right Facebook groups activity								
Radical right groups	**Year**							
	2010	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***
Pro-Russian	0	0	0	0	10447	17033	19257	23413
Other	58	311	604	1164	2997	791	1914	1739
Total	58	311	604	1164	13444	17824	21171	25152

During the course of the Ukrainian crisis, the role of actual military interventions has remained low in comparison to different tools of asymmetric warfare (e.g., information warfare, economic measures, cyber war, and psychological war on all levels), often referred to as hybrid warfare [25]. This cyber war passed national or post-Soviet Union borders more widely and the spread of *fake news* reached the western world. The conflict in the Ukraine re-awakened Russian propaganda. For example, Twitter analyst Lawrence Alexander has identified an increase in bot registration coinciding with the start of the Euromaidan protests on 2013/2014 year in Ukraine and subsequent armed uprisings by pro-Russian militants in Eastern Ukraine in early spring of 2014 [26]. Lawrence's investigation corelates with rise of pro-Russian Facebook groups in 2014. Prior to 2014, on FB there were only radical right groups with low activity, but after 2014, the situation has changed. The activity of the newly created pro-Russian groups started rapidly growing. According to *NATO Strategic Communications Centre of Excellence*, some techniques, such as Russian propaganda techniques in particular, are used for achieving psychological influence and manipulation on social media [27]. One of such techniques is the mass-generated content which is used in order to spread manipulative messages and minimize alternative voices.

To analyze the dynamics of the most relevant topics in the groups, four keywords were chosen, namely, *Lithuania*, *Russia*, *land*, and *sky*. The words *Lithuania*, *land and Russia* were chosen for this experiment based on the previously defined most prominent characteristics of radical right groups. The word *Russia* also was chosen in order to assess and compare the dynamics of topics discussed by pro-Russian and other radical right in relation to the country. The neutral word *sky* was chosen to reveal whether there is any space for neutral topics in the datasets of radical right groups.

TABLE III.

The dynamics of the word *Lithuania* in the posts of radical right groups on Facebook								
Radical right groups	**Year**							
	2010	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***
Pro-Russian	0	0	0	0	2195	5173	4410	5302
Other	51	74	446	657	1890	520	740	857
Total	51	74	446	657	4085	5693	5150	6159

The topic word *Lithuania* is relevant for all the radical right groups "Fig. 3". This word in the posts of FB groups appears more than 22 300 times throughout the whole

period of groups' activity "table 3". In 2014, the topics that mentioned the word *Lithuania* were mostly discussed by newly created pro-Russian groups rather than by other radical right groups. In 2017, both types of radical right FB groups mentioned *Lithuania* in the content of their posts the most frequently if compared to the previous years. *Lithuania* appears 5 302 times in pro-Russian groups and 857 times in radical right groups.

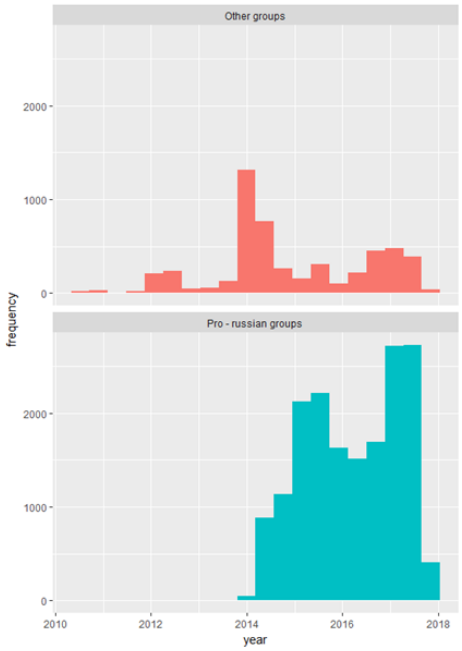


Fig. 3 The dynamics of the word *Lithuania* in the posts of the radical right groups on Facebook.

As was previously mentioned, the increased instances of mentioning *Lithuania* were the most prominent in pro-Russian groups. *NATO Strategic Communication Centre of Excellence* claims that in the period ranging from 1 November 2017 to 31 January 2018, the proportion of bot activity in Twitter was relatively high, with 62 per cent of all tweets mentioning NATO and Lithuania [31]. In other radical right FB groups, Lithuania is mentioned less often as opposed to the pro-Russian groups. The data in the NATO report correlate with the experimental results. The Russian hybrid troll or bot activity campaign has reached the users of social networks in Lithuania, and the experiment shows that this campaign is still being successfully implemented. According to NATO Hybrid trolls (as we have labelled hired, pro-Russian trolls), communicate a particular ideology and, most importantly, operate under the direction and orders of a particular state or state institution. In the context of the Ukraine crisis, the aim of hybrid trolls has been to promote the Kremlin's interests and portray Russia as a positive force against the 'rotten West' and the US hegemony[28].

Russia-related topics seem to be more important to pro-Russian groups than other radical right FB groups (Fig. 4). The word analysis of the FB groups' posts that were split to words shows that from the beginning to the end of 2018, the words *Russia* appeared 16 times more than in other radical right groups. The word count indicates that the word appeared 2 864 times in pro-Russian and 178 times in other radical right groups "table 4".

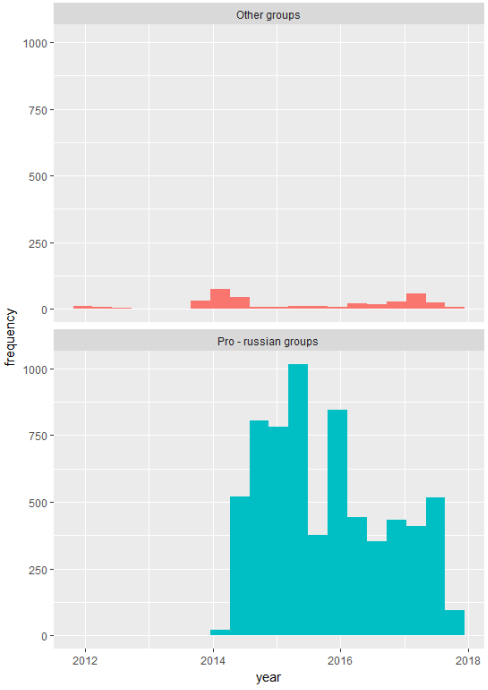


Fig. 4 The dynamics of the word *Russia* in the posts of the radical right groups on Facebook.

The word *Russia* in the topics of pro-Russian groups was most frequently used in 2014 and 2015. This data correlate with Russia's policy and international political crises of 2013 and 2015 – after Russian military intervention to Ukraine, various sanctions were imposed on Russia by the United States, the European Union (EU) and other countries as well as international organizations. In 2015, Russia intervened to Syrian civil war (30 September 2015 – February 2016) and this event correlates with the dynamics of the topics on Russia in pro-Russian FB groups Russian. The members of other radical right groups show no attention to this international crisis, the Russian topic in their FB posts is irrelevant.

TABLE IV.

The dynamics of the word *Russia* in the posts of radical right groups on Facebook								
Radical right groups	**Year**							
	2010	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***
Pro-Russian	0	0	0	0	768	1032	659	405
Other	0	3	6	18	57	16	31	47
Total	0	3	6	18	825	1048	690	452

Creating 'noise' or 'informational fog' around a topic is a strategy used to distract attention from more strategically important events. An important example of this has been the case of the downing of Malaysian air flight MH17. Russian media channels and social media distributed a large volume of messages offering numerous explanations

for why the plane crashed. Another bot campaign was launched to distract the public by offering an *alternative explanation* of the murder of the Russian politician Boris Nemtsov, claiming that he was killed by jealous Ukrainians. Such 'news' were published just a few hours after the attack [1]. The experiment shows that the word *Russia* in the pro-Russian groups became more actively used during the turmoil caused by Russia's policy. This could have affected the results of the trending topics in order to make 'noise' or 'informational fog' around any given topic.

The themes related to *land* are more relevant to the members of both groups. The word count estimations show that from 2010 to 2017, the word *land* appeared 110 times in pro-Russian groups and 141 times in other radical groups "table 5". The assessment of the thematic dynamics of the groups indicate that in 2014, the word *land* was more popular in the posts of other radical right groups than in what was posted by pro-Russian users (Fig. 5).

TABLE V.

The dynamics of the word *land* in the posts of radical right groups on Facebook								
Radical right groups	**Year**							
	2010	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***
Pro-Russian	0	0	0	0	46	15	32	17
Other	0	0	3	41	94	1	1	1
Total	0	0	3	41	140	16	33	18

The word correlates with the Lithuanian land-related political crisis related to the restrictions imposed on foreigners who want to purchase land for agricultural purposes in Lithuania. The referendum by the Republic of Lithuania held on 2014 July was related to the abovementioned restrictions. Prior to the referendum, there were many protests and a rally against land purchase restrictions. These events also ignited debates in the virtual space and affected the topics that were generated in the radical right FB groups.

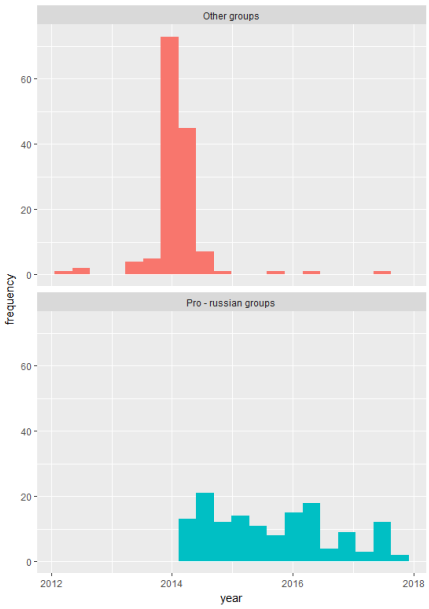


Fig. 5 The dynamics of the word *land* in the posts of the radical right groups on Facebook

In order to compare the content of the posts in radical right FB groups, a neutral keyword *sky* was chosen. The assessment of dynamics show that the word *sky* did not appear in the content produced by the radical right FB groups "

The dynamics of the word *sky* in the posts of radical right groups on Facebook								
Radical right groups	**Year**							
	2010	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***
Pro-Russian	0	0	0	0	1	5	4	13
Other	0	0	0	0	0	0	0	1
Total	0	0	0	0	1	5	4	14

This indicates that the content generated by the members of the radical right groups is similar to the political background. As Veronika Solovian, the administrator of the popular Finnish-Russian website russia.fi, admits, the trolls are commenting on political topics. They are able to attract other participants into arguments, and other users do not necessarily immediately identify them as trolls [29]. The experiment reveals that political topics are indeed relevant for radical right-wing political groups on Facebook. The largest part of the generated political content could be generated by trolls or bots. Therefore, social media provides fertile ground for the dissemination of propaganda and disinformation. The latter indicates that social media can be an effective tool to manipulate people's mind and influence their decisions. Ease of Use

V. Conclusions and Future Work

Facebook developer acc with API requests and R tools set (*Lubridate, Tidytext, ggplot2)* can help to analyze radical right FB groups establishment and themes dynamics. For social and political scientists, the most important result is

that in Lithuania radical right groups on Facebook posts together with nationalism, strong nation and xenophobic ideology also appears topics related to the support for the Russian policy and former communist ideology. The analysis reveals that some radical right groups in addition to the nationalistic ideology manifest pro-Russian and pro-socialism ideology.

Radical right groups on Facebook started to appear in 2010, but the year 2014 was important for the radical right FB groups as new ideology-following radical right groups appeared and was rapidly growing each year. Experiment data correlates with the awakening of Russian propaganda on social media.

The topic word *Lithuania* is relevant for all the radical right groups. This word in the posts of FB groups appears more than 22 300 times throughout the whole period of groups' activity. The increased instances of mentioning *Lithuania* were the most prominent in pro-Russian groups. Russia-related topics seem to be more important to pro-Russian groups than other radical right FB groups and land-related topics is more important to other radical right groups. These topics activity correlates with national or international political crisis: the land-related topics activity reaches its maximum before referendum related to the restrictions imposed on foreigners who want to purchase land for agricultural purposes in Lithuania, the word *Russia* in the topics of pro-Russian groups was most frequently used in 2014 and 2015 while after Russian military intervention to Ukraine, various sanctions were imposed on Russia by the United States, the European Union (EU) and other countries as well as international organizations. The assessment of dynamics show that the word *sky* did not appear in the content produced by the radical right FB groups. This indicates that the content generated by the members of the radical right groups is similar to the political background.

Future plans are to make different kind of radical right groups generated content most frequency words estimations and analyze it dynamics. Future work is also to analyze dynamics of FB groups incoming information and the posting dynamics of most active groups' members.

VI. References

[1] B. Podobnik, M. Jusup, D. Kovac, and H. E. Stanley, "Predicting the Rise of EU Right-Wing Populism in Response to Unbalanced Immigration," *Complexity*, vol. 2017, pp. 1–12, Aug. 2017.

[2] J. Bartlett, "From hope to hate: how the early internet fed the far right | World news | The Guardian," 2017. [Online]. Available: https://www.theguardian.com/world/2017/aug/31/far-right-alt-right-white-supremacists-rise-online. [Accessed: 03-Mar-2019].

[3] D. R. Forsyth and D. R. Forsyth, *Group dynamics*, 2nd ed. Pacific Grove Calif.: Brooks/Cole Pub. Co, 1990.

[4] NATO strategic communications centre of excellence, "Robotrolling 2018/1 | StratCom," *2018*. [Online]. Available: https://www.stratcomcoe.org/robotrolling-20181. [Accessed: 03-Mar-2019].

[5] "Top 20 Facebook Statistics - Updated March 2019." [Online]. Available: https://zephoria.com/top-15-valuable-facebook-statistics/. [Accessed: 03-Mar-2019].

[6] J. Weng, Y. Yao, E. Leonardi, and F. Lee, "Event Detection in Twitter," 2011.

[7] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health."

[8] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," 2007.

[9] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Comput. Human Behav.*, 2014.

[10] J. Park, W. L.-P. of the fifth international A. conference, and undefined 2011, "Revolution 2.0 in Tunisia and Egypt: Reactions and sentiments in the online world," *icwsm.org*.

[11] K. Alaimo, "How the Facebook Arabic Page 'We Are All Khaled Said' Helped Promote the Egyptian Revolution," *Soc. Media + Soc.*, vol. 1, no. 2, p. 205630511560485, Sep. 2015.

[12] Matt Hicks, "(18) Facebook Tips: What's the Difference between a Facebook Page and Group? | Facebook," 2010. [Online]. Available: https://www.facebook.com/notes/facebook/facebook-tips-whats-the-difference-between-a-facebook-page-and-group/324706977130/. [Accessed: 03-Mar-2019].

[13] "Graph API." [Online]. Available: https://developers.facebook.com/docs/graph-api. [Accessed: 03-Mar-2019].

[14] K. Arzheimer, "Electoral Sociology: Who Votes for the Extreme Right and why-and when?"

[15] C. Mudde, *The ideology of the extreme right*. 2018.

[16] J. B.-L. and L. Computing and undefined 1992, "Not unles you ask nicely: The interpretative nexus between analysis and information," *academic.oup.com*.

[17] D. H.-L. and L. Computing and U. 2004, "Delta prime?," *academic.oup.com*.

[18] M. Eder, J. R.-L. and L. Computing, and undefined 2012, "Do birds of a feather really flock together, or how to choose training samples for authorship attribution," *academic.oup.com*.

[19] R. Perera, … S. A.-2010-M. 2010, and undefined 2010, "Twitter analytics: Architecture, tools and analysis," *ieeexplore.ieee.org*.

[20] Q. Zhao, Y. Tian, Q. He, N. Oliver, … R. J.-P. of the 19th, and undefined 2010, "Communication motifs: a tool to characterize social communications," *dl.acm.org*.

[21] G. Grolemund, H. W.-J. of S. Software, and undefined 2011, "Dates and times made easy with lubridate," *academia.edu*.

[22] L. Wilkinson, "ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H.," *Biometrics*, vol. 67, no. 2, pp. 678–679, Jun. 2011.

[23] J. Silge, D. R.-T. J. of O. S. Software, and undefined 2016, "tidytext: Text mining and analysis using tidy data principles in r," *theoj.org*.

[24] H. Wickham, "Simple, Consistent Wrappers for Common String Operations [R package stringr version 1.4.0]."

[25] V. Sazonov, H. Molder, and Muur Kristiina, "COMBINED ANALYSIS PREPARED BY THE NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE RUSSIAN INFORMATION CAMPAIGN AGAINST THE UKRAINIAN STATE AND DEFENCE FORCES," 2015.

[26] A. Lawrence, "Social Network Analysis Reveals Full Scale of Kremlin's Twitter Bot Campaign · Global Voices," 2015. [Online]. Available: https://globalvoices.org/2015/04/02/analyzing-kremlin-twitter-bots/. [Accessed: 03-Mar-2019].

[27] Svetoka Sanda, "Social Media as a Tool of Hybrid Warfare | StratCom," 2016. [Online]. Available: https://www.stratcomcoe.org/social-media-tool-hybrid-warfare. [Accessed: 03-Mar-2019].

[28] NATO, "Internet Trolling as a hybrid warfare tool: the case of Latvia | StratCom," 2016. [Online]. Available: https://www.stratcomcoe.org/internet-trolling-hybrid-warfare-tool-case-latvia-0. [Accessed: 03-Mar-2019].

[29] J. Aro and Y. Kioski, "This is How Pro-Russia Trolls Manipulate Finns Online – Check the List of Forums Favored by Propagandists," 2015. [Online]. Available: https://www.stopfake.org/en/this-is-how-pro-russia-trolls-manipulate-finns-online-check-the-list-of-forums-favored-by-propagandists/. [Accessed: 03-Mar-2019].

Fast Language-Independent Correction of Interconnected Typos to Finding Longest Terms

Using Trie for Typo Detection and Correction

Behzad Soleimani Neysiani
Department of Software Engineering, Faculty of Computer & Electrical Engineering,
University of Kashan,
Kashan, Esfahan, Iran,
B.Soleimani@grad.kashanu.ac.ir
Tel Number: +98-913-960-6471

Seyed Morteza Babamir
Department of Software Engineering, Faculty of Computer & Electrical Engineering,
University of Kashan,
Kashan, Esfahan, Iran,
Babamir@kashanu.ac.ir

***Abstract*— Triagers deal with bug reports in software triage systems like Bugzilla to prioritizing, finding duplicates, and assigning those to developers, which these processes should be automated, especially for huge open source projects. These bug reports must be mined by text mining, information retrieval, and natural language processing techniques for automation processes. There are many typos in user bug reports which cause low accuracy for artificial intelligence techniques. These typos can be detected based on standard dictionaries, but correction of these typos needs human knowledge based on the context of bug reports. It is important which neither Google Translator nor Microsoft Office Word can detect interconnected terms –a common type of typos in bug reports- having more than two meaningful terms. This research provides a novel language-independent approach for fast correction of interconnected typos based on natural language processing and human neural network structure to detect and correct interconnected typos. A new tree-based method proposed for term matching and two algorithms proposed for fast longest term finding in an interconnected typo. A dataset is used including 180-kilo typos based on four famous bug report dataset of Android, Eclipse, Mozilla Firefox, and Open Office projects. Then proposed method evaluated on typos versus the state of the art. The results show the runtime performance of the proposed method is as same as the related works but the average words length is improved and at least more than 57% of typos in the dataset can be classified as interconnected typos.**

Keywords*— *Information Retrieval, Natural Language Processing, Duplicate Detection, Bug Reports, Typo Correction, Lexical Interconnected Typo, Trie

I. INTRODUCTION

Many huge projects, especially open source projects have a large range of analyzers, designers, developers, testers and end users, which after each new release, all of them may find some issues or bugs and/or have some suggestions to improve the software. Software triage systems such as Bugzilla are a software which usually gets these reports online and then the Triagers will deal with these bug reports to evaluate the importance and priority of each report, finding duplicate reports based on their contents, assign bug reports to developers for checking bugs and planning to modifying the project in future [1]. Because of the large amount and volume of bug reports, many researchers have tried to automate these processes since 2004 by artificial intelligence techniques and algorithms[1]. Duplicate bug reports detection is a famous problem in this research area [2, 3]. The algorithms and techniques of duplicate bug report detection such as Term Frequency and Inverse Document Frequency in information retrieval technique need to check the similarity of two bug reports to each other word by word, so the lexical correctness of words and terms is very important for these techniques [4]. There are many typos in bug reports, e.g. more than 50% of bug reports have typos and more that 2.5% of bug reports have more than 50% typos [4]. These typos distort similarity detection process in duplicate detection. It is very important to detect and correct these typos automatically because there are more than 1.5 million typos [4] in Mozilla Firefox, Android, Open Office and Eclipse datasets [5] and about 390 kilo unique typos in those. A scientific semi-dictionary is made for typo detection in bug reports to detect typos automatically [4] including general English words and many scientific words like abbreviations or proper nouns. This semi-dictionary can be made for every language based on some valid reference like computer dictionaries or reference websites.

There are many types of typos in texts such as additional, removal or substitute characters. Interconnected terms are a regular typo in the software context because there are many method or class names in this context which contains interconnected terms like ‘getItemById’ or ‘printAllMembers’. Sometimes these words are camel case and sometimes users typed them and have not any specific case sensitivity. Also, sometimes typists forgot to press space between words, so there will be many interconnected terms in the software bug reports or even other contexts too. Also, it is possible to find some interconnected typos in optical character recognition (OCR) output too. These interconnected terms must be separated otherwise humans and/or computers algorithms and methods like term frequency of information retrieval techniques cannot recognize the text or detect similarities for duplicate bug report

detection problem. The main purpose of this research is to figure out how does correct these typos rapidly.

The organization of the paper is as follows: section 2 explains the literature view and related works. Section 3 describes the methodology of interconnected typo correction, section 4 will discuss about evaluation results in experiments and the section 5 will conclude the research.

II. Literature Review

Typo detection and correction is a regular and an ancient issue in text mining and natural language processing [6, 7]. There are many effort on typo detection and correction in a scientific context like clinical records which uses Shannon's noisy channel model to predict next words based on previous word sequence [8]. In some case there is less previous word sequence like web query, so the log of web query can be used as baseline and maximum entropy model can help for rare queries to conquer the sparseness problem of prior data [9].

Some researchers focus on correction of misspelled typos by different kind of machine learning and natural language processing models e.g. creating a confusion matrix for different type of misspelling like additional or removal or transposal or replaced characters to searching these patterns in terms and predict the correction [10]. Also, phonetic, language and keyboard models can be useful for correction prediction by decision tree as a machine learning based technique [11, 12]. Another approach can be creating a model based on machine learning techniques to detect typos and predict the correction according to context and domain knowledge [13, 14]. String transduction tries to map one string to another and can be used for misspelled typo corrections too [15]. Also, machine learning is used in character scale to typo detection and corrections, but the recall rate is low (about 30%) [16].

Some other researchers focus on using tree structure for typo correction. It is possible to make a tree based on probabilistic model of relation between characters of words which what characters can be come after a special character and in advance mode, after a sequence of characters. So, these models use Bayes theory to make a prediction model on a tree which called Trie and use it for typo correction as the user is typing [17, 18]. The tree structure can be used for grammatical checking and translating too by merging several grammatical trees in a Trie [19]. The simple Trie (without probability) is used for spell checking too [20]. The acyclic deterministic finite automata is a graph with similar structure which can be used for spell checking and typo correction [21]. There are some methods for query in Trie by wild characters too [22]. Trie-based index structure can be used for real-time interaction like search recommendation and query completion [23].

The interconnected terms problem was not important a lot in other contexts and there is no specific method for correction of interconnected terms. As it was tested, the google translate and Microsoft office word just can detect two parts interconnected terms and suggest a correction for them, but if there are more than 2 meaningful terms, they cannot detect and suggest any correction. It shows that even huge companies have not been investigated with this problem. So, a divide and conquer algorithm based on a longest common sequence algorithm have been made as shown in Fig. 1 to find out the meaningful terms in an interconnected term. It is a simple brute force algorithm which will consider all combinations of start and end index of a substring in interconnected term to find a meaningful term. Meaningfully checking needs a dictionary. Luckily a good trustful dictionary for computer context have been made in our last research and can be used for this purpose too.

It's obvious that checking a word in the dictionary is a frequent operation, especially in meaningful word detection so the time complexity of this process is very important. Usually dictionaries sort their terms to use binary search with $\log_2(N)$ time complexity for term checking which N is the number of terms in the dictionary. Also, every word needs to be compared with suspicious meaningful word which complexity of this operation is based on the length of terms even though almost string comparer method use short circuit idea for time reduction, in other words when they find first different character between two words, they will cut the comparison operation. So, the meaningful word detection takes the logarithm (N) operation in worst case to find out the result and it is in worst case usually in this procedure because many substrings are meaningless and they are not in dictionaries.

***Algorithm**: Meaningful Word Finding*
***Input**: a connected term with index of 1 to L*
***Output**: a list of meaningful words with start and end index in connected term*
For I in range of 1 to L
  For J in range of I+1 to L
    If substring of term from I to J is in dictionary
      Put the (I, J, substring) in the output

Fig. 1. Algorithm of finding the meaningful words in an interconnected term

The selected dictionary contains of more than 600,000 terms, so it needs 20 comparing ($\log_2(600{,}000)$) each time. Also, the above algorithm has two for loop which take Combination (n, 2) operations equal to n×(n-1)/2 time complexity and each iteration needs a dictionary term checking, so the total complexity of this algorithm is in the equation (1) which N is the number of terms in dictionaries, L is the average length of each term and *n* is the length of interconnected term. Also this algorithm can be parallel easily by dispatching combinations between some threads or processes and 2 threads or processes can be made at least to parallelize this algorithm which everyone use half combinations.

$$t(N, L, n)_{Alg1} = log_2^N \times L \times \frac{n \times (n-1)}{2} \quad (1)$$

Meaningful substring can be everywhere in interconnected term and have overlap e.g. 'hishe' can be 'hi' and 'she' or 'his' and 'he', so the next step is to find the meaningfulness combination between substrings which have no overlap (e.g. 'his' and 'she' which is not possible according to 's' overlap in main interconnected term). This algorithm use recursive depth first search approach to find all non-overlap combination which

shown in Fig. 2. It take 4 inputs containing the output list of previous algorithm which has the meaningful words, a start index based on the list of meaningful words which show start search index for next combinations, a list of selected index in meaningfulness combination which is considered in current path of depth search, and the last input is the length of interconnected term which can be considered as ac constant in this algorithm. Also, the output of this algorithm is a list of combination too. This final list should be evaluate based on the context of interconnected terms and the best combination is picked semantically. This algorithm will consider all combination of the meaningful words and choose those combination with no overlap, so if there are N words in the meaningful words list, the time complexity of this algorithm equals to 2^N which is exponential and it is a non-polynomial problem, so, it take a long time and it is not suitable for real-time situation like correction suggestion as user is typing in text editors which is very important because if the user look the suggestion and correct this typo, it is no necessary to evaluate the result combination semantically by artificial intelligence techniques and it is enough to sort the output list based on a metric and show the top-10 suggestion to user, then user will pick best one. The average length of words can be a good metric because every much the average length of words be high, the combination contains largest meaningful component in interconnected terms and the possibility of meaningfulness is more.

***Algorithm**: Finding meaningfulness combination between substrings*
***Input**: MW as a list of meaningful words from 1 to n indexes, SI as start index of searching in meaningful words with 0 initial value, SC as selected combination with empty initial value, LCT as length of connected term*
***Output**: a list of meaningfulness combination of words*
If start index is 0
  Consider end point equal to 0
Else
  Consider end point equal to end index in SI-th of MW
  If end point equals to LCT
    Return SC
For each index J which start index of J-th of MW equals to end point
  Consider SCn as new list containing SC plus J as appended value
  Call this algorithm with MW, J, SCn , LCT …
    and put the result in output list if it is not empty

Fig. 2. Algorithm of finding meaningfulness combination between substrings

III. Proposed Method

In the middle procedure of meaningful word finding process, it has been considered that neural networks of human brain look at a word and predict the next letters based on priors and it seems the human brain use a tree like algorithm to finding the correctness of a word. So, a binary like tree proposed to be made for meaningful word checking. This process need 2 steps: creating the tree, parsing the tree for checking the existence of a term in dictionary. Also after making this tree, it was found that this tree can be used to find meaningful words more efficient than brute force algorithm, so in step 3 this tree should be used for finding the meaningful terms. Then these meaningful terms should be checked where which one is more possible in main interconnected term to be meaningful. Thus, there are 4 main steps to separate interconnected terms which shown in Fig. 3 which every step will be explained in next sections with an example.

Fig. 3. The 4 steps of finding meaningful words in an interconnected term

Suppose that there are a dictionary with these words: ‘hello’, ‘book’, ‘help’, ‘his’, ‘hiss’, ‘she’. Also, we consider ‘hellohelphissbookhishel’ as a multiple interconnected term with a typo in last term. Now the process of neural like tree making will be explained for matching the input terms which this tree has called a neural matching tree (NMT).

1.1 Neural Matching Tree Creation Process

This tree is like the binary tree but it have more than two child, so it is an n-ary like tree. It has a root and every word in dictionary, will be appeared as a path below of the root. Every letter in the words will be put in a node in the tree. Also, every node will contain a flag for showing the end of the word and if a node contains a letter which is the end of a word, the flag will be true, otherwise it will be false. Every node can be implemented by a map or dictionary data structure in programming languages. So, for the supposed example, this tree will be like the Fig. 4. In this tree, the flag of end letter of every word is T (true) and has different color.

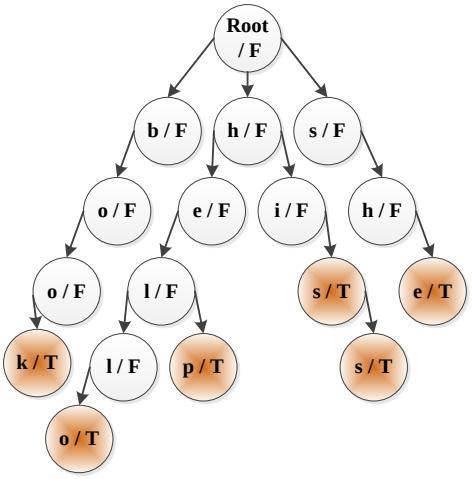


Fig. 4. Neural Match Tree example for supposed dictionary

It is interesting that a path can have multiple final node for example both words 'his' and 'hiss' have same prefix and in this tree, have same path. It should be mentioned that this neural matching tree, can compress the dictionary too. Human mind is like this tree as to when we look at a word, some path in our brain will be activated and we can predict next letters. The procedure of creating NMT is explained in Fig. 5. This algorithm take a list of words of a dictionary and return the root node for NMT. This procedure is very simple and for every word, parse the tree once time from root node and check the child nodes to have the letters of word, otherwise create a child node for each letter as shown in Fig. 3. Also, for the last node which contain the last letter of word, put true for the flag of showing the end of word.

```
Algorithm: Creating Neural Match Tree
Input: a list of words in a dictionary
Output: a root node to a neural match tree
Create a Root node containing no letter with false flag as end character
For each word in words list of dictionary
        Consider n as a node pointed to Root node
        For each letter in word
                If letter is not in child node of n
                        Create a child node for n with letter and false flag
                Consider child node containing the letter as new n
        Put true flag for node n (the last one)
```

Fig. 5. Algorithm of creating neural match tree

Time and memory complexity of NMT are important too. As explained before, the NMT memory is less than a simple dictionary and it can be used for compression too. Also, the creation procedure of NMT need a parse on whole dictionary just one time, so it depends on the number of words in dictionary and the length of each word. If every word has average length of L and there are N words in dictionary (as denoted before), the time complexity will be N×L.

1.2 Neural Matching Tree Using Process

The next step is to use the NMT for checking a new term is valid or not, in other word is it in dictionary or not. This procedure is like the NMT creation process which had described in Fig. 6. It will check every letter of suspicious term is in NMT or not. If the first letter is in child nodes of root node, then the next letter will check with the child node of that selected child node. Unlike the checking process in regular dictionaries which has $Log_2(N)\times L$ time complexity, this process just has L time complexity.

```
Algorithm: Checking the existence of word in NMT as a dictionary
Input: a Root node for NMT and a term for checking
Output: a Boolean value indicating the existence of the … word in the NMT
Consider n as a node pointed to Root node
For each letter in word
  If letter is not in child node of n
    Return False
  Consider child node containing the letter as new n
Return True
```

Fig. 6. Algorithm of checking the existence of a word in NMT

1.3 Finding Meaningful Words in a Interconnected Term by Neural Matching Term

Now, it is time to do the main procedure instead of brute force algorithm in Fig. 1. The first loop step of brute force algorithm cannot be connivance because there may be some lexical mistake in interconnected term and some substring are useless, so every substring maybe meaningful. Second loop in brute force algorithm can be more intelligence based on NMT by cutting the searching existence of substring as soon as find a letter is of substring is not in next node of NMT. The algorithm of finding meaningful words in interconnected term by NMT is shown in Fig. 7. The time complexity of searching in NMT is less than regular dictionaries. So, the time complexity of this algorithm will be L×L.

```
Algorithm: Finding Meaningful Words of a Connected Term by NMT
Input: Root of NMT and a Connected Term with index of 1 to L
Output: the list of meaningful substring of connected term
An empty list for connected terms as output
For I in range of 1 to L
  J = I
  Consider n as a node pointed to Root node
  While J <= L
    If n has a final flag (is the last letter of a word)
      Put substring of connected term from I to J … index into output list
    If letter is not in child node of n
      Break out from while loop
    Consider child node containing the letter as new n
```

Fig. 7. Algorithm of finding meaningful substrings in an interconnected term by NMT

1.4 Finding meaningfulness combination

As mentioned before, recursive algorithm of finding meaningfulness combinations has high time complexity, so, a

new iteration based algorithm has created for this purpose as shown in Fig. 8. There is a new input in this algorithm to limit search based on average word length metric as mentioned and select just top combination with most ranks. This algorithm need a priority queue which can be implemented by heap algorithm as heap queue to contain every combination with its rank. Every time this algorithm choose highest rank combination. If this combination is completed and considered all non-overlap words, the combination will be added to output list, otherwise it will be progressed to choose next word and add to its combination, and calculate the rank again based on new combination. The time complexity of this algorithm depends on number of words in a combination and parameter N which in worst case, if every letter of interconnected term have considered as a meaningful word, the time complexity of this algorithm will be LCT×N and it's polynomial.

***Algorithm**: Finding meaningfulness combination between substrings by NMT*
***Input**: MW as a list of meaningful words from 1 to n indexes, LCT as length of connected term, N as number of top most meaningfulness combination*
***Output**: a list of meaningfulness combination of words*
Consider LS as a list of search states …
with a combination containing 0 as index of selected … words with 0 rank
While LS is not empty and has not N output
Pop highest rank combination in LS as HRC
Choose last index of combination in HRC as end point
For each index J which start index of J-th of MW equals to end point
Consider HRCn as new combination containing HRC combination …
plus J as appended value and calculate rank based on HRCn
If the end point of J-th of MW equals to LCT
Put the combination of HRCn in output list
Else
Put HRCn in LS

Fig. 8. Algorithm of finding meaningfulness combination between substrings

IV. Evaluation Method

The new scientific semi dictionary as word list and unique typo dataset of bug reports have picked for evaluating proposed algorithms [4]. The implementation of proposed algorithms done in Python 3.6 programming language and a Core i5 1.8 GHz computer with 12GB memory having windows 8.1 x64. In the first step, the dataset has analyzed and it was denoted that it has 391,807 suspicious typos but it was detected that 149,749 typos are numeric values which are hexadecimal or have semi-hexadecimal form. Also, some of them was newly devised words and some other have another type of typos. So the terms with length more than 5 characters have been choose for evaluating proposed algorithms which contains 182,402 terms. The evaluation has been done in 3 experiment as follow:

1. Testing algorithms based on explained example with 'hellol654shelphissbookhishel' which also has some digits as interconnected term considering the limited explained dictionary containing these words: 'hello', 'book', 'help', 'his', 'hiss', 'she'.
2. Testing algorithms based on same example with 'hellol654shelphissbookhishel' as interconnected term considering the scientific semi dictionary.
3. Testing algorithms based on selected typos of dataset containing interconnected terms and considering the scientific semi dictionary.

In first test, the NMT creation take 0.04 milliseconds and the process of extracting meaningful words takes 0.45, 1.44 and 0.10 milliseconds for serial and parallel brute force algorithm and NMT based approach respectively. Obviously, NMT algorithm is very fast and have 350 % improvement in time consuming. Then the processes of finding meaningfulness combination have 0.15 milliseconds in both recursive and the new iterative proposed algorithms to extract 'hello l654 help hiss book his hel'.

In second test, 453 milliseconds takes to read the scientific semi dictionary with 610,411 terms from file and the NMT creation take 4,828.07 milliseconds (4.8 seconds). The process of extracting meaningful words takes 0.72, 2.01 and 0.40 milliseconds for serial and parallel brute force algorithm and NMT based approach respectively which shown in Fig. 9 versus first experiment. The processes of finding meaningfulness combination have 233,750.16 and 0.81 milliseconds for recursive and the new iterative proposed algorithm respectively. So the proposed algorithms are very efficient for real-time applications. It is interesting that based on words of this dictionary, one of the final result is 'hel lol 6 54 shel phis sbo ok hishe l' which maybe far from the result of first experiment, but these words were in dictionary (almost as meaningful abbreviations). So it should be considered that this process can be improved to prioritize full words more than abbreviations or special terms, though in some case maybe this result is more useful and it should be selectable.

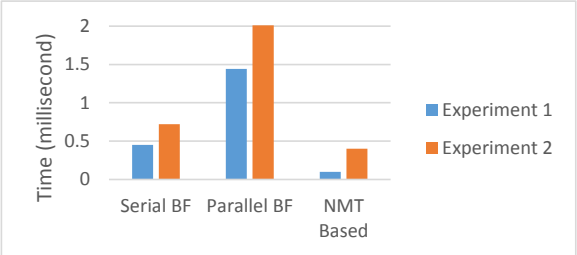


Fig. 9. Time of extracting meaningful words of interconnected terms in experiment 1 and 2

In last test, all typos in the dataset have evaluated and the results save in a file with some statistic about time, words length and words count. Among 182,402 terms, 306 terms have no

meaningful words at all. Other ones (182,096 terms), take 40.132 seconds totally (0.22 milliseconds averagely) to find meaningful words and take 158.25 seconds (about 2.6 minutes) totally (0.86 milliseconds averagely) to find meaningfulness words. Now the results should be manually checked to finding best combination but as an elementary approach for evaluation, the results have been choose as true positive results with at least 2 words and average words length more than 5 characters or having interconnected term with length more than 10 characters and more than 3 characters average words length, which show 104,337 terms (57% of total terms) can be classified as interconnected terms. It shows that interconnected term is one of common typo in software bug reports and should be considered importantly.

V. Conclusion

This research focus on correction of interconnected terms typos by natural language processing based on a reliable word list like a formal dictionary to build an n-ary tree inspired from human neural network to recall the memory. This tree was called Neural Matching Tree (NMT) which is created based on the word list. Then the interconnected term will be parsed based on NMT and the meaningful substrings in interconnected term will be extracted. Then non-overlap combinations of meaningful substrings had picked as correction output words. The proposed algorithms are very simple and their time complexity are negligible. Another achievement of this research is showing that there are many interconnected terms in software context especially bug reports. So the correction of interconnected typos can be useful for other goals like duplicate bug report detection which use information retrieval techniques like term frequency that use the lexical form of words and depends on having non typo in bug reports.

In the next step, many improvement can be used for meaningfulness combination extraction process to achieve the best one between other combinations and also based on the main context. Also, another metrics can be introduced for this purpose instead of average words length which this research had introduced and had used.

References

[1] B. Soleimani Neysiani and S. M. Babamir, "Methods of Feature Extraction for Detecting the Duplicate Bug Reports in Software Triage Systems," presented at the International Conference on Information Technology, Communications and Telecommunications (IRICT), Tehran, Iran, 2016.

[2] B. Soleimani Neysiani and S. M. Babamir, "Improving Performance of Automatic Duplicate Bug Reports Detection Using Longest Common Sequence," in *IEEE 5th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, Tehran, Iran, 2019.

[3] B. Soleimani Neysiani and S. M. Babamir, "New Methodology of Contextual Features Usage in Duplicate Bug Reports Detection," in *IEEE 5th International Conference on Web Research (ICWR)*, Tehran, Iran, 2019.

[4] B. Soleimani Neysiani and S. M. Babamir, "Automatic Typos Detection in Bug Reports," presented at the IEEE 12th International Conference Application of Information and Communication Technologies, Kazakhstan, 2018.

[5] A. Alipour, A. Hindle, T. Rutgers, R. Dawson, F. Timbers, and K. Aggarwal. (2013). *Bug Reports Dataset*. Available: https://github.com/kaggarwal/Dedup

[6] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006, pp. 43-50.

[7] K. Kukich, "Techniques for automatically correcting words in text," *Acm Computing Surveys (CSUR)*, vol. 24, pp. 377-439, 1992.

[8] K. H. Lai, M. Topaz, F. R. Goss, and L. Zhou, "Automated misspelling detection and correction in clinical free-text records," *Journal of biomedical informatics*, vol. 55, pp. 188-195, 2015.

[9] Q. Chen, M. Li, and M. Zhou, "Improving query spelling correction using web search results," in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007.

[10] H. M. Noaman, S. S. Sarhan, and M. Rashwan, "Automatic Arabic spelling errors detection and correction based on confusion matrix-noisy channel hybrid system," *Egypt Comput Sci J*, vol. 40, p. 2016, 2016.

[11] G. A. d. M. Almeida, "Using phonetic knowledge in tools and resources for Natural Language Processing and Pronunciation Evaluation," Master, Universidade de São Paulo, 2016.

[12] G. A. de Mendonça Almeida, L. Avanço, M. S. Duran, E. R. Fonseca, M. d. G. V. Nunes, and S. M. Aluísio, "Evaluating phonetic spellers for user-generated content in brazilian portuguese," in *International Conference on Computational Processing of the Portuguese Language*, 2016, pp. 361-373.

[13] Y. Huang, Y. L. Murphey, and Y. Ge, "Intelligent typo correction for text mining through machine learning," *International Journal of Knowledge Engineering and Data Mining*, vol. 3, pp. 115-142, 2015.

[14] Y. Huang, Y. L. Murphey, and Y. Ge, "Automotive diagnosis typo correction using domain knowledge and machine learning," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013, pp. 267-274.

[15] J. Ribeiro, S. Narayan, S. B. Cohen, and X. Carreras, "Local String Transduction as Sequence Labeling," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1360-1371.

[16] M. Korpusik, Z. Collins, and J. Glass, "Character-based embedding models and reranking strategies for understanding natural language meal descriptions," *Proc. Interspeech*, pp. 3320-3324, 2017.

[17] H. Duan and B.-J. P. Hsu, "Online spelling correction for query completion," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 117-126.

[18] B.-J. Hsu, K. Wang, and H. Duan, "Online spelling correction/phrase completion system," ed: Google Patents, 2012.

[19] K. Oflazer, "Error-tolerant tree matching," in *Proceedings of the 16th conference on Computational linguistics-Volume 2*, 1996, pp. 860-864.

[20] H. Shang and T. Merrettal, "Tries for approximate string matching," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, pp. 540-547, 1996.

[21] S. Deorowicz and M. G. Ciura, "Correcting spelling errors by modeling their causes," *International journal of applied mathematics and computer science*, vol. 15, pp. 275-285, 2005.

[22] N. Ito, "Character-string retrieval system and method," ed: Google Patents, 1997.

[23] P. Fafalios and Y. Tzitzikas, "Type-Ahead Exploratory Search through Typo and Word Order Tolerant Autocompletion," *J. Web Eng.*, vol. 14, pp. 80-116, 2015.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Towards Adoption of Technology-Enhanced Learning: Understanding Its Benefits and Limitations

Vilma Sukackė
Kaunas University of Technology
Kaunas, Lithuania
vilma.sukacke@ktu.lt

***Abstract*—Technology is embedded in all walks of people's lives, including the diverse forms and types of contemporary education. With technology becoming increasingly popular in a variety of contexts, scholars have developed models that allow to explore how technology-enhanced learning (TEL) is adopted. One of the key prerequisites for TEL adoption is its acceptance by the involved users. TEL acceptance can be hindered if technology is implemented without properly exploring its benefits and limitations. They should be also well understood by policy makers who offer top-down guidelines to employ TEL as a means of modernizing educational systems across the globe. Therefore, the present paper highlights the most prominent topics of the debates on implementing TEL when it comes to the two opposite ends of the spectrum, namely technology supporters on the one end and critics on the other.**

Keywords—Technology-enhanced learning, TEL, technology adoption

I. INTRODUCTION

Contemporary teaching and learning, regardless of its different forms (i.e. formal, informal, informal), can no longer be imagined without any use of technology. Technology, especially if it has been developed for educational purposes, is often seen as a means to promote and enable the modernization of educational systems. Both public and scientific discourse have a lot of different opinions about technological interventions in education. Those individuals who oppose using technology for educational purposes and their opponents, individuals who are in favor of technology, are often sarcastically labelled as "technophobes" [1] on the one end of the spectrum and "technoromantics" [2] or "technology enthusiasts" [1] on the other. The representatives of the two stances try to empirically ground their stance, which is achieved by conducting research in the area that is referred to as technology-enhanced learning (TEL). Over the past 50 years, research in the field has been increasingly conducted by representatives of educational sciences, informatics, and psychology.

There are many different terms in scientific literature for defining TEL, which at a first glance seem to be synonymous. However, there is a considerable discrepancy between their operationalization and realization. In scientific discourse, terms such as e-learning, m-learning, technology-based learning, computer-based learning, game-based learning, etc. are still popular and visible [3-5]. Nevertheless, there is a growing number of sources claiming that this area is best reflected by the term "technology-enhanced learning". Although it is not clearly defined in the scientific literature, it could be explained as a form of learning adapted to (1) be accessed in a number of ways by using technology, (2) organizing teaching/learning processes, (3) communicating, (4) collaborating, (5) making the learning process more effective, and (6) performing other type of learning-related tasks [5].

When criticizing the components of the term "technology-enhanced learning" and the lack of their clear operationalization in scientific literature, Bayne [4] emphasizes that the English version of the term does not reflect the word "teach" because technology is primarily used for teaching, not learning. As the author of the present paper supports the latter observation, after providing a brief overview of TEL adoption models, the paper further explores the negative and positive aspects of technology-enhanced teaching /learning by focusing on both, the educators and the learners.

II. ADOPTION OF TECHNOLOGICAL INNOVATIONS IN EDUCATION

As was previously indicated, with the hope to modernize education and increase students' learning outcomes, instructors across the globe implement technology. In many cases, when implementing technology for educational purposes, they engage in educational borrowing, which can be explained as a deliberate adoption of various learning objects (e.g. multimodal instructional content, curriculum, teaching/learning methods, Information and Communication Technology, etc.) from one context to [6]. Adoption of novel technology (or innovation) is a complex process that takes a long time. As can be seen in Fig. 1, there are different stages involved in the adoption process, and some innovations do not become mainstream. To study the process of innovation adoption, Rogers [7] also proposed a Diffusion of Innovation Model (see Fig. 2), which is still widely applied across different disciplines. It can be suggested that the initial stages of implementation are essential to the successful adoption of, in this case, digital learning objects.

It can be suggested that the key component that might lead to adoption of the aforementioned innovations is technology acceptance. One of the most commonly used models for measuring it is the Technology Acceptance Model (TAM), which was presented by Davis in the 1980s [8]. Research shows that, despite the widespread use of technology, even younger educators are reluctant to adopt and deploy technology, they often feel anxiety, and are skeptical about

the educational potential of technology. TAM research reveals similar trends in the adoption of learner technologies. Other additional factors, such as culture and foreign language [9-10] also make the adoption and implementation of technology particularly difficult, thus their lifecycle becomes brief and their true educational potential remains undisclosed.

In order to solve the aforementioned issues and to encourage more adoption of TEL, it is vital to better understand its benefits and limitations. The remaining of the paper discusses the main tendencies that emerge in scholarly debate over the former and the latter.

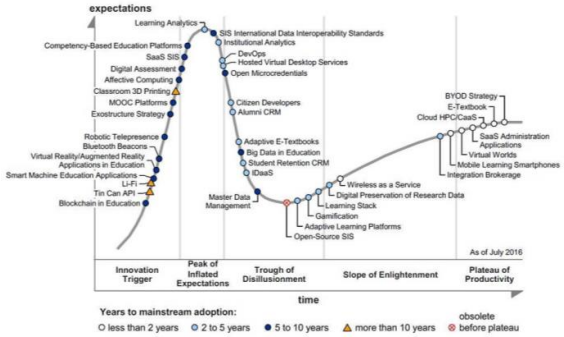


Fig.1. Gartner's Hype Cycle of Technology in Education [11]

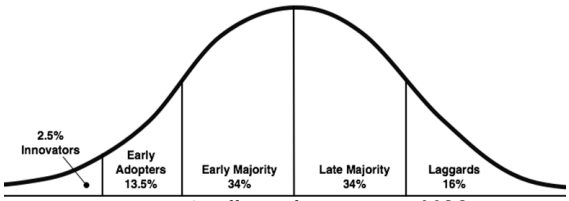


Fig. 2. Roger's Diffusion of Innovations Model [7]

III. LIMITATIONS OF TEL AS DISCUSSED IN SCIENTIFIC DISCOURSE

Although technology-enhanced teaching/learning in its definition suggests a number of benefits, in scientific discourse, there is considerable criticism towards it. Kohn and Hoeffsteder [12] illustrate one of the most common points of criticism when it comes to technology used in education by employing a metaphor of "the caravan effect". They suggest that the travelers (i.e. technology enthusiasts) stop for a short while (i.e., new technology), but travel quickly to find a new source. Other authors suggest that scholars might be trying to solve problems by using technology when technology itself is the root cause of the problems in the first place [13]. There is a lot of criticism for the past debates and studies on technology-enhanced teaching/learning because of their previous focus on technology over the student and his/her learning outcomes. It is more and more emphasized that technology should be a means to achieve the goal, not the goal itself [13]. Other authors note that the breakthrough of technology development is controlled by computer scientists and business needs. Therefore, introducing such technologies into the learning process, especially if they are unexplored by educational scientists, should be more carefully assessed [14].

New technologies that appear in educational contexts are often called disruptive [15]. This is primarily because they become innovations that present challenges for all those who are involved in the teaching/learning process. Despite their innovativeness, disruptive technologies often need to be adapted to specific contexts and needs [6, 16] . Such changes, respectively, require changes from the educators, learners or educational institutions as well.

Many a time, the decision to introduce new technologies in the teaching/learning process is brought top-down, for example, through educational policy decisions or guidelines. In practice, however, such pressure often does not produce any results [17]. Developing and introducing new technologies to educational settings is often costly [17], requires a lot of time and other resources from the individuals who implement them [18].

There is a global trend that the generation of educators, especially university teachers, is aging [19]. It can be assumed that many of them belong to the generation X, which, unlike the generation Y and Z (for the alternative names of the generations, see Fig. 3 below), does not have such good digital literacy skills. It can be argued that such educators also lack the ability to design contemporary courses that include technological interventions, as research overviews suggest that technology is very often used as a new medium to convey the same content [20]. For these reasons, educators without appropriate training have to obtain new qualifications and training, which is highly undesired because it also wastes a lot of time and financial resources. Learners are also reluctant to take on additional independent work. Technology can be implemented to successfully work under the learning paradigm and to direct the whole process at the learner, give him/her a lot of autonomy in deciding what, when, and how to learn. However, previous research reveals that learners are not yet fully prepared for self-directed learning through the use of technology [21].

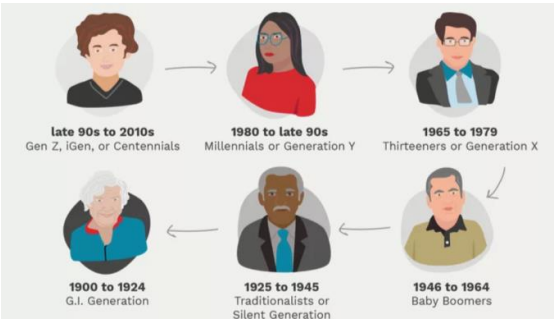


Fig. 3. Different titles of generations (Rosenberg, 2018)

Other challenges related to technology-enhanced teaching/learning can be observed as well. Due to the complexity of the field of technology-enhanced teaching/learning research, a large number of studies are fragmented, explore only certain aspects from the perspective of educational sciences, informatics or psychology. Another shortcoming is that these studies are often carried out in a very specific cultural context with a small, unrepresentative sample, and therefore the reported statistically significant

results should be interpreted with caution [11]. It is also worth noting that there is a lack of longitudinal research that reveals the true potential of technology-enhanced learning, maintaining knowledge and skills compared to the traditional learning methods. Interdisciplinary research that acknowledges and overcomes the aforementioned limitations could lead to more accurate results and deeper understanding of the phenomenon.

IV. BENEFITS OF TEL AS DISCUSSED IN SCIENTIFIC DISCOURSE

Educational sciences distinguish between three paradigms, namely, teaching (also, instruction), interaction, and learning [23]. In the first paradigm, all decisions related to learning content, methods, and other processes are decided upon by the educator. In the second paradigm, there is a closer link between the educator and the learner, and the latter is given more freedom to choose how to learn. The third paradigm gives the learner the most flexibility as s/he has the most freedom, can choose both the learning content and the methods; the role of the educator changes completely, s/he becomes the facilitator, mentor, and consultant of the learning process. Scholarly output of educational sciences emphasizes the need to move to the learning paradigm. This can be achieved with the help of technology.

Despite the criticism and skepticism surrounding the technology-enhanced learning potential, many scholars or education policy makers support such interventions in educational settings. Scientific literature, based on both empirical research and theoretical discussion, identifies a number of technological advantages for both educators and learners. For example, the introduction of new technologies is beneficial to educators' performance. Laurillard et al. [17] observes that new technologies extend the practice of educators and make them more professional.

Technology-enhanced teaching/learning can use a variety of teaching/learning methods and strategies, access multimodal content, or deliver content in an attractive, convenient, easy, and free-of-charge manner. There is a lot of research that reveals the benefits of blended or distance learning [24]. Universities and businesses are also developing Massive Open Online Courses (MOOCs) that are attractive because they are usually free, flexible in terms of time and space, inclusive, as well as suitable for both traditional and non-traditional students. It should be noted that even the most popular MOOCs (e.g. Fig. 4) suffer from high student dropout rates, but MOOC enthusiasts (especially researchers in open universities) continue to conduct research, improve the design of MOOCs to address this challenge. It can be anticipated that the advances in research will improve the performance and importance of MOOCs in life-long learning practices.

Fig. 4. Some of the most popular MOOCs [25]

It is also important to mention that TEL allows a learner to study in relatively authentic conditions, making learning more meaningful. For example, one of the most important aspects of learning a foreign language successfully is authenticity. Learners do not always have the opportunity to travel to the country where the language has the status of a state language, but with the help of technology, virtual simulations, tandems, access to corpora, etc. achieving authenticity is possible. TEL is well-suited to STEAM subjects and various fields of medicine, in which learners, through various simulations or virtual experiments, are enabled to understand and see intricate processes and to also avoid the risk to their own or others' lives. There is also a number of virtual reality tools through which the learners enhance, for example, their professional skills, by performing certain tasks physically [26].

Meaningful learning also takes place during personalized (as well as individualized) learning or collaborative learning. In modern education, neither one nor the other can be imagined without technology [11, 27]. A review of scientific literature reveals that learning which is personalized via technology support, allows the learner to control the pace of learning and track its progress through various systems that adapt to the learner's needs and progress, encourage to solve challenges, and provide feedback quickly [11]. The following benefits of collaborative TEL are also highlighted: better learning, longer retention of acquired knowledge, training of critical thinking, more accurate and creative problem solving, motivation, transfer of learning to other situations, etc. [28].

V. CONCLUSION

Changes in education are inseparable from socio-economic changes in society, driven by technological progress [29]. It can be said that any teaching/learning contexts are inseparable from Information and Communication Technology. There is a lot of criticism when it comes to technology-enhanced teaching/learning in both scientific and public discourse, which encourages all interested parties to reflect on how and why technologies are used for educational purposes. Past errors or limitations of research indicate that it is crucial to appropriately plan technological interventions, to reflect on all important variables, educational factors and goals, and only then to implement technology; more importantly, not to use technology just for the sake of using technology. Only then will it be possible to guarantee what Dror [30] calls qualitative and quantitative changes in teaching/learning.

REFERENCES

[1] J. F. Coget, "Technophobe vs. Techno-enthusiast: Does the Internet Help or Hinder the Balance Between Work and Home Life?", Academy of Management Perspectives, vol. 25, no. 1, pp. 95-96, 2011.

[2] N. Selwyn, "Editorial: In praise of pessimism—the need for negativity in educational technology," BJET, vol. 42, no. 5, pp. 713-718, 2011. https://doi.org/10.1111/j.1467-8535.2011.01215.x

[3] K. Beatty, Teaching & Researching: Computer-Assisted Language Learning. In R. C. Richey, Ed. Routledge Encyclopedia of Terminology for Educational Communications and Technology. New York, NY: Springer, 2010.

[4] S. Bayne, "What's the matter with 'Technology Enhanced Learning'?", Learning, Media and Technology, 2014. DOI: 10.1080/17439884.2014.915851.

[5] M. Teresevičienė, A. Volungevičienė, V. Žydžiūnaitė, L. Kaminskienė, A. Rutkienė, E. Trepulė, and S. Daukilas, Technologijomis grindžiamas mokymas ir mokymasis organizacijose. Vytauto Didžiojo universitetas: Versus Aureus, 2015.

[6] M. H, Romanowski, H. Alkhateeb, and R. Nasser, "Policy borrowing in the gulf cooperation council countries: Cultural scripts and epistemological conflicts," International Journal of Educational Development, vol. 60, pp. 19-24, 2018. DOI: 10.1016/j.ijedudev.2017.10.021

[7] E. M. Rogers, Diffusion of innovations. Simon and Schuster, 2010.

[8] F. Davis, A technology acceptance model for empirically testing new-end-user information systems: Theory and results. Massachusetts, United States: Sloan School of Management, Massachusetts Institute of Technology, 1986.

[9] C. White, "Distance learning of foreign languages," Language Teaching, vol. 39, no. 4, pp. 247-264, 2006.

[10] S. Martin, and I. M. A. Valdivia, "Students' feedback beliefs and anxiety in online foreign language oral tasks," International Journal of Educational Technology in Higher Education, vol. 14, no. 1, p. 18, 2017.

[11] J. S. Groff, "Personalized Learning: The State of the Field & Future Directions," 2017. Available: https://curriculumredesign.org/wp-content/uploads/PersonalizedLearning_CCR_May2017.pdf.

[12] K. Kohn, and P. Hoffstaedter, "Authenticated language learning with do-it-yourself corpora," 13th International CALL Conference, Antwerp, Belgium, 2008.

[13] N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, and S. Barnes, Eds., Technology-Enhanced Learning: Principles and Products. Springer, 2009.

[14] L. Castañeda, and N. Selwyn, "More than tools? Making sense of the ongoing digitizations of higher education," International Journal of Educational Technoly in Higher Education, vol. 15, no. 22, 2018. https://doi.org/10.1186/s41239-018-0109-y

[15] G. Conole, "MOOCs as disruptive technologies: strategies for enhancing the learner experience and quality of MOOCs," RED: Revista de Educacion a Distancia, vol. 50, 1-18, 2016.

[16] B. Janiūnaitė, Edukacinės novacijos ir jų diegimas. Kaunas: Technologija, 2004.

[17] D. Laurillard, M. Oliver, B. Wasson, and U. Hoppe, "Implementing technology-enhanced learning," In Technology-Enhanced Learning, pp. 289-306. Springer, Dordrecht, 2009.

[18] U. C. Okonkwo, "Computer assisted language learning (CALL) software: Evaluation of its influence in a language learning process," UJAH: Unizik Journal of Arts and Humanities, vol. 12, no. 1, pp. 76-89, 2011.

[19] T. Oshagbemi, "The impact of age on the job satisfaction of university teachers," Research in Education, vol. 59, no. 1, pp. 95-108, 1998.

[20] A. Kirkwood, and L. Price, "Technology-enhanced learning and teaching in higher education: what is 'enhanced'and how do we know? A critical literature review," Learning, Media and Technology, vol. 39, no. 1, pp. 6-36, 2014.

[21] V. Morkūnienė, "Studentų mokymasis ir jo vertinimas: mokymosi paradigmos atvejis," 2005. Available: http://alytauskolegija.lt/wp-content/uploads/straipsniai/Morkuniene.pdf

[22] M. Rosenberg, "Generational Names in the United States: Gen X, Millennials, and Other Generation Names Throughout the Years," 2018. Availabe: https://www.thoughtco.com/names-of-generations-1435472

[23] P. Jucevičienė, D. Gudaitytė, V. Karenauskaitė, D. Lipinskienė, B Stanikūnienė, and G. Tautkevičienė, Universiteto edukacinė galia: atsakas 21-ojo amžiaus iššūkiams: monografija. Kaunas: Technologija, 2010.

[24] D. Rutkauskienė, O. Suk, and D. Gudonienė, Eds., ICT enhanced learning: monograph. Kharkiv: Planeta print, 2017.

[25] K. Pearce, "Ultimate guide to Moocs: Take online courses from elite universities," 2013. Available: https://www.diygenius.com/the-ultimate-guide-to-moocs/

[26] E. Staurset, and E. Prasolova-Førland, "Creating a Smart Virtual Reality Simulator for Sports Training and Education," Smart Innovation, Systems and Technologies, vol. 59, 2016.

[27] A. Zmuda, D. Ullman, and G. Curtis, Learning personalized: The evolution of the contemporary classroom. John Wiley & Sons, 2015.

[28] H. Minagawa, "Fellow Language Learners as Producers of Knowledge and Understandings: A Case of a Tertiary Japanese Linguistics Course," Journal of Peer Learning, vol. 10, no. 4, pp. 41-58, 2017.

[29] T. Newell, Five Paradigms for Education: Foundational Views and Key Issues. Springer, 2014.

[30] I. E. Dror, "Technology enhanced learning: The good, the bad, and the ugly," Pragmatics & Cognition, vol. 16, no. 2, pp. 215-223, 2008.

INFORMACINËS TECHNOLOGIJOS • IVUS 2019 • ISSN 2029-249X • eISSN 2029-4824

Numerical analysis of SLSSIM similarity on medical X-ray image domain

Jonas Brusokas, Linas Petkevičius
Institute of Computer Science
Vilnius University
Vilnius, Lithuania
jonas.brusokas@mif.vu.lt, linas.petkevicius@mif.vu.lt

***Abstract*—The X-ray has been adopted and used for various purposes including medical diagnostics. To remove noise created by new low dose X-ray imaging procedures and reduce medical image size, X-ray image reconstruction and lossless compression using deep neural networks are being researched. To enable this, image similarity metrics capable of performing well on X-ray images must be used. In this paper, the requirements for medical X-ray similarity metrics are defined. A new similarity metric is proposed taking to account the quality of structures within different intensity levels. An analysis is given comparing the proposed and other currently known metrics performance on real X-ray images in simulated scenarios.**

***Index Terms*—Image similarity metrics, Low dose X-ray imaging, Medical X-ray images**

I. Introduction

The X-ray ever since its inception has been widely adopted and used for various purposes. One of the key uses of X-ray imaging is in medical diagnostics. It allows for a non-invasive method of diagnosing various bone structure defects, infections, arthritis, and most cancers [1]–[3].

Despite wide usage and rapid technological advances the ionizing radiation emitted during X-ray imaging procedures creates real health risks for patients [4]–[6]. Efforts have been made to reduce the risks associated with radiation exposure by creating new procedures for performing X-ray imaging. These procedures (commonly referred to as low-dose) use lower voltage or amperage settings to reduce the amount of radiation emitted thus reducing the risks [7]. Unfortunately, using these types of procedures results in images having artifacts and noises which can reduce diagnostic suitability [8]. There are cases where it might not be necessary to remove the noises in order to give an accurate diagnosis [9]. In most cases, however, it is imperative to remove or reduce the amount of noise on an X-ray image. Conventional approaches that use defined properties of noise distribution do not yield satisfactory results [10], [11].

There have been some successful attempts in using deep neural networks for image reconstruction tasks, outperforming other approaches in removing various noises from images [12]–[14]. There have also been attempts in using deep neural networks for lossless image compression, to make storing images and using them for calculation more efficient [15]. Furthermore, a significant amount of research has recently been conducted on finding ways of improving X-ray imaging procedures, especially on automated means of disease or abnormality detection [2], [3], [16].

One of the key factors to the accuracy and effectiveness of a deep neural network is the objective function. In image reconstruction and compression image similarity metrics are used as objective functions. They compare two images with each other and produce a scalar result denoting their degree of difference [17]. The similarity metric defines what image properties are being evaluated so selection of an effective metric is crucial.

In this paper a similarity metric for medical X-ray images domain is proposed and compared with other known metrics. In section II essential properties of X-ray images in the medical domain and requirements for a similarity metric are defined. An overview of currently used metrics is made in section III and the definition of the proposed metric is made in section IV. The analysis comparing performance of the metrics is made in section V.

II. Requirements for medical X-ray image similarity metrics

Medical X-ray images are created and used to enable trained radiologists to analyse the human body, determine and diagnose various irregularities and illnesses. X-ray images can be done over any part of the human body and as such contain various structures including bones, tissue, and organs [1], [2]. Any assessment of X-ray image similarity or quality must take into account properties of the images which enable them to be used for accurate and reliable diagnosis.

Through analysis of literature and working with trained specialists in the field of radiography several important requirements for medical X-ray image similarity metrics have been identified:

- R1 The metric must detect emerging noises and artifacts from images created with low dose X-ray procedures. As was previously stated, in most cases to enable low dose procedures, it is important to detect and remove noises from the created images as they hinder diagnosis [8], [13].
- R2 The metric must detect perceptual geometric distortions. They can appear on images during imaging, reconstruction or decompression. From a radiologists perspective

the distortions may cause difficulties or even render the images completely unusable for diagnostics [18].

R3 The metric must take into account the discrepancies of images within specific ranges of brightness intensity levels. Due to the nature of the X-ray, images contain various human body structures including bones, tissue, organs. They are captured as areas of different brightness intensities [1]–[3]. In this paper we will refer to them as sub-levels. Most image quality assessment approaches attempt to evaluate the broader perceptual view of image quality and focus on more visible areas. In medical X-ray images, all structures captured in these images have significant value and greatly contribute to diagnosis [13]. Taking into account reconstruction or decompression accuracy of not only the whole picture but also structures within different intensity ranges is critical.

The proposed metric (in section IV) and following experiments (in section V) will take these properties into account.

III. Selected metrics

Research into image similarity metrics (in some cases referred to as image quality assessment) has witnessed attention and notable progress over the past decades [19]. Image quality assessment is essential in fields that use image processing [20]. The majority of metrics used evaluate the similarity of the distorted image use the complete reference image [19]. In this paper 3 selected general purpose similarity metrics which are potentially usable in the X-ray image domain will be discussed and compared. It is important to note, that for the rest of the paper metric definitions will be used, where metric value 1 means the compared images are equal, and value 0, means they are completely dissimilar.

A. Mean squared error (MSE)

Mean squared error is considered to be one of the most simple and straight-forward similarity metrics. It is computed by averaging the intensity differences of the two compared images [17]. MSE is known for being quite mathematically convenient for optimization purposes. Although the metric does not correlate well with perceived visual quality [21]. Despite its flaws it still remains widely used for many image processing tasks. For the purposes of this paper, this metrics' performance will be compared to others during analysis. The following normalized form of the metric will be used:

$$MSE^*(X,Y) = 1 - \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \ / \ \ M^2 \qquad (1)$$

Here X and Y are compared images, M - maximal pixel intensity, X_i and Y_i - a concrete pixel in the image, n - total number of pixels in image.

B. Structural similarity index (SSIM)

As a response to unsatisfactory perceived visual quality results of MSE and similar metrics a new approach was proposed to construct new similarity metrics. It was based on the idea that images are highly structured and rather than evaluating pixels individually (like in MSE), groups of spatially proximate and related pixels should be taken into account together in order to achieve more accurate representation of perceivable quality [17]. This is based on the assumption that the human visual system is adapted to identify various structures within field of sight and effectively spot any distortions affecting them [17], [22]. A well known metric based on these principles is the Structural Similarity Index (SSIM) [23]. SSIM metric calculates in account three key features of the images: luminance changes, contrast changes and structural changes [17]. In this paper it will be defined and used in its typical form:

$$\text{SSIM}^*(X,Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)} \qquad (2)$$

SSIM has seen wide usage for many image processing tasks [23]. It has served as inspiration and basis for many other similarity metrics supplementing additional features to the standard metric [24]–[28].

C. Weighted SSIM (wSSIM)

Weighted SSIM is a general purpose modification of the SSIM metric mentioned in sub-section III-B. It was created to be used as an objective function in Deep neural networks. It is a composition of SSIM and traditional L_1 loss. Intuitively, SSIM gives the perceptual image assessment and L_1 decreases the metric value for more distorted images and increases for less distorted images [29]. The authors claimed that if used for training, it would put more emphasis on pairs which are performing worse and increase training speed and accuracy. In this paper L_1 will be defined in its normalized form and wSSIM will be defined in the following form:

$$L_1^*(X,Y) = 1 - |X - Y|_{L_1} \ / \ \ (M \cdot N) \qquad (3)$$

$$wSSIM^*(X,Y) = 1 - SSIM^* \cdot (1 - L_1^*) \qquad (4)$$

Here X and Y are compared images, N - number of pixels in the images, M - maximal pixel intensity.

IV. Proposed SLSSIM metric

Substantial research regarding using Deep neural networks as a method for medical image reconstruction and processing is very recent [30], [31]. Although a significant amount of metrics have been developed to suit different domains, there is no definitive similarity metric for comparing medical X-ray images.

As mentioned in section II a metric for medical X-ray image similarity assessment must take into account specific properties of the images. Also, it has become clear from image quality measurement research that handcrafting and tailoring a similarity metric by combining several known metrics can yield significantly better performance than using any single known metric [32], [33].

In this paper a similarity metric for assessing medical X-ray image similarity is proposed. It is called the sub-level structural similarity index (or SLSSIM). As its name implies it is based on the widely used structural similarity index (or

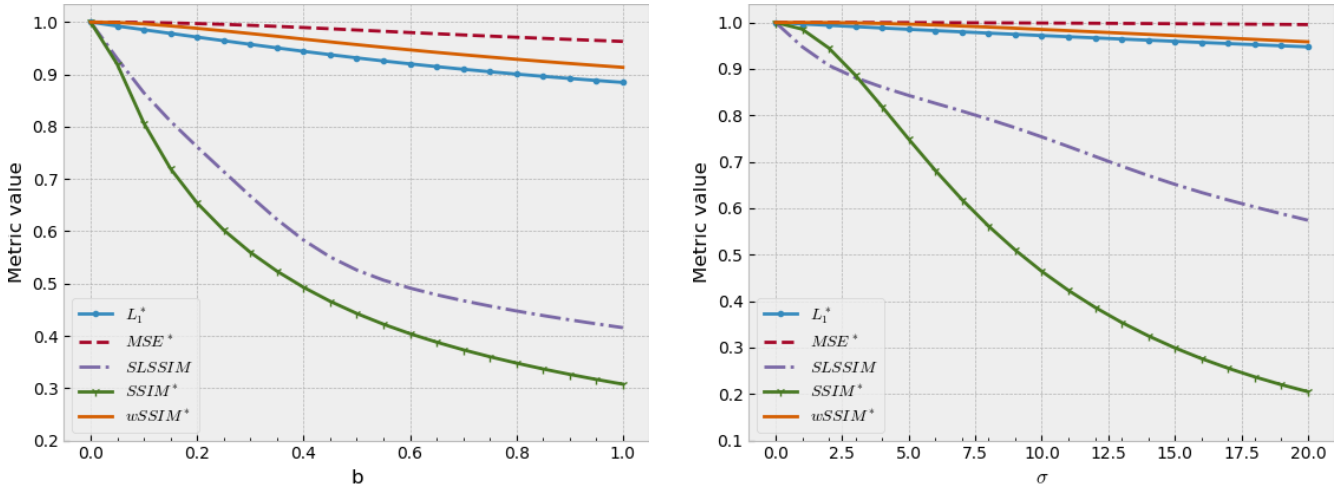


Fig. 1. Results with Poisson Gauss noise augmentation (left), additive Gauss noise augmentation (right)

SSIM). Its main differentiating feature is the consideration of human body structures appearing within different sub-levels of the image (as mentioned in requirement R3). It is achieved by calculating SSIM values on discrete sub-levels and not the entire image. This should enable more strict comparison of the structures appearing within the sub-levels. In this paper sub-levels are mathematically defined as:

$$X^{(i)}, Y^{(i)} \in \left[\frac{iM'}{k}, \frac{(i+1)M'}{k}\right) \tag{5}$$

Here X and Y are images, $X^{(i)}, Y^{(i)}$ are sub-levels of the images, N - number of pixels in image, M' - amount of different values that a pixel can have, k - is number of compared discrete sub-levels. The mathematical definition of SLSSIM is:

$$SLSSIM = \sqrt[k+1]{(1 - L_1^*) \cdot \prod_{i=0}^{k} (SSIM^*(X^{(i)}, Y^{(i)}))} \tag{6}$$

SLSSIM is a root of a product where the members are normalized L_1 calculated from the entire image and SSIM calculated from defined sub-levels. Here, L_1 is used the same way as in subsection III-C, to decrease the metric value further if the image is more distorted and. Also, as all of the product members have values in interval $[0; 1]$ the root is used to prevent a steep downward gradient of the metric. We will be using 8 sub-levels ($k = 8$) for metrics performance analysis.

V. Metrics Analysis

Analysis has been carried out to evaluate performance of the selected metrics in real-life scenarios. Tests were conducted using real X-ray images from openly available medical X-ray image data sets. The data sets used for analysis were:

1) RSNA pediatric image set created by the Radiological society of North America, which contains hand X-ray images [34];
2) NIH image set created by National Institute of Health (of the United States of America) which contains chest X-ray images [35];
3) MURA musculoskeletal image set created by Stanford University which contains elbow, forearm, hand and shoulder X-ray images [36];

For the analysis 3000 images were selected from the data sets (1000 from each) in a random order. All medical images within the data sets were gray-scale, with pixel value ranging within the interval $[0; 255]$.

Image augmentations have been used to simulate noises and distortions in the X-ray images to simulate real-life medical X-ray imaging scenarios. To analyse performance values of metrics were calculated on image pairs that contained reference image – unmodified image from a data set, and augmented image – created from reference image by applying an augmentation. Similarity between the two images was evaluated using the selected and proposed similarity metrics from sections III and IV. The results were aggregated by taking the mean metric value from all the images for each augmentation level.

A. Noise detection results

As defined by requirement R1 in section II a medical X-ray image similarity metric must detect emerging noises. Two image augmentations simulating real life scenarios have been selected to generate augmented images for testing:

- Poisson-Gauss noise augmentation $\mathcal{PG}(a, b)$ – represents artifacts and noises emerging in images obtained via computed tomography X-ray imaging when using X-ray sensors in low dosage configuration [8]. The augmented image was generated using parameters: $a = 31$ and $b = 0.05k, k \in [0; 20], k \in \mathbb{Z}$. Higher b values result in more noise in the image. There are confirmed results stating that noises generated with parameter $b \approx 0.1$ can occur in low dose imaging [8].

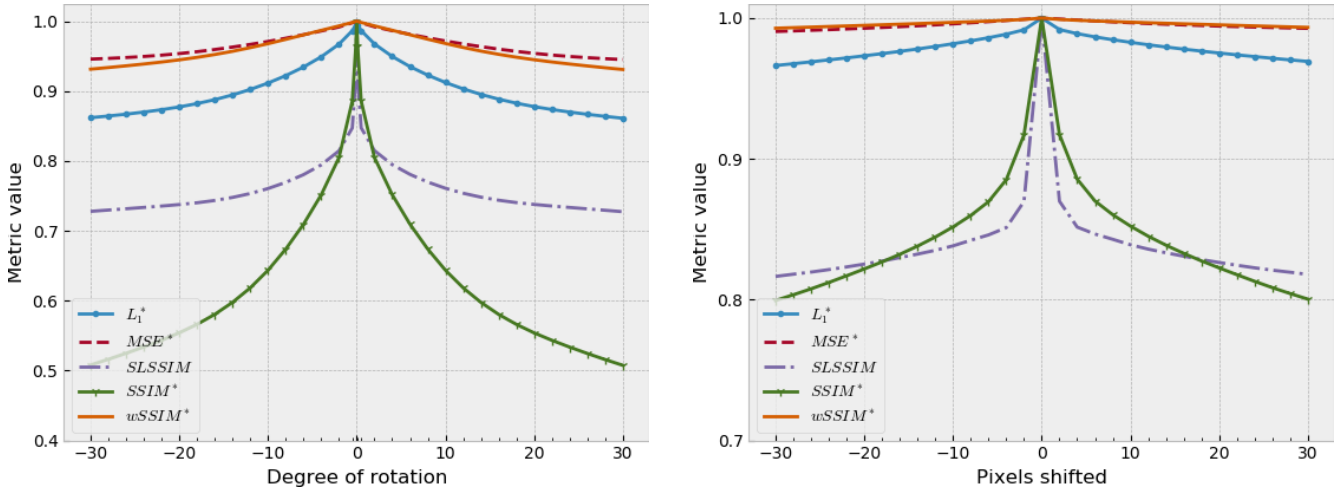


Fig. 2. Results with rotation augmentation (left), Y-axis translation augmentation (right)

- Additive Gauss noise augmentation $\mathcal{N}(\mu, \sigma)$ – represents general noises and artifacts that can appear in X-ray images. Some research points to its applicability in simulating noises emerging from low dose imaging as well [37]. The augmented image was generated using parameters: $\mu = 0, \sigma \in [0; 20], \sigma \in \mathbb{Z}$. Higher σ values result in more noise in the image. Visually, the noise becomes noticeable at $\sigma \approx 5$.

1) Poisson Gauss noise detection results: On the left side of figure 1 similarity metrics ability to detect emerging Poisson-Gauss noise is observed. Metrics L_1^*, MSE^*, $wSSIM^*$ do not detect the noise even when the noise parameter b is high (images are highly disrupted). Whereas, $SSIM^*$ metric and the proposed metric $SLSSIM$ perform well. When noise parameter $b = 0.1$, metric values $SLSSIM \approx 0.78$ and $SSIM^* \approx 0.66$. These results are acceptable, as visually the noise should still allow for accurate diagnosis.

2) Additive Gauss noise detection results: On the right side of figure 1 similarity metrics performance against emerging additive Gauss noise is witnessed. Similarly to the previous results, metrics L_1^*, MSE^*, $wSSIM^*$ do not detect the noise well, the resulting values do not drop below 0.94. However, in this case the difference between the proposed $SLSSIM$ metric and $SSIM^*$ is higher. When noise parameter $\sigma = 5$, metric values $SLSSIM \approx 0.84$ and $SSIM^* \approx 0.76$. Although these values are still acceptable, the rate at which $SLSSIM$ value decreases as the noise level rises is not as representative of the added disruptions as the $SSIM^*$. It can be stated that $SSIM$ is slightly more accurate to detecting additive Gauss noise.

B. Geometric distortion detection results

As defined by requirement R2 in section II a medical X-ray image similarity metric must also detect perceptual geometric distortions. Two widely used image augmentations have been selected to simulate geometric transformations and generate augmented images for testing:

- Translation augmentation – used to evaluate general metric robustness against translation (shifting of the image). Theoretically, translation can occur during reconstruction or decompression. The augmented image was generated by translating the reference image on the Y axis by y pixels, where values $y = \{-30, -28, ..., 26, 28, 30\}$
- Rotation augmentation – also used to evaluate general metric robustness. The augmented image was generated by rotating the reference image by r degrees, where values $r = \{-30, -28, ..., 26, 28, 30\}$

1) Rotation augmentation detection results: On the left side of figure 2 similarity metrics ability to detect rotations is observed. As an emerging pattern, metrics L_1^*, MSE^*, $wSSIM^*$ do not detect the distortions even when the degree of rotation is very high. $SSIM^*$ and $SLSSIM$ still perform well, with $SSIM^*$ once again being slightly more sensitive to the level of distortion.

2) Translation augmentation detection results: On the right side of figure 1 similarity metrics performance in detecting translations is witnessed. Once again, metrics L_1^*, MSE^*, $wSSIM^*$ do not detect distortions well and metrics $SSIM^*$, $SLSSIM*$ do so.

C. Analysis results

As observed in the tests L_1^*, MSE^*, $wSSIM^*$ metrics are not suitable for accurately detecting noises or geometric transformations in medical X-ray images. While tests show that $SSIM^*$ is more sensitive to additive Gauss noise, but $SLSSIM$ has an advantage in being the only metric that specifically measures discrepancies in sub-levels of the image. Both $SSIM^*$ and $SLSSIM$ metrics are potentially suitable for use with medical X-ray images.

VI. Samples of metrics behaviour on X-Ray images

Sample images from the data sets have been displayed to better visualise the impact that augmentations have on the images. Also, metric values are given for each image pair.

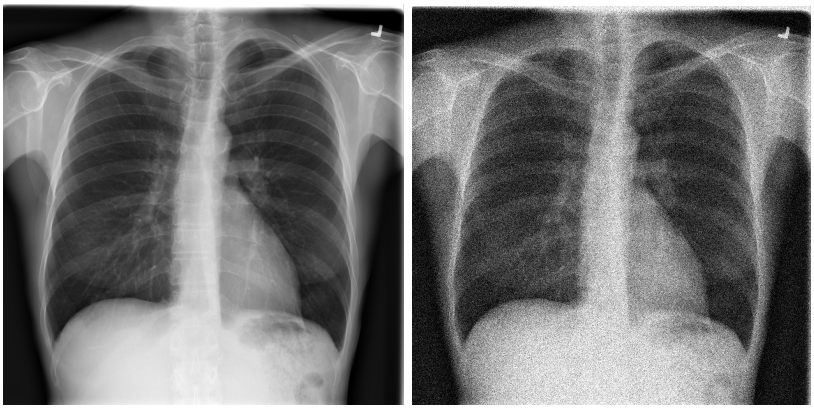


Fig. 3. Sample chest X-ray image from NIH data set (left), same image distorted with Poisson Gauss noise (right)

Fig. 4. Sample wrist X-ray image from MURA data set (left), image distorted with Poisson Gauss noise (center), visualisation of sub-level range $[64; 127]$ of distorted image (right)

In figure 3 a reference X-Ray image from NIH chest X-ray image data set can be seen on the left. An augmented image distorted by Poisson-Gaussian noise, $a = 31, b = 0.08$ can be seen on the right. It simulates the effect that low dose imaging would have on the image [8]. The metric values calculated from reference and augmented images are: $MSE^* = 0.99$, $L_1^* = 0.93$, $SSIM^* = 0.10$, $wSSIM^* = 0.93$ and $SLSSIM = 0.36$.

Similarly, in figure 4 a reference X-Ray image from the MURA musculoskeletal X-ray image data set can be seen on the left. An augmented image distorted by Poisson-Gaussian noise, with same parameters as before can be seen in the center. The image on the right represents a visualisation of the augmented image sub-level of intensity range $[64; 127]$. In this particular case, this sub-level represents the bone structure from the image, with tissue and other matter being barely represented. The metric values calculated from reference and augmented images are: $MSE^* = 0.98$, $L_1^* = 0.88$, $SSIM^* = 0.08$, $wSSIM^* = 0.89$ and $SLSSIM = 0.27$. It is important to note that the metrics values taken from images seen in figures 3 and 4 correspond with the results of the analysis in section V.

VII. Conclusions

Although there are many general purpose image similarity metrics used in various domains, only a few of them are suitable for use in comparing medical X-ray images. Based on the analysis performed on multiple medical X-ray image data sets and using different augmentations, L_1^*, MSE^*, $wSSIM^*$ metrics are ineffective in this domain as metric values never drop below 0.8 even when significant distortions are applied. $SSIM^*$ and the proposed $SLSSIM$ can be used to effectively compare X-ray images as both metrics' value is much lower when comparing heavily distorted images, in some cases going below 0.3. Both metrics perform similarly in tested cases. Unlike $SSIM^*$, $SLSSIM$ takes into account differences between human body structures in the images, which should allow diagnosticians to make more accurate diagnoses. The proposed $SLSSIM$ can be used as a basis of an objective function in medical X-ray image reconstruction or compression tasks.

REFERENCES

[1] U. Eglė, "Diagnostic X-ray hardware," 2011, (In lithuanian).

[2] M. Woźniak and D. Połap, "Bio-inspired methods modeled for respiratory disease detection from medical images," Swarm Evol. Comput., vol. 41, pp. 69–96, Aug. 2018.

[3] M. Woźniak, D. Połap, G. Capizzi, G. Lo Sciuto, L. Kośmider, and K. Frankiewicz, "Small lung nodules detection based on local variance analysis and probabilistic neural network," Comput. Methods Programs Biomed., vol. 161, pp. 173–180, Jul. 2018.

[4] B. V. Daga, V. R. Shah, and S. V. Daga, Radiodiagnosis, nuclear medicine, radiotherapy and radiation oncology. Jaypee Brothers Medical Publishers, 2013.

[5] A. Berrington de González et al., "Projected Cancer Risks From Computed Tomographic Scans Performed in the United States in 2007," Arch. Intern. Med., vol. 169, no. 22, p. 2071, Dec. 2009.

[6] Y. Y. Kim, H. J. Shin, M. J. Kim, and M.-J. Lee, "Comparison of effective radiation doses from X-ray, CT, and PET/CT in pediatric patients with neuroblastoma using a dose monitoring program.," Diagn. Interv. Radiol., vol. 22, no. 4, pp. 390–4, 2016.

[7] S. P. Raman, M. Mahesh, R. V. Blasko, and E. K. Fishman, "CT scan parameters and radiation dose: Practical advice for radiologists," J. Am. Coll. Radiol., vol. 10, no. 11, pp. 840–846, 2013.

[8] S. Lee, M. S. Lee, and M. G. Kang, "Poisson–Gaussian Noise Analysis and Estimation for Low-Dose X-ray Images in the NSCT Domain," Sensors, vol. 18, no. 4, p. 1019, Mar. 2018.

[9] A. Neverauskiene et al., "Image based simulation of the low dose computed tomography images suggests 13 mAs 120 kV suitability for non-syndromic craniosynostosis diagnosis without iterative reconstruction algorithms," Eur. J. Radiol., vol. 105, pp. 168–174, Aug. 2018.

[10] Jiang Hsieh, Computed Tomography: Principles, Design, Artifacts, and Recent Advances, Second Edition, 2009.

[11] L. L. Geyer et al., "State of the Art: Iterative CT Reconstruction Techniques," Radiology, vol. 276, no. 2, pp. 339–357, Aug. 2015.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," J. Mach. Learn. Res., vol. 11, pp. 3371–3408, 2010.

[13] D. Wu, K. Kim, G. El Fakhri, and Q. Li, "A Cascaded Convolutional Neural Network for X-ray Low-dose CT Image Denoising," May 2017.

[14] J. Xie, L. Xu, and E. Chen, "Image Denoising and Inpainting with Deep Neural Networks," pp. 341–349, 2012.

[15] G. Toderici et al., "Full Resolution Image Compression with Recurrent Neural Networks," Aug. 2016.

[16] Q. Ke et al., "A neuro-heuristic approach for recognition of lung diseases from X-ray images," Expert Syst. Appl., vol. 126, pp. 218–232, Jul. 2019.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, E. P. Simoncelli, and S. Member, "Image Quality Assessment: From Error Visibility to Structural Similarity," vol. 13, no. 4, pp. 1–14, 2004.

[18] T. Černius, "Image reconstruction of x-ray images using deep learning," Vilnius University, 2018, (In lithuanian).

[19] K. Xu, X. Liu, H. Cai, and Z. Gao, "Full-reference image quality assessment-based B-mode ultrasound image similarity measure," Jan. 2017.

[20] D. M. Chandler, "Seven Challenges in Image Quality Assessment: Past, Present, and Future Research," ISRN Signal Process., vol. 2013, pp. 1–53, 2013.

[21] A. B. Watson and Bernd, Digital images and human vision. MIT Press, 1993.

[22] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?," in IEEE International Conference on Acoustics Speech and Signal Processing, 2002, p. IV-3313-IV-3316.

[23] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," J. Vis. Commun. Image Represent., vol. 22, no. 4, pp. 297–312, May 2011.

[24] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," Neurocomputing, vol. 257, pp. 104–114, Sep. 2017.

[25] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar Wavelet-Based Perceptual Similarity Index for Image Quality Assessment," Jul. 2016.

[26] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in 2012 19th IEEE International Conference on Image Processing, 2012, pp. 1473–1476.

[27] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," IEEE Trans. Image Process., vol. 23, no. 2, pp. 684–695, Feb. 2014.

[28] Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," IEEE Trans. Image Process., vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[29] N. Johnston et al., "Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks," Mar. 2017.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015.

[31] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image Reconstruction is a New Frontier of Machine Learning," IEEE Trans. Med. Imaging, vol. 37, no. 6, pp. 1289–1296, Jun. 2018.

[32] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," IEEE Trans. Image Process., vol. 27, no. 1, pp. 206–219, Jan. 2018.

[33] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," 2015, p. 93940K.

[34] "RSNA Bone Age data set from Kaggle." [Online]. Available: https://www.kaggle.com/kmader/rsna-bone-age. [Accessed: 01-Mar-2019].

[35] "NIH Chest X-rays data set from Kaggle." [Online]. Available: https://www.kaggle.com/nih-chest-xrays/data. [Accessed: 01-Mar-2019].

[36] "Stanford University MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs." [Online]. Available: https://stanfordmlgroup.github.io/competitions/mura/. [Accessed: 01-Mar-2019].

[37] Hanying Li, T. Toth, S. McOlash, Jiang Hsieh, and N. Bromberg, "Simulating low dose CT scans by noise addition," in 2002 IEEE Nuclear Science Symposium Conference Record, 2002, vol. 3, pp. 1832–1834.

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Media Analysis that Reflects the Spread of anti-Christian Opinion

Monika Briedienė
Vytautas Magnus University
Lithuania
monika.briediene@vdu.lt

Valdas Kilpys
Vytautas Magnus University
Lithuania
valdas.kilpys@gmail.com

Tomas Krilavičius
Vytautas Magnus University,
Baltic Institute of Advanced Technology
Lithuania
t.krilavicius@bpti.lt

***Abstract*— This paper presents an interdisciplinary research on the dissemination of anti-Christian information in Lithuanian media (period: 2016.01.01-2019.01.01). The object of the research is the main directions of expression of anti-Christian verbal aggression. With this study, we analyze the fundamental problem: how to detect and evaluated verbal aggression directed against Christians beliefs. There are two areas of research in this place: computerized data collection, processing and expert of theology evaluation of data and results. Such a study has not been conducted in Lithuanian media so far, it has only a few matches in general as well. The results allowed us to detect a portal and a topic that contained the most aggressive texts.**

Keywords*— *verbal aggression, anti-Christian propaganda, media framing, Lithuanian language media

I. INTRODUCTION

The spread of today's information space has become a major challenge not only in terms of mutual communication. Due to the ever-increasing number of electronic texts, various natural language processing works are becoming increasingly important. Virtual reality, as equivalent to unseen existence, slowly, but surely intrudes into human consciousness and changes a human being and his whole value system, including faith matters. That is why this research becomes very important to Christians community. The problem of work is reflected in the fact that information passed through verbal aggression against faith has become inevitable in the present world. Therefore, having an aspiration of faith, which is inevitably linked to the aspect of personality, the believer may also encounter a rather unfriendly information space, which has an impact on faith and is the object of exploration of this article in the most general sense.

Many researches [1], [2], [3] underline the influence of media aggression on the formation of personal opinion and human behavior. In this context, studies are necessary, on the basis of which the main forms of aggression that arise from virtuality and are directed against faith could be identified. The scheme for achieving this goal is illustrated below (see Figure 1). However, the performance of studies is inseparable from the necessity to identify "reference points". After structured interviews with experts (respondents), topics (areas) affected by verbal aggression are in this order: (1) Ethical aspects of artificial insemination, (2) The ethics of priests and monks (property issues, celibacy, etc.), (3) Influence of the Church on political state processes. The analysis of the resulting material will enable systematic coverage of the totality of approaches that exist in specific information resources and, after its operationalization, purify the resistance points for the part of empirical research.

The research of the main trends of verbal aggression by impact on the congregation permits to make the following conclusion: verbal aggression is generally expressed indirectly by describing a "hero" by aggressive text author. Usually a churchman becomes a target. According to the areas that face verbal aggression, the relevant keywords were formulated: 2-3 keywords have been assigned to each topic.

The ultimate goal of this research can be achieved by performing the following intermediate tasks: (1) a related work analysis (see Section II), (2) qualitative research (see Section III), (3) a construction of the representative corpus containing Lithuanian texts (see Section IV), (4) an analytical selection of the most promising methods, evaluation of the obtained results (see Section V) (5) conclusions (recommendations) and our further research plans (see Section VI).

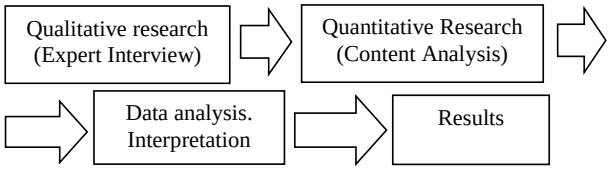


Fig. 1 Visualization of projected directions of analysis

II. RELATED WORKS

Due to the interdisciplinarity of work, any similar study was not identified. However, the basic idea of the work is to find texts in online media that form the reader's opinion. The importance of this topic is justified by several authors [4], [5]. The relevance and impact of the study is justified by the article [6]: "Our findings illuminate how variation in media attentiveness conjoin to determine whether and the degree to which no fundamentalists feel antagonistically toward Christian fundamentalists." Other study [7] on Christian rights talks about morality and identity politics. It becomes clear that the question of the influence of the Church is very important for the state. Under consideration is question how much the Church influenced legislation.

In our study we focus on verbal aggression in Lithuanian media against Christian beliefs. Aggression in the media topic [2], [3] is important for many researches. Some study [8] provides a meta-analysis review of the experimental effects of media violence on viewers' aggression in unstructured social interaction.

Annotated data are required for qualitative media analysis. Text mining methods are specify by several authors in book [9]. According to our research, we chose to harvest the web. Selected tool simply turns websites into structured data.

III. QUALITATIVE RESEARCH

Expert interviews (qualitative research) are the first stage of empirical research. Its main task is to identify the most vivid

forms of verbal aggression and fields by using the data obtained. In subsequent chapters (with specific problem fields that are most attacked in new media) discuss in detail and contextually the frequency, strength, and contextuality of aggression in the new media.

The essential statement that emerged from the first part of the empirical study is that none of the 9 experts (respondents) claimed that the Lithuanian-language news portals dominate the positive attitude towards the Church, faith, and believers.

The focus of this part of the empirical research is on the topics/areas most affected by verbal aggression in the new media. According to the results, after the semi-structured interviews with experts/respondents, the topics/areas affected by verbal aggression were as follows:

a) Ethical aspects of artificial insemination;

b) The ethics of priests/monks (property issues, celibacy, etc.);

c) Influence of the Church on political state processes.

It was found that the first and second for most of the experts were equally important and did not lead to more ambiguity. However, the interlocutors could not clearly agree on the third item and the topic "Influence of the Church on Political State Processes" in the field of research was only a marginal gain

IV. DATA SET

We have performed data analysis of three biggest and most popular (according to total number of real users [10]) online media sources in Lithuania: Delfi (1)[1], Lrytas (2)[2], 15min.lt (3)[3] (selected period: 2016.01.01-2019.01.01). Each case required different approaches and tools, so in this study we use OutWit Hub [4] for finding, collecting (scrapping) and organizing all necessary data. OutWit Hub [11] dissects web pages into their different elements (contacts, links, images, data, news, etc.). As the program knows how to navigate from page to page in sequences of results, it can automatically extract quantities of information objects and organize them into usable collections. Such a method was used to select the required information: article title, date, number of comments.

The selected data consist of 999 articles and annotated video or reportage (see Figure 2). Some a picturesque examples of selected articles title: "31-year-old priest, known as Ben Lyris in Facebook: my priest's life as an anonymous alcoholic", "After the sex scandal, the priest offered to address the pope: and it is not clear to the devil where the truth lies", "Vatican Radio Journalist: The Pope is the best in the world to buy" and etc.

For finding useful articles in web portals we used lexical units selected by the expert of theology: 2-3 keywords were chosen for each of the three topics named above (for accurate distribution see Figure 3 and Figure 4). Firstly, the phrases were used in the form specified by the expert, lately we added data which was found with modified keywords. Lemmatization, diacritics ignorance and distance size between phrase words were used as methods for modification. It leads to contain data that are more useful.

Next step was data cleaning. There was a part of articles (about 5 %) that overlaps through the same topics several times. To avoid inaccuracies in the final results, such texts have been removed and left only the first option found.

After all data preprocessing, the expert of theology indexed all articles by evaluating relevance by keyword and topic. The use of the Likert scale in this case is not expedient, as it is best suited for exploring a clear object with positive and negative statements. Meanwhile, in this case, we are dealing with much more complicated, multi-faceted, different from conventional marketing research objects.

Low index articles were called inappropriate (see Figure 3). This information influenced the final calculations of the level of aggression. It should be noted that a database gathered with one tool is not suitable for our investigation and needs to be redesigned in a manually.

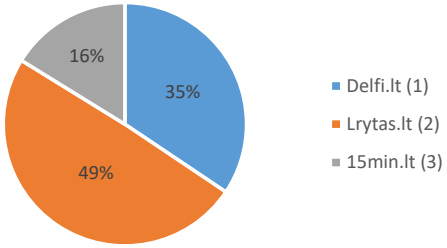


Fig. 2 Distribution of articles between portals

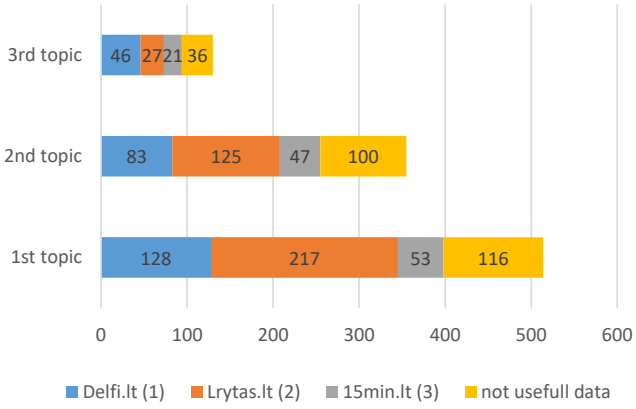


Fig. 3 Number of articles in topics by portal

[1] https://www.delfi.lt
[2] https://www.lrytas.lt
[3] https://www.15min.lt
[4] Downloaded from https://www.outwit.com/

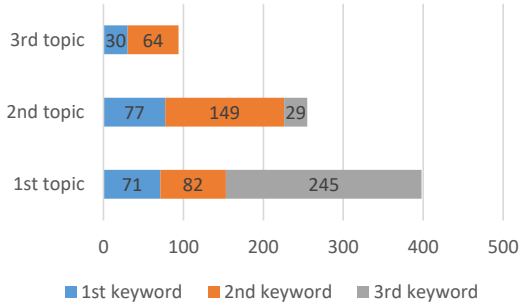


Fig. 4 Number of articles by topic founded by different keywords

V. METHODOLOGY, EXPERIMENTS AND RESULTS

We have conducted qualitative and quantitative content analysis. Researchers [12] most often analyze patterns of content quantitatively using statistical methods, also use qualitative methods to analyze meanings of content within texts. The combination of methods is particularly common in interdisciplinary research; this study is not an exception either.

Research on the dissemination of anti-Christian information was done with prepared set of collected data. Necessary data was analyzed with R[5] program [13]. The same tool was used for graphical analysis, too.

Our results answers to the question: what is seen from the main Lithuanian media in a particular topic by ordering results by time scale (see Figure 5, Figure 6, Figure 7). Most of the information is reflected in the spikes in articles discussing certain topic. After collecting all of these spikes (periods of interests), they can be compared to the available context data. Relevant context was identified by experts of theology. Certain world events are repeating every year or are recurring periodically, others are one-off but significant.

There were two context data lists made: one for the first topic and another for the second and the third topics together. The data representing the first thesis (Ethical aspects of artificial insemination) shows that the largest number of articles against Christians was written in December 2016. At that time, topics of artificial insemination, embryo storage and donation were actively discussed in Lithuania, and the Seimas of the Republic of Lithuania adopted new legislation. Other articles frequency overlaps in web portals are also strongly related to political situation. In this way, the opinion of the community is strongly forming, influencing the opinion of the individual. The data about the second and third topics is very similar (see Figure 8). Most articles were published in September 2018 - at that time Pope visited Lithuania. Other coincidence is June 2016: these facts supported the theory that at the beginning of the admission to the seminary, the number of articles directed against Christians is increasing.

The intenseness of the impact of verbal aggression depends on the attitudes of the acceptor, but the publishing frequency of articles remains important. Statistical analysis has shown that aggressive texts are presented to readers more than every two days (see Figure 9).

The level of aggression in the articles was evaluated by an expert of theology. After analyzing the data, it is determined that the most aggressive texts are written by web portal Lrytas (see Figure 10), the most aggressive topic is Ethical aspects of artificial insemination.

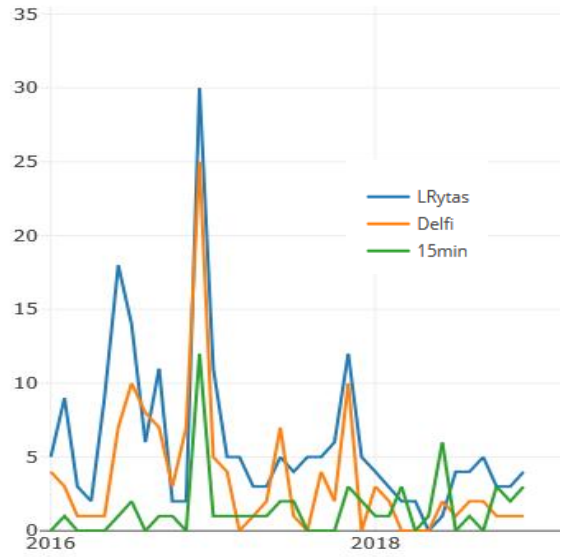


Fig. 5 Number of articles in the first topic by portals

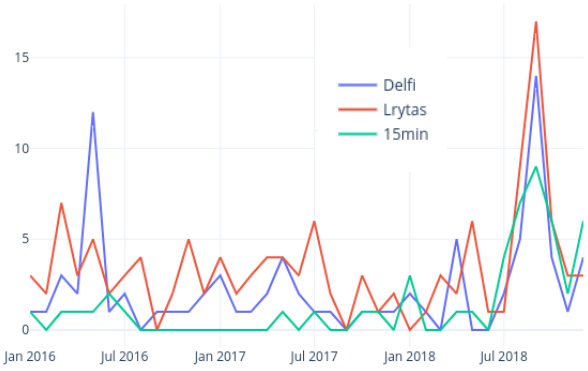


Fig. 6 Number of articles in the second topic by portals

[5] Downloaded from https://www.r-project.org/

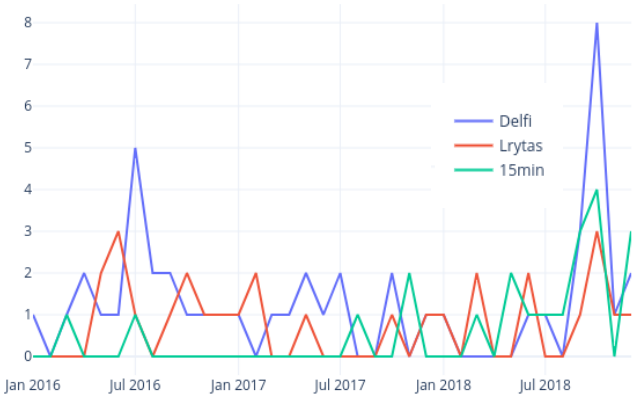


Fig. 7 number of articles in the third topic by portals

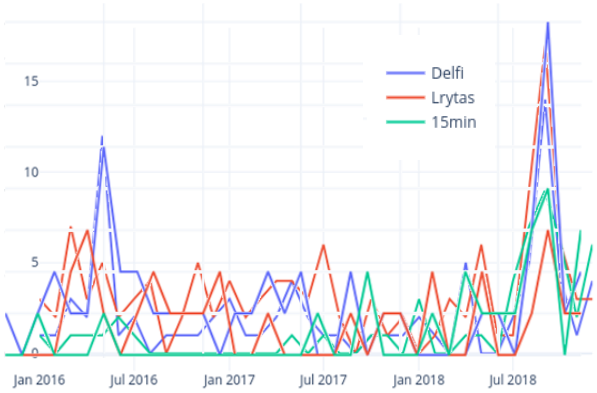


Fig. 8 Data similarity between the second and third topics

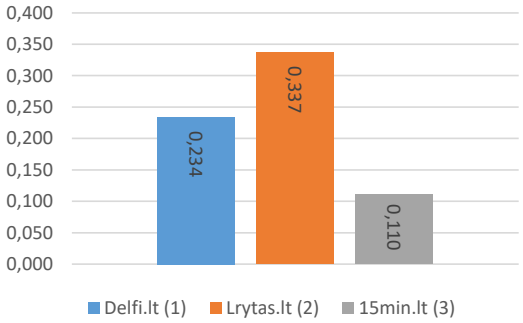


Fig. 9 Frequency of portals writing during the period

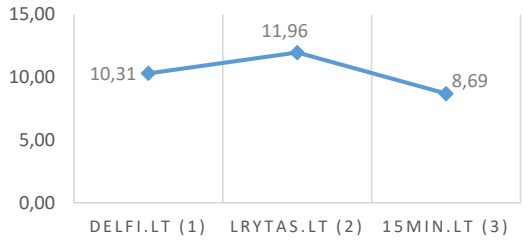


Fig. 10 Average of aggression levels in articles

VI. Conclusion and Future Works

After all investigations done, we can assert that (1) the most aggressive texts are written by web portal Lrytas (https://www.lrytas.lt), (2) the most aggressive topic in Lithuanian media is Ethical aspects of artificial insemination, (3) aggressive texts are presented to readers more than every two days, (4) articles frequency overlaps in web portals are also strongly related to political phenomena.

In this work we offer and test the chosen methods for a specific field of research. This media analysis reflects the spread of anti-Christian opinion in 2016-2018 years in three most popular Lithuanian web portals. These research findings are the result of knowledge generated through multiple theological research and data science methods. These methods are not entirely new, but the application is very specific and quite unique in this position.

In principle, the proposed method is also applicable to other (non) denominations of faith, but the basic condition for choosing it should be related to the amount of material and a certain rule in the new media for the type of verbal aggression under investigation.

Evaluation of the theoretical possibilities of overcoming the aforementioned aggression trends is not the main objective of this paper. It is rather implied trend of further studies. On the other hand, having considered the fact that the positivity of media is related with its contribution to the common goodness, spiritual rest, education, the formation of the strategy of aggression overcoming is undoubtedly an important task nowadays.

Our future plans are to develop another data collection method that will reduce manual work. Also create an automatic tool for evaluating aggression in the media is important. Another important area is the automatic detection of verbal aggression against Christians in the online media.

References

[1] C. A. Anderson, K. Suzuki, E. L. Swing, Ch. L. Groves, D. A. Gentile, S. Prot, Ch. Pan Lam, A. Sakamoto, Y. Horiuchi, B. Krahé, M. Jelic, W. Liuqing, R. Toma, W. A. Warburton, Xue-Min Zhang, S. Tajima, F. Qing, P. Petrescu, "Media Violence and Other Aggression Risk Factors in Seven Nations," *Personality and Social Psychology Bulletin*, vol. 43, no. 7, pp. 986-998, 2017.

[2] J. Cantor, B. J. Wilson, "Media and Violence: Intervention Strategies for Reducing Aggression," *Media Psychology*, vol. 5, no. 4, pp. 363-403, 2009.

[3] S. M. Coyne, J. Archer, "Indirect aggression in the media: A content analysis of british television programs," *Aggressive behavior*, vol. 30, no. 3, 2004.

[4] M. C. Nisbet , T. P. Newman, Framing, the Media, and Environmental Communication, The Routledge Handbook Of Environment And Communication, 2015.

[5] D. Scheufele, "Framing as a theory of media effects," *Journal of Communication*, vol. 49, no. 1, pp. 103-122, 1999.

[6] L. Bolce, G. De Maio, "A Prejudice for the Thinking Classes: Media Exposure, Political Sophistication, and the Anti-Christian Fundamentalist," *American Politics Research*, vol. 36, no. 2, 2008.

[7] M. S. Miceli, "Morality Politics vs. Identity Politics: Framing Processes and Competition Among Christian Right and Gay Social Movement Organizations," *Sociological Forum*, vol. 20, no. 4, p. 589–612, 2005.

[8] W. Wood, F. Y. Wong, J. G. Chachere, "Effects of media violence on viewers' aggression in unconstrained social interaction," *Psychological Bulletin*, vol. 109, no. 3, pp. 371-383, 1991.

[9] C.C. Aggarwal, C.X. Zhai, Mining text data, Springer, 2012.

[10] "Gemius Baltic," 2017 m. pradžia augino svetainių populiarumą, 15 02 2017. [Online]. Available: https://www.gemius.lt/visos-naujienos/id-2017-m-pradzia-augino-svetainiu-populiaruma.html. [Accessed 01 03 2019].

[11] W. t. O. H. H. Center, "OutWit Hub Light, Pro, Expert & Enterprise Editions v7.x," 30 01 2018. [Online]. Available: https://www.outwit.com/downloads/release/7.0/outwit-hub-help%20v7.0.pdf. [Accessed 01 03 2019].

[12] S. C. Woolley, P. N. Howard, " Computational propaganda worldwide: Executive summary," *Working Paper No. 2017.11*, 2017.

[13] R. C. Team, "R: A language and environment for statistical," R Foundation for Statistical Computing, Vienna, Austria, 2013.

Aligning agile software development with enterprise architecture framework

Karolis Noreika
Kaunas, Lithuania
E-mail: knoreikos@gmail.com

Saulius Gudas
Institute of Data Science and Digital Technologies
Vilnius University
Vilnius, Lithuania
E-mail: saulius.gudas@mii.vu.lt

Abstract. The effectiveness of internal processes is a key in modern day economy for companies of all sizes. This also includes the effectiveness in software development management and its alignment with business goals both short and long term. But it is not always easy to align IT development with organizational goals. This paper suggests a method for aligning modern software development approaches with enterprise architecture frameworks.

Keywords: agile software development, business and IT alignment, enterprise architecture framework, TOGAF

I. Introduction

Agile approach is a popular software development methodology. Agile approach currently is being adopted to business strategy execution, decision making to achieve strategic goals. Companies are "going agile" [1] in order to improve productivity of software teams as well as business teams making business decisions. "Going agile" is a big organizational change. It means that employees in all levels of organization will need to adapt to the new way of working, which is getting the results of their daily duties evaluated much faster than in the traditional way of working. However, when "going agile", the overall goals of the organization are not always supported with an organizational change. There are researches that emphasize the importance of supporting the agile way of working from organizational perspective (provide appropriate physical atmosphere, work environment that encourages creativity) [2]. The gaps between business and IT strategies appear. It might result in not sufficient quality of software products, that are not in line with overall goals of the organization both short and long term.

Eventually the misalignment becomes so significant that organizations get into a position when there is no way back – either decommission the system or accept the extremely costly support of it (i.e. mainframe systems in financial institutions).

Enterprise architecture is a well-defined practice for conducting enterprise analysis, design, planning, and implementation, using a comprehensive approach at all times, for the successful development and execution of strategy [3]. The agile methodology life cycle could be structured and aligned with TOGAF life cycle, which is a standard for enterprise architecture development. TOGAF is a framework – a detailed method – for designing, planning, implementing, and governing an enterprise architecture [4], [10].

This paper proposes a methodology for how agile methodology life cycle can be aligned with TOGAF enterprise architecture framework.

II. Agility in Business Management and Software Development

Agile approach allows business representatives to see the value of the software product being developed faster compared to traditional software development. Traditional or "waterfall" software development dates back to around 1970ties when the development of large enterprise IT systems was started to be described in a scientific way [5].

The waterfall methodology utilizes the idea that each phase in software development is sequential and cannot repeat. The agile methodology promotes the idea of repeated and iterative steps, which are explained in Figure 1 below.

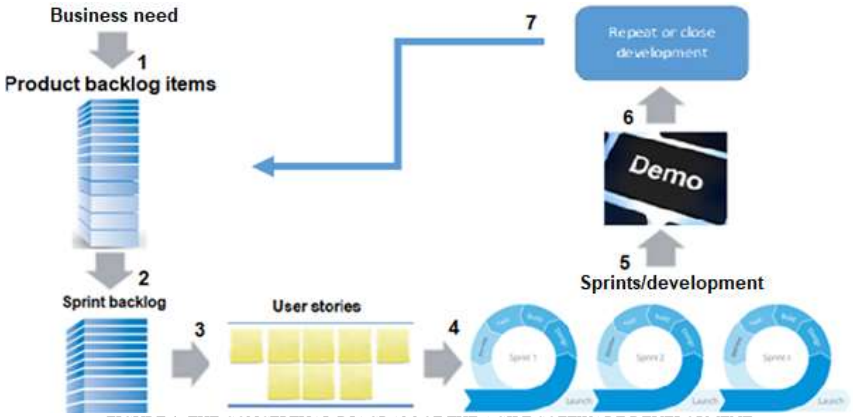


FIGURE 1. THE CONCEPTUAL DIAGRAM OF THE AGILE SOFTWARE DEVELOPMENT

Figure 1 explains how the software product is being developed after the business need is received. It contains the whole agile life cycle which is organized by having different information/knowledge flows. The detailed description of each of the elements is in table 1 below.

TABLE 1. DESCRIPTION OF AGILE LIFE CYCLE ELEMENTS

Step/ element no.	Step/element name	Step/element description
1	Data flow no. 1	The data flow, containing the information needed for understanding the business need, problem. It is transformed into product backlog item (element no. 2)
2	Product backlog items	List of features and requirements that the solution should have once completed.
3	Data flow no. 2	An incoming data flow for the next element in the life cycle – sprint backlog (element no. 4). It contains the features that software should contain after development iteration – sprint.
4	Sprint backlog	List of features that will be developed in the next sprint. Sprint is a time frame with a list of features described and approved by business and IT representatives.
5	Data flow no. 3	Once the high level features to be developed in the next sprint are agreed upon – the details must be clarified to the level needed in order to accomplish the business needs. This data flow contains the sprint backlog items or features explained in smaller pieces of information or requirements – user stories.
6	User stories	The detailed requirements are worked on – analyzed by the development team so that both IT and business representatives understands the problem each user story would solve.
7	Data flow no. 4	The detailed user stories are placed in some software tool that would allow keeping track of the progress of development of user stories.
8	Sprints/ development	This is the most beneficial part of the agile life cycle. It is a method of constantly developing small part of the overall software solution and getting the feedback fast. Each sprint consists of: a) Design – designing the user interface, business rules placed in the solution. b) Build (develop) – coding, styling, working on the solution from development perspective. c) Test – test the developed solution against the requirements. d) Review – review the solution test findings and decide what to improve. e) Launch – after the items that were agreed to be developed at the end of sprint are verified against the solution itself and the found changes that were necessary to do are done, the project team decides should the solution at current stage be deployed into production (or live) environment, where it could be already used by business representatives. Note: agile promotes the approach that the project team should be able to continue the iterative development for indefinite amount of time. Therefore, the number of sprints with the same phases as mentioned above could continue indefinitely.
9	Data flow no. 5	The information gathered from the business need at the beginning of the project and throughout development phase aggregated to prepare the demo of the solution.
10	Demo	The demo for the solution is a system presentation conducted to all relevant stakeholders.
11	Data flow no. 6	The decision after the demo whether the solution should be included into production environment or should the development continue with taking the next set of requirements made.
12	Repeat or close development	As Agile methodology describes – self-organizing team should be capable of keeping the accepted efficiency for product development indefinitely. This means that if business managers decide – the team should be able to repeat the whole cycle indefinite amount of times until the repetition does not increase the value significantly. If the decision is made to close the project – the agile life cycle is completed.
13	Data flow no. 7	If it is decided to continue development – the next set of requirements is taken from the product backlog items list (element no. 2) and the agile life cycle is repeated.

There are a lot of details and techniques how agile life cycle should be managed to achieve the best efficiency [6], [7], [8], but this is not a subject of this paper.

However, running a successful business is not only about doing software development in an agile way. Often IT development is ahead of business decisions where to come up with a suitable software solution development teams needs quick decisions by business that might be applicable across multiple projects in same business area (i.e. store related documents in single repository, have same classification of them, etc.) The business side in the enterprises is also starting to take decision based on agile methodology, although it is often perceived as a part of startup culture – i.e. not something established and large organizations would do. Table 2 below represents the comparison of agile and traditional approach on business decision making in agile and traditional ways.

TABLE 2. AGILE AND TRADITIONAL DECISION MAKING COMPARISON

Methodology	Flexibility	Risk	Adapting to change	Amount of data needed to make decision
Agile (including variations)	Higher	Higher	Faster	Smaller
Traditional	Lower	Lower	Slower	Larger

Agile methodology could be applied to business decision making by mapping the agile phases to decision making process – i.e. limit the information needed to make decision could be compared to sprint backlog. Having a deadline for a decision could be understood as the date for demo. Adjust to new information on the decision could be understood as review part of sprint.

III. IDENTIFICATION OF GAPS BETWEEN BUSINESS AND IT STRATEGIES

A. Business and IT alignment model

The business and IT alignment model was created by Henderson and improved by Venkatraman to represent business strategy alignment with IT strategy thus providing analysis method aimed for competitive advantage [9]. Figure 2 below represents the strategic alignment model.

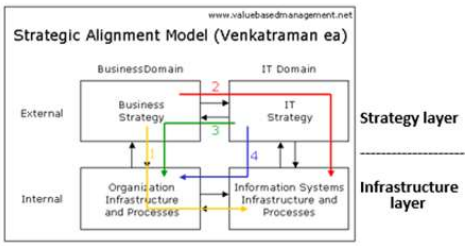


FIGURE 2. THE BUSINESS AND IT ALIGNMENT MODEL

There are four domain alignment perspectives where each focuses on different aspect of alignment between the business and IT alignment, i.e.:

1) Strategy execution – business strategy is the driver for organization design changes and the logic of IT infrastructure. In this perspective, the top management of the organization dictates the strategy of the company and the IT management is the strategy implementer.

2) Technology potential – business strategy is the driver for change, however it is closely aligned with IT strategy as wel,l therefore the IT systems are more aligned with IT strategy and also business strategy.

The top management should provide the vision of the technology to articulate the logic and choices to IT strategy what would best support the chosen business strategy. The role of IT manager in this perspective should be of the technology architect – the IT manager should efficiently and effectively design and also implement information system infrastructure that is consistent with the IT strategy. This alignment perspective could be used for aligning business and IT strategy along with IT systems in an agile way as vision is also one of key aspects to have for the agile development teams to be successful and self-organizing.

3) Competitive potential – focuses on utilizing emerging IT capabilities to impact new products and services also to influence key attributes of strategy (distinct competences) as well as form new relationships (business governance). This perspective also allows the changing of business strategy via emerging IT capabilities. The role of the management is of business visionary who dictates how emerging IT competences and functionality would impact the business strategy. The role of IT manager is of the one who identifies and interprets the trend in the IT environment to assist the business managers to understand the potential opportunities and threats from an IT perspective and handle them accordingly.

4) Service level – this perspective focuses on building world class IT team. Therefore, the role of IT manager is also of a business leadership with tasks of making the internal business succeed with the operating guidelines from top management.

B. TOGAF

TOGAF is framework for designing, planning, implementing, and governing an enterprise information technology architecture. The TOGAF standard includes a content framework to drive the Architecture Development Method (ADM). TOGAF is an iterative process model (enterprise architecture development life cycle) supported by best practices and a re-usable set of existing architecture assets. TOGAF supports Capability-Based Planning of enterprise architecture [10].

The TOGAF framework is presented in Figure 3 below.

Enterprise architecture development life cycle (defined in TOGAF) could be used for analysis of the agile software development approach.

The TOGAF life cycle in Figure 3 was transformed to a schematic view of a table (Figure 4) in which the columns represent the phases of TOGAF enterprise architecture development life cycle and agile methodology life cycle and the activities in the intersecting sections – the phases of agile development (design, build, test and deploy) [11]. This approach could be used into applying agile way of working for building up and aligning with enterprise architecture implementation that TOGAF provides.

Although companies can change strategies quickly, they then face the big slowdown of executing one or several strategies. For enterprise architects, this has traditionally meant defining a new target state, comparing it with the current state, and then developing a road map. But this multistep process is now perceived as taking too long — by the time EA has all of these documented and approved, the business will have moved on [12].

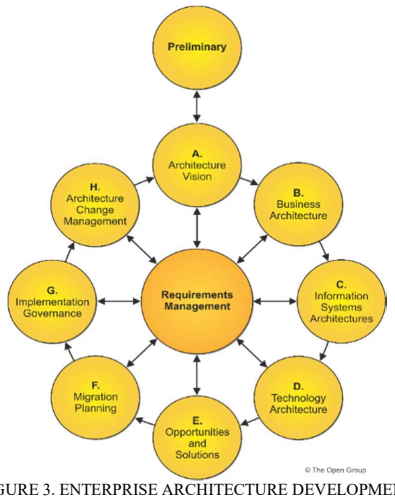


FIGURE 3. ENTERPRISE ARCHITECTURE DEVELOPMENT LIFE CYCLE TOGAF (https://www.opengroup.org/togaf)

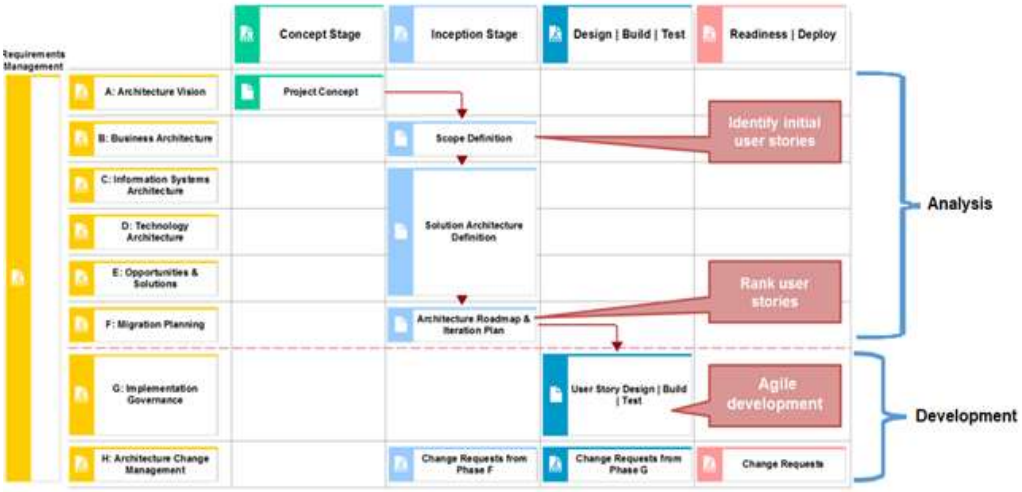


FIGURE 4. TOGAF ENTERPRISE ARCHITECTURE FRAMEWORK AND AGILE METHODOLOGY ALIGNMENT MODEL

IV. ALIGNING AGILE LIFE CYCLE WITH ENTERPRISE ARCHITECTURE FRAMEWORK

It is a common belief that TOGAF and also other large enterprise architecture frameworks are “waterfall”. This is a common misinterpretation largely due to these models encompassing all related IT activities and not specific. But basically all these enterprise architecture frameworks are sets of tools, similar like agile where one also should choose the tools and methods suitable for each specific case. A problem in large organizations is that there are different levels of maturity of agile of different teams. Business representatives (also called stakeholders, subject matter experts or in agile – product owners) represent the business only fragmentally – whenever there is a question regarding IT and business alignment – it is solved on ad-hoc basis, but a long term IT and business strategy should be capable of answering these questions on a higher – strategic – level which is orchestrated by using the TOGAF methodology.

The idea behind mapping TOGAF to agile life cycle is use the strategic vision that TOGAF provides by using its framework and utilize the benefits of agile continuous

improvement and "inspect and adapt" approach. The suggested mapping is displayed in Figure 5.

When TOGAF is used for overall overview on the enterprise architecture and agile is used for project's iterations, the business gets benefit from even faster deliveries and projects are aligned with business goals at all times.

V. Case Study

Large enterprises often combine the IT infrastructure „in-house" together with outsourcing it. It could be only storing part of data or all the data of the enterprise. These decisions are made according to IT strategy mostly and not always these decisions are aligned with business strategy. As the technology advances to the cloud based solutions more and more companies are concerned about the safety of the data in the cloud based systems. Combining these concerns with the agreed service level agreements provided by external vendors not being maintained for enterprises to run their operations smoothly (i.e. important IT system being outsourced is not working part of the day due to agreed service level agreement breached) might lead to decisions to insource the IT and IS infrastructure. But the cost of such decisions is very dependent on the level of alignment between IT and business strategy and the less is the alignment, the bigger are the costs. Whenever an enterprise is faced with such decision, it is very important to keep the alignment between business and IT strategies moving on.

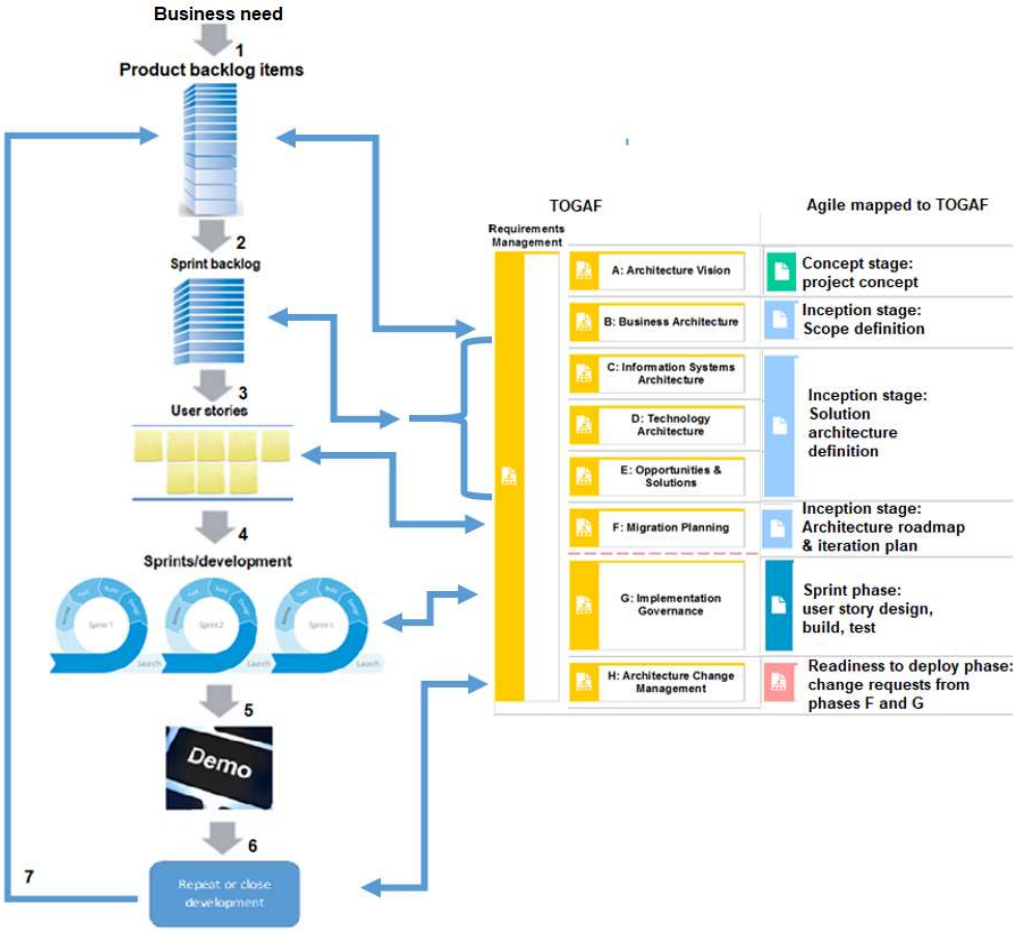


FIGURE 5. THE SIMPLIFIED MAPPING OF TOGAF TO AGILE LIFE CYCLE

By using the TOGAF enterprise architecture framework and agile methodology alignment suggested in chapter 4, the company facing such decision might significantly reduce the cost and impact of the migration from outsource provider to in-house solution by constantly aligning the enterprise architecture which TOGAF describes with the constant feedback, "inspect and adapt" approach that agile promotes.

The case where such suggestion was made was about large enterprise moving over 2000 servers of different purposes from outsource to in-house. When using the suggested method

of enterprise architecture framework TOGAF being aligned with agile methodology the implementation of the change could have taken at least 10 % less effort both in terms of cost and time needed for the change as the comparison of activities by using only agile methods and the suggested method showed. Also it is worth noting that this situation could have been avoided if all the tools and methodologies mentioned in this paper were used: IT and business alignment model for overseeing the potential IT infrastructure decisions, TOGAF for overseeing enterprise architecture and the TOGAF and agile alignment model suggested by this paper which helps to see the potential gaps between agile software development and business strategy much faster.

VI. Conclusions

The agile way of working is something that in some enterprises is new but others are already very far away in implementing this approach into daily decision making process both business and software development. These decisions need to be constantly aligned with the overall business strategy to have the effective enterprise run smoothly. Therefore, it is very important to align the enterprise architecture of the organization with agile approach to make the most benefit of enterprise architecture framework like TOGAF, which provides the tools to ensure business and IT alignment whereas agile provides the speed and the possibility to adapt to changes. Method, suggested in this paper, supports utilization of those mentioned benefits from both tools and allows to improve not only software development process which agile supports, but also keep the alignment between IT and business strategy by constantly keeping IT projects aligned with business strategy which TOGAF supports to make sure right solutions are developed and aligned with long term goals of the enterprise. The proposed approach could be further improved through the use on different types of organizations (i.e. financial, trade, manufacturing) and adapting it in a generalized way for further usage.

References

[1] M. Pikkarainen, J. Haikara, O. Salo, P. Abrahamsson, J. Still, Pikkarainen, M., Haikara, J., Salo, O. et al. “Empir Software Eng” (2008) 13: 303. https://doi.org/10.1007/s10664-008-9065-9

[2] Crawford, Broderick & León de la Barra, Claudio & Soto, Ricardo & Misra, Sanjay & Monfroy, Eric. (2013). “Agile Software Development: It Is about Knowledge Management and Creativity”. ICEIS 2013 - Proceedings of the 15th International Conference on Enterprise Information Systems. 2. 10.5220/0004447802650272.

[3] Federation of EA Professional Organizations, "Common Perspectives on Enterprise Architecture," Architecture and Governance Magazine, Issue 9-4, November 2013 (2013).

[4] Dirk Draheim, Gerald Weber “Trends in Enterprise Application Architecture” 2nd International Conference, TEAA 2006, Berlin, Germany, November 29 - December 1, 2006, Revised Selected Papers

[5] Winston Royce, “Managing the Development of Large Software Systems”, Proc. Westcon, IEEE CS Press, 1970, pp. 328-339

[6] Jeffrey Verret “Implementing Agile Methodology: Challenges and Best Practices” University of Oregon (2018)

[7] Darrell K. Rigby, Jeff Sutherland, Hirotaka Takeuchi “Embracing Agile” Harvard Business Review 2016, pp.40–48

[8] Zamudio, Lizbeth & Aguilar, José & Tripp-Barba, Carolina & Misra, Sanjay. (2017). A Requirements Engineering Techniques Review in Agile Software Development Methods. 683-698. 10.1007/978-3-319-62404-4_50.

[9] Henderson, John and Venkatraman, N.” Strategic alignment: A model for organization transformation via information technology” Working Paper 3223-90. Massachusetts Institute of Technology, 1990, 458 p. ISBN 9781245057264.

[10] The TOGAF® Standard, Version 9.2 https://www.opengroup.org/togaf

[11] BPM utils.com Enterprise Agile solution Delivery Framework 2017

[12] Henry Peyret “EA Methodologies Enlarge To Address The New Business Landscape”, 2013
http://entreprise-agile.com/ForresterEA.pdf

INFORMACINËS TECHNOLOGIJOS ● IVUS 2019 ● ISSN 2029-249X ● eISSN 2029-4824

Lithuanian news clustering using document embeddings

Lukas Stankevičius
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
lukas.stankevicius@ktu.edu

Mantas Lukoševičius
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania
mantas.lukosevicius@ktu.lt

***Abstract*—A lot of research of natural language processing is done and applied on English texts but relatively little is tried on less popular languages. In this article document embeddings are compared with traditional bag of words methods for Lithuanian news clustering. The results show that for enough documents the embeddings greatly outperform simple bag of words representations. In addition, optimal lemmatization, embeddings vector size, and number of training epochs were investigated.**

Keywords—document clustering; document embedding; lemmatization; Lithuanian news articles.

I. Introduction

The knowledge and information are inseparable part of our civilization. For thousands of years from news of incoming troops to ordinary know-how could have meant death or life. Knowledge accumulation throughout the centuries led to astonishing improvements of our way of live. Hardly anyone could persist having no news or other kinds of information even throughout the day.

Despite information scarcity centuries ago, nowadays we have the opposite situation. Demand and technology greatly increased the amount of information we can acquire. Now one's goal is to not get lost in it. As an example, the most popular Lithuanian news website each day publishes approximately 80 news articles. Add other news websites not only from Lithuania but the entire world and one would end up overwhelmed to read most of this information.

The field of text data mining emerged to tackle this kind of problems. It goes "beyond information access to further help users analyze and digest information and facilitate decision making" [1]. Text data mining offers several solutions to better characterize text documents: summarization, classification and clustering [1]. However, when evaluated by people, the best summarization results currently are given only 2-4 points out of 5 [2]. Today the best classification accuracies are 50-94% [3] and clustering of about 0.4 F1 score [4]. Although achieved classification results are more accurate, the clustering is perceived more promising as it is universal and can handle unknown categories as it is the case for diverse news data.

After it was shown that artificial neural networks can be successfully trained and used to reduce dimensionality [5], many new successful data mining models had emerged. The aim of this work is to test how one of such models – document to vector (Doc2Vec) can improve clustering of Lithuanian news.

II. Related work on Lithuanian language

Articles on Lithuanian language documents clustering suggest using K-means [4], spherical K-means [6] or Expectation-Maximization (EM) [7] algorithms. It was also observed that K-means is fast and suitable for large corpora [7] and outperforms other popular algorithms [4].

[6] considers Term Frequency / Inverse Document Frequency (TF-IDF) as the best weighting scheme. [4] adds that it must be used together with stemming while [6] advocates to do minimum and maximum document frequency filtering before applying TF-IDF. These works show that TF-IDF is significant weighting scheme and it could be optionally tried with some additional preprocessing steps.

We have not found any research on Lithuanian language regarding document embeddings. However, there are some work on word embeddings. In [8] word embeddings using different models and training algorithms were compared after training on 234 million tokens corpus. It was found that Continuous Bag of Words (CBOW) architecture significantly outperformed skip-gram method while vector dimensionality showed no significant impact on the results. This implies that document embeddings like word embeddings should follow same CBOW architectural pattern. Other work [9] compared traditional and deep learning (with use of word embeddings) approaches for sentiment analysis and found that deep learning demonstrated good results only when applied on the small datasets, otherwise traditional methods were better. As embeddings may be underperforming in sentiment analysis it will be tested if it is a case for news clustering.

III. Text Clustering Process

To improve clustering quality some text preprocessing must be done. Every text analytics process consists „of three consecutive phases: Text Preprocessing, Text Representation and Knowledge Discovery" [1] (the last being clustering in our case).

A. Text preprocessing

The purpose of text preprocessing is to make the data more concise and facilitate text representation. It mainly involves tokenizing text into features and dropping the ones considered less important. Extracted features can be words, chars or any *n*-gram (contiguous sequence of *n* items from a given sample of text) of both. Tokens can also be accompanied by the structural or placement aspects of document [10].

The most and least frequent items are considered uninformative and dropped. Tokens found on every document are not descriptive and they usually include stop words such

as "*and*", "*to*". On the other hand, too rare words are insufficient to attribute to any characteristic and due to their resulting sparse vectors only complicate the whole process.

Existing text features can be further concentrated by these methods:

- stemming;
- lemmatization;
- number normalization;
- allowing only maximum number of features;
- maximum document frequency – ignore terms that appear in more than specified documents;
- minimum document frequency – ignore terms that appear in less than specified documents.

It was shown that the use of stemming in Lithuanian news clustering greatly increased clustering performance [4].

B. Text representation

For the computer to make any calculations with the text data it must be represented in numerical vectors. The simplest representation is called "Bag Of Words" (BOW) or "Vector Space Model" (VSM) where each document has counts or other derived weights for each vocabulary word. This structure ignores linguistic text structure. Surprisingly, in [11] it was reviewed that "unordered methods have been found on many tasks to be extremely well performing, better than several of the more advanced techniques", because "there are only a few likely ways to order any given bag of words".

The most popular weight for BOW is TF-IDF. Recent study [4] on Lithuanian news clustering have shown that TF-IDF weight produced the best clustering results. TF-IDF is calculated as:

$$tfidf(w,d) = tf(w,d) \cdot log\frac{N}{df(w)} \quad (1)$$

where:

- $tf(w,d)$ is term frequency, the number of word w occurrences in a document d;
- $df(w)$ is document frequency, the number of documents containing word w;
- N is number of documents in the corpus.

One of the newest and widely adopted document representation schemes is Doc2Vec [12]. It is an extension of the word-to-vector (Word2Vec) representation. A word in the Word2Vec representation is regarded as a single vector of real number values. The assumption of Word2Vec is that the element values of a word are affected by those of other words surrounding the target word. This assumption is encoded as a neural network structure and the network weights are adjusted by learning observed examples [13]. Doc2Vec extends Word2Vec from the word level to the document level and each document has its own vector values in the same space as that for words [12].

C. Text clustering

There are tens of clustering algorithms to choose from [14]. One of the simplest and widely used is *k*-means algorithm. During initialization, *k*-means algorithm selects *k* means, which corresponds to *k* clusters. Then algorithm repeats two steps: (1) for every data point choose the nearest mean and assign the point to the corresponding cluster; (2) recalculate means by averaging data points assigned to the corresponding cluster. The algorithm terminates, when assignment of the data points does not change after several iterations. As the clustering depends on initially selected centroids, the algorithm is usually run several times to average over random centroid initializations.

IV. The Data

A. Articles

Article data for this research was scraped from three Lithuanian news websites: the national *lrt.lt* and commercial websites *15min.lt* and *delfi.lt*. Articles URL's were scraped from sitemaps in *robots.txt* files in websites. Total of 82793 articles (26336 from *lrt.lt*, 31397 from *15min.lt* and 25060 from *delfi.lt*) were retrieved spanning random release dates of 2017 year.

Raw dataset contains 30338937 tokens from which 641697 are unique. Unique token count can be decreased to:

- 641254, dropping stop words;
- 635257, normalizing all numbers to a single feature;
- 441178, applying lemmas and leaving unknown words;
- 41933, applying lemmas and dropping unknown words;
- 434472, dropping stop words, normalizing numbers, applying lemmas and leaving unknown words.

Each article has on average 366 tokens and on average 247 unique tokens. Mean token length is 6.51 characters with standard deviation of 3.

While analyzing articles and their accompanying information, it was noticed that some labelling information can be acquired from article URL. Both websites have categorical information between the domain and article id parts in URL. Total of 116 distinct categorical descriptions were received and normalized to 12 distinct categories as described at [4]. Category distributions are:

- Lithuania news (20162 articles);
- World news (21052 articles);
- Crime (7502 articles);
- Business (7280 articles);
- Cars (1557 articles);
- Sports (5913 articles);
- Technologies (1919 articles);
- Opinions (2553 articles);
- Entertainment (769 articles);
- Life (944 articles);
- Culture (3478 articles);
- Other (9664 articles, which do not fall into previous categories).

It is clearly visible that category distribution is not uniform. The biggest categories are "Lithuanian news" and "World news" taking up to 49 % of all articles.

B. Words

Lithuanian word data was scraped from two semantic information databases: *morfologija.lt* and *tekstynas.vdu.lt/~irena/morfema_search.php*. The latter website has more accurate information, including word frequency while the first is very large and was observed having some mistakes. Therefore, these two databases were merged prioritizing words from the second one. Resulting word database contained 2212726 different word forms including 72587 lemmas.

V. Clustering Evaluation

The main evaluation metrics can be acquired by confusion matrix, depicted in Table I. Here for true and predicted conditions we get counts of following types:

- TP (true positives). The true condition is positive and the predicted condition is positive.
- TN (true negatives). The true condition is negative and the predicted condition is negative.
- FP (false positives). The true condition is negative but the predicted condition is positive.
- FN (false negatives). The true condition is positive but the predicted condition is negative.

If it would be a classification task, then we would know real classes and just simply get percentage of them predicted accurately. However, in the clustering process nor we know actual class, nor we have a meaning of returned predicted class. We must rely an additional information - label of our news article category, given by the editor of the news website. This way we make assumption that clusters we want to achieve are similar to categories of articles. There indeed must be a reason, some similarity between articles, why they were put in the same category. The only drawback of our approach is that having high number of documents would require many pair calculations. Based on chosen condition, confusion matrix elements are as following:

- TP – pairs of articles have same category label and are predicted to be in the same cluster.
- TN – pairs of articles belong to different categories and are predicted to be in different clusters.
- FP – pairs of articles belong to different categories but are predicted to be in the same cluster.
- FN – pairs of articles having same category label but are predicted to be in different clusters.

We will use F1, as the one widely used, and MCC, as more robust, evaluation scores:

$$F1 = 2\frac{precision \cdot recall}{precision + recall} \tag{2}$$

$$precision = \frac{TP}{TP+FP} \tag{3}$$

$$recall = \frac{TP}{TP+FN} \tag{4}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

MCC score ranges from -1 (total disagreement) to 1 (perfect prediction), while 0 means no better than random prediction. F1 score varies from 0 (the worst) to 1 (perfect).

VI. Experiments

To ensure that experiments are as reproducible as possible, each experiment was repeated 50 times and confidence interval of each resulting clustering scores calculated. In each repetition distinct number of articles were randomly (each time) selected from the dataset. However, for the same number of documents this repeated random pickup would be the same (if we were to have another experiment with same number of documents then these 50 samplings of articles would be the same). This ensures that we evaluate as much data as possible while keeping the same subset for different experiments.

All experiments were carried out using only articles from the 10 biggest categories. For each of them equal number of articles were sampled. Only variables associated with dataset loading, text preprocessing and representation phases were varied. Actual clustering was done using *k*-means algorithm.

In all experiments the following actions and parameters were used if not specified otherwise:

- used 1500 articles;
- vocabulary pruned to maximum of 10000 words;
- 0.95 maximum document frequency (BOW);
- 0.05 minimum document frequency (BOW);
- Distributed Bag of Words (DBOW) architecture of Doc2Vec model used;
- Doc2Vec method trained on same articles to be clustered (not all corpus);
- window size of 5 words (Doc2Vec models);
- 20 training epochs (Doc2Vec models);
- 200 vector size (Doc2Vec models);
- minimum word count of 4 (Doc2Vec models);
- all number normalized to "#NUMBER" feature;
- words with known lemma lemmatized;
- words in stop word list dropped from documents;
- unigrams used (feature as a single word).

A. Number of articles and preprocessor method experiment

In this experiment dataset size and preprocessor method were varied to determine how the two are correlated. Tried text representations include BOW and Doc2vec with distributed bag of words variation. It was also examined how well Doc2Vec would perform if trained on all the 82793 articles.

B. Reducing words to lemmas experiment

This experiment investigated 3 scenarios:

1) lemmas are not used;
2) words for which lemmas could be found were replaced with them and other words discarded;
3) same as 2 but unknown words remained.

Another parameter, namely maximum number of features, solves similar issues as lemmatization. Due to this reason several values of maximum number of allowed features were tried.

C. Training epochs and embedding vector size experiment

In this experiment two parameters for Doc2Vec were optimized: training epochs (from 5 to 100) and vector size

(from 5 to 400). Distributed bag of words version of Doc2Vec was used.

D. Clustering articles from a defined release interval

In this experiment the best configurations for BOW and Doc2Vec will be tried on articles released in one week from 2017-04-28 to 2017-05-04 dates, covering total of 1001 articles. Both models with same articles will be run 50 times and the best run selected. Doc2Vec is trained on same articles used for clustering using maximum number of 40000 features and vector size of 52.

The best resulting clusters will be analyzed with the same BOW workflow as documents but reducing features only with 0.8 maximum and 0.1 minimum document frequencies. 10 words with the biggest TF-IDF weights will be selected as representative of each cluster.

VII. Results and Analysis

A. Number of articles and preprocessor method experiment

Experiment results are shown in Fig. 1. The best recorded MCC score is 0.403 (0.464 for F1) for Doc2Vec, distributed bag of words variation trained on all corpus and clustering 3000 articles. It is clearly visible that all text representation models are better with higher number of documents. When clustering a small number of documents we can observe that BOW model outperforms Doc2Vec if the latter is trained only on documents that are later used for clustering. However, starting with 300 documents Doc2vec outperforms BOW model. This shows that Doc2Vec model depends on how many documents it is trained on as the model trained on all corpus has the biggest MCC score of 0.201 when clustering 100 articles. However, advantage of training on all corpus instead of only documents to be clustered quickly diminishes as the number of clustering documents approaches 700.

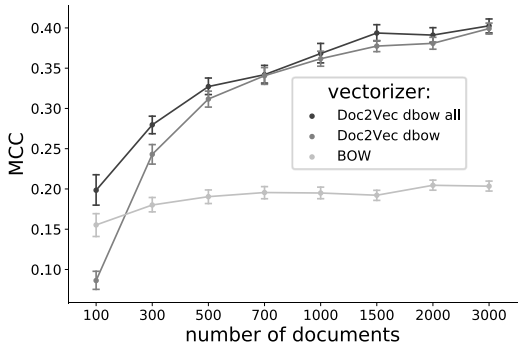


Fig. 1. MCC score dependency on text representation method and number of documents used in clustering.

B. Reducing words to lemmas experiment

Experiment results are depicted in Fig. 2. It was observed that converting known words to lemmas gives MCC score boost both for BOW and Doc2Vec models. The highest increase of MCC score (from 0.122 to 0.221 for 10000 maximum features) for BOW representation is observed then after lemmatization non-lemmatized words are dropped. On the other hand, Doc2Vec representation yields higher MCC score increase then non-lemmatized words are left (from 0.356 to 0.401 for 40000 maximum number of features). It is clearly visible that both vectorization methods benefit from lemmatization.

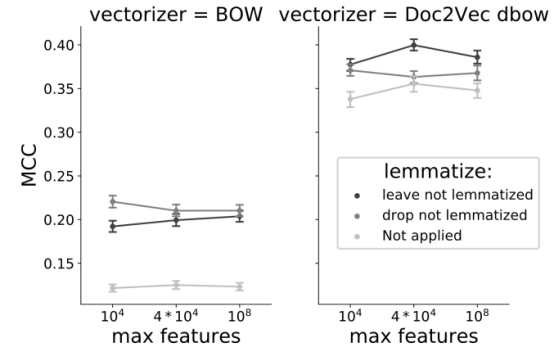


Fig. 2. MCC score dependency on how words are changed to their lemma with or without constrain of maximum features.

C. Training epochs and embedding vector size experiment

Clustering results for several epochs and vector sizes are depicted in Fig. 3. The highest average MCC score was recorder for vector size of 150 and 20 epochs at 0.381. It is interesting to note that increasing number of training epochs to 100 reduces MCC to 0.316. This reduction is observer for all vector sizes and could be explained as overfitting. On the other hand, only 5 epochs give poor results with maximum MCC of 0.133 for vector size of 10 and it should be regarded as underfitting. With optimal number of training epochs being 20, there are many vector sizes (from 20 to 400) yielding very similar MCC results. This shows that small vector sizes such as 20 are enough to train 1500 articles dataset for 20 epochs for good text representation.

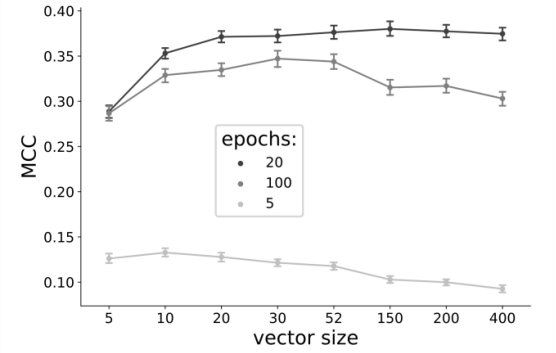


Fig. 3. MCC score dependency on vector size and number of training epochs in Doc2Vec distributed bag of words representation clustering

D. Clustering articles from defined release interval

The best Doc2Vec model trained on a small corpus outperformed the best BOW model (MCC 0.318 and 0.145, F1 0.415 and 0.282). Cluster features and statistics of Doc2vec model are depicted in Table I. It shows that model performs reasonably well and can distinguish:

- very small (1.9 % of all articles) distinct weather forecast category (cluster Nr. 5);
- classical categories as culture, sports, and crime (clusters Nr. 3, 8 and 10);
- hot topics as university reform, Brexit and current political scandals (clusters Nr. 1, 4 and 8).

TABLE I. CLUSTERS STATISTICS

		Category label										
Cluster Nr.	***Number of articles in cluster***	***Other***	***Crime***	***Culture***	***Lithuania news***	***Technologies***	***Opinions***	***World news***	***Entertainment***	***Sports***	***Business***	***Most descriptive features and their translation to English***
1.	40	11	0	0	**24**	0	3	0	0	0	2	universitetas, mokslas, eur, mokykla, studija, pertvarka, akademija, rektorius, vu, kokybė // university, science, eur, school, study, transformation, academy, rector, vu (Vilnius University), quality
2.	87	27	0	2	**35**	3	**15**	3	0	0	2	muzika, alkoholis, kultūra, ntv, filmas, visuomenė, maistas, namas, liga, lelkaitis // music, alcohol, culture, ntv, film, society, food, house, illness, lelkaitis (surname of a person)
3.	118	**29**	1	**40**	18	4	1	4	**16**	2	3	koncertas, teatras, muzika, rež, biblioteka, festivalis, džiazas, kultūra, paroda, muziejus // concert, theater, music, dir, library, festival, jazz, culture, exhibition, museum
4.	106	8	0	0	16	0	1	**80**	0	0	1	es, brexit, derybos, le, pen, may, macronas, partija, th, politinis // es, brexit, talks, le, pen, may, macron, party, th, political
5.	19	0	0	0	16	0	0	2	0	0	1	laipsnis, šiluma, temperatūra, naktis, debesis, debesuotumas, lietus, įdienojus, pūs, termometrai // degree, heat, temperature, night, cloud, clouds, rain, be broad daylight, will blow, thermometers
6.	184	1	0	0	16	5	0	**160**	0	0	2	jav, korėtis, raketa, korėja, branduolinis, putinas, jungtinis, pajėgos, karinis, sirijos // usa, korėtis, rocket, korea, nuclear, putin, united, forces, military, syrian
7.	120	11	1	0	37	4	9	10	0	0	**48**	įmonė, seimas, įstatymas, mokestis, savivaldybė, kaina, šiluma, asmuo, projektas, pajamos // company, parlament, law, tax, municipality, price, heat, person, project, income
8.	79	4	1	1	**67**	0	1	0	0	2	3	seimas, pūkas, partija, teismas, komisija, konstitucija, pirmininkas, įstatymas, apkalti, taryba // parlament, pūkas (surname of a person), party, court, commission, constitution, chairman, law, impeachment, board
9.	64	0	0	0	0	0	0	0	0	**64**	0	rungtynės, taškas, žaidėjas, čempionatas, ekipa, rinktinė, įvartis, pelnyti, pergalė, raptors // match, point, player, championship, team, team, goal, win, victory, raptors (name of basketball club)
10.	184	13	**67**	2	27	3	0	68	0	0	4	policija, automobilis, vyras, vairuotojas, pranešti, įtariamas, sulaikyti, žūti, teismas, asmuo // police, car, man, driver, report, suspected, detained, die, court, person

VIII. CONCLUSIONS

In this work BOW and Doc2Vec text representation methods were compared. Our research shows that Doc2Vec greatly outperforms BOW model. Clustering weeks' worth of data the highest MCC scores are 0.318 versus 0.145. However, for Doc2Vec method to outperform BOW when clustering less than 300 articles, it must be trained on a much larger dataset. We estimated optimal embedding vector size large enough starting with 20 and optimal number of training epochs around 20. Analysis of words conversion to their lemmas showed that lemmatization of words is beneficial for both BOW and Doc2Vec representations.

REFERENCES

[1] Aggarwal CC, Zhai C, editors. Mining text data. Springer Science & Business Media; 2012 Feb 3.

[2] Liu L, Lu Y, Yang M, Qu Q, Zhu J, Li H. Generative adversarial network for abstractive text summarization. In Thirty-Second AAAI Conference on Artificial Intelligence 2018 Apr 29.

[3] Liu G, Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing. 2019 Feb 1.

[4] V. Pranckaitis and M. Lukoševičius, Clustering of Lithuanian news articles. Proceedings of the IVUS 2017, pp. 27-32.

[5] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science. 2006 Jul 28;313(5786):504-7.

[6] Mackutė-varoneckienė, Aušra; Krilavičius, Tomas. Empirical study on unsupervised feature selection for document clustering. In Human Language Technologies – The Baltic Perspective 2014. p. 107-110.

[7] Ciganaitė, Greta, Aušra Mackutė-Varoneckienė, and Tomas Krilavičius. Text documents clustering. Informacinės technologijos. XIX tarpuniversitetinė magistrantų ir doktorantų konferencija" Informacinė visuomenė ir universitetinės studijos"(IVUS 2014): konferencijos pranešimų medžiaga, 2014, p. 90-93. 2014.

[8] Kapočiūtė-Dzikienė, Jurgita, and Robertas Damaševičius. Intrinsic evaluation of Lithuanian word embeddings using WordNet. Computer Science On-line Conference. Springer, Cham, 2018.

[9] Kapočiūtė-Dzikienė, Jurgita, Robertas Damaševičius, and Marcin Woźniak. Sentiment analysis of Lithuanian texts using traditional and deep learning approaches. Computers 8.1 (2019): 4.

[10] Aker A, Paramita M, Kurtic E, Funk A, Barker E, Hepple M, Gaizauskas R. Automatic label generation for news comment clusters. In Proceedings of the 9th International Natural Language Generation Conference 2016 (pp. 61-69).

[11] White L, Togneri R, Liu W, Bennamoun M. Sentence Representations and Beyond. In Neural Representations of Natural Language 2019 (pp. 93-114). Springer, Singapore.

[12] LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. In: International conference on machine learning. 2014. p. 1188-1196.

[13] MIKOLOV, Tomas, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[14] Charu C. Aggarwal , Chandan K. Reddy, Data Clustering: Algorithms and Applications, Chapman & Hall/CRC, 20

Daiktų interneto objektų identifikavimas

Raimundas Savukynas

Vilniaus universiteto Matematikos ir
informatikos instituto doktorantas
Vilnius University, Institute of Mathematics
and Informatics, Doctoral Student
Akademijos g. 4, LT-08663 Vilnius
El. paštas: raimundas.savukynas@mii.vu.lt

Daiktų interneto resursai ir teikiamos paslaugos yra pasiskirsčiusios informaciniame tinkle, todėl būtina turėti mechanizmą leidžiantį identifikuoti ir atrasti išmaniuosius įrenginius, jų teikiamas paslaugas ir resursus. Paprastai identifikacija susideda iš vardų priskyrimo resursams ir resursų adresavimo mechanizmo tam, kad atrasti ir pasiekti nutolusius daiktų interneto resursus. Šiuolaikiniuose tinkluose naudojami DOI (angl. Digital Object Identification) ir URI (angl. Universal Resorce Identifier), kurie identifikuoja resursą, jo tipą bei aibę lygiaverčių vardų. Jie leidžia unikaliai identifikuoti fizinius objektus, esančius bet kurioje pasaulio vietoje panaudojant EPC (angl. Electronic Product Code) technologiją. EPC leidžia DOI ir URI panaudoti kartu su RFID (angl. Radio-Frequency Identification). RFID technologija skirta objektų žymėjimui ir sekimui, paremta radijo dažnio signalo panaudojimu objekto žymoje esančios informacijos įrašymui ir nuskaitymui. Šiame straipsnyje apžvelgti ir palyginti RFID, URI, DOI daiktų interneto objektų identifikavimo metodai.

Pagrindiniai žodžiai: *daiktų internetas, objektų identifikavimas, ateities internetas, interneto paslaugos, išmanieji įrenginiai.*

Įvadas

Išmaniųjų įrenginių apsuptis ir internetas tapo įprastu reiškiniu mūsų gyvenime. Daugumoje gyvenimo sričių ir ūkio sektorių, valdymas ir stebėjimas vyksta naudojant internetą, o taip pat įvairius skaitmeninius įrenginius sąveikaujančius per tinklą. Duomenys tinklu perduodami ne tik naudojant „žmogus ir žmogus" arba „žmogus ir kompiuteris" sąveikas, bet ir per jutiklius tarpusavyje sąveikaujant įrenginiams, kurie dar yra vadinami interneto daiktais. Taip susiformavo daiktų internetas (angl. *Internet of Things – IoT*), kuris yra nauja tinklų konfigūracija, apimanti fizinių objektų komunikavimą ir objektų bei žmonių sąveiką internete. Daiktų internetas yra vadinamas kita interneto evoliucijos pakopa, kur daiktai tampa aktyvūs verslo, informacijos ir socialinių procesų dalyviai, galintys komunikuoti ir sąveikauti tarpusavyje, o taip pat su juos supančia išmaniąja aplinka besikeisdami duomenimis, autonomiškai reaguoti į fizinio pasaulio įvykius bei įtakoti aplinką, atliekant įvairius veiksmus ir teikiant paslaugas. Išmaniųjų aplinkų ir daiktų interneto teikiami privalumai apima komforto, saugumo, energetinių išteklių optimalaus panaudojimo ir daugelį kitų paslaugų, kurios gerokai pagerina gyvenimo kokybę. Daiktų interneto panaudojimo galimybės neapsiriboja vien tik teikiamomis paslaugomis žmonėms, bet leidžia pritaikyti įvairias technologijas pramonėje, prekyboje, transporte ir netgi skaitmeninėje realybėje, kur tarpusavyje komunikuoja ir sprendimus priima išmanūs programiniai agentai (Gubbi *et al*. 2013).

Daiktų internetas yra pasaulinis daiktų, turinčių sąsajas su jutikliais, valdikliais, programine įranga bei gebančių surinkti, apdoroti ir keistis įvairia informacija tinklas, kurio veikimas grindžiamas šiuo metu esamų ar visiškai naujai vystomų informacijos ir komunikacijos technologijų tarpusavio sąveika. Daiktų interneto kontekste daiktu laikomas fizinio arba informacijos pasaulio objektas, kuris gali būti identifikuotas ir įjungtas į komunikacinius tinklus. Kiekvienas objektas turi elektroninę žymą (angl. *Electronic Tag*) ir nuolat prijungtas prie duomenų bazės bei tinklo, todėl tokius objektus galima nesudėtingai kontroliuoti ir valdyti. Fiziniai objektai paprastai egzistuoja fiziniame pasaulyje, o informacija apie juos gaunama per stimuliuojamus ir sujungiamus jutiklius. Toks fizinis objektas, kuris įgauna komunikavimo savybes, gali jungtis į daiktų internetą. Fiziniu objektu gali būti mus supančios aplinkos dalis, kuri fiziškai sąveikauja su kitais daiktais, žmonėmis ar aplinkomis. Informacijos objektai paprastai egzistuoja informaciniame pasaulyje ir gali būti saugomi, tvarkomi ir naudojami. Toks informacijos objektas veikia išmaniojoje aplinkoje, naudoja išmaniąsias sąsajas ir turi unikalius identifikacijos atributus, kurie leidžia jį skaitmeniškai personalizuoti. Informacijos objektu gali būti skaitmeninis elementas turintis konkretų tikslą, sudarytas iš duomenų eilučių ir galintis atlikti nustatytus veiksmus. Fizinį objektą informacijos pasaulyje gali atstovauti vienas ar daugiau informacijos daiktų, o informacijos objektas gali egzistuoti ir be jokio kartu susijusio fizinio daikto (Espada *et al*. 2011).

Daiktų internetas yra suprantamas kaip informacinei visuomenei skirta pasaulinė infrastruktūra, teikianti šiuolaikines paslaugas, sujungiant objektus jau esamų ir naujai besivystančių informacinių bei komunikacinių technologijų pagrindu. Identifikuojant, renkant duomenis, juos apdorojant ir naudojantis komunikacijų galimybėmis, daiktų internetas leidžia visapusiškai panaudoti objektus įvairioms paslaugoms, užtikrinant aukštus saugumo ir privatumo reikalavimus. Daiktų interneto pagalba tikimasi sujungti į visumą naujausias technologijas kaip šiuolaikinė mašina – mašinai (angl. *Machine to Machine* – M2M) komunikacija, autonominiai tinklai, duomenų analizė ir sprendimų priėmimas, saugumo ir privatumo apsauga, debesų kompiuterija su moderniomis pojūčių ir paleidimo technologijomis. Pagrindiniai daiktų interneto požymiai (Nitti *et al.* 2016):

- sujungiamumas: viskas gali būti sujungta su pasaulio informacijos ir komunikacijų infrastruktūra;
- susiejamumas: daiktų internetas gali teikti su objektais susijusias paslaugas, atsižvelgiant į apribojimus kaip privatumo apsauga ir semantinė darna tarp fizinių ir informacijos objektų;
- heterogeniškumas: daiktų interneto objektai gali būti nevienalyčiai, nes remiasi absoliučiai skirtingomis technologinėmis platformomis ir tinklais;
- dinamiškumas: objektų būsena gali keistis dinamiškai, nes jie užmiega ir atsibunda, įjungiami arba išjungiami, o taip pat keičiasi jų vieta ir greitis;
- mastas: valdomų ir tarpusavyje komunikuojančių objektų skaičius bus daug kartų didesnis nei prie dabartinio interneto prijungtų įrenginių skaičius.

Daiktų interneto kūrimas priklauso nuo daugelio inovacinių technologijų, bet pirmiausia nuo jutiklių, mobilių įtaisų, bevielio tinklo, nanotechnologijų ir standartinio interneto. Daiktų internetą galima laikyti ateities technologijų vizija, kai realaus pasaulio objektai tampa interneto sudėtine dalimi, kur kiekvienas objektas yra tiksliai identifikuojamas, nustatoma jo vieta ir būsena, o pats objektas pasiekiamas tinkle. Šios naujos paslaugos ir jų išmanumas kelia daiktų interneto technologijoms aukšto lygio reikalavimus (Kraijak *et al.* 2015):

- identifikuotas ryšys: daiktų internetas turi užtikrinti komunikaciją tarp objekto ir daiktų interneto, paremtą pagrindiniais daikto identifikatoriais, o taip pat apimti galimą skirtingų daiktų identifikatorių įvairiarūšiškumą, kurie būtų tvarkomi vienodai;
- autonominis aprūpinimas: daiktų internetas turi suteikti paslaugas, kurias būtų galima pateikti užfiksuojant, perduodant ir automatiškai apdorojant objektų duomenis, kurie grindžiami taisyklėmis ir sukonfigūruoti operatorių, o tuo pačiu priklauso nuo automatinės duomenų sintezės ir jų gavybos;
- interoperabilumas: daiktų internetas turi apimti sąveiką tarp įvairiarūšių ir išskirstytų sistemų, naudojantis skirtinga informacija ir paslaugomis;
- autonominis tinklas: daiktų internetas turi įgalinti savarankišką tinklą, siekiant prisitaikyti prie skirtingų taikomųjų sričių, komunikacijos aplinkų, didelio įrenginių skaičiaus ir skirtingų jų tipų;
- vietos nustatymas: daiktų internetas turi naudoti objektų buvimo vietos nustatymo galimybes, nes komunikacijos ir paslaugos priklauso nuo objektų padėties informacijos, kurią gauti gali varžyti šalies įstatymai, teisės aktai arba saugumo reikalavimai;
- saugumas: daiktų internetas turi palaikyti privatumo apsaugą, nes kiekvienas objektas prijungtas prie daiktų interneto sukelia dideles grėsmes saugumui, kurios susijusios su užfiksuotų duomenų apie jų savininkus arba vartotojus konfidencialumu, autentiškumu, duomenų ir paslaugų vientisumu;
- kokybiškos ir saugios paslaugos: daiktų internetas turi garantuoti aukštos kokybės ir didelio saugumo paslaugas, susijusias su žmogaus kūnu, nes jos remiasi paslaugomis teikiamomis fiksuojant, perduodant ir apdorojant duomenis, kurie parodo žmogaus statines savybes ir dinaminę elgseną;
- savaiminis diegimas: daiktų internetas turi turėti savaiminio diegimo technologiją, kuri įgalintų automatinę kartą arba semantika grindžiamą konfigūraciją, vientisam tarpusavyje susietų objektų sujungimui su programomis ir jų reikalavimais;
- valdymas: daiktų internetas turi nustatyti pavaldumą, siekiant užtikrinti patikimas tinklo operacijas, nes objektus valdančios programos paprastai dirba automatiškai be tiesioginio žmonių dalyvavimo, tačiau jų visi veiklos procesai turi būti prižiūrimi ir kontroliuojami iš skirtingų vietų.

Viena iš svarbiausių daiktų interneto problemų yra laikoma daiktų interneto objektų identifikacija, kuri suteikia galimybę identifikuoti objektus ir nuskaityti jų informaciją be prisilietimo. Daiktų interneto objektų identifikacijos problema nėra nauja ir vienalytė, o greičiau tam tikras problemų mazgas, todėl jų sprendimui yra sukurta gausybė identifikavimo būdų ir nuolatos ieškoma naujų bei efektyvesnių (Madakam *et al.* 2015).

Šio straipsnio tikslas yra apžvelgti ir palyginti egzistuojančius daiktų interneto objektų identifikavimo metodus. Straipsnyje pirmiausia aptariama dažniausiai šiam tikslui naudojama daiktų interneto architektūra. Paskui nagrinėjami pagrindiniai daiktų interneto elementai, objektų identifikavimo protokolai ir standartai, skirti išskirstytoms aplinkoms. Toliau apžvelgiami daiktų interneto objektų identifikavimo metodai. Pabaigoje pateikiama šių metodų palyginimo lentelė.

Daiktų interneto architektūra

Prie interneto prijungtų daiktų skaičius nuolat didėja ir prognozuojama, kad iki 2020-jų metų bus apie 50 milijardų sąveikomis susietų fizinių objektų. Toks daiktų interneto objektų mastas glaudžiai sujungia realų pasaulį su skaitmeniniu informacinių technologijų pasauliu, kuris grindžiamas automatinės identifikacijos, buvimo vietos realiame laike, jutiklių ir valdiklių technologijomis. Daiktų interneto pagalba turi būti galima susieti daugybę įvairių rūšių objektų per internetą, todėl reikalinga lanksti sluoksninė architektūra (Chen *et al.* 2014).

Iki šiol pasiūlytos daiktų interneto architektūros dar nėra tapusios etaloniniu modeliu, tačiau yra nuolat vystomi pasauliniai projektai mėginantys sukurti bendrąją architektūrą, remiantis įvairiomis mokslinėmis analizėmis ir pramonės poreikiais. Viena iš tokių pasiūlytų daiktų interneto architektūros modelių yra penkių sluoksnių architektūra 1 pav. susidedanti iš verslo, taikomųjų programų, paslaugų valdymo, objektų abstrakcijos ir objektų lygmenų (Wu *et al.* 2010; Khan *et al.* 2012).

1 pav. Daiktų interneto architektūra

Verslo lygmuo valdo daiktų interneto sistemos veiklą ir paslaugas. Jo pagrindinė atsakomybė yra kurti verslo modelius, diagramas, struktūrines schemas remiantis gautais duomenimis iš taikomųjų programų lygmens, o taip pat projektuoti, analizuoti, įgyvendinti, vertinti, stebėti ir kurti su daiktų interneto sistema susijusius elementus. Šis lygmuo leidžia palaikyti sprendimų priėmimo procesus, kurie yra grindžiami didžiųjų duomenų (angl. *Big Data*) analize, o taip pat stebėti ir valdyti žemiau esančius lygmenis. Verslo lygmuo išlygina kiekvieno sluoksnio išvestį su numatytu išėjimu tam, kad padidinti teikiamų paslaugų ir išlaikyti vartotojų privatumą (Muhic *et al.* 2014).

Taikomųjų programų lygmuo suteikia vartotojui tam tikras paslaugas (pvz., gali suteikti vartotojui reikiamus temperatūros ir oro drėgmės matavimus). Jis taip pat turi galimybę suteikti aukštos kokybės išmaniąsias paslaugas, siekiant patenkinti įvairius vartotojų poreikius. Šis lygmuo apima tokias rinkas kaip išmanieji namai, pastatai, medicina ir t. t. (Perera *et al.* 2013).

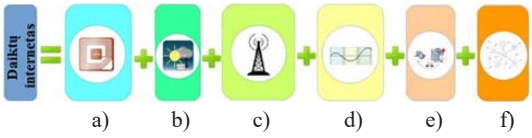
Paslaugų valdymo lygmuo arba tarpinės programinės įrangos lygmuo sujungia paslaugas su vartotoju remiantis adresais ir pavadinimais. Jis leidžia daiktų interneto taikomųjų programų programuotojui dirbti su įvairiais objektais be konkrečios techninės įrangos platformos. Šis sluoksnis apdoroja gautus duomenis, priima sprendimus ir pristato reikalingas paslaugas per tinklo protokolus (Liu *et al.* 2012).

Objektų abstrakcijos lygmuo perduoda objektų lygmenyje sukurtus duomenis į paslaugų valdymo lygmenį per saugius kanalus. Šis lygmuo suderina prieigą prie įvairių objektų duomenų naudodamas bendrą kalbą ir procedūras. Duomenys gali būti perduodami per tokias technologijas kaip RFID, 3G, GSM, UMTS, WiFi, Bluetooth, IR, ZigBee ir kt. Šiame lygmenyje tvarkomi duomenų valdymo procesai, debesų skaičiavimai ir kitos funkcijos (Mitton *et al.* 2011; Hemalatha *et al.* 2015).

Objektų lygmuo parodo daiktų interneto fizinius jutiklius, kuriais surenkama ir apdorojama informacija. Jis apima jutiklius ir valdiklius tam, kad atlikti įvairias užklausas apie būseną, drėgmę, slėgį, svorį, vietą ir pan. Standartiškai savaiminio diegimo mechanizmams yra reikalingas sąvokos lygmuo tam, kad sukonfigūruoti įvairius objektus. Sąvokos lygmuo skaitmenizuoja ir persiunčia duomenis į objekto abstrakcijos lygmenį per saugius kanalus. Šiame lygmenyje yra priimami daiktų interneto didieji duomenys (Kuyoro *et al.* 2015).

Daiktų interneto elementai

Daiktų interneto elementai 2 pav. padeda geriau suprasti tikrąją daiktų interneto reikšmę ir funkcionalumą, kurį užtikrina identifikavimas, fiksavimas, komunikacija, skaičiavimai, paslaugos ir semantika (Shah *et al.* 2016).



2 pav. Daiktų interneto elementai:
a) identifikavimas, b) fiksavimas, c) komunikacija,
d) skaičiavimai, e) paslaugos, f) semantika

Identifikavimas yra svarbus daiktų internete, norint pavadinti ir suderinti paslaugas su jų užklausomis. Adresuojant daiktų interneto objektus reikia diferencijuoti tarp objekto identifikatoriaus ir jo adreso. Objekto identifikatorius nurodo savo vardą jutikliui ir jo adresas per komunikacijos tinklą perduodamas į nurodytą vietą. Skirtumas tarp objektų identifikavimo ir adresavimo yra būtinas, nes identifikavimo būdai nėra visur vienodi. Identifikavimas yra naudojamas siekiant suteikti objektams tinkle aiškią tapatybę (Choudhary *et al.* 2016).

Fiksavimas daiktų internete reiškia duomenų rinkimą iš tarpusavyje susietų objektų tinkle ir jų siuntimą atgal į duomenų saugyklą, duomenų bazę arba debesį. Surinkti duomenys yra analizuojami norint atlikti konkrečius veiksmus, kurie yra grindžiami reikalingomis paslaugomis. Daiktų interneto jutikliai gali būti išmanūs priėmimo elementai, valdikliai arba nešiojami fiksavimo įrenginiai (Mačiulienė *et al.* 2012).

Komunikacija daiktų internete jungia tarpusavyje įvairius objektus tam, kad būtų galima teikti konkrečias išmaniąsias paslaugas. Dažniausiai daiktų interneto mazgai turi veikti naudodami mažai energijos esant komunikacijos ryšiams su galimais praradimais ir triukšmais (Miettinen *et al.* 2017).

Skaičiavimai daiktų internete yra atliekami dėka mikrovaldiklių ir taikomosios programinės įrangos. Daugelis programinės įrangos platformų yra naudojamos teikti įvairias daiktų interneto funkcijas, kur tarp jų operacinės sistemos yra labai svarbios, nes veikia visą įrenginio aktyvacijos laiką. Debesų platformos taip pat sudaro kitą svarbią daiktų interneto skaičiavimų dalį. Jos suteikia sąlygas išmaniesiems objektams siųsti savo duomenis į debesis, didiesiems duomenims būti tvarkomiems realiu laiku, o tiesioginiams vartotojams gauti naudos iš žinių, kurios buvo išgautos iš didžiųjų duomenų (Granjal *et al.* 2015).

Paslaugos daiktų internete gali būti skirstomos į tapatybės nustatymo, informacijos surinkimo, bendrąsias ir universaliąsias paslaugas. Tapatybės nustatymo paslaugos yra pagrindinės ir svarbiausios, nes kiekviena programa, kuri turi perkelti realaus pasaulio objektus į informacijos pasaulį privalo juos identifikuoti. Informacijos surinkimo paslaugos surenka ir apibendrina pradinius sensoriaus matavimus, kurie turi būti apdoroti ir perduoti į daiktų interneto programas. Bendrosios paslaugos veikia informacijos surinkimo paslaugų viršuje, naudoja išgautus duomenis priimant sprendimus ir atitinkamai reaguoja. Universaliosios paslaugos siekia suteikti bet kuriuo metu bendrąsias paslaugas, kai tik jos yra kam nors reikalingos. Pagrindinis daiktų interneto programų tikslas yra pasiekti universalių paslaugų lygį, tačiau tai nėra lengva, nes yra dar daug problemų, kurios turi būti netolimoje ateityje išspręstos (Jia *et al.* 2012).

Semantika daiktų internete reiškia visapusišką gebėjimą išmaniai išgauti žinias iš skirtingų mechanizmų tam, kad suteikti reikalingas paslaugas. Žinių gavyba apima išaiškinimą, resursų naudojimą, modeliavimo informaciją, duomenų atpažinimą ir informacijos analizavimą, norint suprasti teisingo sprendimo prasmę suteikiant tikslias paslaugas. Tokiu būdu semantika reprezentuoja daiktų interneto intelektą, siunčiant poreikius į reikiamą resursą (Gazis *et al.* 2015).

Daiktų interneto objektų identifikavimo protokolai

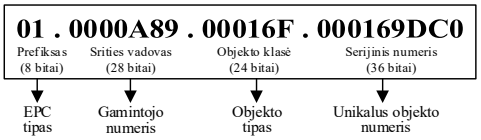
Šiuo metu egzistuoja visa aibė standartizuotų technologijų, kurios yra plačiai paplitusios, gerai ištirtos ir įvertintos įvairių kriterijų požiūriu, todėl jas suskirstysime pagal daiktų interneto elementus ir pateiksime 1 lentelėje (Al-Fuqaha *et al.* 2015).

1 lentelė. Daiktų interneto elementų technologijos

Daiktų interneto elementai		**Technologijos**
Identifikavimas	Vardas	EPC, uCode
	Adresas	IPv4, IPv6
Fiksavimas		Išmanūs jutikliai, montuojami jutikliai, nešiojami jutikliai, valdikliai, etiketės
Komunikacija		RFID, NFC, UWB, Bluetooth, Z-Wave, BLE, IEEE 802.15.4, BLE, LTE-A, WiFi, WiFiDirect
Skaičiavimai	Aparatūra	SmartThings, Intel Galileo, Cubieboard Arduino, Gadgeteer, Phidgets, Raspberry Pi, BeagleBone
	Programos	Contiki, TinyOS, LiteOS, Riot OS, Android, Nimbits, Hadoop
Paslaugos		Tapatybės nustatymo, informacijos surinkimo, bendrosios, universalios
Semantika		RDF, OWL, EXI

Elektroninis produkto kodas (angl. *Electronic Product Code* – EPC) yra unikalus identifikavimo numeris, kuris skirtas identifikuoti fizinius objektus. EPC ne tik leidžia identifikuoti objektus, bet ir gauti informaciją apie juos per informacijos paslaugų sistemą. Naudojant EPC identifikuotiems objektams priskiriami unikalūs serijos numeriai, todėl kiekvieno objekto EPC yra skirtingas. Toks būdas leidžia unikaliai, tiksliai ir konkrečiai identifikuoti atskirus objektus (An *et al.* 2012).

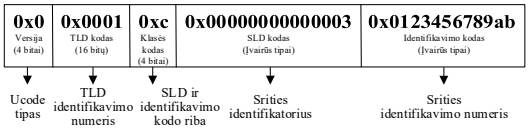
EPC struktūrą 3 pav. sudaro 64 bitų arba 96 bitų ilgio informacija, kurioje yra EPC tipas, gamintojo numeris, objekto tipas ir unikalus objekto numeris. EPC 64 bitų kodas palaiko apie 16 tūkstančių gamintojų su skirtingais numeriais, apima nuo 1 iki 9 milijonų įvairių objektų tipų ir saugo 33 milijonus unikalių serijos numerių kiekvienam objekto tipui. EPC 96 bitų kodas palaiko apie 268 milijonus gamintojų su skirtingais numeriais, apima nuo 16 milijonų įvairių objektų tipų ir saugo 68 milijardus unikalių serijos numerių kiekvienam objekto tipui (Miorandi *et al.* 2012; Guo *et al.* 2013).



3 pav. EPC struktūra

Ucode yra unikalus identifikavimo numeris, kuris skirtas identifikuoti objektus, vietas ir konceptus realiame pasaulyje. Pats Ucode nesusietas su identifikuojamu resursu ir sugeneruojamas visiškai nepriklausomai. Jo priskyrimu rūpinasi aukščiausio lygio srities (angl. *Top Level Domain* – TLD) ir antro lygio srities (angl. *Second Level Domain* – SLD) organizacijos, o klientai norėdami prieiti prie resurso, kreipiasi į specialią duomenų bazę, kuri pagal kodą grąžina nuorodą į resursą. uCode tinka identifikuoti, bet kokio tipo resursus ir visiškai nepriklauso nuo taikymo srities (Gigli *et al.* 2011).

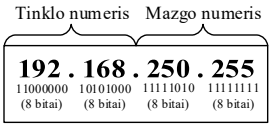
Ucode struktūrą 4 pav. sudaro 128, 256 ar 512 bitų ilgio informacija, kurioje yra versija, TLD kodas, klasės kodas, SLD kodas ir identifikavimo kodas. Versija nurodo Ucode tipą, kuris žymimas dvejetainiu skaičiumi. TLD kodas apibrėžia rezervuotą vietą, kuri skirta unikalių identifikatorių žymoms ir Ucode numeriams. Klasės kodas nusako ribą tarp SLD ir identifikavimo kodo. SLD kodas yra apibūdinamas kaip žemesnio lygio Ucode vietos domenas ir identifikuoja kiekvieną sritį. Identifikavimo kodas nustato individualų tapatybės nustatymo numerį (Lee *et al.* 2011; Sun *et al.* 2015).



4 pav. Ucode struktūra

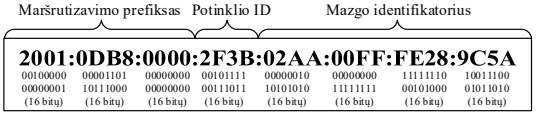
Interneto protokolas (angl. *Internet Protocol* – IP) yra taisyklių rinkinys, reglamentuojantis daiktų interneto objektų adresavimą, duomenų skaidymą į paketus prieš siunčiant ir surinkimą juos atsiuntus, o taip pat jų maršrutų parinkimą. IP kartu su perdavimo valdymo protokolu (angl. *Transmission Control Protocol* – TCP) sudaro TCP/IP protokolą, kuris yra pagrindinis interneto protokolas. Taip pat jis yra TCP/IP rinkinio protokolas adresuojantis ir maršrutizuojantis duomenis, kuris atsako už vieno duomenų paketo gabenimą iš mazgo į kitą mazgą. IP tiksliai apibrėžia, kad kiekvienas objektas turi unikalų numerį, kuris yra vadinamas interneto protokolo adresu (angl. *Internet Protocol Address*). Šiuo metu plačiausiai paplitusi ketvirtoji protokolo versija IPv4, tačiau naujausia šeštoji protokolo versija IPv6 ją palaipsniui keičia (Atzori *et al.* 2010).

IPv4 naudoja 32 bitus adresui užrašyti, o tai apima apie 2^{32} įvairių interneto objektų. IPv4 struktūrą 5 pav. sudaro 4 dešimtainiai skaičiai nuo 0 iki 255, kurie tarpusavyje atskirti taškais. IPv4 adresą sudaro tinklo numeris ir mazgo šiame tinkle numeris. Tinklo numeris skirtas tinklui pažymėti ir naudojamas maršrutizuojant paketus. Mazgo numeris yra priskiriamas konkrečiam tinklo objektui (Salman *et al.* 2014).



5 pav. IPv4 struktūra

IPv6 naudoja 128 bitus adresui užrašyti, o tai apima apie 2^{128} įvairių interneto objektų. IPv6 struktūrą 6 pav. sudaro 8 šešioliktainiai skaičiai, kurie tarpusavyje atskirti dvitaškiais. IPv6 struktūroje yra išskiriama tinklo ir mazgo dalis, kur 64 bitai naudojami tinklo daliai ir maršrutizavimui, o likusieji 64 bitai skirti mazgo identifikavimui (Jara *et al.* 2012).



6 pav. IPv6 struktūra

Daiktų interneto objektų identifikavimo standartai

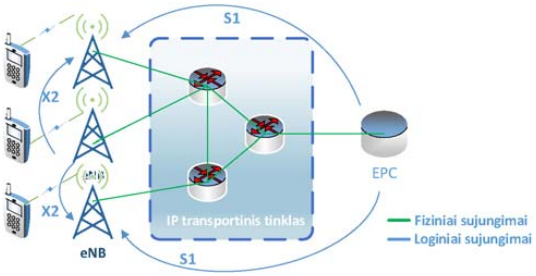
Daugelis daiktų interneto objektų identifikavimo standartų yra sukurti siekiant palengvinti taikomųjų programų programuotojų ir įvairias paslaugas teikiančių organizacijų darbus. Šie standartai tarpusavyje skiriasi duomenų sparta, veikimo nuotoliu, energijos sąnaudomis, saugumo lygiu, naudojamų dažnių ir kt. parametrais. Skirtingos tarptautinių standartų kūrimo organizacijos palaiko daiktų interneto objektų identifikaciją kaip pasaulinis žiniatinklio konsorciumas (angl. *World Wide Web Consortium* – W3C), interneto inžinerijos darbo grupė (angl. *Internet Engineering Task Force* – IETF), elektroninių produktų kodų įmonė (angl. *Electronic Product Code Company* – EPCglobal), elektrotechnikos ir elektronikos inžinierių institutas (angl. *Institute of Electrical and Electronics Engineers* – IEEE) ir Europos telekomunikacijų standartizacijos institutas (angl. *European Telecommunications Standards Institute* – ETSI). Šių tarptautinių organizacijų sukurti daiktų interneto standartai suklasifikuoti į atskiras programų, paslaugų teikimo, infrastruktūros, įtakos grupes ir pateikti 2 lentelėje (Bandyopadhyay *et al.* 2011).

2 lentelė. Daiktų interneto standartai

Programų standartai		**DDS**	**CoAP**	**AMQP**	**MQTT**	**MQTT-SN**	**XMPP**	**HTTP REST**	
Paslaugų teikimo standartai		mDNS				DNS-SD			
Infrastruktūros standartai	Maršruto lygmuo	RPL							
	Tinklo lygmuo	6LoWPAN				IPv4/IPv6			
	Ryšio lygmuo	ZigBee							
	Objektų lygmuo	LTE-A		EPC global		IEEE 802.15.4		Z-Wave	
Įtakos standartai		IEEE 1888.3			IPSec			IEEE 1905.1	

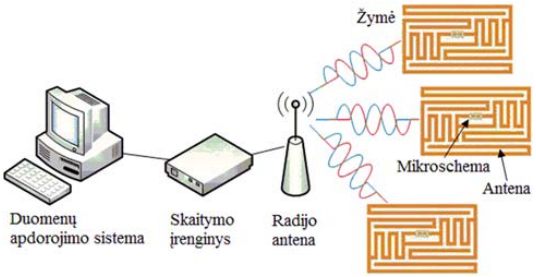
Mobiliojo ryšio standartas LTE-A (angl. *Long Term Evolution-Advanced* – LTE-A) apima daugumą tinklo komunikacijos protokolų, kurie tinka mašinų grupės komunikavimui (angl. *Machine Type Communications* – MTC), daiktų interneto infrastruktūrai ir išmaniesiems miestams. Objektų lygmenyje LTE-A standartas naudoja ortogonalų dažnių dalijimąsi daugialype prieiga (angl. *Orthogonal Frequency Division Multiple Access* – *OFDMA*), kuria kanalo dažnių juostos plotis yra padalijamas į mažesnes juostas dar vadinamas fizinių šaltinių blokais. Taip pat LTE-A standartas naudoja daugelio sudedamųjų nešlių (angl. *Component Carrier* – CC) skleidimo spektrų techniką, kuri leidžia turėti iki 5 nešančiųjų dažnių su 20 MHz dažnių juostos pločiu. Tokios perdavimo spartos turėtų būti pasiektos išlaikant esamą LTE standarto radijo dažnio spektrą be neigiamo poveikio vartotojų įrenginiams (Kocakulak *et al.* 2017).

LTE-A tinklas 7 pav. sudarytas iš dviejų pagrindinių sudedamųjų dalių. Pirmoji sudedamoji dalis yra pagrindinis tinklas (angl. *Core Network* – CN), kuris kontroliuoja mobiliuosius įrenginius ir susijęs su IP paketų srautais. Antroji sudedamoji dalis yra radijo prieigos tinklas (angl. *Radio Access Network* – RAN), kuris reguliuoja bevielio ryšio, radijo prieigos, vartotojo ir valdymo plokštumų protokolus. RAN susideda iš radijo valdymo įrangos (angl. *Integrated Network Controller* – INC), mažų ekonomiškų galinių tinklo įrenginių ir atsparių išorinėms lauko sąlygoms bazinių stočių, kurios yra tarpusavyje sujungtos X2 sąsajomis. RAN ir CN tarpusavyje yra sujungtos S1 sąsajomis. Mobilūs arba MTC įrenginiai gali prisijungti prie bazinių stočių tiesiogiai arba per MTC vartus. Jie taip pat turi tiesioginį ryšį su kitais MTC įrenginiais (Palattella *et al.* 2016).



7 pav. LTE-A tinklo struktūra

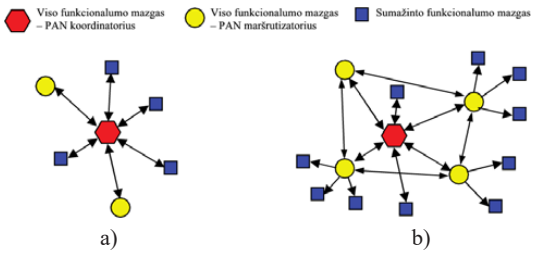
EPCglobal standartas aprašo EPC naudojimą žymėse ir brūkšniniuose koduose, kurie talpina įvairią informaciją apie nutolusius objektus. EPCglobal sistemą 8 pav. sudaro duomenų apdorojimo sistema, skaitymo įrenginys, radijo antena, žymė su įmontuota mikroschema ir nedidele antena, kurią galima pritvirtinti prie objekto, o informaciją nuskaityti radijo dažnio atpažinimo aparatūra. Nuotolinis žymės identifikavimas suteikia plačias sistemos pritaikymo galimybes gamybos, prekybos, logistikos ir kitose įmonėse (Pardal *et al.* 2010).



8 pav. EPCglobal sistemos struktūra

IEEE 802.15.4 standartas skirtas belaidžių jutiklių tinklams (angl. *Wireless Sensor Network* – WSN) sujungti, kurie yra nedidelių funkcinių galimybių. Dėl šio standarto charakteristikų, tokių kaip mažos energijos sąnaudos, nedidelė duomenų perdavimo sparta, aukštas pranešimų pralaidumas yra naudojamas IoT, M2M ir WSN. Jis suteikia patikimą ryšį, veiksmingumą tarp skirtingų platformų, valdomumą esant dideliam skaičiui mazgų, aukštą saugumo lygį, šifravimo ir autentifikavimo paslaugas. IEEE 802.15.4 standartas palaiko skirtingų dažnių kanalų juostas ir naudoja tiesinės sekos spektro sklaidos (angl. *Direct Sequence Spread Spectrum* – DSSS) technologiją. Naudojamų dažnių kanalais objektų lygmuo perduoda ir priima duomenis per tris dažnių juostas: 250 kbps prie 2,4 GHz, 40 kbps prie 915 GHz ir 20 kbps prie 868 GHz. Aukštesni dažniai ir platesnės dažnių juostos suteikia didelį pralaidumą ir mažą uždelsimą, nes žemesni dažniai suteikia geresnį jautrumą ir padengia didelius atstumus (Roman *et al.* 2011).

IEEE 802.15.4 standartas palaiko viso funkcionalumo (angl. *Full Function Device* – FFD) ir sumažinto funkcionalumo (angl. *Reduced Function Device* – RFD) tinklo mazgus. RFD mazgas paprastai priima ir siunčia jutiklio duomenis. FFD mazgai gali būti asmeninės erdvės tinklo (angl. *Personal Area Network* – PAN) koordinatoriais, maršrutizatoriais arba paprastais mazgais. PAN tinklo koordinatorius yra pagrindinis tinklo mazgas, kuris kuria tinklą, priima į jį kitus mazgus ir sinchronizuoja jų veikimą. Maršrutizatoriai gali surinkti duomenis iš kitų mazgų ir perduoti kitam maršrutizatoriui ar koordinatoriui. IEEE 802.15.4 standartas 9 pav. numato žvaigždinę (angl. *Star*) ir iš mazgo į mazgą (angl. *Peer-Peer*) tinklo topologijas. Žvaigždinės topologijos tinkle mazgai komunikuoja tiesiogiai su koordinatoriumi, todėl jame nenaudojami maršrutizatoriai. Topologijos iš mazgo į mazgą tinkle naudojami maršrutizatoriai, kurie yra komunikavimo tarpininkai, galintys persiųsti vienų mazgų pranešimus kitiems mazgams (Li *et al.* 2015).



9 pav. IEEE 802.15.4 tinklo topologijos:
a) žvaigždinė, b) iš mazgo į mazgą

Mažos galios belaidžio ryšio komunikacijos Z-Wave standartas yra skirtas namų automatizacijos tinklams (angl. *Home Automation Networks* – HAN) ir nuotoliniam objektų valdymui. Šis standartas yra naudojamas žemo duomenų perdavimo reikalaujančiose vietose, kuriose veikia mažą energijos kiekį vartojantys radijo siųstuvai ir imtuvai. Z-Wave tinklas 10 pav. veikia 900 MHz dažnio juostose ir suteikia 40 kbps duomenų perdavimo greitį (Tan *et al.* 2014).

10 pav. Z-Wave tinklo struktūra

Daiktų interneto objektų identifikavimo metodai

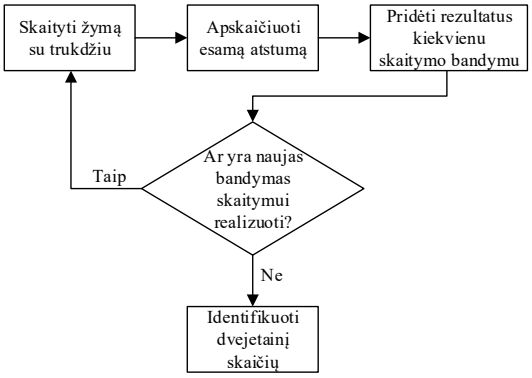
Daiktų internete įvairūs objektai yra pasiskirstę tam tikroje aplinkoje, todėl būtina turėti tam tikrą mechanizmą, kuris leistų identifikuoti ir atrasti vienas nuo kito nutolusius daiktų interneto įrenginius. Daiktų interneto objektų identifikavimo metodai leidžia vartotojams interaktyviai sąveikauti su asmeniniais daiktais turinčiais identifikatorius ir gauti šiuolaikines paslaugas. Prie tokių yra priskiriami radijo dažnio identifikavimo (angl. *Radio Frequency Identification* – RFID), bendro resursų identifikavimo (angl. *Uniform Resource Identification* – URI) ir skaitmeninio objekto identifikavimo (angl. *Digital Object Identification* – DOI) metodai (Meidan *et al.* 2017).

RFID metodas yra naudojamas daiktų interneto objektams identifikuoti ir sekti. Jį sudaro dvejetainio kodo skaitymo bandymų skaičius, fiksuotų bandymų dvejetainis kodas, dvejetainis žymos kodas, kuris yra saugojamas duomenų bazėje. RFID metodas nuskaito duomenis saugomus objekto žymoje naudodamas tris pagrindinius dažnių diapazonus (Kim *et al.* 2014):

- 13,56 MHz aukšto dažnio diapazonas (angl. *High Frequency* – HF), pasižymintis trumpu žymės duomenų nuskaitymo atstumu;
- 860 – 960 MHz ultra aukšto dažnio diapazonas (angl. *Ultra High Frequency* – UHF), galintis nuskaityti žymės duomenis vidutiniu atstumu;
- 2,45 GHz super aukšto dažnio diapazonas (angl. *Super High Frequency* – SHF), išsiskiriantis dideliu žymės duomenų nuskaitymo atstumu.

RFID metodas 11 pav. prasideda nuo žymos su trukdžiu skaitymo. Nuskaičius tokią žymą yra gaunamas dvejetainis kodas, kuris gali būti neteisingas dėl aplinkos trukdžių. Gauto su trukdžiais dvejetainio kodo atstumo skirtumas, apskaičiuojamas atsižvelgiant į žymos kodą. Kiekvienu žymos skaitymo bandymu, gautas atstumo skirtumas yra pridedamas. Sprendimas atlikti didesnį skaičių bandymų yra ankstesnės duomenų apdorojimo posistemės konfigūracijos rezultatas su galimomis parinktimis (Escobar *et al.* 2015):

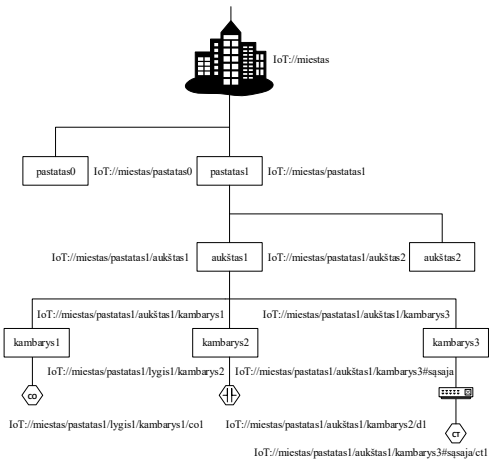
1. Realizuoti fiksuotą skaičių bandymų tam, kad būtų galima identifikuoti objekto žymą didelių trukdžių aplinkoje, kur skaitymų bandymų skaičius nėra optimizuotas konkrečiai aplinkai.
2. Realizuoti fiksuotą skaičių bandymų grindžiamų trukdžių specifinėje aplinkoje tipu ir lygiu, kur objekto žyma bus nuskaitoma.
3. Dinamiškai nustatyti bandymų skaičių grindžiamų apskaičiuotais skirtumų atstumais, kur bandymai pasibaigia kai skirtumas tarp minimalios sumos ir likusių sumų pasiekia iš anksto numatytą toleranciją.



11 pav. RFID metodas

URI metodas yra naudojamas fizinių arba informacijos daiktų interneto objektų identifikavimui. Jį sudaro simbolių seka iš amerikietiškos informacijos mainų koduotės standarto. URI metodas susideda iš daiktų interneto simbolių ir tęsiamas po dvitaškio. Du pasviri brūkšniai nurodo kelią, kuris yra koduojamo objekto loginis adresas. Kelio seka yra atskiriama pasvirais brūkšniais ir toliau aprašomos fizinių objektų erdvinės vietos. Sąsaja yra išskirta nuo ankstesnės dalies grotažyme bei atsakinga už jutiklių su internetu sujungimą. Objekto identifikacinis numeris nurodo fizinio objekto vardą ir gali būti sudarytas iš įvairių raidžių, kurių derinys turi būti unikalus (Ma *et al.* 2015).

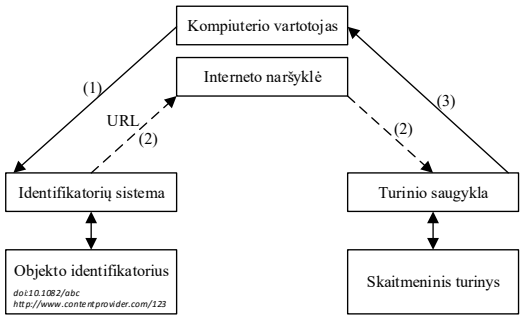
URI metodas 12 pav. charakterizuoja visame tinkle esančius fizinius objektus pagal medžio hierarchiją. Aikštė nurodo medžio šakninį mazgą, toliau esantis sluoksnis reiškia pastatą, o visi kiti sluoksniai žymi pastato grindis ir kambarius. Pirmame kambaryje įrengtas dūmų jutiklis, antrame kambaryje sumontuotas drėgmės jutiklis, o trečiame kambaryje esantis vandens nuotėkio jutiklis yra prijungtas prie tinklo (Wen *et al.* 2016).



12 pav. URI metodas

DOI metodas yra naudojamas daiktų interneto objektams identifikuoti skaitmeninėje erdvėje. Jį sudaro simbolių seka iš priešdėlio ir priesagos, kurie tarpusavyje yra atskirti pasvirusiu brūkšniu. Priešdėlis yra katalogo ir registravimo kodų tarpusavio derinys, kuris nurodo identifikuojamo daiktų interneto objekto savininką. Priesaga sudaryta iš identifikatoriaus simbolių sekos ir nustato daiktų interneto objekto informaciją, kuri priešdėliui yra visada skirtinga. Priešdėlio ir priesagos kombinacija neturi simbolių sekos ilgiui apribojimų, o taip pat ir kitoms sudedamosioms dalims, kurios yra visuotinai unikalios (Park *et al.* 2011).

DOI metodas 13 pav. susideda iš trijų etapų, kuriuos įvykdžius yra gaunami objekto identifikavimo rezultatai. Pirmame etape kompiuterio vartotojas per interneto naršyklę siunčia užklausą identifikatorius saugančiai sistemai. Ši sistema patikimai saugo nuolatinius objektų identifikatorius su jungtiniais duomenimis, kuriuos galima automatiškai pakeisti į bendrus resursų adresus (angl. *Uniform Resource Locator* – URL) su papildomais metaduomenimis. Antrame etape grąžinamas atsakymas į pateiktą kompiuterio vartotojo užklausą su tiksliais objekto identifikavimo duomenimis. Trečiame etape interneto naršyklei su identifikatorių sistemos pateiktais duomenis arba URL suteikiama prieiga prie informacijos, kurią atsiunčia turinio saugykla pagal kompiuterio vartotojo užklausą (Danilov *et al.* 2016).



13 pav. DOI metodas

Atliekant daiktų interneto identifikavimo metodų palyginimą yra naudojama stiprybių, silpnybių, galimybių, grėsmių (angl. *Strengths, Weaknesses, Opportunities, Threats* – SWOT) analizė, kurios rezultatai pateikti 3 lentelėje.

3 lentelė. RFID, URI, DOI metodų palyginimas

Metodo savybės	RFID metodas	URI metodas	DOI metodas
Stiprybė	Centralizuotas	Paprastas	Tikslus
Silpnybė	Nestandartizuotas	Išplėstas	Nederantis
Galimybė	Prisitaikantis	Trumpinamas	Plečiamas
Grėsmė	Neuniversalus	Nesaugus	Nepatikimas

Išvados

1. Daiktų internetas yra neatskiriama ateities interneto dalis, kuri apibrėžiama kaip pasaulinė ir dinaminė tinklo infrastruktūra, gebanti save konfigūruoti, grindžiama standartiniais ir suderinamais komunikacijos protokolais, kur fiziniai ir informacijos objektai turi identifikacines savybes, fizinius atributus, informacines personalijas, naudoja išmaniąsias sąsajas ir tiesiogiai komunikuoja su kitais objektais.

2. Daiktų interneto architektūra turi turėti lanksčias sąsajas, kurios užtikrintų paskirstyto tipo daiktų interneto resursų tarpusavio sąveiką, saugumą ir patikimumą, palaikytų universalius daiktų interneto objektų identifikacijos protokolus ir standartus.

3. Daiktų internetas apibrėžia išmanią aplinką, kurioje fiziniai daiktai įgauna tarpusavio komunikavimo galimybes ir tiesiogiai įjungti į vientisą informacinį tinklą, kuriame daiktai tampa aktyvūs verslo procesų dalyviai, todėl siekiant, kad daiktų interneto objektai galėtų tarpusavyje komunikuoti yra reikalingas daiktų interneto objektų identifikavimo metodas.

4. Daiktų interneto identifikacijos protokolų ir standartų apžvalga parodė, kad šias technologijas galima suskirstyti į centralizuotas identifikacijos technologijas, naudojančias centrinę duomenų bazę identifikuojamų objektų metaduomenims saugoti ir paskirstytas identifikacijos sistemas, kuriose informacija apie daiktų interneto objektų teikiamas paslaugas ir resursus yra paskirstyta daiktų interneto mazguose.

5. Daiktų interneto objektų identifikavimo metodų apžvalga ir palyginimas parodė, kad prieigos kontrolės panaudojimas leidžia apsaugoti daiktų interneto objektų duomenis nuo nesankcionuoto naudojimo ir leidžia informaciją pasiekti tik įgaliotiems vartotojams.

Literatūra

AL-FUQAHA, A.; GUIZANI, M.; MOHAMMADI, M.; ALEDHARI, M. (2015). Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications, *International Journal of IEEE Communications Surveys and Tutorials* 17(4): 2347–2376.

AN, J.; GUI, X. L.; HE, X. (2012). Study on the Architecture and Key Technologies for Internet of Things, in *Proc. of the 7th International Conference on Electrical and Computer Engineering (ICECE),* December 20–22, 2012, Dhaka, Bangladesh, 329–335.

ATZORI, L.; IERA, A.; MORABITO, G. (2010). The Internet of Things: A Survey, *International Journal of Computer and Telecommunications Networking* 54(15): 2787–2805.

BANDYOPADHYAY, D.; SEN, J. (2011). Internet of Things: Applications and Challenges in Technology and Standardization, *International Journal of Wireless Personal Communications* 58(1): 49–69.

CHEN, S.; XU, H.; LIU, D.; WANG, H. (2014). A Vision of IoT: Applications, Challenges, and Opportunities with China Perspective, *International Journal of IEEE Internet of Things* 1(4): 349–359.

CHOUDHARY, G.; JAIN, A. K. (2016). Internet of Things: A Survey on Architecture, Technologies, Protocols and Challenges, in *Proc. of the 2nd International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, December 23–25, 2016, Jaipur, India, 1–8.

DANILOV, K. N.; KULIK, V. A.; KIRICHEK, R. V. (2016). Review and Analysis of Methods of Identification and Authentication of the Internet of Things Devices, *Science Journal of Information Technology and Telecommunications* 1(3): 49–57.

ESCOBAR, L. H.; SMITH, N. R.; CRUZ, J. R. R.; BARRON, L. E. C. (2015). Methods of Selection and Identification of RFID Tags, *International Journal of Machine Learning and Cybernetics* 6(5): 847–857.

ESPADA, J. P.; MARTINEZ, O. S.; BUSTELO C. P. G.; LOVELLE, J. M. C. (2011). Virtual Objects on the Internet of Things, *International Journal of Interactive Multimedia and Artificial Intelligence* 1(4): 24–30.

GAZIS, V.; GORTZ, M.; HUBER, M.; LEONARDI, A.; MATHIOUDAKIS, K.; WIESMAIER, A.; ZEIGER, F.; VASILOMANOLAKIS, E. (2015). A Survey of Technologies for the Internet of Things, in *Proc. of the 11th International Conference on Wireless Communications and Mobile Computing (IWCMC),* August 24–28, 2015, Dubrovnik, Croatia, 1090–1095.

GIGLI, M.; KOO, S. (2011). Internet of Things: Services and Applications Categorization, *Journal of Advances in Internet of Things* 1(2): 27–31.

GRANJAL, J.; MONTEIRO, E.; SILVA, J. S.; (2015). Security for the Internet of Things: A Survey of Existing Protocols and Open Research Issues, *International Journal of IEEE Communications Surveys and Tutorials* 17(3): 1294–1312.

GUBBI, J.; BUYYA, R.; MARUSIC, S.; PALANISWAMI, M. (2013). Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions, *Journal of Future Generation Computer Systems* 29(7): 1645–1660.

GUO, Z. J. (2013). Research on Key Technologies and Application of Internet of Things, in *Proc. of the 25th Chinese Control and Decision Conference (CCDC)*, May 25–27, 2013, Guiyang, China, 2797–2801.

HEMALATHA, D.; AFREEN, B. E. (2015). Development in Radio Frequency Identification Technology in Internet of Things, *International Journal of Advanced Research in Computer Engineering and Technology* 4(11): 4030–4038.

JARA, A. J.; ZAMORA, M. A.; SKARMETA, A.; (2012). Glowbal IP: An Adaptive and Transparent IPv6 Integration in the Internet of Things, *International Journal of Mobile Information Systems* 8(3): 177–197.

JIA, X.; FENG, Q.; FAN, T.; LEI, Q. (2012). RFID Technology and its Applications in Internet of Things (IoT), in *Proc. of the 2th International Conference on Consumer Electronics, Communications and Networks (CECNet)*, April 21–23, 2012, Yichang, China, 1282–1285.

KHAN, R.; KHAN, S. U.; ZAHEER, R.; KHAN, S. (2012). Future Internet: The Internet of Things Architecture, Possible Applications and Key Challenges, in *Proc. of the 10th Conference on Frontiers of Information Technology (FIT)*, December 17–19, 2012, Islamabad, Pakistan, 257–260.

KIM, K. J.; HONG, S. P. (2014). A Study on the Development Method for Trust-Based Activation in Internet of Things, *International Journal of Contemporary Engineering Sciences* 7(31): 1715–1721.

KOCAKULAK, M.; BUTUN, I. (2017). An Overview of Wireless Sensor Networks Towards Internet of Things, in *Proc. of the 7th Annual Computing and Communication Workshop and Conference (CCWC),* January 9–11, 2017, Las Vegas, Nevada, USA, 1–6.

KRAIJAK, S.; TUWANUT, P. (2015). A Survey on Internet of Things Architecture, Protocols, Possible Applications, Security, Privacy, Real-World Implementation and Future Trends, in *Proc. of the 16th International Conference on Communication Technology (ICCT),* October 18–20, 2015, Hangzhou, China, 28–31.

KUYORO, S.; OSISANWO, F.; AKINSOWON, O. (2015). Internet of Things (IoT): An Overview, in *Proc. of the 3th International Conference on Advances in Engineering Sciences and Applied Mathematics (ICAESAM),* March 23–24, 2015, London, United Kingdom, 53–58.

LEE, G. M.; CRESPI, N. (2011). Internet of Things for Smart Objects: Ubiquitous Networking between Humans and Objects, in *Proc. of the 4th International Conference on Advanced Infocomm Technology (ICAIT), July 11–14, 2011, Wuhan, China, 588–592.*

LIU, J.; XIAO, Y.; PHILIP. C. L. (2012). Authentication and Access Control in the Internet of Things, in *Proc. of the 32nd International Conference on Distributed Computing Systems Workshops,* June 18–21, 2012, Macau, China, 588–592.

LI, S.; XU, L. D.; ZHAO, S. (2015). The Internet of Things: A survey, *International Journal of Information Systems Frontiers* 17(2): 243–259.

MA, R.; LIU, Y.; SHAN, C.; ZHAO, X. L.; WANG, X. A. (2015). Research on Identification and Addressing of the Internet of Things, in *Proc. of the 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing,* November 4–6, 2015, Krakow, Poland, 810–814.

MAČIULIENĖ, M. (2012). Power Through Things: Following Traces of Collective Intelligence, *Research Journal of Social Technologies* 4(1): 168–178.

MADAKAM, S.; RAMASWAMY, R.; TRIPATHI, S. (2015). Internet of Things (IoT): A Literature Review, *International Journal of Future Computer and Communication* 3(5): 164–173.

MEIDAN, Y.; BOHADANA, M.; SHABTAI, A.; GUARNIZO, J. D.; OCHOA, M.; TIPPENHAUER, N. O.; ELOVICI, Y. (2017). ProfilIoT: A Machine Learning Approach for IoT Device Identification Based on Network Traffic Analysis, in *Proc. of the 32nd ACM SIGAPP Symposium On Applied Computing (SAC)*, April 4–6, 2017, Marrakech, Morocco, 506–509.

MIETTINEN, M.; MARCHAL, S.; HAFEEZ, I.; SADEGHI, A. R.; ASOKAN, N.; TARKOMA, S. (2017). IoT Sentinel: Automated Device-Type Identification for Security Enforcement in IoT, in *Proc. of the 37th IEEE International Conference on Distributed Computing Systems (ICDCS)*, June 5–8, 2017, Atlanta, Georgia, USA, 53–58.

MIORANDI, D.; SICARI, S.; DE PELLEGRINI, F.; CHLAMTAC, I. (2012). Internet of Things: Vision, Applications and Research Challenges, *Archival Journal of Ad Hoc Networks* 10(7): 1497–1516.

MITTON, N.; RYL, D. S. (2011). From the Internet of Things to the Internet of the Physical World, *Journal of Comptes Rendus Physique* 12(7): 669–674.

MUHIC, I.; HODZIC, M. I. (2014). Internet of Things: Current Technological Review, *Journal of Periodicals of Engineering and Natural Sciences* 2(2): 1–8.

NITTI, M.; PILLONI, V.; COLISTRA, G.; ATZORI, L. (2016). The Virtual Object as a Major Element of the Internet of Things: A Survey, *Journal of IEEE Communications Surveys and Tutorials* 18(2): 1228–1240.

PALATTELLA, M. R.; DOHLER, M.; GRIECO, A.; RIZZO, G; TORSNER, J.; ENGEL, T.; LADID, L. (2016). Internet of Things in the 5G Era: Enablers, Architecture and Business Models, *IEEE Journal on Selected Areas in Communications* 34(3): 510–527.

PARDAL, M. L.; MARQUES, J. A. (2010). Towards the Internet of Things: An Introduction to RFID Technology, in *Proc. of the 4th International Workshop on RFID Technology (IWRT), June 8–12, 2010, Madeira, Portugal, 30–39.*

PARK, S.; ZO, H.; CIGANEK, A. P.; LIM, G. G. (2011). Examining Success Factors in the Adoption of Digital Object Identifier Systems, *Journal Electronic Commerce Research and Applications* 10(6): 626–636.

PERERA, C.; ZASLAVSKY, A.; CHRISTEN, P.; GEORGAKOPOULOS, D. (2013). Context Aware Computing for The Internet of Things: A Survey, *International Journal of IEEE Communications Surveys and Tutorials* 16(1): 414–454.

ROMAN, R.; ALCARAZ, C.; LOPEZ, J.; SKLAVOS, N. (2011). Key Management Systems for Sensor Networks in the Context of the Internet of Things, *International Journal of Computers and Electrical Engineering* 37(2): 147–159.

SALMAN, M. A. (2014). On Identification of Internet of Things, *International Journal of Sciences: Basic and Applied Research* 18(1): 59–62.

SUN, Y.; BIE, R.; THOMAS, P.; CHENG, X. (2015). Theme issue on advances in the Internet of Things: identification, information, and knowledge, *Journal of Personal and Ubiquitous Computing* 19(7): 985–987.

SHAH, S. H.; YAQOOB, I. (2016). A Survey: Internet of Things (IoT) Technologies, Applications and Challenges, in *Proc. of the 4th IEEE International Conference on Smart Energy Grid Engineering (SEGE),* August 21–24, 2016, Oshawa, Canada, 381–385.

TAN, J.; KOO, S. G. M. (2014). A Survey of Technologies in Internet of Things, in *Proc. of the 10th IEEE International Conference on Distributed Computing in Sensor Systems*, May 26–28, 2014, Marina Del Rey, California, USA, 269–274.

WEN, Y.; JINLONG, W.; QIANCHUAN, Z. (2016). Physical Objects Registration and Management for Internet of Things, in *Proc. of the 35th Chinese Control Conference,* July 27–29, 2016, Chengdu, China, 8335–8339.

WU, M.; LU, T. J.; LING, F. Y. (2010). Research on the Architecture of Internet of Things, in *Proc. of the 3rd International Conference on Advanced Computer Theory and Engineering,* August 20–22, 2010, Chengdu, China, 484–487.

IDENTIFICATION OF OBJECTS OF INTERNET OF THINGS

Raimundas Savukynas

S u m m a r y

The resources of the Internet of Things and services are globally distributed in information network, so it is necessary to have a mechanism which allows the identification and discover of smart devices, their services and resources. Typically, the identification consists of the assigning names for the resources and the mechanism of the addressing resources in order to discover and access the remote resources of the Internet of things. Modern networks use the Digital Object Identification (DOI) and Universal Resource Identifier (URI), who identify a global resource, its type and a set of equivalent names. They uniquely allows to identify physical objects, which are located anywhere in the world, using the Electronic Product Code (EPC) technology. EPC allows DOI and URI also to use with the Radio-Frequency Identification (RFID). RFID technology is used for marking and tracing items, which is based on the use of the frequency signal of the radio in the object tag for the information recording and reading. The purpose of this article is to review and compare RFID, URI, DOI identification methods of objects of the Internet of Things.

Keywords: *internet of things, objects identification, future internet, internet services, smart devices.*

INFORMACINĖS TECHNOLOGIJOS. XXIV tarpuniversitetinės tarptautinės magistrantų ir doktorantų konferencijos „Informacinė visuomenė ir universitetinės studijos“ (IVUS 2019) medžiaga, 2019 m. balandžio 25 d., Kaunas, Lietuva / INFORMATION TECHNOLOGY. Proceedings of the XXIV International Master and PhD Conference “Information Society and University Studies” (IVUS 2019), 25th April, 2019, Kaunas, Lithuania. – Kaunas: Vytauto Didžiojo universiteto leidykla, 2019. – 210 p., iliustr.

ISSN 2029-249X
ISSN 2029-4824 (elektroninis leidinys)

INFORMACINĖS TECHNOLOGIJOS

XXIV tarpuniversitetinės tarptautinės magistrantų ir doktorantų konferencijos „Informacinė visuomenė ir universitetinės studijos“ (IVUS 2019) medžiaga, 2019 m. balandžio 25 d., Kaunas, Lietuva

INFORMATION TECHNOLOGY

Proceedings of the XXIV International Master and PhD Students Conference “Information Society and University Studies” (IVUS 2019), 25th April 2019, Kaunas, Lithuania

Editor Rūta Užupytė
Graphic designer Rasa Švobaitė
Approved for printing 2013 04 12. A run of 50 copies. (No of P. 210)
Order No. K13-031.
Published: Vytautas Magnus University Publishing Office, S. Daukanto g. 27, LT-44249 Kaunas. SL 1557

Redaktorė Rūta Užupytė
Leidinio dailininkė Rasa Švobaitė
Pasirašyta spausdinti 2013 04 12. Tiražas 50 egz. 13,25 leidyb. apsk. l. (210 p.)
Užsakymo Nr. K13-031
Išleido: Vytauto Didžiojo universiteto leidykla, S. Daukanto g. 27, LT-44249 Kaunas. SL 1557