REZEKNE HIGHER EDUCATION INSTITUTION
FACULTY OF ENGINEERING

ISSN 2256-070X

ENVIRONMENT. TECHNOLOGY. RESOURCES

Proceedings of the 10th International Scientific and Practical Conference
June 18-20, 2015

VOLUME III

Rezekne
2015

ENVIRONMENT. TECHNOLOGY. RESOURCES: Proceedings of the 10th International Scientific and Practical Conference June 18-20, 2015. *Volume III.* Rezekne: Rēzeknes Augstskola, 2015. p 240.

Recommended for publication by the Scientific Council of Rezekne Augstskola on June, 3rd, 2015.

Regular scientific collected papers include reports of the 10th International Scientific Practical Conference "Environment.Technology.Resources." The themes of the papers reflect the modern dynamic changes in the complex interplay between humanity and nature. The results of the studies included in the collected papers, show that environmental science changes its vector from a passive nature protection and environmental decontamination to an active prevention of sources of the negative effects through using the latest solutions in technologies, mechanics, computer sciences, rationalising the complex use of resources, including at the same time natural, as well as technogenic energy and material resources.
Three volumes of collected papers represent the research in environmental protection, biology and ecological farming, energy, materials science, mechanics, metalworking, laser technologies, mathematical modelling, electrical engineering, environmental economics and management, computer technologies and modelling socio-technical systems, intellectual decision support systems, environmental education and sustainable development processes.

The author of the paper takes resposibility for the content of the paper.

ISSN 2256-070X

Conference Organizing Committee:

Chairman:

Dr.habil.geol. Gotfrids Noviks - Latvia, Rezekne Higher Education Institution

Co - Chairman:

Dr.sc.ing. Edmunds Teirumnieks - Latvia, Rezekne Higher Education Institution

Members:

Dr.habil.sc.ing. Dagnija Blumberga - Latvia, Riga Technical University
Dr.habil.dat. Juris Borzovs - Latvia, University of Latvia
Dr. Xiao Chen - China, Hefei University
Dr.-Ing. Dr. h.c. Horst Exner - Germany, Mittweida University of Applied sciences
Dr.sc.ing. Egils Ginters - Latvia, Vidzeme University of Applied Sciences
Dr.sc.ing. Peteris Grabusts - Latvia, Rezekne Higher Education Institution
DSc. Dimitar Grekov – Bulgaria, Agricultural University - Plovdiv
Dr.habil.sc.ing. Janis Grundspenkis - Latvia, Riga Technical University
DSc. Raycho Ilarionov - Bulgaria, Technical University of Gabrovo
DSc. Hristo Ivanov Beloev - Bulgaria, "Angel Kanchev" University of Ruse
Dr.habil.chem. Maris Klavins - Latvia, University of Latvia
Dr. Linas Kliucininkas - Lithuania, Kaunas University of Technology
Dr.sc.ing. Eriks Kronbergs - Latvia, Latvia University of Agriculture
Dr. Walter Leal - Germany, Hamburg University of Applied Sciences
PhD Eng. Valeri Markov Mladenov – Bulgaria, Technical University of Sofia
Dr.sc.ing. Andris Martinovs - Latvia, Rezekne Higher Education Institution
Dr.sc.ing. Vladislav Morozov - Russia, Geophysical Centre of the Russian Academy of Sciences
Dr.geogr. Olgerts Nikodemus - Latvia, University of Latvia
Dr.habil.sc.ing. Igor Plokhov - Russia, Pskov State University
Vilnis Rantins - Latvia, Chairman of the Management Board of the Association of Mechanical Engineering and Metalworking Industries
Dr.sc.ing. Artis Teilans - Latvia, Rezekne Higher Education Institution
Tekn. Dr., Dr.eng. h.c. Ralejs Tepfers - Sweden, Chalmers University of Technology
Dr.-Ing. Josef Timmerberg - Germany, Jade University of Applied Sciences
Dr.sc.ing. Olegs Uzga - Rebrovs - Latvia, Rezekne Higher Education Institution
PhD Pepo Yordanov – Bulgaria, Technical University of Sofia Branch Plovdiv
Dr. Juri-Rivaldo Pastarus - Estonia, Tallinn University of Technology

Reviewers:

Dr.sc.ing. Ludmila Aleksejeva - Latvia, Riga Technical University
Dr.sc.ing. Janis Andersons - Latvia, University of Latvia
PhD Andres Annuk - Estonia, Estonian University of Life Sciences
Dr.habil.sc.ing. Dagnija Blumberga - Latvia, Riga Technical University
Dr.sc.ing. Vitalijs Bolsakovs - Latvia, Riga Technical University
Dr.habil.sc.comp. Juris Borzovs - Latvia, University of Latvia
Dr. Rene Brunsch - Germany, Apromace Data Systems GmbH
Dr.sc.ing. Edgars Cubars - Latvia, Rural Support Service
Dr. Kevin Curran - Northern Ireland, University of Ulster
Dr.eng. Hristina Deneva - Bulgaria, Technical University of Gabrova
Dr.eng. Margarita Deneva - Bulgaria, Technical University of Sofia
PhD Vitalij Denisov – Lithuania, Klaipeda University
Dr. Maija Dimitrivo - Bulgaria, Agricultural University - Plovdiv
Dr.oec. Sandra Ezmale - Latvia, Rezekne Higher Education Institution
Dr.sc.ing. Ainars Galins - Latvia, Latvia University of Agriculture
Dr.sc.ing. Egils Ginters - Latvia, Vidzeme University of Applied Sciences
Dr.sc.ing. Vladimirs Gonca - Latvia, Riga Technical University
Dr.sc.ing. Aleksandrs Gorbunovs - Latvia, Riga Technical University
Dr.sc.ing. Peteris Grabusts - Latvia, Rezekne Higher Education Institution
Dr.sc.ing. Julija Gusca - Latvia, Riga Technical University
Dr.sc.comp. Anita Jansone - Latvia, Liepaja University
Dr.agr. Imants Jansons - Latvia, Latvia University of Agriculture
Dr.sc.ing. Aivars Kakitis - Latvia, Latvia University of Agriculture
Dr.paed. Ilmars Kangro - Latvia, Rezekne Higher Education Institution
PhD Atis Kapenieks - Latvia, Riga Technical University
Dr.paed. Aivars Kaupuzs - Latvia, Rezekne Higher Education Institution
Dr. Lars Kempt - Germany, Technical University Chemnitz
Dr.habil.chem. Maris Klavins - Latvia, University of Latvia
Dr.sc.ing. Olga Kononova - Latvia, Riga Technical University
Dr.biol. Georgs Kornilovs - Latvia, Institute of Food Safety, Animal Health and Environment BIOR
Dr. Sveta Kostadinova - Bulgaria, Agricultural University - Plovdiv

Mg.chem. Inese Bernane - Latvia, Rezekne Higher Education Institution

III

INFORMATION TECHNOLOGIES

CONTENT

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 11-16

Latvian language as a code in different communication channels

Linda Bajarune[1], Andris Ozols[2]
[1]Liepaja University,
[2]Riga Technical University

Abstract. **This paper is dedicated to analyze of Latvian language as a code in such literary communication channels like press, poet, prose, legal literature. Calculations for zero-order, first-order, second-order and third-order Shannon entropy have been made and also corresponding values of redundancy and compression coefficients have been determined. All the calculations are done with a self-made computer program. Different communication channels of Latvian language are compared mutually and also Latvian language is compared with English and Russian as codes.**

Keywords: **compression coefficient, Latvian language, redundancy, Shannon entropy.**

I INTRODUCTION

Nowadays as we speak we are saying that we live in the age of information. Information concept usually associated with the two objects - the source of information and the consumer's existence. The information is very difficult to define. Accurate and in all cases acceptable definition of information is not even created. Information processes and studies are very frequent occurrence. But Latvian language so far had not yet been analyzed from the point of view of communication theory.

Natural language is one of the main ways how to communicate. As the system of signs it is a tool to send and receive information. With its system of signs, symbols and rules of their combinations, connections and typesetting, language is a unique communication code and according to that other nonverbal codes are being used.

From its beginning language also has been a tool for information storage and makes the structure and navigation system for this stored information.

In space overloaded with information where modern technologies allows very tight interaction between people whose location is very far from each other information flows in such enormous speed. Approach how the language has been researched has to be changed and also rules for natural language have to be made than language could develop not like some abstract tool for abstract communication, but as rich and creative instrument that can be used to store and transmit information about this new and fast changing reality. That is why not only research about usage of language has been made but also it is important to pay attention to its statistic properties [1].

In 1948 C.Shannon in his article "A Mathematical Theory of Communication" developed information theory revealing the most important aspects of communication systems. The two main concepts of this theory are the concepts of probability and coding. In his theory C.Shannon introduces so called structural information. Shannon completely ignores whether the text is important, correct, incorrect or irrelevant. Similarly, questions about information senders and recipients are excluded. It is also irrelevant whether the text is logical and meaningful or letters are selected at random. Here appears a paradox - randomly selected letters provide the maximum information, whereas the text with a greater meaning and linguistic diversity corresponds to smaller information value [2].

From the point of view of communication theory, any language is a code. How informative is this code is characterized by its Shannon entropy. If diversity of language is greater, Shannon entropy also is higher.

A natural language can be considered as a complex system since the succession of its symbol units (letters, syllables or words) inside a text obeys some rules (grammatical or syntactical), which, however, are of probabilistic nature allowing the insertion of randomness in the text structure [12].

In this article, Latvian language has been studied as code of literary communication channels such as the press, poetry, prose and legal literature. It was a challenge to do something that has been never done before – to calculate the entropy of Latvian language, compare it with other languages and try to analyze it.

II MATERIALS AND METHODS

Entropy is a quantitative measure of uncertainty in thermodynamics and information theory. Entropy concept is used in various information optimal

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.182

encoding problem studies. The concept of structural information which is used in communication theory describes how much randomness is in a random event, how probable it is. In a communication theory this event is called a message. It is assumed that an ensemble of messages (e.g., letters, ciphers, pixels, etc.) is transmitted over the communication channel. These messages are determined by the code which is used. Each message i is characterized by its probability, p_i. The sum of all probabilities equals to one.

Shannon entropy is the average amount of information contained in a message Entropy is a quantity which depends only on the statistic nature of the information source expressed in message probabilities.

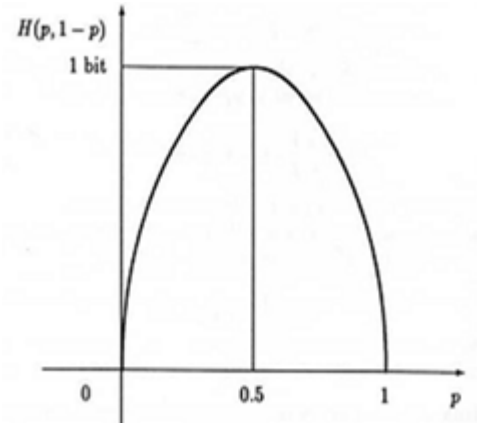

Fig.1. Entropy of two-message ensemble versus probability of one message.

In our article we consider entropy of an arbitrary message ensemble $X = (X_1, X_2, X_3, X_4....X_m.)$ which is alphabet and messages are letters.

Entropy characteristics:

1) $H(X) \geq 0$
2) Entropy is additive for two independent message ensembles X and Y
   $H(X + Y) = H(X) + H(Y)$,
3) Entropy is a limited quantity:
   $H \leq H_{max} = \log_2 m$.

We shall consider ensembles as Markov sources where the probability of each message depends on the appearance of previous messages. The appearance of a certain letter in the text generally depends on the previous letters .Thus we have to use the conditional entropies of different orders.

Zero-order entropy $H(0)$ does not take into account any interdependence of messages (they are assumed to be independent and with equal probability $1/m$) and it is calculated by the following formula

$$H(0) = \log_2 m \quad \text{bits/symbol,} \qquad (1)$$

where m is the number of letters in the alphabet and also the space between words, because entropy is being calculated to the written language. It can be shown that $H(0)=H_{max}$.

First-order entropy $H(1)$ also does not take into account the interdependence of messages, but in this case their probabilities are not equal:

$$H(1) = \sum_{i=1}^{m} p_i \log_2 p_i \quad \text{bits/symbol,} \qquad (2)$$

where p_i – probability of the message I (letter). The first-order entropy dependence on the message probability is shown in Fig.1 for the special case of the ensemble of two messages with probabilities p and $1-p$**.** It is seen that entropy maximum is achieved in the case of equal message probabilities, i.e., in this case $H(1)=H(0)=H_{max}$ as stated before.

Second-order entropy is calculated by formula

$$H(2) = \sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{j|i} \log_2 p_{j|i} \quad \text{bits/symbol,} \qquad (3)$$

where pj | i - j conditional probability of the letter, if the former has been the letter i. To calculate the second-order entropy must take into account what the symbol stands before the symbol, or all possible combinations of the two symbols.

Third order entropy is calculated by formula

$$H(3) = -\sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{j|i} \sum_{k=1}^{m} p_{k|j,i} \log_2 p_{k|j,i} \qquad (4)$$

It is possible to calculate and report the source of redundancy, if entropy has been calculated.

$$\rho = 1 - \frac{H(A)}{H_{\max}(A)} = 1 - \frac{H_n(A)}{H_0(A)} \qquad (5)$$

With information redundancy understands the duplicate or collateral data activation in system data blocks which the withdrawal does not detract from the adequacy of an array of real objects they describe. [5] Redundancy in information theory is the number of bits that are used to send the message minus the fair amount of information in bits. Data compression is a way to exclude unwanted redundancy, but if the message shall be carried out in a noisy channel with a limited capacity, then the redundancy is desirable.

So redundancy in our language is the words that we say, but even without these words, the information is comprehensible. About that we can make sure every time when we send text messages, trying to say as much as possible with the least possible symbols to convene only one message - words, letters are omitted or even written without spaces, but the text still is understandable.

Txt wtht vwls. TextWithoutSpacing. –if we can read and still understand this text, which is written without vowels and without spacing, that means that vowels or spaces are redundant for this message.

But redundant words or letters are because we can understand the information in a noisy place. For example, where is salt package? If in the moment when that being said drives noisy large car, word salt

can sound like some other word, than redundant word – package – will help to understand the real meaning of the sentence. This example is very simple, but in most of cases if in the sentence one word is missing, we can guess this word. If one letter is missing, clearly we can guess this letter.

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Or rather...According to a researcher (sic) at Cambridge University, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter be at the right place. The rest can be a total mess and you can still read it without problem. This is because the human mind does not read every letter by itself but the word as a whole [6].

If 50% of language is redundant than it is possible to save 50% investment what needs to send electronic message in this language. Something similar happens when the file is compressed. If only there is noise somewhere in transmission and one of the symbols from compressed file is destroyed, then it is not possible to repair original file.

In the digital coding redundancy plays an important role in using encryption with numbers of even ones. For letter A in binary system stands 01000001. So, to transmit letter A we need 8 bits to send these 8 symbols. But if the line is with interferences and we receive combination with mistake – 010000?1, we cannot tell anymore which this letter is. It can be A if the missing symbol is 0, but it can be C if the missing symbol is 1. Of course in the normal context would not have problems to understand it, but if redundancy would be already used and file is compressed? In that case we should add parity bit. That would be another redundant bit, but with it would be possible to solve the problem. If the sum of numbers is even, than 0 is added, but if odd, than number 1 is received. So if we receive 010000?10, added 0 tells us that we should receive A (0100001), but if we receive 010000?11, then added number 1 tells us that we should receive letter C (01000011).

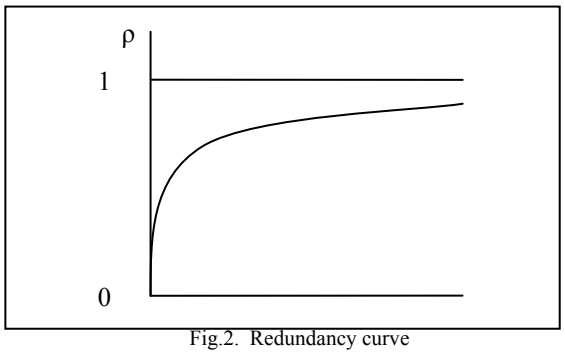

Fig.2. Redundancy curve

If really noisy channel is expected, it is possible to come to an agreement to send parity bits after every 4 bits. It could seem unnecessary to send redundant bits, which do not need. But if we are compressing text form 10 000 to 8000 symbols, excluding redundant symbols, for every transmitted sign we should add also parity bite – it would be 8000 parity bits. 8000 parity bits are equivalent 1000 symbols, which mean that it is more profitable [7].

Redundancy is closely related to compression coefficient. Figure Nr.2 shows the curve of redundancy. It calculates by formula:

$$r = \frac{H(A)}{H_{\max}(A)} = \frac{H_n(A)}{H_0(A)} = 1 - \rho \qquad (6)$$

In language as a message can consider:

1) Letter
2) Word
3) Sentence

In Latvian language sound conforms to letter, that is why letters are considered as code combination, which conforms to the message- sound.

Letters are the basic of all language, so the letters are chosen as code combination. Language is code which is used to transmit information, and letters are code combinations for information coding.

In this paper first, second and third order entropy results for three information channels - press, poet, prose, legal literature. - are given.

To make calculations computer program was made. Program was based on web page, with php coding language, and also little bit of html language. All the calculations was made with php script and MySQL data base management system [8].

The principle how program is working

1) Analyzed text is written or copied in the input field. After adding text, have to press button "Add"
2) Program replaces all capital letters with small letters, takes off all the punctuation and replaces spacing with symbol „_"
3) Text is processed forming combinations of three letters, two letters and one letter. At first in the data base program writes all the new combinations, but if the combination is found repeatedly, then the number of combination is increasing by one.
4) When analysis of the text is done, the number of combinations is placed in formulas to calculate first, second and third order of entropy. Results H(1), H(2) and H(3) are shown on the screen.

III RESULTS AND DISCUSSION

Fig.3 shows average results of entropy from all analyzed .texts.

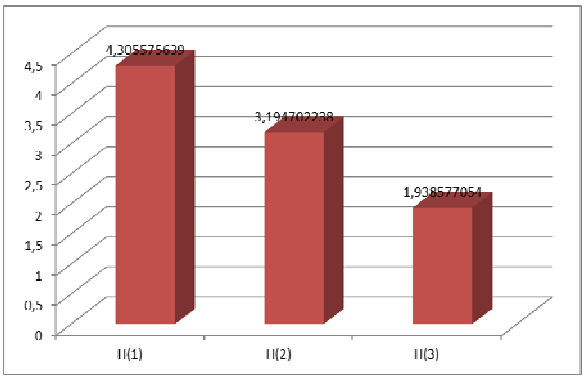

Fig.3. Average entropy in bits.

Table 1 shows entropy of Latvian language calculated by the author of this paper, entropy of English language calculated by C.E.Shannon and entropy of Russian language given by professor A.Ozols in his lectures [9]. Entropy of Latvian language is the largest, because Latvian language has more letters in the alphabet, but the value of entropy decrease faster, because combinations repeats more rarely and the compression also is the greatest.

TABLE I

COMPARISON OF ENTROPY VALUES

H(N)	Latvian	English	Russian
H(0)	5.0875	4.75	5
H(1)	4.3056	4.07	4.05
H(2)	3.1947	3.36	3.52
H(3)	1.9386	2.77	

Using values of entropy it is possible also to calculate redundancy and compression coefficient.

$$\rho(H1)=1-\frac{H(A)}{H_{\max}(A)}=1-\frac{H_n(A)}{H_0(A)}=1-\frac{4,3056}{5,0875}$$

$$=0,1537=15,4\%$$

$$\rho(H2)=1-\frac{H(A)}{H_{\max}(A)}=1-\frac{H_n(A)}{H_0(A)}=1-\frac{3,1947}{5,0875}$$

$$=0,3720=37,2\%$$

$$\rho(H3)=1-\frac{H(A)}{H_{\max}(A)}=1-\frac{H_n(A)}{H_0(A)}=1-\frac{1,9386}{5,0875}$$

$$=0,6189=61,9\%$$

Redundancies corresponding to the entropies of three orders are shown in Figure nr.4.

Comparing redundancy of Latvian language with English language and Russian language, Latvian language has the highest redundancy. It is because written Latvian language is understandable also without signs of lengthening, cedillas and without vowels. Higher entropy is possible to achieve if diversity is higher, which means combinations repeats differently, also that increases the value of entropy. If the most frequently used combination in text is dropped out, it would be possible to understand the text anyway.

TABLE II

VALUES OF REDUNDANCY

H(N)	Latvian	English	Russian
H(1)	15,4%	14,3%	19%
H(2)	37,2%	29,2%	29,6%
H(3)	61,9%	41,7%	

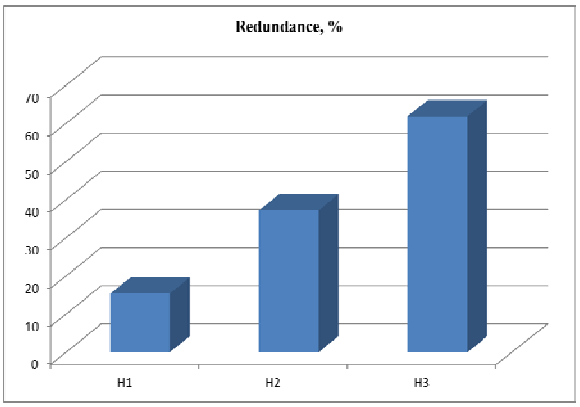

Fig.4. Redundancies corresponding to the entropies of three orders.

To define what would be the most exact redundancy, at which order of entropy it should be calculated, authors had to calculate average length of word. That is 4.52 and was calculated manually from one of the chosen texts of prose. Rounding up it shows that 5th order of entropy should be calculated. Text breakdown by word lengths, see Figure No.5.

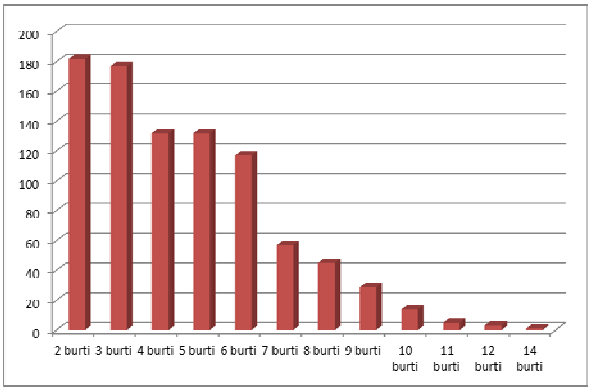

Fig.5. Word frequency versus word length.

Compression coefficient results:

At H(1) $r=\frac{H(A)}{H_{\max}(A)}=\frac{H_n(A)}{H_0(A)}=1-\rho=0,8463$

At H(2) $r=\frac{H(A)}{H_{\max}(A)}=\frac{H_n(A)}{H_0(A)}=1-\rho=0,628$

At H(3) $r=\frac{H(A)}{H_{\max}(A)}=\frac{H_n(A)}{H_0(A)}=1-\rho=0,3811$

Results also are displayed in Figure nr.6.

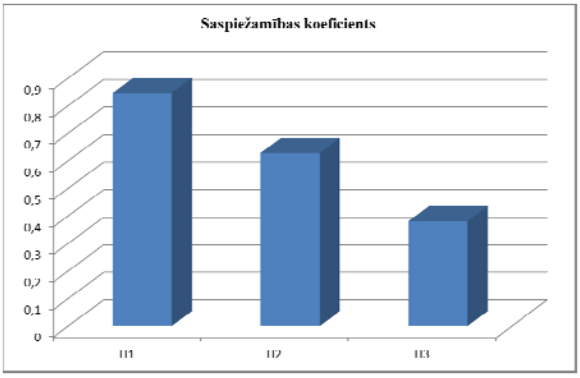

Fig.6. Compression coefficient

Average values of entropy in different information channels are summarized in Figure nr.7. First order entropy shows that the highest entropy is for press channel, followed by legal literature, then prose and the last poetry. However, the second order entropy the smallest is for legal literature but the highest for press and prose. Second-order entropy shows one letter dependence before standing letter. The higher the entropy, the greater is the appearance probability for two-letter combination. The third order entropy calculates sum of probabilities, if the appearance of one letter is dependent on before standing two letters. The smallest is the third order entropy for legal literature, followed by poetry, but the highest entropy is for prose.

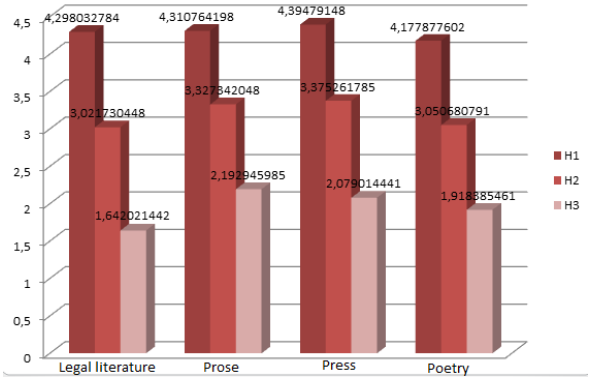

Fig.7. Average values of entropy in different information channels

IV CONCLUSIONS

To calculate the most precise results entropy of Latvian language was calculated from different information sources. All the calculations were made with web based computer program which was tested and entropy for one sentence was calculated manually.

Entropy till the third order was calculated and the results were compared for different information channels and analyzed. After these results value of average entropy was determined.

TABLE III
AVERAGE ENTROPY

H(N)	Latvian language
H(0)	5.0875
H(1)	4.3056
H(2)	3.1947
H(3)	1.9386

By defining average length of words, author made conclusion that to calculate the most precise redundancy of Latvian language, results for fifth-order entropy would need. We also can admit from the calculated average length of word that value of fifth-order entropy would be the sufficient exact to characterize Latvian language.

Comparing the entropy of the Latvian language with other languages, you can see diversity of the results.

Trend can be observed that when entropy order increases, entropy value descends slower. It can be observed due to the estimated H8 in Russian language. With the increase of entropy order ties between the letters of the word also decreases.

The probability that combinations with such a large number (more than five) will appear more than once in order not to give the entropy value equal to zero is very small. Except in cases where in the text is discussed specific topic where one word with a large number of symbols repeats (such as the recipe book of potato dishes). Calculating entropy after H5 should be interesting as a message to choose the word. But in that case definitely very long and different texts should be selected.

Entropy has an unpredictable value in its every order. However, it can be analyzed and explained. Entropy can be used as characteristics of a text both by doing calculations and making experiments. Entropy can characterize the source of information – type of a text or an author.

The larger the variety is in the text, the higher the entropy. Entropy is not a constant value. It can change by time because people use new words. Complexity and variety of the language increases consequently. According to the results, entropy changes in different sources.

It can be concluded that colloquial speech would reach the higher entropy value. The assumption is further reinforced with calculated entropy for one of the prose works. Entropy value was at its highest, because colloquial speech was used.

V REFERENCES

[1] C.E.Shannon "A Mathematical Theory of Communication", The Bell System Technical Journal,Vol. 27, pp. 379–423, 623–656, July, October, 1948

[2] W. Gitt "Information, Science and Biology" Journal of Creation 10(2):181-187, August 1996. Available: https://answersingenesis.org/genetics/information-theory/information-science-and-biology/ [Acessed Jan 20, 2015]

[3] The Electronic Frontier Foundation “Data compression home page” [Onine]. Available: http://www.data-compression.com/index.shtml [Accessed: Jan. 20, 2015]

[4] S.Haykin “Communication Systems” 4th edition – USA: McMaster University, „John Wiley & Sons, inc.”, pp.816, 2001.

[5] Latvian Academy of Sciences “Terminology Commission data base AkadTerm” [Online] Availabe http://termini.lza.lv/term.php?term=redundance&list=&lang=LV&h=yes [Accessed: Jan. 20, 2015]

[6] Keith Rayner, Sarah J. White, Rebecca L. Johnson, and Simon P., Raeding Wrods With Jubmled Lettres There Is a Cost Liversedge, *Psychological Science*, 2003, 17(3), 192-193

[7] University of Texas at Austin “Compuer science Home page” [Onine] Available https://www.cs.utexas.edu/~eberlein/cs337/errorDetection3.pdf [Acessed Jan 20, 2015]

[8] Society Homo Culturalis “HC.lv - portal" [Online] Avaiable http://web.hc.lv/kods/php-mysql/raksti/izstrades-vides-sagatavosana-uz-windows/ [Acessed Dec.10, 2008]

[9] Ozols A., “Signal transmission theory” Lecture synopsis, Riga Technical university, Latvia, 2009.

[10] “Datu kompresija kļūst arvien aktuālāka”, *Sakaru pasaule* , Riga, 2(50),2008.

[11] .R.Ospanova.”Calculating Information Entropy of Language Texts”, World Applied Sciences Journal 22 (1): 41-45, 2013.

[12] C. Papadimitriou, K. Karamanosa, F.K. Diakonos, V. Constantoudis, H. Papageorgiou. “Entropy analysis of natural language written texts, Contents lists available at ScienceDirect Physica A.Greece, 2010 [Available: journal homepage]: www.elsevier.com/locate/physa

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 17-21

Application of Project Management Information Systems in Efficiency Improvement of Quality Management System

Solvita Berzisa
Information Technology Institute, Riga Technical University, Kalku 1, Riga, Latvia

Abstract. **Project management information systems (PMIS) and quality management system (QMS) are two components in the project oriented organization that helps to achieve required quality of the project product. QMS define quality framework and PMIS helps to ensure quality framework requirement related to the projects. The objective of paper is to evaluate and demonstrate PMIS options for efficiency QMS development and maintenance. QMS requirements are identified according to ISO 9001:2008 standard and PMIS options of efficiency improvement are evaluated according to ISO 9004:2009 self-assessment tool.**

Keywords: **quality management system, project management information system, PMIS configuration, ISO 9001.**

I INTRODUCTION

For increase of competitiveness and productivity and improve performance an organization are introducing information systems (IS) and quality management systems (QMS) [1], [2]. QMS ensures that products or services are always consistently supplied, meeting customer and applicable regulatory requirements and seeking to enhance customer satisfaction [3]. QMS are certified according to standard (ISO 9001 [4], CMMI [5] etc.) to approve QMS quality to the customers. IS can support and influence QMS processes [1] and it effectivity and efficiency.

In cases when product or service realization has been organized in the projects one of the main IS in organization is project management information system (PMIS). PMIS ensure wide range of functionality for different kind of projects and organizations and also different modification and customization options. Join design and integration of PMIS (or IS) and QMS provides increase of QMS efficiency [3].

The objective of this paper is to evaluate and demonstrate PMIS options for efficiency QMS development and maintenance. Requirements of QMS according to ISO 9001:2008 are summarized and the PMIS configuration options according to these requirements are demonstrated.

This paper is structured as follows. The second section has description of the PMIS functionality and configuration options. QMS structure and requirements has been described in Section 3. Definition of the PMIS configuration requirements according to the QMS requirements is presented in Section 4. Evaluation of efficiency improvement possibilities by using the PMIS configured according to the QMS requirements is summarized in the discussion section and conclusions are provided in the last section.

II PMIS FUNCTIONALITY AND CONFIGURATION

PMIS is a standardized set of automated tools and techniques used in project management for planning, execution, management and closing of the project, as well as for collecting, combining and distributing project information [6]. PMIS provides a wide range of functions directly supporting PM [7], as well as tools for its configuration and modification. With configuration is defined the most appropriate PMIS configuration depending on project situation [8] [9]. Project situation requirements for PMIS have been identified according:

- Project classification [10] according project type, product, size, organization, management / planning approaches and related guidance's.
- Project environment and specific requirements.
- Enterprise environment factors [6] that includes enterprise available project management applications, government, industrial and quality standards etc.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.173

- Organizational process assets [6] that includes processes and procedures, for example, QMS (detail review of QMS in Section III.) and corporate knowledge base.

Definition of the PMIS configuration requirements must include the following information [8]:

1) Data entities or work items used in project;
2) Attributes or data fields of each data entity;
3) Processes or workflows related to the data entities.

Definition of the PMIS configuration requirements can also include records of data entity, for example, risk lists, reports, metrics etc.

III QMS OVERVIEW AND REQUIREMENTS

QMS is process oriented framework [1] and ensures controlled performance of organization processes with target to deliver products or services according to customer and applicable regulatory requirements [3].

QMS according to ISO 9001 distinguish three types of processes: management, realization and support. QMS processes and its interaction within organization is being defined in the ISO 9001:2008 standard [11] and ISO 9004:2009 standard [12].

ISO 9001:2008 [11] defines quality policy, objectives, manual, management responsibility, resource management, documented procedures, work instructions, documents, records, measurements, analysis and improvement requirements. Most part of documented work instructions, documents and records have been related to realization processes.

ISO 9004:2009 [12] is guidance that describes how QMS according to ISO 9001:2008 can achieve sustained success with effective and efficient quality management approach. This guidance describes activities need to be done to ensure that organization processes and practices are effective and efficient. The self-assessment tool is used in the guidance for evaluation performance of an organization and degree of maturity of QMS [12]. The self-assessment tool evaluate maturity level of QMS key elements: management for the sustained success of organization; strategy and policy; resource management; process management; monitoring, measurement, analysis and review; improvement, innovation and learning [12]. The self-assessment can be done also for detailed elements.

Ensuring of QMS according to ISO 9001:2008 consists of two parts: adoption and maintenance. ISO 9001:2008 requirements have been introduced during adoption of QMS. Continues improvement is one of important part of QMS maintenance [13]. ISO 9004:2009 self-assessment tool [12] helps in identification of QMS gaps and possible improvements.

IV QMS REQUIREMENTS TO PMIS CONFIGURATION

In cases when product or service realization has been organized in the projects these requirements have been integrated in project processes, plans, metrics and records. The projects also need to collect measurement of metrics and key performance indicators (KPI) about processes and products for QMS analysis and improvement planning.

QMS according to ISO 9001:2008 defines common quality related requirement for the PMIS configuration for all projects in organization. The QMS requirements [11] and the PMIS configuration requirements relation is shown in Table I. The QMS requirements to the PMIS configuration are divided in two groups: product realization and measurements.

Main QMS requirements have been related to the project product realization and are identified from work instructions / procedures and its associated documents and records. In PMIS these requirements are implemented as data entities and workflows. As examples in Table I have been shown one ISO 9001 mandatory documented procedure 'Corrective actions' and three product realization processes: 'Requirement specification', 'Risk management' and 'Change management'.

Other set of the QMS requirements have been related to measurement of product and process quality. QMS defines quality metric and KPI. PMIS ensures definition of metrics and KPI and store of measurements values. Measurement values are analyzed centrally using reporting options of PMIS tools (for example MS Project Server reports [14]), data warehouse [15] or business intelligence tools (for example, also MS Project Server BI center [14]). Examples of metrics and KPI have been given in Table I.

The QMS requirements identify only part of the PMIS configuration requirements that gives quality baseline for the project. From the other project situation factors come the project specific PMIS requirements depending on project classification, environment, tools etc. As result other data entities, workflows and metrics can be included in the PMIS configuration and also modification in QMS defined PMIS configuration requirement can be done by including additional workflow statuses and data entity attributes.

V DISCUSSION

Application of PMIS contributes two areas of effective and efficient QMS development and maintenance: processes (also product realization) and monitoring.

Application of PMIS helps to improve effectivity and efficiency of QMS in following areas:

- Planning and control of organization product realization and other processes. During

definition of the PMIS configuration requirements processes, process input/output, interactions, required records and measurements are identified. In PMIS integrated processes have been documented and partially automatized with data entity forms, workflows, review evidence records etc.

- Monitoring, measuring, analyzing, reviewing and reporting. Part of monitoring and measurement activities has been automatized by using PMIS. During definition of the PMIS configuration requirements have been identified metrics and KPI that need to be evaluated. Monitoring result help to assess and understand the organization current performance.
- Information distribution that helps to keep interested parties informed about progress against plans. PMIS collected data are easy available for reports.

TABLE I

QMS VS. PMIS CONFIGURATION REQUIREMENTS

QMS REQUIREMENT	PMIS CONFIGURATION REQUIREMENT	EXAMPLE			
		EXAMPLE 1 : CORRECTIVE ACTIONS RELATED TO THE PRODUCT	EXAMPLE 2: REQUIREMENT SPECIFICATION	EXAMPLE 3: RISK MANAGEMENT	EXAMPLE 4: CHANGE MANAGEMENT
Work instructions and procedures	Different kind of processes together with related data entities.	Data entity: *Issue* Process: *Issue* status workflow	Data entity: *Requirement*, *Requirement specification* (or project documents with type requirement specification), *Requirement specification review* (or document review) Process: *Requirement* management status workflow, *Requirement specification* (or project document) status workflow	Data entity: *Risk* Process: *Risk* status workflow	Data entity: *Change request* Process*: Change request* status workflow
Documents	Forms of documents – data entity with attributes.	Attributes of *issue*	Attributes of *requirement*, (including requirement traceability); Template of *requirement specification*; Checklist of *requirement specification review*	Attributes of *risk*	Attribute of *change request*
Records required in ISO 9001	Records are information inserted in data entity form or status transaction in data entity workflow. All required records need to be include configuration with data entity or workflow status	ISO 9001 requires record "Nature of the product nonconformities and any subsequent actions taken, including concessions obtained" PMIS ensure this requirement with *issue* data entity, attributes and workflow	ISO 9001 requires record "7.2.2. Results of the review of requirements related to the product and actions arising from the review" PMIS ensure this requirement with *requirement* workflow status "7.5.3. The unique identification of the product, where traceability is a requirement" PMIS ensure this requirement with *requirement* and *requirement specification* data entity attribute		ISO 9001 requires record "7.3.7. Results of the review of design and development changes and any necessary actions" PMIS ensure this requirement with *change request* data entity, attributes and workflow

Metrics	*Metrics* data entity with defined metric records *Measurement* data entity	For example "*Issues* in product / project / project phase"	For example "Count of *requirements* in product / project / project phase" "Count of *changed requirements* in product / project / project phase" "Count of *requirements changes* in product / project / project phase" "Count of corrections in requirement specification during review"	For example "*Risks* (identified, removed / occurred) in product / project / project phase" "Reviews to *risk"* "Count of *risk* occurrence"	For example "*Changes* in product / project / project phase"
KPI	*KPI* data entity with defined metric records	For example "Critical issues per product / project / project phase" (Target value: zero)	For example "Count of corrections in requirement specification during review per product / project / project phase" (the more the better , for example target value: 10)	For example "Count of not reviewed *risk* occurrence per product / project / project phase" (Target value: zero)	

According to the self-assessment tool [12] maturity level achievement of following QMS elements are easier by using PMIS:

- Process planning and control – key processes are defined and managed, interactions are defined and systematically measured (Level 2).
- Monitoring – monitoring process is performed periodically (Level 2), process capabilities are monitored (Level 3).
- KPI – formal set of defined key indicators (Level 2), main conditions for success are identified and tracked by indicators (Level 3), management decisions are supported by reliable data (Level 3), data is available to show progress of KPI over time (Level 4), strategy and objectives is monitored (Level 4).
- Analysis – statistical tools are used (Level 2, 3).

Application of PMIS only helps to easier achieve maturity levels because PMIS can be implemented only QMS defined requirements. PMIS is only tool for QMS implementation that helps to create it more efficient. To achieve each level the first is required improvements in QMS and then in PMIS and other organization IS.

VI CONCLUSION

PMIS is one of IS that is used for QMS support and implementation in the project oriented organization. This paper demonstrates PMIS options for efficiency QMS development and maintenance. PMIS contributes two areas of QMS: project product realization and monitoring of product quality and processes. Application of PMIS helps to develop QMS effective and efficient according to ISO 9001:2008 but also is need other organization IS that contributes other areas of the QMS requirements.

Limitation of this research is that only QMS according to ISO 9001:2008 for project oriented product realization has been reviewed and analyzed.

This is background research about PMIS options in the QMS requirement implementation. Future research directions are QMS integration with portfolio management systems and business intelligence tools for the project quality data analytics and the quality process monitoring.

VII REFERENCES

[1] S. G. G. Ferreira and J. Á. Carvalho, "Articulation between information systems and quality management systems," in *Creating Global Competitive Economies: A 360-Degree Approach - Proceedings of the 17th International Business Information Management Association Conference, IBIMA 2011*, 2011, vol. 4, pp. 2111–2113.

[2] L. Y. Fok, W. M. Fok, and S. J. Hartman, "Exploring the relationship between total quality management and information systems development," *Inf. Manag.*, vol. 38, no. 6, pp. 355–371, Jul. 2001.

[3] P. R. da Cunha and A. D. de Figueiredo, "Quality Management Systems and Information Systems: Getting More than the Sum of the Parts," *AMCIS 2005 Proc.*, 2005.

[4] ISO, *ISO 9000 quality management*. .

[5] "CMMI Institute." [Online]. Available: http://cmmiinstitute.com/. [Accessed: 18-Mar-2015].

[6] Project Management Institute, *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)—Fifth Edition*. PMI, 2014, p. 618.

[7] S. Bērziša, "The Baseline Configuration of Project Management Information System," *Sci. J. Riga Tech. Univ. Comput. Sci.*, vol. 5, no. 40, pp. 59–65, 2009.

[8] S. Bērziša, "Application of Knowledge and Best Practices in Configuration of Project Management Information Systems," PhD thesis, Riga Technical university, 2012.

[9] S. Bērziša and J. Grabis, "Knowledge Reuse in Configuration of Project Management Information Systems: A Change Management Case Study," in *Proceedings of the 15th IEEE International Conference on Intelligent Engineering Systems 2011*, 2011, pp. 51–56.

[10] S. Bērziša, "Project Management Knoeldege Retrieval: Project Classification," in *Enviroment. TEechnology. Resources: Proceedings of the 8-th International Scientific and Practical*

Conference June 20-22,2011. Volume II, 2011, vol. I, pp. 33–39.

[11] ISO, *ISO 9001:2008 - Quality management systems -- Requirements*.

[12] ISO, *ISO 9004:2009 - Managing for the sustained success of an organization -- A quality management approach*. .

[13] I. Mustapha, A. Jusoha, and K. M. Nora, "A review on quality management systems maintenance framework based on process based management, Knowledge quality and knowledge self-efficacy," *J. Teknol.*, vol. 72, no. 4, pp. 7–12, 2015.

[14] "Project Server 2013 - PWA Sample BI Reports." [Online]. Available: http://epmcentral.com/articles/ps13/ps13samplereports.php. [Accessed: 20-Mar-2015].

[15] N. Gojgic, A. Veljovic, M. Zahorjanski, L. Stanojevic, and N. Cvijovic, "Implementation of data warehouse in analysis of QMS inconsistencies," *Metal. Int.*, vol. 17, no. 3, pp. 5–10, 2012.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 22-28

Neurons vs Weights Pruning in Artificial Neural Networks

Andrey Bondarenko, Arkady Borisov, Ludmila Aleksejeva
Riga Technical University, *Faculty of Computer Science and Information Technology, Decision Support Systems Group.*
Address: Mezha str. ¼ - 463, Riga, LV-1048, Latvia

Abstract. **Artificial neural networks (ANN) are well known for their good classification abilities. Recent advances in deep learning imposed second ANN renaissance. But neural networks possesses some problems like choosing hyper parameters such as neuron layers count and sizes which can greatly influence classification rate. Thus pruning techniques were developed that can reduce network sizes, increase its generalization abilities and overcome overfitting. Pruning approaches, in contrast to growing neural networks approach, assume that sufficiently large ANN is already trained and can be simplified with acceptable classification accuracy loss.**

Current paper compares nodes vs weights pruning algorithms and gives experimental results for pruned networks accuracy rates versus their non-pruned counterparts. We conclude that nodes pruning is more preferable solution, with some sidenotes.

Keywords: **artificial neural networks, generalization, overfitting, pruning.**

I INTRODUCTION

Artificial neural networks has been successfully applied in many different areas to solve problems of classification and regression. ANN's, specifically multi-layer feed-forward artificial neural networks trained using error back-propagation give good results and recent advances in deep learning proved to give unprecedented classification accuracy [1]-[2]. Main problem with ANN's traiing is choosing hyper-parameters which can severely influence model performance. Choosing architecture with insufficient amount of neurons can give unsatisfying classification rates, while choosing too much neurons will badly influence training time and will cause overfitting. To overcome such problems two approaches exist: growing neural networks [3][4] and training excessively large network with subsequent pruning. In current paper we utilize second approach to overcome overfitting of trained ANN's, thus rise its generalization abilities as well as prepare previously trained artificial neural networks for rules extraction. We provide results of experiments with both nodes and weights pruning approaches.

Current paper is structured as follows: section two gives overview of pruning methods, section three describes used algorithm, section four presents results of experiments and sections five and six hold discussion and conclusion.

II PRUNING METHODS OVERVIEW

Exist different approaches to ANN prunning. Both neurons themselves (thus all incomming and outgoing weights) and specific weights can be prunned. Paper [5] provides overview, we briefly present main ideas mentioned there along with other methods not listed in the source paper. We can divide all pruning algorithms into two main categories nodes/weights removal based on sensitivity analysis and penalty term based methods that utilize penalty term to remove 'unused' / least important weights. Some algorithms combine both approaches, while some cannot be easily added to one or the other family of methods. Sensitivity analysis relies on calculation of influence of specific node or weight.

- *Sensitivity analysis based methods*

Sensitivity method from [6] by Mozer and Smolensy calculates error with unit removed and without it being removed, thus deleting least important units. Instead of calculating error directly they use derivative calculated during error back-propagation to approximate it. Segee and Carter in [7] have found that small variance in weights incoming into neuron is signaling that subject neuron can be safely removed.

Karnin in [8] describes method for weights pruning, which does not requires specific sensitivity calculation phase. All necessary data about weights updates are stored during training. This makes this approach unusable in case one wants to prune already trained

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.166

ANN. Nevertheless such approach is perfectly usable in case one have control over training of ANN. The main idea is sensitivity analysis of weights and their removal if they have too small sensitivity. Here is the

$$S_{ij} = -\sum_{n=0}^{N-1} \frac{\partial E}{\partial w_{ij}} \Delta w_{ij}(n) \frac{w_{ij}^f}{w_{ij}^f - w_{ij}^i} \quad (1)$$

Where w^f is specific weight value, *0* is its value after its pruning, $E(w^f)$ is error with given weight enabled and $E(0)$ is error when this weight is pruned. Finally S_{ij} is sensitivity of weight between nodes *i,j*. Instead of computing sensitivity value directly (which would lead to sensitivity estimation phase) authors propose to estimate it using sum of all changes of weight during training:

$$S_{ij} = -\frac{E(w^f) - E(0)}{w^f - 0} w^f \quad (2)$$

Rudy Setiono [9] Describes rather simple pruning algorithm which uses simple heuristics to find weights to be pruned. It assumes that one have ANN with single hidden layer (although approach can be generalized to multiple hidden layers), And afterwards removes input-hidden and hidden-output weights if their values are not satisfying specific constant. Actually there are two constants used, both of them should be set-up manually.

Rudy Setiono as well described rules extraction algorithm [10] called N2FPA which uses simple estimations of effect of removal of neurons in the network. Neurons are removed one by one. In case error worses significantly pruning stops. This is the method which was used and slightly modified in current paper.

Le Cun et.al in [11] describe method call *Optimal Brain Damage* (OBD) which measures "saliency" of a weight by estimating second derivative of the error with respect to the weight. They made couple of assumptions after which compute such derivatives during modified error back-propagation. One drawback of such method is necessity of storage of Hessian matrix. After one weight is pruned, retraining is done to find another weight to prune.

Optimal Brain Surgery from [12] (OBS) goes one step further in comparison to OBD, it utilizes inverse Hessian matrix to calculate optimal weight to be deleted, but at the same time it solves optimization problem which as result gives remaining weights updates necessary to lower network error. Such approach allows simultaneous update of all remaining weights thus retraining is not required. OBS is one of the best methods for pruning. As well as OBD it should hold Hessian matrix thus requires additional memory.

- *Penalty based methods*

Penalty term methods are utilizing weight decay / penalty term in one way or another to force neural network during its training get rid of unnecessary weights.

Chauvin [13] uses cost function with specific term which poses average energy expended by weights, as well there is a modification with additional magnitude of weights term which penalizes large weights and large amount of weights.

Weigend et. Al [14]-[16] minimizing specific cost function with additional term penalizing network complexity as a function of the weights magnitudes relative to the defined constant w0. Choosing such constant should be done via trials/errors.

Ji et.al [17] propose another penalty term based pruning approach based on modified error function which tries to minimize number of hidden nodes and weights magnitudes. The limitation of proposed method is that it assumes single hidden layer ANN with one input and out *linear* output node. Method assumes retraining after each removed weight.

- *Weight decay methods*

Plaut et. Al [18] proposes simple cost function which decays weights. Cost function specifics tends to fact that algorithm favors nodes with lot of small weights in contrast to node with single large connection. Nowlan and Hinton [19] describe more complex cost function with penalty term which models the probability distributions of weights as mixture of Gaussians.

- *Interactive pruning*

Sietsma and Dow [19] describe interactive method in which designer inspects network and marks nodes to be pruned. Algorithm provides several heuristics to determine candidates for removal. Authors have shown on training problems that their method is capable of finding relatively small networks with good accuracy in comparison to large trained networks which were not able to find solution.

TABLE I

ERROR RATES WITH STANDARD DEVIATIONS. MEAN PRUNED NODES/WEIGHTS COUNTS.

	MLP train avg. (std.dev)	MLP test avg.. (std.dev	Pruned Weights train avg.. (std.dev)	Pruned Weights test avg.. (std.dev)	Pruned Nodes train avg.. (std.dev)	Pruned Nodes test avg.. (std.dev)	Pruned Weights (std.dev) Pruned Counts (std.dev)
Haberman (10-fold X-validation)	25.99% (0.0098)	26.78% (0.0439)	24.39% (0.0102)	**24.91**% (0.0550)	24.98% (0.0101)	26.17% (0.0371)	54.9 (42.0) 23.8 (10.7)
Ionosphere (10-fold X-validation)	10.83% (0.0013)	10.83% (0.0115)	4.21% (0.0117)	10.25% (0.0346)	4.55% (0.0116)	**9.22%** (0.0350)	34.1 (26.8) 34.3 (16.0)
Monks-1 (train/test)	21.51% (0.0158)	32.74% (0.0138)	0.83% (0.0268)	**1.81%** (0.0566)	6.83% (0.0718)	13.22% (0.1201)	45.9 (17.3) 22.4 (16.3)
Monks-2 (train/test)	38.46% (0.0194)	36.04% (0.0167)	12.47% (0.1795)	12.21% (0.1759)	11.26% (0.1751)	**10.25%** (0.1597)	16.8 (11.8) 20.1 (12.7)
Monks-3 (train/test)	6.56% (0.0)	**2.88%** (0.0022)	5.16% (0.0182)	3.45% (0.0108)	3.33% (0.0068)	5.76% (0.0075)	32.4 (28.1) 29.3 (9.1)
Parkinsons (10-fold X-validation)	24.58% (0.0023)	24.61% (0.0190)	14.83% (0.0307)	16.38% (0.0581)	14.30% (0.0154)	**15.57%** (0.0671)	10.5 (21.3) 8.3 (18.6)
Pima (10-fold X-validation)	23.93% (0.0083)	24.56% (0.0436)	21.64% (0.0079)	23.74% (0.0398)	22.12% (0.0055)	**23.05%** (0.0322)	56.0 (34.3) 22.7 (3.3)
WDBC (10-fold X-validation)	4.16% (0.0033)	4.33% (0.0251)	1.83% (0.0026)	**2.63%** (0.0238)	1.77% (0.0028)	2.93% (0.0218)	23.3 (9.2) 18.5 (11.0)
WPBC (10-fold X-validation)	0% (0.0)	**0%** (0.0)	0% (0.0000)	0.17% (0.0091)	0% (0.0000)	**0%** (0.0000)	153.6 (140.6) 50.0 (0.0)

- *Auto-pruning methods*

Next discussed approach is auto-pruning method called *lprune* [20] by Lutz Prechelt. It proposes to prune at each step all weights not satisfying specific formula controlled by parameter *lambda.* Experiments showed that this parameter should be adaptive, algorithm to support dynamic adjustment is proposed. According to author proposed methods overcomes OBD and OBS in terms of accuracy and simplicity of pruned ANN

Another auto-pruning method [21] by William Finoff et. al utilizes modified cost function, does not requires full training of ANN and uses dynamic adjustment of penalty term. Similar to OBS this method performs dynamic topology adjustments.

- *Other methods*

Kruchke [22] describes Local Bottlenecks method in which neurons "compete" with each other to survive. Magnitudes of vectors determine degree to which neuron participates in modeling target function, this is treated as neuron gain. In case gain is zero, neuron is not participating in classification task and can be removed. In case two neurons have parallel or anti-parallel weights vectors they are redundant and can be removed as well. Method utilizes specific parameter which should be tuned carefully.

Same author proposes another method called Distributed Bottlenecks [22][23] which puts constraints on weights rather than deletes them. This server as sort of dimensionality reduction. Such approach makes weight vectors that are farther apart than average to become more farther from each other and vectors that are closer than average to become more closer. Again method uses special constant which should be chosen manually.

III PROPOSED ALGORITHM

We utilized algorithm described by Rudy Setiono in [10] (part of N2FPA rules extraction method.) In essence we are using nodes/weights pruning. We operate on trained ANN, on each pruning iteration we try to determine neuron or weight which needs to be removed. For all weights or neurons in input and hidden layers we calculate classification error for network operating without them. (This essentially means we are setting activations of pruned neurons to zero, or nullifying weights.) When neuron/weight, which upon removal gives network with smallest cost function is found it is removed. In our case this is neuron/weight which after removal gives network with smallest error classification rate. Afterwards network is retrained. If accuracy drops, remained the same or have risen over a small amount (we used tolerance equal to 2.5% - i.e. we are ok with error growth for this amount, then neuron/weight is really pruned. If error rises significantly candidate neuron/weight is left intact and new search for pruning candidate is initiated. In case error rises we are Retraining gives chances to get simpler network with high generalization and good classification rates, which can be observed looking into table 1.

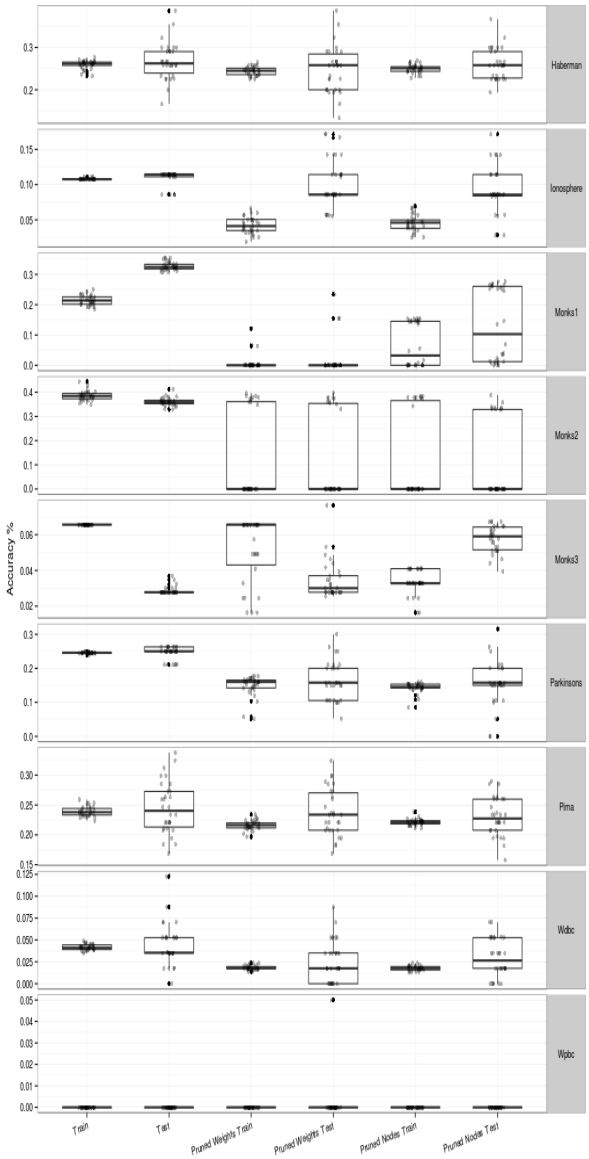

Fig. 1. Classification error.

This algorithm was described in our previous work [24]. Although in current papers we used a slightly modified version with two main changes. First of all in new version we have pruned not only hidden neurons but input layer as well – this basically worked as a feature selection. And secondly when algorithm encounters rise in error it saves ANN previous state, thus in case of several consecutive trials and failures to retrain network and get better accuracy (or at least not too bad – according to error tolerance) algorithm restores last known good ANN state (before significant rise of the error).

Below you can find pseudo code of weights pruning algorithm (with slight modifications in regards to algorithm published in [24]):

```
Inputs:

maxIter –        determines maximum count of
prunning iterations

maxPrunedNodes – maximum amount of nodes to
be pruned

errorRiseTol -   determines acceptable error
rise

maxFallbacks -   in case neurons are pruned
and then reverted – how many times before we
quit?

Program:

iter = 1
while(iter<maxIter ||
prunedWeights<maxPrunedWeights)
  for all not pruned weights in all layers
    if (lastWeightInLayer()) continue;
    removeWeight(n)
    cost = testNetwork()
    if (cost > largestKnownCost)
      largestKnownCost = cost
      indexOfPrunedWeight =
getIndexOfPrunedWeight
      prunedWeights = prunedWeights + 1
    end
  end

  S = saveNetworkState()
  retrainNetwork()

  classError = testNetwork()
  errRatio = classError/smallestClassError
  if (errRatio > 1 + errorRiseTol)
    revert pruned weight
    fallbacksCounter = fallbacksCounter + 1
    prunedNodes = prunedNodes - 1
  else
    ///leave pruned neuron as is
    fallbacksCounter = 0
  end
  if (fallbacksCounter >= maxFallbacks)
    this = restoreNetworkState(S)
    break
  end
  iter = iter + 1
end
```

Here one can notice hyper-parameters listed in the beginning, m*axIter* -controls maximum possible amount of pruning iterations, *maxPrunedNeurons* – controls maximum amount of neurons to be pruned. We need both parameters as neurons after pruning can be restored, thus some iterations will not result in network pruning. Although they will leave network with weights adjusted during retraining. Apart from that two other hyper-parameters are: *errorRiseTol* –

which controls maximum rise of error (we used classification error) after removal of neuron which will not cause pruned neuron reversal/restore. Thus let's say in case error have risen for 5% in case of neuron removal in comparison to best/lowest known error rate and our parameter is 0.075 we will leave network intact, but if it is below 0.05 network will get back pruned neuron. Finally *maxFallbacks* controls how many attempts algorithm makes in pruning neurons and reverting them back consequently before termination. Thus if this parameter is equal to 10, then in case of ten subsequent iterations neuron is pruned, but then restored due to high rise in error algorithm terminates.

ANN itself is trained using Cross-entropy cost function equipped with penalty term (weights decay). Below is cost function:

$$F(w,v) = -\left(\sum_{i=1}^{k} \sum_{p=1}^{C} l_p^i \log S_p^i + (1 - l_p^i)(1 - \log S_p^i) \right) \quad (3)$$

Where k – is the number of patterns, l_p^i = 0 or 1 is the target value for pattern x^i at ouput unit p, p = 1,2,…,C. C is the number of output units/neurons. S_p^i is the output of the network at unit p:

$$S_p^i = \sigma\left(\sum_{m=1}^{h} \sigma\left(\left(x^i\right)^T w^m \right) v_p^m \right) \quad (4)$$

Here to simplify things a bit we provide formulas for single hidden layered neural network, but in reality for our experiments we utilized two hidden layers. x^i is an n-dimensional input pattern, i=1,2,…,k. w^m is an n-dimensional vector of weights for the arcs connecting the input layer and the m-th hidden unit, m = 1,2,…,h. v^m *is a* C-dimensional vector for the weight connecting the m-th hidden unit and the output layer. The activation function is sigmoid function with domain (-1, +1):

$$\sigma(y) = \frac{1}{1 + e^{-y}} \quad (5)$$

Finally for all our weights we are applying weight decay factor 0.0001. This is quite simple approach in comparison to other described in theoretical part, but still it does the job. Cross-entropy was chosen as it is capable of dealing with problems of error derivative platoe better than standard round mean square error (RMSE) [25]. Apart from this we utilized Stochastic Gradient Descent batch training. Batch size was chosen to be 20.

IV EXPERIMENTS

In our experiments we have utilized three 10-fold cross-validation, but for some test-sets like monk's train and test data are already provided thus there we utilized thirty runs to get averaged results. We decided to utilize two hidden layers neural networks so that some networks will be able to utilized this to their advantage, in case one of the layers is not needed we will be able to see this after pruning will be finished – such layers should have small amount of intact neurons in one of the layers. For our experiments we utilized well known UCI [26] data sets: Monks-1, Monks-2, Monks-3, Ionosphere, Haberman, Pima diabetes, WDBC, WPBC and Parkinsons.

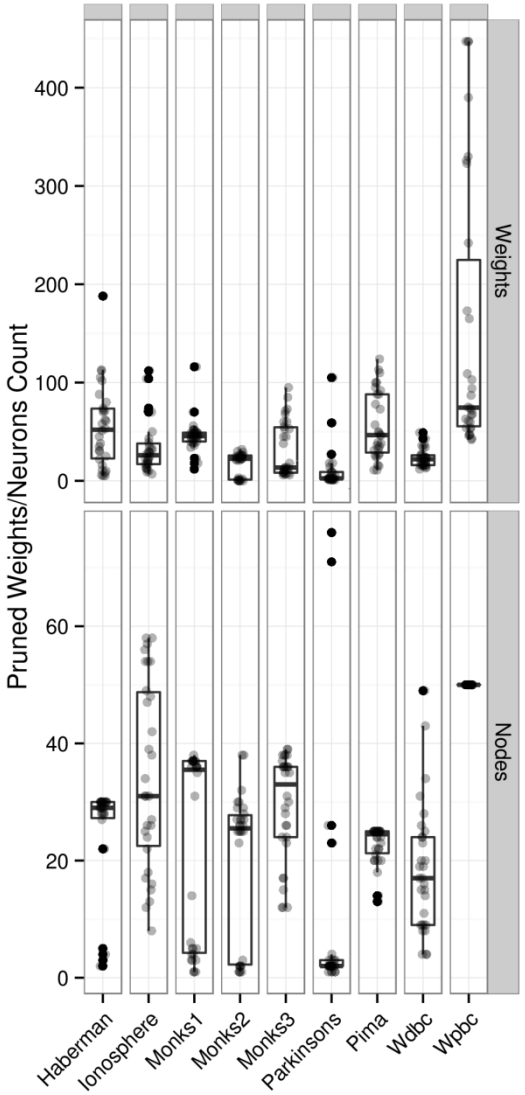

Fig. 2. Counts of pruned weights/nodes

Some of mentioned problems utilize only categorical variables - like monks. In such cases we have transformed input data into binary format thus instead of 5 inputs we used 17. In other cases the only transformation applied was rescaling of data into [-1, 1] region. Data sets are binary classification problems; we utilized two output neurons to represent solution of the network. Data sets themselves are pretty small

ranging from about 150 to 500 entries. Table 1 holds classification accuracy for all data sets. It contains average classification rate with standard deviation for both train and test cases along with 10 x-validation folds or runs for non-pruned and pruned neural networks. Last table column holds networks hidden layers structures before and after pruning.

For the pruning algorithm we used 0.025 as an *errorRiseTol* tolerance level, usually around 50 (depends on total amount of neurons, it should be around 60-70% of that) as *maxIter* iterations. MaxIter count should be larger than maximum amount of neurons to be pruned (which was always equal to neurons count in hidden layers minus 2 – we cannot prune all neurons from all layers.) In all cases we decided to utilize 2 layer hidden neuron networks (with error-backpropagation) trained using cross-entropy cost (error) function and stochastic gradient descent as learning algorithm. All cases were executed using 10-fold cross validation except monk's data sets – they already are divided into training and testing sets

Figure 1 as well as Table 1 present classification errors for all 30 experiments across all datasets. One can note that all in all weights pruning when run on test data is giving classification accuracy almost equal to nodes pruning. Single exception is Monks-2 dataset, where weights pruning perform better. Now looking on figure 2 one can observe that counts of pruned nodes/weights greatly varies between runs. This suggests large amount of local minima. One way of dealing with this can be pre-training of ANN using DBN.

All in all training set classification graphs show decrease of train set classification error.

V DISCUSSION

As one can notice in many cases acquired ANN models are significantly smaller than initial networks. Exceptions are wdbc and parkinsons data sets where we can see ~50% drop in neurons counts. Both of them have rather complex structure thus require bigger models (in comparison to other data sets). Looking at weights pruning approach – almost always networks sizes are at least 70%-80% of the original size (with some exceptions). Looking at accuracy rates we can say there is a parity between weights and nodes pruning. Although in context of rules extraction neurons pruning is more preferable approach as it is producing smaller networks and it is less time consuming (there are much less neurons than weights in network). As we already noted algorithm have auto-stopping criteria allowing it to perform several trials before deciding to stop. Used algorithm assumes training of neural network with removed neuron/weight, while afterwards in case of unsatisfactory results removed neuron is returned back to the ANN. Diligent reader can note that there exists several possibilities in regards to how and which neuron should be returned back into neural net. We used same neuron, but we have not explored possibilities of adding random neuron. As we already mentioned in case algorithm fails to get pruned network, it restores it's last state – before it started to fail with removal of weight/node.

Another point to mention is interesting behavior of algorithm on Monks-2 dataset observing network errors in Figure 1 one can note that in half of all cases algorithm was able to get near zero error rate. At the same time weights pruning performed much better than nodes pruning at Monks-3 testing data set, which have 5% noise in training data.

VI CONCLUSIONS

In current paper we presented improved algorithm for pruning artificial neural networks via nodes/weights pruning along with experimental data (UCI classification data sets were used) showing that both types of pruning simplify network structure, but nodes pruning does better job. While weights pruning can give better results at cost of more complex ANN structure and higher computational time. Of course if some data set have complex structure – in such cases both approaches will end up with only slightly pruned neural network – but with better generalization value. When algorithm was applied to UCI datasets in many cases it produced much simpler ANN models with only a few neurons/or couple of dozens of weights while having slightly worse or in some cases better classification accuracy rates. Such 'simpler' models are faster to execute and are better candidates for knowledge extraction. Further research directions are exploration of other techniques for returning neuron back after retraining phase. Another area of future research can be dealing with local minima which causes early stopping during pruning. It is interesting to see causes of such early stopping – are they caused entirely by poorly trained ANN and can they be overcome or not.

VII REFERENCES

[1] A.-Krizhevsky, I. Sutskever, G. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012.

[2] G. Hinton, L. Deong, D. Yu, G. Dahl and others, "Deep Neural Networks for Accoustic Modelling in Speech Recognition". *IEEE Signal Processing Magazine*, November, 2012.

[3] X. Qiang, G. Cheng, Z. Wang, "An Overview of Some Classical Growing Neural Networks and New Developments", *IEEE, Education Technology and Computer (ICETC)*, 2nd International confernece Vol.3. 2010.

[4] V. Chaudhary, A.K. Ahlawat, R.S. Bhatia, "Growing Neural Networks using Soft Competitive Learning". International Journal of Computer Applications (0975-8887) Volume 21-No.3, May 2011.

[5] R. Reed, “Pruning Algorithms – A Survey”, IEEE Transactions on Neural Networks, Vol.4., No.5., September 1993.

[6] M. C. Mozer and P. Smolensky, "Skeletonization: A Techique for Trimming the Fat From a Network via Relevance Assessment," in Advances in Neural Information Processing, pp. 107-115, (Denver 1988), 1989.

[7] B. E. Segee and M. J. Carter, "Fault Tolerance of Pruned Multilayer Networks," in *Proc. Int. Joint Conf. Neural Networks*, Vol. 2., (Seattle), pp.447-452, 1991.

[8] E. D. Karnin, "A Simple Procedure For Pruning Back-Propagation Trained Neural Networks", *IEEE Trans. Neural Networks*, Vol. 1., No. 2, pp.239-242, 1990.

[9] R. Setiono and H. Liu, "Understanding Neural Networks via Rule Extraction," IJCAI, 1995.

[10] R. Setiono and W. H. Leow, "Pruned Neural Networks for Regression" in *PRICAI 2000 Topics in Artificial Intelligence, Lecture Notes in Computer Science* Vol. 1886, 2000, pp. 500-509.

[11] Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal Brain Damage," in *Advances in Neural Information Processing (2),* D.S. Touretzky Ed. (Denver 1989), 1990, pp. 598-605.

[12] B. Hassibi, D. G. Stork, G. J Wolf, "Optimal Brain Surgery and General Network Pruning.”

[13] Y. Chauvin, “A Back-Propagation Algorithm With Optimal Use of Hidden Units” *Advances in Neural Information Processing*, (1) D.S. Touretzky ed. (Denver 1998), 1989, pp. 519-526.

[14] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Back-Propagation, Weight Elimination and Time Series Prediction," in *Proc. 1990 Connectionist Models Summer School,* D. Touretzky, J Elman, T. Sejnowsky, and G. Hinton, Eds., 1990, pp. 105-116.

[15] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by Weight-Elimination Applied to Currency Exchange Rate Prediction," *in Proc. Int. Joint Conf. Neural Networks*, vol. I, (Seattle), 1991, pp.837-841.

[16] A. S. Weigend, D. E. Rumelhart and B. A. Huberman, "Generalization by Weight-Elimination With Application to Forecasting," *in Advances in Neural Information Processing (3) R. Lippmann, J. Moody, and D. Touretzky, Eds., 1991, pp. 875-882*.

[17] C. Ji, R. R. Snapp, and D. Psaltis, "Generalizing Smoothness Constraints From Discreet Samples," *Neural Computation, Vol. 2, No. 2, 1990, pp.188-197*.

[18] D. C. Plaut, S. J. Nowlan, and G E. Hinton, “Experiments on Learning by Back Propagation,” *Tech. Rep. CMU-CS-86-126*, Carnegie Mellon Univ., 1986.

[19] Sietsma, and R. J. F. Dow, "Neural Network Pruning – Why and How" Proc. of the IEEE *International Joint Conference on Neural Netoworks*, 1989, pp. 325-333.

[20] L. Prechelt, “Adaptive Parameter Prunning in Neural Networks,” International Computer Science Institute, March. 1995.

[21] W. Finnoff, F. Hergert, and H. G. Zimmermann, “Improving Model Selection by Nonconvergent Methods”, *Elsiever Neural Networks*, Vol. 6, Issue 6, 1993, pp.771-783.

[22] J. K. Kruschke, “Creating Local and Distributed Bottlenecks in Hidden Layers of Back-Propagation Networks,” in *Proc. 1988 Connectionist Models Summer School*, D. Touretzky, G. E. Hinton, and T. Sejnowsky, Eds., 1988, pp 120-126.

[23] J,. K. Kruschke, “Improving Generalization in Back-Propagation Networks with Distributed Bottlenecks,” *in Proc. Int. Joint Conf. Neural Networks*, Washington DC, Vol. 1, 1989, pp.443-447.

[24] A. Bondarenko, A. Borisov, “Neural Networks Generalization and Simplification via Prunning”, Scientific Journal 2014 of Riga Technical University, 2014.

[25] P. Golik, P. Doetsch, and H. Ney, “Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison”, in *Interspeech*, pp. 1756-1760, Lyon, France, August 2013.

[26] K. Bache, M. Lichman, (2013), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science (last accesed Sept 15, 2014).

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 29-35

Computer Programming Aptitude Test as a Tool for Reducing Student Attrition

Juris Borzovs, Laila Niedrite, Darja Solodovnikova
University of Latvia, Faculty of Computing.
Address: Raina blvd 19, Riga, LV 1586, Latvia

Abstract. **The stable trend to lose from one-third to half of students in the first study year of computing studies motivated us to explore, which methods are used to determine in advance such applicants, who have no change to overcome the first study year. Initially, a research about the factors influencing the attrition in Faculty of Computing at the University of Latvia was conducted. The research revealed that the trend of non-beginning studies might indicate the wrong choice of the study field and possible lack of understanding of what is programming by enrolled students (applicants as well as pupils).**

The study provides the review of the situation with programming aptitude tests in the world, which could serve as one of the solutions to the dropout reduction. An action plan is proposed, which is based on the exploration of students and evaluation of activities already conducted at the Faculty of Computing of the University of Latvia to reduce dropout (School of Young Programmers, Compensative Course in High School Mathematics, Mentoring programs). Moreover, the supplementation of these activities by one of the existing programming attitude tests (or a combination of several tests) or a necessity to develop a new similar test is considered.

Keywords: **Aptitude test, attrition rate, computer science education, data analysis.**

I INTRODUCTION

In recent years the observed practically stable trend to lose from one-third to half of students in the first year of computing studies motivated us to explore, whether the world has not found indeed a method, how to determine in advance such applicants that have no chance to overcome the first study year.

On the one hand, the dropout students and the teaching staff have wasted their resources. On the other hand the Ministry of Science and Education and experts that are evaluating the study programs frequently associate the high dropout rate with a low effectiveness of the implementation of the study program (another matter – whether it is reasonable) and ask what is done to reduce the attrition.

To reduce the attrition, it was decided initially to perform a research about the factors influencing the attrition in Faculty of Computing at the University of Latvia [1]. Our study investigated students enrolled into the computer science bachelor and programmer professional programs in one year (2013). It were originally assumed, that following factors could have a potential impact on attrition:

1. High school grades (admission score)
2. Compensative course in high school mathematics
3. Intermediate grades for core courses
4. Prior knowledge in programming.

The results of the study in more detail can be found in [1]. Further, a short summary of the main conclusions is given to justify the choice of appropriate solutions to reduce the attrition.

There exists a large group of students, who in fact do not begin studies and who are later expelled. The majority of expelled students drop out in the 1st semester of the 1st year. These students include both students with weaker and with quite good high school grades. Since the 1st year dropout consists mostly of students, who have not really begun studies, so the high school grades do not have significant effect on dropout, but they may have an impact on the further study process.

The hypothesis that a programming background is an important factor influencing dropout was rejected, since we found out that the ratio between students with and without prior knowledge remained the same at the start and end of the first study year. This was concluded based on the data from self-assessment questionnaires. Also there is still a large number of students, who do not really begin studies, which could be explained by the wrong choice of a study program and insufficient insight into "what is programming", what the program developers are doing at work, and during the self-assessment test the knowledge in an informatics at school was possibly incorrectly treated as knowledge in programming.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.175

It is not necessary to prove, that every man has a different capability, however, in the society an opinion dominates, that every „normal" student is capable to acquire the basics of the most important sciences, e.g. language, mathematics etc. Therefore, they are included into the school curriculum. There is still an open question, whether in this set of skills and knowledge, also the programming should be included.

Even in most of high schools the programming is not on the list of mandatory subjects, and so it is in the whole world, despite the recent actions (e.g. [2]).

It is not proved, however, there is a widespread opinion that the programming is not among skills that should be taught mandatory, because a significant part of "normal" people can not learn them.

In the 1st study year it is meant usually by „programming" just writing a computer program in some programming language. Certainly, such limited understanding is quite far from all methods of developing computer programs [3], and includes not more than 20% of tasks in professional software development. However, it can be hardly imagined that software developer can be without these program coding skills.

According to the research results [4], universities should prepare an action plan to reduce the attrition, which, at the same time, does not lower the quality of studies [5] or admission requirements [6].

In the following part of the paper, a review about the programming aptitude tests is given, that can be used to reduce attrition. We will propose an action plan that is based on investigation of students and on analysis of already undertaken actions to lower the attrition rate ("The School of Young Programmers" [7], Compensative course in high school mathematics, Mentoring program). We will also outline considerations about supplementation of these actions with existing programming aptitude tests or combination of them, also the necessity to develop new targeted tests will be discussed.

II MATERIALS AND METHODS

Attempts to pre-select the most appropriate students for programming are at least 50-60 years old [8]-[10]. Searching *'computer programming aptitude test'* in *Google Scholar*, at least 60 publications for the period from 1960 until 2014 are obtained, but in neither of them, including the youngest publications, „silver bullet" unfortunately was not found.

All existing approaches to computer programming aptitude determination are conditionally divisible into two parts: the ones based on the psychological tests (for instance, [11]-[18]) and the ones based on solving specifically designed non-programming tasks (for instance, [19]-[21]).

In practice, hybrid tests [22] are often used, which contain elements to check the „logical reasoning, numerical problem solving, pattern recognition, ability to follow complex procedures and attention to detail". So, in addition to problem solving, such tests include also evaluation of various specific personality traits.

There also exist tests, which are used to self-asses the existing programming knowledge. Programming simulation includes „pseudocode, control structures (e.g. loops), look-up tables, sets, arrays, boolean true/false, looping and other programming structures".

If the time comes, when an adequate idea of programming and one's skills in this area will be obtained before applying to the university computer science study program, everything written above will become obsolete. While this is not the case, the most promising psychological [16], [18] and problem solving, for example, [22] self-test summary should be offered to the prospective students.

Psychological Tests

In one of the studies [16], Myers-Briggs personality types detection test is used to determine, which types of college students are doing better in the programming introductory course.

Myers-Briggs test is available, for example, at this website [23]. The test determines a total of 16 different personality types, 4 aspects are evaluated:
1) general attitude: Extraverted vs. Introverted;
2) perception function: Sensing vs. Intuition;
3) judging function: Thinking vs. Feeling;
4) perception – judging domination: Judging vs. Perceiving.

The study [16] concludes that „sensing students performed better on programming assignments than intuitive students, and that judging students achieved higher programming averages than perceptive students". Commenting on the findings of these types, it can be noted that *Sensing* means that an individual relies on a specific, topical information, but *Intuition* means that a person relies on his or her vision of the world. On the other hand, people with judging dominance perceive the world as an ordered structure that follows the set of laws, in contrast to the perceiving people, who perceive the world as a structure that can take different forms and results.

Another study [18] describes how results of various tests performed by students, correlate with their programming ability. It was determined that 2 tests SQ [24] and EQ [25] used together, showed the best correlation results. A test „Cambridge Personality Questionnaire" (SQ - Systemizing Quotient - test) was used, which determines how easily an individual understands object systems. The second test used in the study, was „*The Cambridge Behaviour Scale*" (*Empathy Quotient test - EQ*), which describes how easily an individual understands human emotions. A difference between the two test results: SQ - EQ. ($r = .67$) revealed the correlation with programming results. Separately, each of these tests showed significantly poorer correlation.

The authors of the study explained the obtained results with the fact that people with SQ significantly higher than EQ prefer dealing with ordered systems in everyday life, instead of dealing with people, who they do not understand. The authors also suggest that a large SQ – EQ value indicates the aptitude for studies that require great effort and practice that definitely corresponds to the programming studies.

Problem Solving Tests

One of the further examined examples [22] describes a test that already exists at the university, while the other shows the possibility to realize a test in a modern environment, attractive and habitual for pupils [26].

The test of the University of Kent includes questions about 1) „Logical thinking and problem solving", which essentially evaluates the ability to prevent problems arising during the development of IS, 2) „Pattern and syntax recognition", which essentially tests the ability to discern differences and pay attention to details, to verify the information, 3) „Ability to follow complex procedures", which assesses the ability to organize, process events, objects in a logical sequence, according to some regulations, assess the impact of an action on the future.

The test includes 26 questions. Answers are assigned points and groups are defined by the number of points, but the selection of specific questions and division into groups is unclear, since the test is based on the practical experience of its usage. This gives an insight into the test, which has already been implemented and used in practice at the university, however, the usage of such test in our case should be further explored and evaluated.

The second example, „Aptitude and Logical Reasoning" [26] test, which is implemented as a mobile application and is available in Google Play, is not primarily intended to measure the programming aptitude, but helps to improve problem solving skills and prepare for various exams and similar tests. There are many such mobile applications, which suggests that both the test and the its implementation environment are topical and demanded among users.

III RESULTS AND DISCUSSION

According to the conclusions of the primary study of the distribution of students [1], a majority of students really do not begin studies, therefore, activities targeting both „early" dropped out students and applicants are considered. Several ideas are planned to be studied further, for example, the possibility for students to assess themselves their suitability for a chosen field of study, by offering a variety of self-assessment tests - personality tests, logic tests, mathematics tests. For example, the contents of the latter test could be based on the compensative course in high school mathematics, which is already conducted at the Faculty of Computer Science. It could be implemented as a computerized test for self-assessment that candidate students could take before they choose the study program.

For example, the contents of the latter test could be based on the compensative course in high school mathematics, which is already conducted at the Faculty of Computer Science. It could be implemented as a computerized test for self-assessment that candidate students could take before they choose the study program. The groundwork in this direction has been started in the project "School of Young Programmers" of the Faculty of Computer Science at the University of Latvia [7]. The purpose of this project is to promote the comprehension about programming in high schools.

A detailed insight into both the study about students, the analysis of existing activities aimed at reducing the dropout and the new action plan follows.

Results of the Initial Survey - Reasons for the Choice of the Faculty of Computing

The reasons for the students' choice of the Faculty of Computing of the University of Latvia could be determined from the questionnaire completed by prospective students beginning their studies. In 2013 235 questionnaires were processed. Questionnaires are anonymous, summarized results are available. It is allowed to mark several reasons for the choice, and the most marked reasons are shown in the Table1.

I like everything related to computers	127	54,0%
Potentially high salary in the future	108	46,0%
I think that there is a high quality of study	98	41,7%
I have been programming and I like it	97	41,3%
I hope to learn a profession where I could easily get a job	89	37,9%
I am interested in mathematics	91	38,7%

TABLE 1.

REASONS FOR THE CHOICE OF THE FACULTY OF COMPUTING

The reasons with a positive impact on the further study process were identified as „I have been programming before and I like it" and „I am interested in mathematics". Many students also marked the reasons that can indicate misunderstanding of the profession, such as „I like everything related to computers", i.e. underestimation of the true nature of the program is possible, because the idea of computing may be influenced by the use of computers according to the contents of a high school informatics course. Another large part of students choose the reasons that could indicate even a lack of interest for the study content („potentially high salary in the future" and „I hope to learn a profession where I could easily get a job").

Since there is no information about the respondents of the questionnaire, the precise analysis of the connection between reasons and dropout is not possible, but conclusions can be drawn that before enrolment, students underestimate that the curriculum includes mathematical courses (less than a half of students are interested in mathematics), and that they will have to program. 41.3% of new students selected the answer „I have been programming and I like it", but for other students, it is unknown, what the outcome of their familiarization with programming will be, and it would be better to offer the opportunity to get an idea of programming earlier - before enrolment into the study program. While the programming subject is not mandatory in the high school curriculum for everyone, it would be good to create such opportunity.

The table does not include the reasons marked by less than 20% of respondents, including the influence of parents and friends, the opportunity to easily get into the group financed by the state, which are also risky reasons with high potential of dropouts, because these answers do not indicate students' interest and idea of the profession.

School of Young Programmers

The School of Young Programmers (SYP) [7] was created to promote the comprehension about computing [Fig.1]. The lectures available in the online environment allow to understand what a programming is. In 2014 at the lectures of the School of Young Programmers, a modern, intuitive programming language Scratch designed especially for schools, was studied. Scratch is a project of the lifelong learning group "Lifelong Kindergarten" of the Massachusetts Institute of Technology (USA). Although this language is easy to learn, it is possible to create quite complex programs in Scratch 2.0. "The School of Young Programmers" is offered by the Faculty of Computing of the University of Latvia with the support of the Fund of the University of Latvia and JSC "4finance".

Fig.1. E-environment of the School of Young Programmers

SYP offers tutorials of the programming language Scratch, program examples, video materials and environment [27], which allows pupils, who have not studied programming at high school, to learn programming by themselves in a short time. *Scratch* can be used at different levels of difficulty for pupils of different ages, therefore, it can also be used to evaluate, whether a pupil is interested in and is good at programming, before making a decision about studying at the Faculty of Computing or elsewhere.

Mentoring Program

The mentor program is designed at the University of Latvia to support a freshmen at the beginning of their study, because they have to contend with a large amount of new information and settle into an unfamiliar environment. The program works at almost all faculties, including the Faculty of Computing [28]. At the Faculty of Computing, the program is implemented by the Student Authority, in collaboration with the Alumni of the Faculty of Computing, the Student Council of the University of Latvia, Accenture Latvia and FranklinCovey. The participants of the program - mentors, who are senior students, have the opportunity to participate in free training to get new experience useful for their career. Freshmen, in turn, have the opportunity to get an advisor, who can provide information, advice and support. The mentoring program at the Faculty of Computing is designed to help students to earn a diploma, providing support directly in the first year when the dropout is the highest.

The activity of the mentoring program is described by the following statistics: 42 mentors and 66 freshmen applied for the program. Among freshmen, 80.3% of those who took part in the mentoring program, registered for the second semester, but only 63.21% of students who did not have mentor, registered for the second semester. So, among mentored freshmen, there are better results, however, different conclusions could be drawn: either the program really helped first-year students, or more active, interested students, who would cope with the studies either way, registered for the program.

Current Situation with the Compensative Course

The Compensative Course in High School Mathematics was introduced in 2009 with the purpose to expand knowledge of students of the Faculty of Computing, to improve the situation with the study process both for teachers and for students. Students are able to get knowledge on topics, which were not promptly acquired in high school, and the absence of this knowledge complicates the understanding of the university course content. The course also allows teachers to work with the audience with more homogeneous knowledge.

The course content includes topics that were identified during interviews with teachers and taking into account teachers' proposals about the essential topics of their courses, which should be known already from high school.

„Compensative Course in High School Mathematics" is mandatory for the students of the Faculty of Computing with the admission examination score less than 700 (1000 is the maximum possible). The course is taught by means of lectures on the the planned course topics. The course is conducted in the 1st and 2nd semesters of the first study year and the course topics are divided into the groups: Algebra and Functions (in the fall semester), as well as Trigonometry and Geometry (in the spring semester). Students have to take a test on each of these groups of topics and when they get an assessment „passed" for each of the two groups in the corresponding semester, the final test for the semester is obtained. Students are allowed to take each of these tests repeatedly several times. The e-learning environment Moodle is used to inform students about their results. However, the tests themselves have not been implemented in the e-learning environment. If a student passes a test faster than at the end of the semester (if a student is sure that he or she has acquired unclear issues, then he or she can take tests more quickly), then a student is allowed not to attend this course anymore. This course was offered to 122 students out of 254 students enrolled in 2013.

The Compensative Course in High School Mathematics directly represents both the necessary training in mathematics and students' personal qualities, such as motivation, commitment and perseverance, because a student does not have to acquire a new difficult content, but has to learn deeper topics, which have already been studied at high school, to avoid potential problems in further studies.

It should be recalled that this course targets those 122 students with admission score of less than 700, which includes grades for high school mathematics. Therefore, analysing students by the mathematics knowledge aptitude indicator, it is more important to analyse further exactly the weakest, less suitable, group of students, to plan further support activities. However, as a study of dropout showed [1], there is also a high dropout rate among students with more than 700 points, which is why the pre-enrolment activities aimed at prevention of dropout due to other reasons, including the wrong choice, are recommended to be conducted among this group of students as well.

Therefore, some of the results are described at a glance, which could be used to distinguish separate student subgroups that could benefit from different supportive activities to reduce dropout. The full results of the dropout study can be found at the paper [1].

Results of a study on the outcomes of the Compensative Course in High School Mathematics allowed to distinguish 3 groups of students: „do-nothing students" - 15 students (100% of them dropped out), which did not attempt to pass any of the tests, the 2nd group consisting of 74 students who sooner or later passed the tests (30% of them dropped out) and „failers" - 33 students who failed to pass the tests (88% of them dropped out). Conclusions about students were made based on the available information about the history of passing the course. Among „test failers", there are mostly the students who have quickly given up (attempted to pass the tests only once), but more than a half of those, who passed the tests, are the students, who attempted to pass the tests several times. These data indicate that the result depends not so much on the level of high school knowledge, as on other reasons – interest or motivation (or lack of them) to work.

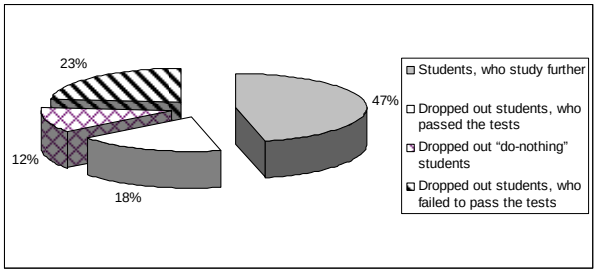

Fig.2. Groups of test takers

So the following groups of students can be distinguished (see Fig.2):

1) Students, who study further – high school knowledge is acceptable after taking the course
2) Dropped out students, who passed the tests – high school knowledge is acceptable after taking the course, the motivation has been sufficient to pass the course, however, students do not continue studies
3) Dropped out „do-nothing" students – wrong choice of the program and lack of motivation, no data on the knowledge in this course
4) Dropped out students, who failed to pass the tests – lack of motivation, high school knowledge is not acceptable as the course was not passed.

Proposed Plan for Dropout Reduction

The Table 2 shows the evaluation of the above-mentioned groups, which activities and when should be planned.

According to the characteristics of each group, the time when activities should be conducted, as well as the type of the support activities, is identified. The existing compensative course in high school mathematics should be retained at the beginning of studies for students with admission scores less than 700 points. Since the previously given distribution into groups is possible only after the fact of dropping out, then, as a preventive measure to reduce dropout caused by other reasons, such as difficulties in settling

in the university environment, difficulties in orientation in the study process, a mentoring program should be developed and popularized at the beginning of studies.

TABLE 2.
ACTIVITIES FOR DROPOUT REDUCTION

Groups	When the activity should be offered?	Activity
Students, who study further	At the beginning of studies	Existing compensative course in mathematics
Dropped out students, who passed the tests	At the beginning of studies	Mentoring program, existing compensative course in mathematics
Dropped out "do-nothing" students	Before enrolment	Aptitude tests
Dropped out students, who failed to pass the tests	Before enrolment	Aptitude tests

On the other hand, to prevent the enrolment of students inappropriate for the programmer profession, which correspond to the latter two groups, profession aptitude tests should be offered, as well as educational and promotional activities for pupils, explaining what is programming and which knowledge is required for a programmer, where one of the tools would be the promotion of the School of Young Programmers.

Our proposals in the area of the professional aptitude test are the following:

1) To implement the assessment test of the compensative course in mathematics as an online test to evaluate the knowledge in mathematics,
2) For the evaluation of programming skills, if perspective students have such skills, the supplementation of materials of the School of Young Programmers with a set of exercises organized in the form of a test, where the execution of these exercises would not exceed a certain time limit. The test would serve as an initial insight into the aptitude measurement and attraction of interest, and the rest of the materials available in the e-environment could be then used for the in-depth understanding.
3) Psychological tests [16], [18], the description of which has already been given above,
4) Problem solving tests [22], [26] - a combination of the best examples.

The proposed package should be promoted before the enrolment.

IV CONCLUSIONS

There are various tests for programming aptitude measurement, at the same time, the specificity of the national education system and the university should be taken into account to evaluate the plan for the most appropriate activities to reduce dropout.

In this paper, the action plan, which, in addition to the programming aptitude tests implemented in a self-assessment form, also includes other activities that should be used both before and after the enrolment and require the involvement of the faculty staff, was proposed.

One of the most important further tasks is the combination of the best examples of the problem solving tests – including both the content and the form – by offering content in the Latvian language as well as conducting the content selection or novelty. The further tasks are related to this test and approbation of the whole action plan and determination of aptitude criteria.

V REFERENCES

[1] J. Borzovs, L. Niedrīte, and D. Solodovņikova, "Factors affecting attrition among first year computer science students: the case of University of Latvia", 2015, submitted for publication.
[2] Computing at School, [Online]. Available: http://www.computingatschool.org.uk/ [Accessed: Mar.17., 2015]
[3] D.Bricklin, "Why Johnny can't program", August 2002, [Online], Available: http://www.bricklin.com/wontprogram.htm [Accessed: Mar.17., 2015]
[4] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates", Computers and Education, vol. 53(3), pp. 563-574, 2009.
[5] L. Paura and I. Arhipova, "Cause Analysis of Students' Dropout Rate in Higher Education Study Program", Procedia-Social and Behavioral Sciences, vol. 109, pp. 1282-1286, 2014.
[6] L. Grebennikov and M. Shah, "Investigating attrition trends in order to improve student retention", Quality Assurance in Education, Vol. 20 (3), pp. 223 – 236, 2012.
[7] "School of Young Programmers" [Online]. Available: http://www.df.lu.lv/nacstudet/jauno-datoriku-skola/, [Accessed: Mar.17., 2015].
[8] T.C. Rowan, "Psychological tests and selection of computer programmers", Journal of the Association for Computing Machinery, Vol. 4, pp. 348-353, 1957.
[9] W.J. McNamara, J.L. Hughes, "Manual for the revised Programmer Aptitude Test", New York: international Business Machines Corporation, 1959.
[10] A.W. Stalnaker, "The Watson-Glaser Critical Thinking Appraisal as a predictor of programming performance", Proceedings of the Third Annual Computer Personnel Research Group, pp. 75-77, 1965.
[11] E. Spranger, " Types of Men: The Psychology and Ethics of Personality", Johnson Reprints, New York, 1966.
[12] C.K. Capstick, J.D. Gordon, and A. Salvadori, "Predicting performance by university students in introductory computing courses", ACM SIGCSE Bulletin Vol. 7(3), pp. 21-29, 1975.
[13] R.E. Mayer, J.L. Dyck, and W. Vilberg, "Learning to program and learning to think: What's the connection?" Commun. of ACM Vol.29 (7), pp. 605-610, 1986.
[14] D.A. Scanlan "The mental abilities associated with programming aptitude", CSC'88 Proceedings of the ACM sixteenth annual conference on computer science, p.737, 1988.
[15] G.E. Evans and M.G. Simkin, "What best predicts computer proficiency?" Commun. of ACM Vol.32(11), pp. 1322–1327, 1989.
[16] C. Bishop-Clark and D.D. Wheeler, "The Myers-Briggs personality type and its relationship to computer programming", Journal of Research on Computing in Education, Vol. 26(3), pp. 358-370, 1994.
[17] C.C. Cegielski, J. Dianne, and D.J. Hall, "What makes a good programmer?", Commun. of the ACM, Vol. 49(10), pp. 73-75, 2006.

[18] S. Wray, "SQ Minus EQ can Predict Programming Aptitude", Proceedings of the PPIG 19th Annual Workshop, pp. 243-254, 2007.

[19] M. Tukiainen and E. Mönkkönen, "Programming aptitude testing as a prediction of learning to program", In: J. Kuljis, L. Baldwin, and R. Scoble (Eds). Proc. PPIG 14, pp. 45-57, 2002

[20] S. Dehnadi, R. Bornat "The camel has two humps", 2006, [Online], Available: http://wiki.t-o-f.info/uploads/EDM4600/The%20camel%20has%20two%20humps.pdf [Accessed: Mar.17., 2015]

[21] T. Lorenzen, H.-L. Chang, "MasterMind©: a predictor of computer programming aptitude", ACM SIGCSE Bulletin, Vol. 38 (2), pp. 69-71, 2006.

[22] University of Kent, Computer Programming Aptitude Test [Online], Available: http://www.kent.ac.uk/careers/tests/computer-test.htm [Accessed: Mar.17., 2015]

[23] Carl Jung's and Isabel Briggs Myers' typology test, [Online], Available: http://www.humanmetrics.com/hr/you/personalitytype.aspx [Accessed: Mar.17., 2015]

[24] Autism Research Centre, Cambridge, "Cambridge Personality Questionnaire" (SQ test). Available: http://www.autismresearchcentre.com/arc_tests, [Accessed: Mar.17., 2015]

[25] Autism Research Centre, Cambridge, "The Cambridge Behaviour Scale" (EQ test), Available: http://www.autismresearchcentre.com/tests/eq_test.asp, [Accessed: Mar.17., 2015]

[26] Android Apps on Google Play: Aptitude and Logical Reasoning, [Online]. Available: https://play.google.com/store/apps/details?id=com.madguy.aptitude.lr [Accessed: Mar.17., 2015]

[27] I. Gorbāns, "Programming environment for every students – Scratch or basics of programming in few hours", E-book, 2014, [Online]. Available: http://skolas.lu.lv/mod/book/view.php?id=29857 , [Accessed: Mar.17., 2015]

[28] Mentoring program of Faculty of Computing at University of Latvia, [Online]. Available: http://www.df.lu.lv/zinas/t/28406/ [Accessed: Mar.17., 2015]

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 36-42

Factors Affecting Attrition among First Year Computer Science Students: the Case of University of Latvia

Juris Borzovs, Laila Niedrite, Darja Solodovnikova
University of Latvia, Faculty of Computing.
Address: Raina blvd 19, Riga, LV 1586, Latvia

Abstract. **The purpose of our study was to identify reasons for high dropout of students enrolled in the first year of the computer science study program to make it possible to determine students, who are potentially in risk. Several factors that could affect attrition, as it was originally assumed, were studied: high school grades (admission score), compensative course in high school mathematics, intermediate grades for core courses, prior knowledge of programming. However, the results of our study indicate that none of the studied factors is determinant to identify those students, who are going to abandon their studies, with great precision. The majority of the studied students drop out in the 1st semester of the 1st year, and the dropout consists mostly of those, who do not really begin studies. Therefore, one of the main conclusions is such that the planned activities of informing about the contents of the program should be carried out, and the perspective students should be offered a possibility to evaluate their potential to study computer science before choosing a study program.**

Keywords: **Attrition rate, computer science education, data processing, data analysis.**

I INTRODUCTION

In recent years, there has been a competition among applicants of the undergraduate programs of the Faculty of Computing at the University of Latvia. Besides, 90% of all entered students have identified the program as their 1st priority. This confirms that there has been a real selection, which identified relatively good and motivated students. However, the considerable number of students that annually abandon studies especially in the first study year (an average of 30% after the first semester, close to 50% over the 1st study year) and comparison with other countries encourages to investigate the causes of attrition.

One of the indicators used to analyze tertiary education, which also includes university level education, is “completion rate”. It is defined as the ratio between the number of students in the study program, which obtained a degree in the reference year, and the number of students enrolled in the study program so many years ago, as the length of the full-time study program [1]. The “completion rate” of the university programs from 23 countries participating in OECD study varies, ranging from 93% in Japan and ending with indicator below 60% in several countries, including Sweden (49%, the worst indicator in the study).

According to OECD method of calculation in 2013, using data about the number of students, who obtained a degree in 2013 [2], and the number of students enrolled in 2009 [3], the “completion rate” of the computer science bachelor program of the University of Latvia was 47%. In comparison with the OECD study, the result is below the worst indicator, but it should be noted that OECD indicators characterize education at the national level, moreover, research shows that indicators of the engineering and information technology programs are worse than in other fields. However, this example and the situation in the first year computer science program mentioned at the beginning shows that there is room for improvement. The results of the other countries studies confirm that in different countries there are diverse factors affecting attrition and their combinations, therefore, local studies about causes for attrition in Latvia are justified.

In OECD studies a related indicator “non-completion rate” is also used. The average value of this indicator in all studied countries [1] was 30%. In other reports, the authors of the paper [4] use "dropout" concept, which is stronger, because in the calculation of the "non-completion rate” such students who fall behind the planned study schedule (for example, take a break in studies) are also taken into account, but the dropout examines only those leaving the studies without a degree. In many educational

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.174

studies the concept of drop-out is used, which is why exactly dropout will be analyzed in our study.

In terms of regularities, which the researchers worldwide are unanimous about, the dropout percentage varies depending on the academic year, besides the highest dropout is in the first study year and then it decreases [5]. For example, in the USA four-year colleges and universities, 38% of all dropouts are exactly in the first academic year, and 29% - in the second academic year [5]. The dropout rate of the second study year is closely related to the results of the first study year, so the first study year is the most important in terms of impact on further studies, including prevention of drop-outs. Therefore, the 1st study year has been selected for our study. Our study investigates students enrolled into the computer science bachelor and programmer professional programs in one year (2013) and analyzes what have these students been doing over time, including how many of them have dropped out of the University of Latvia and what is the ratio of these study leavers among enrolled students, i.e. „dropout percentage".

II MATERIALS AND METHODS

The purpose of our research is to analyze a study model of the first year students enrolled in 2013 into the computer science bachelor program of the University of Latvia in order to evaluate, when the "dropout" takes place in the first year, what are the characteristics of "dropping out" students, what are the mutual influence of these characteristics, as well as to create a model to make it possible to identify students, who are potentially in risk.

As data sources, similar studies often use data from university management information systems, using data supplied by a student during enrolment, for example, age, gender, high school grades, etc., as well as study results [6]-[7]. The second data source is the survey [7]-[9], which is conducted beginning studies and leaving them, for example, to assess student's academic, psychosocial and cognitive characteristics. Studies tend to include both the data about all academic years during several years [6] and the data about the first year [8]-[9]. Besides, studies tend to analyze all university programs, or separate faculties, for example information technology programs [7], [10]-[11].

Data used in our study come from multiple data sources. Admissions data are obtained from the management information systems of the University of Latvia. These include personal data and high school exam grades. To be admitted into the Faculty of Computing at the University of Latvia, grades of the centralized exams in math (75% of the admission score) and Latvian language (25% of the admission score) are used in the competition. The maximum total score is 1000 points. The outstanding applicants, who have additional points for participation in the national math or informatics Olympiads, are considered separately. Since the admissions data are also stored in the data warehouse of the University of Latvia, several previously existing reports were used for the analysis of the distribution of applicants based on the admission score (without integration with other data).

To analyze the students' study process, data about intermediate grades (tests, homework, practical tasks and other grades) for several courses were used. These were the two compulsory courses "Algebra" and "Programming". The first one represents mathematics, the other is a specialty course. Data were available from Excel tables provided by the teachers of the courses.

The study also includes data about the "Compensative Course in High School Mathematics", which is compulsory to all students of the Faculty of Computing, whose admission score is below a certain number of points (700 out of 1000 maximum possible score). These data were obtained from the e-learning environment Moodle grade book data, which stores intermediate grades. In this case, these are data on whether a student has passed 2 tests and how many times he or she has taken the tests.

Data from two surveys were also used in the study. The first survey was carried out in 2013. 235 newly admitted students took part in it beginning their studies. The survey was anonymous, this is why only aggregated results were available. In our study we used aggregated answers to questions about students' prior knowledge in programming. The second of the surveys was carried out in January 2015 during the session. All the students who continued to study (out of those enrolled in 2013, i.e. "survivors") were interviewed. The same question of the first survey about the prior knowledge in programming was asked, but this time respondents were identified. The question contained several possible answers, such as "I studied programming in a school programming course", "I studied programming myself," etc., as well as the opportunity to answer that "I did not study programming before ".

In studies of other authors, statistical methods are used the most often [6]-[7], [10]-[12]. Sometimes, data mining [13] and data warehouse [10] are used in studies. The latter is mainly used for the data preparation, integration and supply of data for systematic analysis.

In our study, the infrastructure of the data warehouse of the University of Latvia was used as an environment for many data source integration, since the part of data was already available in it. The data warehouse is implemented using Oracle database and Microstrategy tool for reporting. MS Excel tool was used for data analysis and results visualization for the purpose of this study.

III RESULTS AND DISCUSSION

In our study we examine the first study year. We pay more attention to the first semester, when the dropout percentage is the biggest throughout the whole study process (see Table 1 and Fig. 1).

TABLE 1.

DYNAMICS OF THE NUMBER OF STUDENTS DROPPING OUT

Changes of study status in time	Number
Students enrolled in both programs in total	254
Dropout after the 1st semester	85
Dropout after the 2nd semester	39
Total dropout after the 1st study year	124

The fact of dropping out is detectable only at the end of the semester, rather than in relation to the time dimension (months) as in [11], because in our case, the withdrawal often occurs because the student does not register for the next study semester. Very few students write applications for the withdrawal of their own volition. This is why the possibility to carry out a survey and to determine the true causes of dropout is also limited. In addition to the examination of all students enrolled in 2013, we found out that for the analysis of dropout, it is necessary to distinguish and examine the subset of students, who actually do not start studies (however, they begin studies formally, since they sign the contract), because there are no intermediate results about fulfilled study excercises.

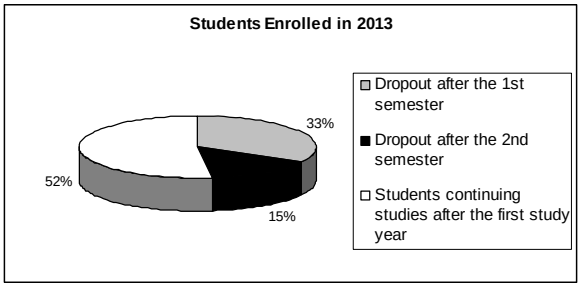

Fig.1. Distribution of enrolled students after the 1st study year

We studied a number of influencing factors, which were originally assumed to have a potential impact on attrition:

1) High school grades (admission score)
2) Compensative course in high school mathematics
3) Intermediate grades for core courses
4) Prior knowledge in programming

School Grades (Admission Score) and Dropout

The first factor that was analyzed, was the „high school grades", which were analyzed by the total score (taking into account both the mathematic and the Latvian language exams), because this score is used in the entry competition, as well as the mathematics exam was analyzed separately, which could have different results assuming that part of applicants are Latvian non-native speakers. The hypothesis of the study is that because of the weak school knowledge, it might be difficult to study. To analyze the total admission score, it was divided into intervals of 100 points, starting with 500, because the smallest score of 2013 applicants, who beat the competition, was 573.5 points.

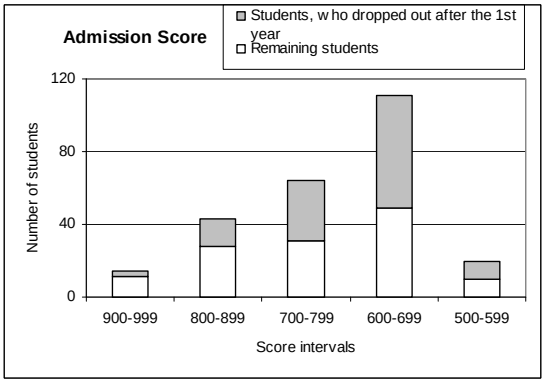

Fig.2. Distribution of remaining students and students, who dropped out, by admission score

The results of the study (see Table 2 and Fig. 2) show that the score at intervals of up to 800 points, which covers 78% of enrolment, the ratio between dropout and "survivors" is close to 50%. Slightly better results are among very good students (22% of the total number of students, score interval 800-1000), but the dropout is also present. Since it is not confirmed that majority of students with worse admission score drop out, but students with better admission score do not, then there is a basis for further research.

TABLE 2.

DYNAMICS OF THE NUMBER OF STUDENTS, WHO DROPPED OUT

Admission Score	Total Remaining Students	Dropout	Total Enrolled	Dropout (%)
900-999	11	3	14	21,43
800-899	28	15	43	34,88
700-799	31	33	64	51,56
600-699	49	62	113	54,87
500-599	10	10	20	50
Total	130	124	254	48,82

It is possible to distinguish and analyze students admitted outside of the competition, who were granted extra points for winning national Olympiads in mathematics or informatics (additional 1000 points for each Olympiad). In the following table (Table 3), it can be seen that none from the 6 outstanding students dropped out.

TABLE 3.
NUMBER AND STATUS OF ENROLLED WINNERS OF OLYMPIADS

Admission Score	Remaining Students	Dropout
2900-2999	1	0
2800-2899	1	0
1900-1999	1	0
1800-1899	3	0

Another hypothesis that was studied was the grades for high school mathematics exam as the dropout indicator.

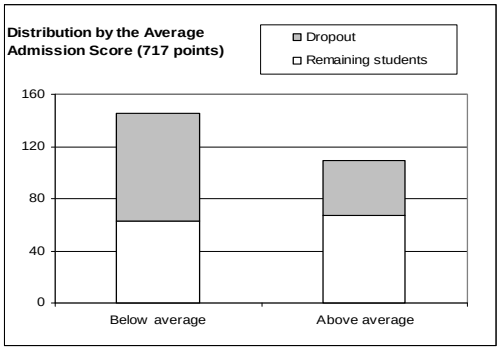

Fig.3. Remaining students and students, who dropped out, by the average admission score

A dropout analysis was carried out, taking into account only math admission examination scores and taking into account the total number of points. In this study, the number of students, who dropped out and the number of remaining students above and below the average number of points were compared: average admission score was 717 (Fig. 3), but taking into account only the math exam, the average score was 566 (Fig. 4). As shown in the charts, the significant changes in the distribution are not observed, therefore, the total score was used for further research, which officially served as the grounds for admission to the program.

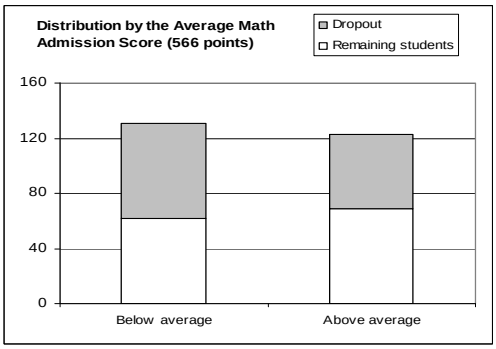

Fig.4. Remaining students and students, who dropped out, by the average math admission score

Compensative Course in High School Mathematics

The 2nd factor that also allows to analyze the impact of the students' school mathematics knowledge on the dropout is the „Compensative Course in High School Mathematics" organized by the Faculty of Computing. Students must participate in the course, learn topics of high school mathematics selected according to the needs of the program, and pass 2 tests. It is allowed to try to pass the tests several times.

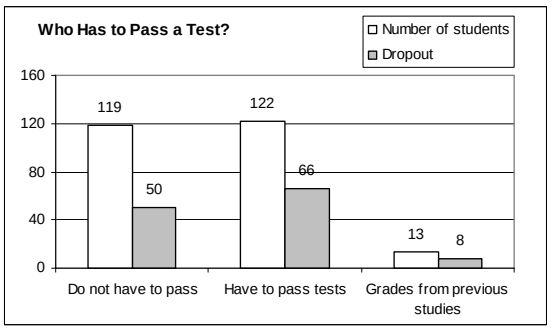

Fig.5. Student distribution in Compensative Course in High School Mathematics

Students, who have more than 700 points, do not pass tests and receive an automatic grade for both tests. So this group shows the same situation observed before, i.e. there is a big dropout (42%) among students with good knowledge of mathematics. The composition of all students and the dropout in comparison with the total number in each group, can be seen in the chart (Fig. 5).

For further research, those students are essential, for whom the effect of the compensative course can be examined in more detail; it is for those, who have to pass the tests (Fig. 6). The dropout rate is higher (54% of the total number), but the history of passing the course tests allows us to find out more information about the students passing the tests.

In this course (as it is also shown further in the paper in relevance to other subject analysis), the group can be distinguished that includes students, who do not do anything (there is no attempt to pass the tests) - 15 students, all of them were subsequently expelled.

Others can be divided into 2 groups. One group includes those students, who learn the course, sooner or later pass the tests (the chart shows only the results and does not reflect, how many times students try to pass the tests). A conclusion about this group can be made that these students meet the necessary requirements for knowledge of high school mathematics, and 30% of these students drop out.

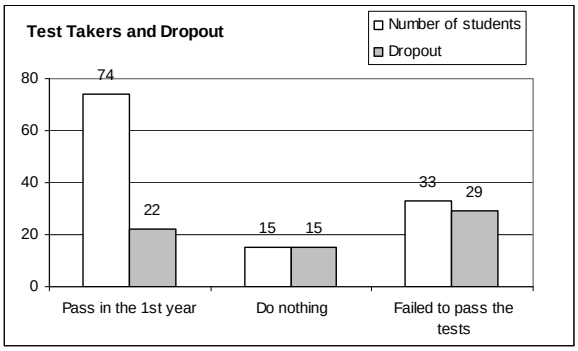

Fig.6. Outcomes of test takers and dropout

However, the group of students, who "failed to pass the tests" on the one hand confirms students' poor knowledge of high school mathematics, but the course also did not help (although, there are no data on whether the course was attended, or students just tried to pass the tests), and over time, students are expelled (at the time of data analysis, only 4 out of 33 students still had „student" status, so there was still a hope that they pass this debt). On the other hand, analyzing the data about repeated test passing by those students who "failed to pass the tests", it can be concluded that only 8 (out of 33) students tried to pass the tests repeatedly. Another ratio is observed among those, who passed the tests - 33 students are such, who passed the test on the first go, the other 41 students tried to pass the tests repeatedly, 6 of them did it more than 2 times. These data indicate that the outcome depends not so much on the school knowledge, but the motivation to work.

Intermediate grades for core courses

The 3rd factor, whose impact on dropout has been studied, was two core courses taught in the 1st semester. One of these two courses was "Algebra" that represented a mathematics course, and the second course was "Programming" that represented a specialty course. Intermediate grades for both courses were evaluated and an obvious group of "do-nothing" students was revealed. The first consecutive test was analysed for these students in both courses, whether there was or was not any grade for it (i.e. whether a student tried to write the test). Analysing other intermediate grades, there was also not more than one (typically, none) submitted assignment of other type (homework, practical work tasks, etc.). There are 45 students, who do nothing in both subjects (in the 1st semester). The number of "do-nothing" students separately for each course is slightly higher (Fig. 7). For example, there are 58 dropped out "do-nothing" students in the 1st semester Algebra course, but if we check whether any of the "do-nothing" group has "survived" (passed the debt or intend to do so later), then there are 2 such students. In Programming course, there are 49 dropped out "do-nothing" students. For comparison, it should be noted that the total dropout in the 1st fall semester is 85 students. It can be seen that the majority of them are the "do-nothing" students of the Algebra course (68% of the fall semester dropout). A conclusion can be drawn that non-writing the 1st test in Algebra is a sign that the student will drop out, and here we do not mean unsatisfactory grades, but rather non-writing the test at all.

We analysed the above-mentioned 45 students who did nothing in both core courses, and investigated what they did in compensative course in high school mathematics. We found out that 13 students failed to pass the tests (as research showed, these were mostly the students, who tried to pass the tests once and did not make an efforts to re-pass the test), 15 students did nothing in this course as well and only 1 student passed the test, while 16 students got an automatic grade (i.e. admission score was higher than 700). A conclusion can be drawn that there is a group of students that practically do not begin studies, i.e. 18% (45 out of 254 students, who began their studies). To conclude, a total of 33% of students were expelled in the 1st semester, i.e. a little more than a half of these students had not begun their studies at all. Speaking about the reasons for non-beginning studies, quite a large ratio of "do-nothing" students (16 students) among those with an admission score of more than 700 points rather indicate other reasons, which are not related to grades.

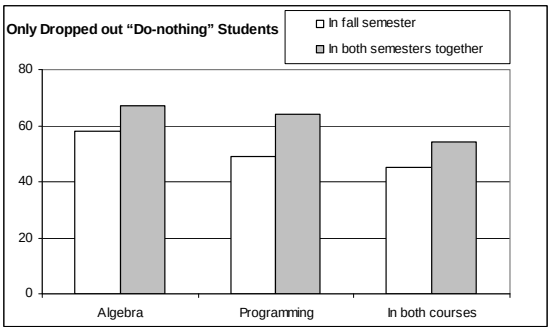

Fig.7. Dropped out "do-nothing" students – comparison by courses

Prior Knowledge in Programming

The 4th factor that was studied, was the impact of no prior knowledge in programming on dropout. Two surveys were carried out. The organization and data collection of them have already been described in the paper. The results of the surveys are given in the Table 4.

22% out of 235 students indicated that they did not have any prior knowledge in programming in the 1st initial survey before the start of the studies, but in the repeated survey of the "survivors" after the first year, 24% out of 114 respondents indicated the same answer. It can be concluded that the initial assumption that students with no prior knowledge in programming are at greater risk to drop out, is invalid, because the

initial ratio among students with and without prior knowledge remained the same.

TABLE 4.

RESULTS OF SURVEYS ABOUT PRIOR KNOWLEDGE IN PROGRAMMING

Questions	1st survey (%)	2nd survey (%)
I studied programming in a high school programming course	43	50
I took part in the informatics Olympiad	14	25
I studied at the training centre „Progmeistars”	4	5
I studied at the programming courses	7	8
I participated in the competition „Bebr@s”	4	2
I studied programming myself	49	48
I learned programming at workplace	13	12
I did not have any prior skills	22	24
Other	4	13

IV CONCLUSIONS

Looking at similar studies that analyze exactly engineering or information technology fields of study, since there is usually one of the highest dropout percentage, [10], [12], we have concluded that dropout is mostly affected by high school final grades, student's average grade during studying at a university, as well as selected program.

Engineering study programs were also studied in Latvia, at Latvian University of Agriculture [11] (34.4% drop out rate in the 1st study year), and results are similar that the main causes of dropout are poor knowledge of high school graduates and selected field of study.

The main conclusions of our research are such that there exists a large group of students, who in fact do not begin studies and who are later expelled. Failure to begin studies refers to the number of courses. In this study, among all analysed courses, dropped out students were best characterised by the fact that they did not write the 1st test in Algebra.

The majority of expelled students drop out in the 1st semester of the 1st year. These students include both students with weaker and with quite good high school grades. The exceptions are the applicants with excellent grades, where the situation differs with a smaller dropout. Since the 1st year dropout consists mostly of students, who have not really begun studies, so the high school grades do not have significant effect on dropout, but they may have an impact on the further study process.

The hypothesis that a programming background is an important factor influencing dropout was also rejected, since we found out that the ratio between students with and without prior knowledge remained the same at the start and end of the first study year.

The research results show that universities should create an action plan to reduce dropout [10], but the University policy should not jeopardize the quality of studies (reduced study [11] or admission [8] requirements lower the level of graduates).

Student characteristics (profile) are different, but these differences are not taken into account in the university requirements (environment, process, etc.) ([10], for example, speak about the group of students above 25 years of age, who have more difficulties to adapt to the university requirements, which leads to dropout). It is necessary to carry out adapted activities for „early dropped out students” and „late dropped out students” [12]. For example, further data analysis is required for late dropped out students, to search for individual solutions promptly, but for early dropped out students, additional information should be provided to help adapt to the study environment.

Definitely, a sufficient information about the provided programs should be offered to potential students before they begin studies [6], [8], and qualified advice about the choice of courses should be given after the start, to reduce the dropout of students, for whom "the course was not such as they were expecting" [8].

The further research should focus on the analysis of grades for other courses in the 1st and later study years, to assess the impact of the outcomes of the 1st year on the further study process and dropout (academic debts, change of financing because of academic debts, etc.).

Since our study concluded that a majority of students really do not begin studies, the activities targeting both “early dropped out students” and candidates are considered. In relation to this issue, several ideas could be further studied, for example, the possibility for students to assess themselves their suitability for a chosen field of study, by offering a variety of self-assessment tests - personality tests, logic tests, mathematics tests. For example, the contents of the latter test could be based on the compensative course in high school mathematics, which is already conducted at the Faculty of Computing. It could be implemented as a computerized test for self-assessment that candidate students could take before they choose the study program.

Similarly, it is planned to help get more insight into "what is programming", because, although the study showed that a lack of prior knowledge in programming did not directly cause a dropout, there is still a large number of students, who do not really begin studies, which could be explained by the wrong choice of a study program. Therefore, it is necessary to offer a way to help get an idea about the profession by means of an interactive test. The groundwork in this direction has been started in the project “School

of Young Programmers" of the Faculty of Computing at the University of Latvia [14]. The purpose of this project is to promote the comprehension about programming in high schools.

A more detailed analysis of the planned activities, as well as a comparison with studies of authors from other countries and proposals to reduce the dropout rate is described in a separate paper and the activities are planned to implement in the near future.

V REFERENCES

[1] OECD, "How many students drop out of tertiary education?", in Highlights from Education at a Glance 2010, OECD Publishing, 2010.

[2] Overview about higher education in Latvia, Available: http://izm.izm.gov.lv/registri-statistika/statistika-augstaka/2013.html , [Accessed: Mar.17., 2015].

[3] Overview about higher education in Latvia, Available: http://izm.izm.gov.lv/upload_file/Registri_statistika/Pielikums_2009.xls , [Accessed: Mar.17., 2015]

[4] OECD, "Education at a Glance 2004: OECD Indicators", OECD Publishing, 2004.

[5] V. Tinto,. "Completing college: Rethinking institutional action", University of Chicago Press, 2012.

[6] F. Belloc, A. Maruotti, and L. Petrella, "University drop-out: an Italian experience", Higher Education, vol. *60*(2), pp. 127-138, 2010.

[7] K. McKenzie and R. Schweitzer, "Who Succeeds at University? Factors predicting academic performance in first year Australian university students", Higher Education Research & Development, vol. 20(1), pp. 21-33, 2001.

[8] L. Grebennikov and M. Shah, "Investigating attrition trends in order to improve student retention", Quality Assurance in Education, Vol. 20 (3), pp. 223 – 236, 2012.

[9] A. L. Reynolds and M. J. Weigand, "The relationships among academic attitudes, psychological attitudes, and the first-semester academic achievement of first-year college students", Journal of Student Affairs Research and Practice, vol.47 (2), pp. 175–195, 2010.

[10] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates", Computers and Education, vol. 53(3), pp. 563-574, 2009.

[11] L. Paura and I. Arhipova, "Cause Analysis of Students' Dropout Rate in Higher Education Study Program", Procedia-Social and Behavioral Sciences, vol. 109, pp. 1282-1286, 2014.

[12] N. Gury, "Dropping out of higher education in France: a micro-economic approach using survival analysis", Education Economics, vol.19(1), pp. 51-64, 2011.

[13] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting School Failure Using Data Mining", In Educational Data Mining, pp. 271-276, 2011.

[14] "School of Young Programmers" http://www.df.lu.lv/nacstudet/jauno-datoriku-skola/, [Accessed: Mar.17., 2015].

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 43-46

Instrument of determination and prediction of public opinion using IPTV statistic data

Vjačeslavs Dubovskis
Rezeknes Augstskola, Faculty of Engineering

Abstract. **Recent tendencies show an increase of information pressure on Latvia. At the moment, Latvian political figures and government persons have recognized that the country takes a hybrid war. There is a need for a new instrument of determination and prediction of public opinion. The new instrument should allow: profiling audience by identification and prediction of the rate of dissemination of information, forecasting the depth of information, forecasting how public opinion is formed, identifying and previse the adoption of the reported information. The results of the work will create a tool that a variety of institutions will be able to apply it. These institutions may be television channels, government agencies, research organizations, and others. The tool will not only accurately determine the current picture of situation, but also with high accuracy to make forecasts about the state of public opinion.**

Keywords: **Digital multimedia broadcasting, Channel Popularity, IPTV, Network Measurement, Modeling.**

I INTRODUCTION

Recent tendencies show an increase of information pressure on Latvia. The society is exposed to the polarization of information. On the one hand, there is Russia, who does it by using government media resource. On the other hand, at the same time, correcting the situation of public opinion, Latvia and Europe.

The public opinion is used with some groups to destabilize targeted regions even further [1]. These groups have become masters of the "battle of narratives" – a struggle beyond the physical elements of a conflict in which the manipulation of the media, the use of the internet, and the integration of information operations with a strategic communication program are as important as weapons systems or even success on the battlefield [7].

The instruments used to determine the sign of the hybrid war are clear. They are based on the identification of signs. But the levels of invasion of the information are measured by outdated methods as sociological research. This method does not provide accurate data, since it is based just on a respondent answers, which can be known to be true. Sometimes the error exceeds more than 50%. This is due to the emotional impact when the respondent does not want to disclose the real opinion [2]. Such researches do not provide operational data. They also do not show the dynamics of objective indicators. The perception of information differentiation is extremely low [3].

There is a need for a new instrument of determination and prediction of whole public opinion. This instrument can be a model of public opinion changes.

The new instrument should consist from:

- profiling the audience with the possibility of excreting by separate groups / classes;
- identifying and predicting the speed of propagation of information;
- forecasting the depth of information;
- forecasting the possible public opinion;
- identify and anticipating the adoption of the reported information.

II METHODOLOGY OF THE RESEARCH

Many technological advances were produced by trying to meet user needs and expectations in such a widespread media. For example, the large number of users that concurrently watch TV sparked the use of IP multicast in major Internet TV (IPTV) designs to save on the transmission cost [6].

IPTV television is one of the most successful platforms for filling model of relevant data, as anonym data are coming from the Middleware software system, what accurately shows the interaction of audience with the incoming information. IPTV System Architecture is shows on Figure 1.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.181

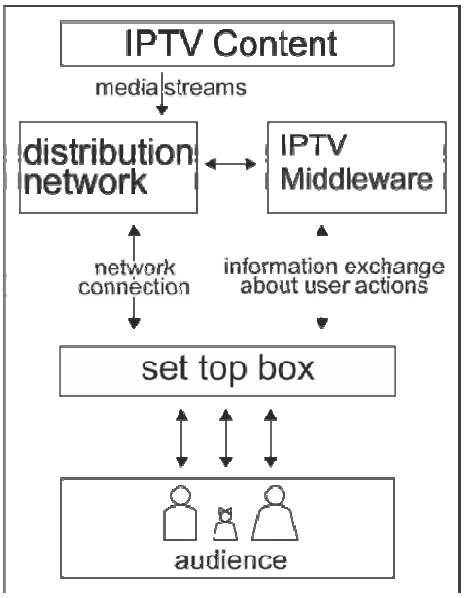

Fig. 1. IPTV System Architecture.

A. Statistical module

To protect the identity of the IPTV network subscribers, individual set top boxes were assigned a non-identifiable ID number for purposes of this research. The authors did not have access to subscriber's identity or address of individual set top boxes.

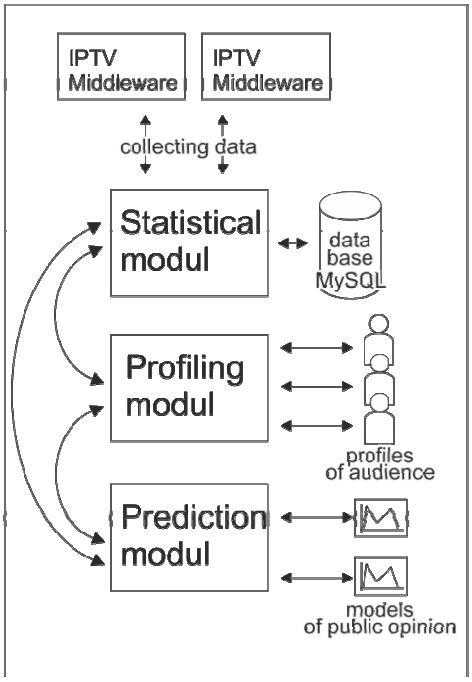

Fig. 2. The model of public opinion changes architecture.

The model of public opinion changes will consist of the following modules:

1. The module of collecting and storing of IPTV Middleware;
2. The module of processing and displaying the collected data;
3. Module of audience profiling;
4. Prediction module.

The model of public opinion changes architecture is shows on Figure 2.

IPTV Middleware displays and saves in the database of all the basic interaction of user with the interface. Middleware system is responsible for communication with every Set Top Box [4]. Then the data is written to the MySQL database where millions of records of watched TV channels and other user actions are placed [5].

Statistics system has internal scripts and a MySQL database which is responsible for storing all user actions data [5].

The main table in this database is table STAT which contains fields that are described in Table I.

Statistics system is collecting data from tree IPTV operator's Middleware. These operators work in all major cities and many districts of Latvia. This will give an opportunity to provide an accurate data for the whole country. The number of users covered by the statistics system data is over 3000. The number of analyzed channels is 120. According to statistical standards it may give an error less than 2% [3].

The collected statistical data module (can be used as the output processing module). It will:

1. Identify and display statistics of popular TV channels.
2. Identify and display statistics of popularity of individual programs.
3. Show general statistics data for TV viewing.
4. Identify and display statistics of TV viewing among profiled audience (specific groups of viewers).

Model of processing and output of data uses the collected data to MySQL database. Processing and displaying takes place by using the written code in PHP. Code will run on the WEB server.

Module of profiling audience will separate a group of viewers by given attributes. Information based on the collected data can be:

1. The generalized (universal) profile of audience;
2. Allocated some of the most common profiles encountered of the auditory.

The profiling of the audience will be done by using tag (classes). Tags are included in EPG. EPG displays the name of the program and its genre. Using the statistics of viewed programs and classifying them by genre, each user can be classified as a profile.

TABLE I
TABLE STAT IN STATISTICS DATABASE

The name of field	The meaning of field	Type of the variable
EVENT_ID	This field signs what type of IPTV content is used for statistics – 1 is for TV, 2 is for Catch-Up TV, 3 is for Time Shift	INTEGER
CHANNEL_TAG	The name of TV channel	VARCHAR2
SUBSCRIBER_ID	Unique number for each subscriber	INTEGER
DEVICE_ID	Unique number for each Set Top Box (IP address)	INTEGER
TIMESTAMP	The time when TV content starts	DATETIME
EPG	Electronic program guide information	VARCHAR2
EPG_GENRE	Genre of content	VARCHAR2
INSERT_DATE	Date and time when the record is inserted into MySQL database	DATETIME
STB_STAT	Set top box status (1 – on, 0 – off)	INTEGER

B. Profiling module

Genres tables:

- News from Russia;
- News from Latvia;
- News from Ukraine;
- Other News;
- Science & Technology;
- Educational program;
- Entertainment – female's;
- Entertainment – men's;
- Sport;
- Fight;
- Cinema- melodrama;
- Movie- drama;
- Movie- thriller/action;
- Comedy;
- Documentary;
- Children's/Family;
- Erotic.

Profile consists from points of each genre.
The main user profiles:

- Active viewer;
- Aggressive;
- News Consumer;
- Seeker of Knowledge;
- Cinema addict;
- Sports fan;
- Children;
- Housewife;
- Balanced profile.

For example, a client who spends about 12 hours watching TV, will be profiled as an "active audience" and viewer who generally watch mostly news programs, can be classified as "consumer news", and the client who watch mainly scientific programs – as "a seeker of knowledge".

Profiling module audience will also be set up as a WEB application that does not require working with it to install extra additional software.

C. Prediction module

One of the most important elements of the instrument will be the prediction possibility. The main task is to build a model of distribution, depth and perception of the proposed ideas. It will be the module of public opinion changes as shows on Figure 3.

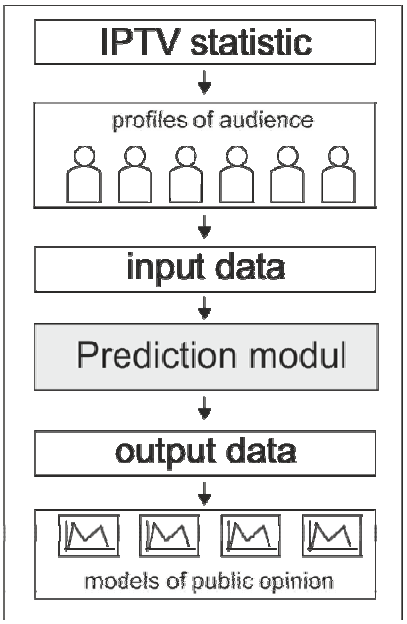

Fig. 3. The module of public opinion changes architecture.

His foundation is a mathematical algorithm. Which, using a collected data from other modules and based on them, creates the requested model.

Predictions module makes model based on the data received from the other modules. Including in the model the following factors:

- Information events;
- Seasonality;
- Distance from political events;
- And others.

Information event - it is a fact that can influence public opinion. These may be for example, natural disaster, major holidays, loud statements of politicians, demonstrations and more.

Seasonality – this is time of year. Since the seasons can change a person's mood, it is necessary to explore and suggests in the model how the public perceives the different events depending on the time of year.

Political events, such as elections, the inauguration, the resignation, the fall can affect the mood of the public.

Considering the listed and other relevant factors, the instrument must predict how the public will react to a variety of international and local events, and how to develop this reaction.

Among the reactions can be identified:

- Taking;
- Not acceptance;
- Indifference;
- Disgust.

Each reaction will be expressed in percent, as all people do not react the same way. For each profile, the audience is the most common reaction.

III RESULTS AND DISCUSSION

The results of the work will create a instrument that a variety of institutions will be able to apply it. These institutions may be:

- Television channels;
- Government agencies;
- Research organizations;
- Military forces;
- Others.

The instrument will not only accurately determine the current picture of situation, but also with high accuracy to make forecasts about the state of public opinion.

The instrument allows tracking the public mood and planning of media companies to inform the public.

Precise knowledge of the dynamics and trends of information dissemination through television can facilitate the work of government agencies related to public awareness. This will save money and improve the quality of work.

IV SUMMARY

At this moment, in Latvia, there are no similar methods used for determining public opinion, which significantly complicate the definition of a real picture of the happening processes in the society. Created model can help to solve this problem.

In the face of possible provocations and hybrid war, especially increases the relevance of the instrument.

The main problems in the future work may include:

- Lack of resources;
- Legal aspects;
- Lack of interest on the part of the organizations;
- Technical complexity.

Given the potential these challenges, it is necessary to involve additional experts to analyze the ideas contained in the article.

V REFERENCES

[1.] C. WILLIAMSON, United States Army War College, Carlisle Barracks, 2009, pp.15-16.

[2.] MD LeCompte, JP Goetz, Problems of reliability and validity in ethnographic research, Review of educational research, 1982, pp.31-35.

[3.] L Jussim, Social perception and social reality: A reflection-construction model, Psychological review, 1991, pp. 54-73.

[4.] Y. J. Won, Mi-Jung Choi, Byung-Chul Park, J. W. Hong, Hee-Won Lee, Chan-Kyu Hwang, Jae-Hyoung Yoo, “End-User IPTV Traffic Measurement of Residential Broadband Access Networks”, IEEE Network Operations and Management Symposium Workshops, NOMS, 2008, pp. 95-100.

[5.] A.Tanovic, I. Androulidakis, F. Orucevic, Analysis of IPTV Channels Watching Preferences in Bosnia and Herzegovina, ETASR - Engineering, Technology & Applied Science Research Vol. 1, No. 5, 2011, pp. 105-113.

[6.] Cha, M., Rodriguez, P., Crowcroft, J., Moon, S., & Amatriain, X., Watching a Nation Watch IPTV, IEEE Transactions on Multimedia, 2007, pp. 1672–1687.

[7.] U.S. Joint Forces Command, The Joint Operating Environment: Challenges and Implications for the Future Joint Force, 2009, pp.39.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, *47-53*

CONDITIONS FOR SUCCESSFUL DEVELOPMENT OF ELECTRONIC COMMERCE IN LATVIA

Sandra Ežmale
Rezeknes Augstskola, Faculty of Engineering,
Address: Atbrivosanas aleja 115, Rezekne, LV-4601, Latvia.

Abstract. **The electronic commerce has become an integral part of the global economy during the last decade. The most obvious indication of the importance of electronic commerce in the world economy is rapidly increasing Internet use in trade and other sectors. The successful growth of e-commerce promotes the use of other electronic media, as well as it is perspective for development of territories located in backcountry and remote regions. In the result of the Internet and electronic commerce technological development there can be observed changes also in the business environment. In these circumstances successful strategies applied in different levels and sectors for the implementation of the most modern information technology achievements, as well as for the development of professionals with appropriate knowledge, expertise and skills, are becoming increasingly important for the development of electronic commerce in Latvia. The article will view the current situation, will explain the main limiting factors and advantages for successful development of the electronic commerce in Latvia.**

Keywords: **strategy, electronic commerce, policy framework.**

I INTRODUCTION

The electronic commerce has become an integral part of the global economy during the last decade. The most obvious indication of the important role of electronic commerce in the world economy is rapidly increasing Internet use in trade and other sectors. ICT and e-commerce are powerful tools that can improve productivity, increase efficiency, facilitate international trade, promote investments and contribute to the development of countries and regions.

Electronic commerce is often regarded as buying and selling using the Internet, but e- commerce involves much more than electronically mediated financial transactions between organizations and customers. Kalakota and Whinston [2] refer to a range of different perspectives for e- commerce:

- A communication perspective – the delivery of information, products/ services or payment by electronic means.
- A business process perspective – the application of technology towards the automation of business transactions and workflows.
- A service perspective – enabling cost cutting and at the same time increasing the speed and quality of service delivery.
- An online perspective – buying and selling products and information online [2].

Yürekli argues that e-commerce is production, advertisement, sale and distribution of goods and services on open network environment (internet) or via closed network environment (intranet). E-commerce has four parties as follows: enterprises, consumers, citizens and government. Moreover, interrelation of these parties in electronic environment forms four types of e-commerce. E-commerce, which brings innovations to economic structure in terms of commercial relations, affects many aspects of economic and social life. Working conditions, changes in consumer preferences, facilities in shopping, procedures in performing public duties, production costs, operational and market structure, competition, productivity, prices and employment can be given as the examples of the affected areas in parallel with the volume of e-commerce development [12]. In information society e-commerce helps to use resources more efficiently and reduce costs, providing significant advantages in price and enhancing competition and efficiency.

In these circumstances successful strategies applied in different levels and sectors for the implementation of the most modern information technology achievements, as well as for the development of professionals with appropriate knowledge, expertise and skills, are becoming increasingly important for the

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.177

overall development of electronic commerce in Latvia.

The main goal is to view current situation and explain the main limiting factors and advantages for successful development of the electronic commerce in Latvia. This research deals with two tasks: (1) to investigate the development tendencies and limiting factors of electronic commerce in Latvia (enterprises, consumers, citizens and government); (2) to determine what strategies can be derived from the strategies of national level for the development of electronic commerce in order to bring a competitive advantage to e-businesses.

II DEVELOPMENT TENDENCIES OF ELECTRONIC COMMERCE IN LATVIA

E-commerce: government, citizens, consumers and enterprises

In the period of the year 2009 to 2013 in Latvia there have been carried out diverse measures in the sphere of e-commerce in order to improve the efficiency of public administration and to provide convenient service receipt: there have been established electronic one-stop shops, there has been carried out the optimization of ICT infrastructure in the public administration, there has been promoted wider applicability of secure e-signature and there have been introduced electronic identification cards, furthered e-services, as well as developed public infrastructure necessary for provision of e-services.

For example, by the year 2013 approximately 240 services have been electronized – there have been developed e-services in health, education, welfare and social services, as well as a range of services for entrepreneurs (electronic registration of a company and other registry services, State Environmental Service e-services, State Labour Inspectorate e-services, etc.). In the national portal www.latvija.lv in 2013 the inhabitants had access to 61 e-services, but in 2012 the number of e-service operations exceeded 1 million times. The country has an electronic procurement system, which has a constantly growing turnover (for example, in 2012 it was 18.27 million LVL, which is 37% more than in 2011). There have also been introduced electronic identity cards, which also includes the tools of identification and electronic signature necessary for obtaining e-services [7].

There can be observed significant improvement in indicators of Latvia among other EU Member States, e.g., the use of the internet or search for the information about goods and services in Latvia is close to the average European level (see Table 1).

The inhabitants of Latvia exceed the average EU indicators in reading newspapers online, using Internet bank, uploading a self-constructed content and acquiring online courses [7].

TABLE 1
PROPORTION OF INTERNET USAGE IN LATVIA AND IN THE EU (28 COUNTRIES), 2009 - 2013

Indicators[1]	Latvia					EU (28 countries)
	2009	2010	2011	2012	2013	
Individuals using the Internet by place of use	50	56	57	62	70	72
Individuals who have never used the Internet	31	29	27	24	22	20
Individuals who regularly use the Internet	61	62	66	70	71	72
Individuals who use the Internet for interaction with public authorities	30	40	41	47	35	41
Individuals having ordered goods or services for private use via the Internet	19	17	20	27	32	47
Level of Internet access households	58	60	64	69	72	79

1 % of the total number of individuals aged 16 to 74
Source: EUROSTAT data

According to the data provided by Central Statistical Bureau of Latvia, in 2013 35% of the population used the Internet for collaboration with state and public institutions. Almost all of them searched for information in the websites of state and public institutions (34%). In 2013 the most popular activity in cooperation with state or public institutions was the submission of income tax return, made by 19% of users, but 23% of Internet users said that they had chosen other activities to cooperate with state and public institutions. In 2013 the proportion of people who downloaded and submitted application forms electronically to the public authorities was 14.3%, while there were 13% who had sent them [8]. Thus, Latvia has not yet achieved the objectives set out in the flagship initiative “A Digital Agenda for Europe" of the document “Europe 2020: A strategy for smart, sustainable and inclusive growth” - 25% of inhabitants who have downloaded and sent application forms electronically to the national authorities.

As it can be seen in Table 1, in comparison with the average level of the European Union Member States (EU 28) there is a relatively low proportion of the

population of Latvia who use the Internet for ordering goods or services - in 2013 only 32% (EU 28 - 47%), however, compared to 2005 it has increased by 27%. The survey data provided by the Central Statistical Bureau of Latvia show that in 2013 most people used the internet for buying clothes and sporting goods (40% of online shoppers) and event tickets (37% of online shoppers). There is no significant difference in IT use in regard to gender, but it is revealed in the analysis of shopping habits, such as a product or services ordered online by men and women. Women purchased clothes and sporting goods on the Internet more often than men (45% of online shoppers, or 10 percentage points more than men). By contrast, among men the most popular goods were electronics (prevalence of 20 percentage points), hardware (prevalence of 12 percentage points), shares, insurance policies and other financial services (prevalence of 9 percentage points). Women also like buying household appliances (prevalence of 5 percentage points), event tickets (prevalence of 4 percentage points) and food products (prevalence of 4 percentage points) [8].

Latvia among the Baltic States is between Estonia and Lithuania in terms of IT use, but the opportunity to sell goods or services online on the Internet is used quite little by the residents of Latvia, lagging behind the EU average indicators by 17 percentage points. Only 3% of Internet users in Latvia apply to doctor's appointment in the institution's website, lagging behind the EU average level by 8 percentage points. A particular example is the Lithuanian and Latvian internet users' activity in using the Internet for telephone calls or video calls, overtaking the EU average level for respectively 36 and 26 percentage points. In 2013 one-third (32%) of the citizens of Latvia used mobile Internet out of home or work. The residents of Latvia most often used their phones or smartphones for accessing the Internet (23%). A laptop was also a popular device among population to access the Internet (16%) [8].

By contrast, the survey data that was obtained at the end of the year 2012 indicates the reasons why the Internet usage among the population of Latvia does not develop so rapidly. The survey data shows that people have a low level of awareness of services available electronically, and distrust in the Internet environment [11]. However, according to the Central Statistical Bureau's data the main reason why people do not use the Internet to submit completed forms to the state, local government or public service providers is the lack of skills or knowledge - it was mentioned by 3.6% of internet users of the definite group in the last 12- month period. It can be concluded that there is still low proportion of the population of Latvia who use e-services, as well as low usage of ICT opportunities offered in entrepreneurship.

The main consumers' barriers to adaptation of the internet include:

- no noticeable benefits;
- luck of trust;
- security problems;
- lack of skills;
- cost.

These factors mean that there is a significant group in each national population that does not plan to use the Internet at all [2].

E-commerce is definitely one of the industries that has developed most rapidly and has a great potential. This is evidenced by the research carried out by European e-commerce organization in 2014, which proves that Eastern European countries have the largest e-commerce revenue growth - +47.3%, while for the rest of the European country groups the maximum increase is 22.7%. Viewing Europe as a whole, the e-commerce market is worth more than 300 billion Euros and the Internet economics makes on average 3.5% of total GDP. However, it must be noted that the Baltic countries are far behind the rest of European countries [5]. For example in 2013 e-commerce in Latvia accounted for 0.67% of total GDP.

TABLE 2
THE USE OF E- COMMERCE IN ENTREPRENEURSHIP IN LATVIA AND EU (28) FROM 2009 TO 2013

Indicators	2009	2010	2011	2012	2013	EU (28 countries)
Share of enterprise turnover on e-commerce	5	7	6	7	8	14
Enterprises having purchased online (at least 1%)	8	9	14	14	13	18
Enterprises having received orders online (at least 1%)	4	6	9	7	7	14
Enterprises using software solutions	8	11	11	13	12	19 (2012)
Enterprises with fixed broadband access	61	66	82	86	91	90

Source: EUROSTATE data

Analyzing data on e-commerce in entrepreneurship, it can be concluded that during the last five years in Latvia there has been a tendency for improvement in many parameters (see Table 2), but still it lags behind the average level of the European Union Member States (28). In 2013 7% of companies of Latvia received orders via the Internet (EU average indicator - 14%) and 7% of companies used e-

commerce in entrepreneurship (EU average indicator - 14%).

TABLE 3
THE USAGE OF COMPUTERS AND INTERNET IN ENTERPRISES OF LATVIA

Indicators	2007	2009	2011	2013
Computer usage	93,3	93,5	95,4	97,4
Internet usage	83,7	86,8	92,2	94,2
Internet website	37,7	42,1	53,4	55,7

According to the data provided by Central Statistical Bureau of Latvia, the computer and the Internet were most frequently (99% or more) used by medium and large companies (with a staff of 50 and more people). By contrast, 97.1% of smaller companies (number of employees from 10 to 49) used computers, while 93.3% - the Internet. This can be explained by the fact that companies of this size quite often used external services for accounting and other documentation. In 2013 the Internet was most frequently used by companies whose basic activity is provision of information and communication services, repair and maintenance of computer and communication equipment, as well as manufacturing of computers, electronic and optical products. The use of the Internet in these sectors has reached 100%. By contrast, the organizations that used the Internet the least were administrative and support service organizations, as well as catering services. In these sectors the Internet was used respectively 90.6% and 89.0% of the companies. In 2013 56% of the companies had their own website, which is 3 percentage points more than in 2012. The proportion of companies, which have their own website, has increased along with the increase in staff. In 2013 51.6% of small companies, 74.5% of medium-sized enterprises and 92.1% of large companies had their website [9].

According to the data provided by the Central Statistical Bureau, in 2013 the broadband Internet was used by 91.4% of the companies. The most popular type of the Internet connection in enterprises is a DSL connection that is used by 56.5% of the companies, but in 2013 the benefits of mobile broadband were used by 32.7% of the enterprises, but other mobile internet connection - 23.0% [9].

E-skills

In Latvia like in other European countries e-skills have become an essential necessity of life and if it is impossible to access or use ICT, it can cause a major obstacle to social integration and personal development. Digital literacy is one of the eight major skills of each individual that is essential in a knowledge-based society. In the Digital Agenda for Europe there is also emphasized the importance of teaching EU citizens to use ICT and digital media [3]. Insufficient level of e-commerce skills among entrepreneurs and the population is as a barrier to economic growth, competitiveness and employment in Latvia and the EU as a whole. According to the EU research it is expected that in 2015 90% of all jobs in every sector will require technical skills, including e-skills. At the same time only 27% of EU citizens have a high level of ICT skills, and 25% - a mid-level skills (In Latvia respectively- 31% and 30%) [5]. Only 53% of the working population in Europe believe that their ICT skills are sufficient to labor market needs [4]. According to the data of the Central Statistical Bureau in 2013 in Latvia computer users' best skills were copying and moving files and folders (82.4% of computer users), as well as using copying and insertion tools (77.8%). By contrast, only 5.9% of computer users had the skills to write a computer program using a specialized programming language (see Table 4).

TABLE 4
COMPUTER LITERACY IN LATVIA IN 2013

Skills	% of computer users of the definite group
Copy or move files and folders	82,4
Use Copy and insertion tools	77,8
Use basic arithmetic formulas in spreadsheet	58,8
Compress files	50,4
Connect and install new devices	51,1
Write a computer programme using a specialized programming language	5,9
Move files from other devices to your computer	62,2
Create electronic presentations with presentation software	41,1
Install new or replace old operating system	18,6
None of the activities mentioned above	42,1

Source: Latvia CSP data

In 2013 almost all Internet users (97%) knew how to use search engines to find information, 76% of users were able to send e-mails with attached files, but 9% of Internet users were able to create a website. It is interesting to note that regardless of relatively high level of Internet skills, people have quite light – minded attitude towards the security of their personal data - less than one in five Internet users (19%) has ever changed the web browser security settings [8].

In March 2013 there was organized "The week of e-skills 2013" for the fourth consecutive year with the goal to improve e-skills of the population of Latvia, during which 42,690 inhabitants of Latvia took the advantage of the opportunity to acquire information and communication technologies. The week of e-skills was organized by LIKTA together with the Ministry of Environmental Protection and Regional Development and major partners - "Lattelecom" and JSC "RIX Technologies". It was held throughout the country with the participation of 114 regions and cities. Overall there were offered 855 events for

residents and entrepreneurs. In 2013 the Ministry of Economy, Ministry of Education, Ministry of Welfare, Ministry of Environmental Protection and Regional Development, the Latvian Information and Communications Technology Association and other NGOs signed a cooperation memorandum on the development of "E-skills partnership" in Latvia within the framework of the following action: providing education that meets labour market requirements; attraction of young people to ICT industry; interactive and modern learning process and digital content; educating society about the need for digital literacy, e-inclusion.

Background for technological change and development in Latvia

Since 50-ies of the 19th century, academic researchers have been actively discussing the idea that technological change and development is the most powerful force for economic growth. Globalisation has significantly increased the competitive pressure on enterprises in many sectors. Together with other factors like the emergence of new, lower – cost producers, fast- changing demand patterns, increased market fragmentation and shortened product life cycles, the pressure has been forming business environment during the last decades. In such environment innovation becomes crucial for the long – term competitiveness and survival of enterprises. One of the key mainsprings of innovations applying e-business technologies is the increasing of availability of knowledge required to facilitate the innovation process [10], especially in the sphere of information and communication technologies.

During the recent years in Latvia like in other European countries, there can be observed a growing lack of ICT professionals - in Europe every fourth is ICT vacancy, and it is expected that in 2015 Europe will have a shortfall of around 700 000 ICT specialists [4]. The lack of ICT professionals is hindering the growth of the sector, improvement of international competitiveness, increase of export and attraction of foreign investors. As it can be seen in Table 5, the number of ICT professionals in Latvia has increased, but still it is not sufficient. According to the data provided by LIKTA there is lack of ICT specialists in Latvia and in the whole Europe and according to LIKTA members' and the ICT sector companies' forecasts they are ready to attract 1000 new specialists a year, while the number of newly trained specialists is determined by the total number of bachelors and engineers, because the person can continue master's and doctoral studies only after previously acquired education, but technicians' (college) level of education allows to carry out rather simple ICT operations. This means that in 2012 there were 688 newly trained specialists (653 specialists with a bachelor's degree and 35 engineers), which does not satisfy the demand of the national economy.

TABLE 5

PREPARATION OF ICT PROFESSIONALS IN LATVIA

Year/ the obtained degree	**2000**	**2002**	**2004**	**2006**	**2008**	**2010**	**2012**
Technicians	22	38	68	150	178	220	247
Bachelors	511	501	578	690	585	612	653
Engineer	100	156	148	103	38	51	35
Masters	98	183	373	311	311	287	295
Doctors	0	2	6	10	11	4	21
Total	**731**	**880**	**1173**	**1264**	**1123**	**1174**	**1251**

Source: Informācijas sabiedrības attīstības pamatnostādnes 2014-2020.gadam [7]

In order to ensure a stable national economic growth, it is necessary to expand and improve ICT practitioners' and professionals' skills, i.e. the digital skills, with a particular focus on preparation of skilled ICT specialists according to labour market requirements, development of professional ICT programmes and ICT education for adolescents.

In Latvia there is also one of the lowest funding for research and development among the EU Member States, which according to the Central Statistical Bureau in 2011 was 0.70% of GDP (in comparison - in 2010 the EU average indicator was 2.0% of GDP). It also affects the performance of Latvia in various international assessments. According to the World Intellectual Property Organization (WIPO), Cornell University and INSEAD's report "The Global Innovation Index 2014: The Human Factor in Innovation", Latvia takes the 34th position among all countries, Estonia has got 24th place, but Lithuania – 39th position. Latvia ranks in the 32nd position among the group of countries with high income, in European region it ranks in the 22nd position, but innovation efficiency index ranks in the 32nd place. Latvia has received the lowest evaluation in the categories as follows: "Human capital and research" (34.1 points) and "Knowledge & technology output" (36.8 points) [13, 14]. In general, it can be concluded that the lack of ICT professionals and low level of funding for research and development in the future could seriously affect the national growth, technological change and development in Latvia, which is an essential factor in the development of electronic commerce.

The most essential problems that influence the development of e-commerce and ICT in Latvia are as follows:

- lack of skills,
- low trust;
- insufficient investment in infrastructure and networks;
- fragmented digital market;
- lack of interoperability;

- insufficient research and development.

E-commerce strategies at national level

The government and the private sector are key partners in the development of e-commerce in any country. The business sector should lead, while governments should support and facilitate (by promoting national e-commerce policy framework and strategies) the development and use of e-commerce and ICT. The national e-commerce and ICT strategies must be carried out according to the economic and social context of each country. At the same time it is important to take into account the international experience and best practices to ensure interoperability and harmonization with the developing global e-commerce and ICT systems.

All relevant factors, for example, awareness of society, access to ICT, physical infrastructure, legal framework, quality of human resources, security and privacy aspects are important for successful national strategies. The key factors are:

- e-leaderships;
- national ICT framework and implementation mechanisms;
- legal framework;
- addressing Digital Divide;
- capabilities of individuals and society, qualified human resources, as well as research and development;
- enhancing SME entrepreneurship.

Basing on the e-commerce development tendencies, main limiting factors and advantages for successful development of the electronic commerce in Latvia, there have been identified key strategies for the development of electronic commerce and ICT at national level:

- the formation of positive environment for e-commerce development: policy and legal framework, infrastructure e.g.;
- further development of e-government: services and e-procurement;
- collaboration between public, academic and private sectors;
- building of human capacity: the awareness at all levels, training and education, development, development of professionals with appropriate knowledge, expertise and skills;
- promotion of interoperable standards and provision of information resources;
- development of sector/ region specific e-strategies.

III CONCLUSION

There can be observed significant improvement in indicators of Latvia among other EU Member States, e.g., the use of the internet or search for the information about goods and services in Latvia is close to the average European level. The inhabitants of Latvia exceed the average EU indicators in reading newspapers online, using Internet bank, uploading a self-constructed content and acquiring online courses.

Analysing data on e-commerce in entrepreneurship, it can be concluded that during the last five years in Latvia there has been a tendency for improvement in many parameters, but still it lags behind the average level of the European Union Member States (28).

Digital literacy is one of the eight major skills of each individual that is essential in a knowledge-based society.

In order to ensure a stable Latvian economic growth, it is necessary to expand and improve ICT practitioners' and professionals' skills, i.e. the digital skills, with a particular focus on preparation of skilled ICT specialists according to labour market requirements, development of professional ICT programmes and ICT education for adolescents. In Latvia there is one of the lowest funding for research and development among the EU Member States. In general, it can be concluded that the lack of ICT professionals and low level of funding for research and development in the future could seriously affect the national growth, technological change and development in Latvia, which is an essential factor in the development of electronic commerce.

The most essential problems that influence the development of e-commerce and ICT in Latvia are as follows:

- lack of skills,
- low trust;
- insufficient investment in infrastructure and networks;
- fragmented digital market;
- lack of interoperability;
- insufficient research and development.

Basing on the e-commerce development tendencies, main limiting factors and advantages for successful development of the electronic commerce in Latvia, there have been identified key strategies for the development of electronic commerce and ICT at national level: (1) policy and legal framework, infrastructure e.g.; (2) further development of e-government; (3) collaboration between public, academic and private sectors; (4) building of human capacity; (5) promotion of interoperable standards and provision of information resources; (6) development of sector/ region specific e- strategies.

IV REFERENCES

[1] AKMAI'S State of the Internet. Q 1 2014 Report, Volume 7, Number 1. 44 p. Available: http://www.akamai.com/dl/akamai/akamai-soti-q114.pdf?WT.mc_id=soti_Q114, [Accessed October 2, 2014].

[2] Chaffey, D. (2007) E- Business and E- Commerce Management. Third Edition. FT Prentice Hall. 663 p.

[3] Digital Agenda: Commission outlines action plan to boost Europe's prosperity and well-being.European Commission - IP/10/581 19/05/2010. Available:

http://europa.eu/rapid/press-release_IP-10-581_en.htm?locale=en, [Accessed October 22, 2014].

[4] Digital Agenda Scoreboard 2012, Available: https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/KKAH12001ENN-PDFWEB_1.pdf, [Accessed October 22, 2014].

[5] European B2C E-commerce Report 2014. 63.p. Available: www.ecommerce-europe.eu, [Accessed October 24, 2014].

[6] IDC White Paper Post Crisis: e-Skills are needed to drive Europe's innovation society, 2009. Available: http://ec.europa.eu/enterprise/sectors/ict/files/idc_wp_november_2009_en.pdf, [Accessed November 12, 2014].

[7] Informācijas sabiedrības attīstības pamatnostādnes 2014.–2020.gadam. (2013) Available: *http://www.varam.gov.lv/lat/darbibas_veidi/e_parv/InfSab/?doc=12668* [Accessed November3, 2014].

[8] Informācijas un komunikāciju tehnoloģiju lietošana mājsaimniecībās 2013.gadā. Informatīvais ziņojums. Available: http://www.csb.gov.lv, [Accessed October 22, 2014].

[9] Informācijas un komunikāciju tehnoloģiju lietošana un e-komercija uzņēmumos 2013.gadā. Informatīvais ziņojums. Available: http://www.csb.gov.lv/sites/default/files/nr_37_informacijas_un_komunikacijas_tehnologiju_lietosana_un_e-komercija_uznemumsos_2013_13_00_lv.pdf, [Accessed October 24, 2014].

[10] Kar, M. (2009) Product and Service Innovation. In: E-Business: Management Perspectives. Edited by Reynolds, J. Online resource centre, Oxford. (270- 3030) 496 p.

[11] Sabiedriskās domas izvērtējums par e-pakalpojumu pielietojumu Latvijā. 2012.gada decembris. Available: http://www.varam.gov.lv/lat/publ/petijumi/pet_Eparv/?doc=14321 [Accessed October 25, 2014].

[12] Savrul, M., Cüneyt Kılıç, C. E-Commerce As An Alternative Strategy In Recovery From The Recession. Procedia Social and Behavioral Sciences 24 (2011) 247–259. Available: http://www.sciencedirect.com/science/article/pii/S1877042811015837 [Accessed October 24, 2014].

[13] The Global Innovtion Index: The Human Factor in Innovation. (2014) Cornell University, INSEAD, and WIPO. Fontainebleau,Ithaca, and Geneva. 429 p. Available:: http://www.wipo.int/edocs/pubdocs/en/economics/gii/gii_2014.pdf, [Accessed October 27, 2014].

.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 54-64

The Myths about and Solutions for an Android OS Controlled and Secure Environment

Imants Gorbans[1, a], Ivans Kulesovs[1, b], Uldis Straujums[1, c], Jānis Buls[2, d]
[1]Faculty of Computing, University of Latvia,
Raiņa bulvāris 19, Riga, LV-1586, Latvia
[2]Faculty of Physics and Mathematics, University of Latvia,
Raiņa bulvāris 19, Riga, LV-1586, Latvia
[a]Imants.Gorbans@lu.lv, [b]Ivans.Kulesovs@gmail.com, [c]Uldis.Straujums@lu.lv,
[d]Janis.Buls@lu.lv

Abstract. **Android is today's most popular mobile operating system for both smartphones and tablets. This fact creates many risks which are not fully recognized. Even advanced users often naively think that by using antivirus software, a firewall, encryption and updates, as well as avoiding potentially risky sites and applications they will be secure. This list is not exhaustive, but nevertheless, in most cases, each item in it only provides the illusion of security. The authors have summarized and pointed out several actual Android security issues and have proposed a number of possible solutions.**

Practical experience as well as direct testing reveals that some Android applications may contain malware. The harmful characteristics of an application often become visible only after it has been run a few times, after an update, or after harmful web content has been downloaded and shown by the application. It has been observed that applications often try to get unauthorized or inattentively authorized access to user data and to send it outside the device.

The situation with Android applications is getting more and more out of control. The authors have proposed a solution for overcoming security issues, while respecting the latest Google solutions. The target group of the proposal is users who use a smartphone or tablet both for private and corporate needs, i.e. a Bring Your Own Device (BYOD) case.

The authors point out and compare four possible Android technical administration solutions based on the unified model for a BYOD case. The authors also propose changes to Android architecture to enhance its security. A look at the mobile operating system, as a web server, has been proposed. Such a principle allows the implementation of a number of security principles taken from web servers solutions.

Keywords: **Android, mobile computing, security, BYOD, smartphones, ICT.**

I INTRODUCTION

Most of the publicly available security manuals for users are quite simple. That is why they usually push users even further away from reality. The advice about asking an ICT specialist for support doesn't always help either, because specialists are often only partly informed. We will be focussing on Android OS, because it dominates about 81% of the world's new smartphone market [1]. We will take a look at the myths or security expectations and into the solutions which may possibly reduce security risks. The authors will focus on users of Android devices who use smartphones or tablets for both private and corporate needs, i.e. a Bring Your Own Device (BYOD) case.

The oldest threat to mobile communication is the interception of calls, but nowadays new threats are coming to the foreground for smartphones. Smartphones already have most of a PCs' functionality, and therefore – the same issues. But smartphones still do not have comparable security solutions to PCs. The everyday user can still only partly reduce the threat of call interception, because it depends mostly on the service provider. But he can choose the right phone with OS and SIM supporting LTE encryption [2] and make important calls only in the 4G network. The call from the phone to the tower will then go only through an encrypted channel.

Another solution is to use additional encryption software at both ends of the call. But such a solution also has an additional technical and legal complexity

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.184

that is outside the scope of this publication. It is worth mentioning that it is the tower which decides on the call encryption mechanism. It means that the mobile service provider can completely disable the encryption [3], without the phone user even being notified. This mostly ends the discussion on the topic of voice calls.

It is possible to get even more information through a smartphone's internet connection than from calls. And there is not enough best practice or verified guidelines on configuring and administering Android devices for a BYOD case. That is why the symbiosis of these two issues is the main topic of this publication.

It is often said that Android is secure, because it is Linux. But, unfortunately, this is not the full Linux in terms of security solutions. There are more than a hundred academic publications and even more technical articles and user manuals on Android security. The number is quite large, because Android began in 2008, but its security issues were only recognized in 2010.

There are more than 18,000 different models of Android devices in the world [4]. OEM and mobile service providers increase the fragmentation of Android even more. The devices running the same Android version are different in most cases, because OEM and mobile service providers do modifications to the OS. Google allows and even encourages this in order to get new ideas for future development.

This policy creates issues for businesses if they want to choose Android as the main platform for their corporate smartphones. According to Gartner's study, less than 10% of companies plan to use Android devices for business purposes [5]. The lack of focus on security, as well as limited user management capabilities, are mentioned as the main reasons for the decision. But the position could change in the future. In the beginning, Android creators did not think about security. The main focus was on the maximal functionality and availability [6]. The situation in the security field is getting better in quite a rapid manner. This is already visible with the most current versions of Android (4.4 and 5).

II THREATS

A. Common Threats to Mobile Security

There are four main types of malware types for mobile devices [7], [8]:

1) Data thieves are the most popular malware. They try to get information about the OS version, product ID, IMEI number, and IMSI number of the infected device. This information can be used for more direct attacks in the future.
2) Rooting-capable malware infects the device in order to get the administrative permission. This allows remote users to access the device's RAM and other resources, e.g. the microphone.
3) Phishing malware. The SMS, or MMS, or email is sent to the infected device and the owner gets subscribed to pay services when opening it or following the link. Of course, the user is not notified. Calls to high-cost services can also possibly be made, or sensitive data sent to third parties while the user thinks that he is communicating with trusted sources. The phishing problem is especially real on Android because of the openness of the platform resulting in easier creation of malware. But, smartphones and tablets are usually connected to purchase and/ or payment systems.
4) Mobile spyware monitors a variety of information which is stored on an infected device like the current location, stored SMS or emails. This type of malware also sends data to third parties through any available channel in the same way as data thieves do. But, spyware concentrates mostly on personal user data retrieval.

The malware can appear in Google Play despite all preventive measures being taken [9]. It is known that there are fake app stores that replicate very similar content. These stores are made by cyber-criminals in order to fool users and provoke them into installing the malware. The fake stores can appear on a device after some harmful app installation or after some illegal Android version update. There are also apps that fake internet banking apps or try to retrieve user financial information [10], [11]. Statistics shows that 79% of all malware attacks are focused on Android OS [8]. Even if the malware does not steal sensitive enterprise data, it is not appropriate on the device used for business, because the device can deny service when it is needed for fulfilling some critical business task.

According to unofficial, but reliable information, about 44% of Android devices still use Android versions from 2.3.3 to 2.3.7 that have significant security vulnerabilities [8]. These vulnerabilities have already been removed in the latest Android versions. Mobile devices that are older than two years cannot receive the security updates, because often manufacturers do not support them. Many manufacturers already drop the development and support smartphone OS after 12-18 months of release. The only option is to install a newer Android version unofficially, if possible. But this option can often break the warranty. Google does not recommend that manufacturers produce devices with previous Android versions a year after the next Android version is available.

ICT specialists and advanced users can do unauthorized modifications to mobile devices, e.g. "rooting" of Android devices and "jail breaking" of iOS devices. These modifications are done in order to

get rid of operating system limitations. This allows the adding of extra features to the device and to install apps which are not allowed. This also changes the security management principles on the device increasing security vulnerability risks e.g. fake app stores.

Unauthorized modification can also be done to unlock the device from the specific carrier or to install additional security features like a firewall. But in most cases this is done specifically to install apps which are not allowed or/ and pirate apps. This increases the risk of infecting a device with malware, as soon as the built-in app check process is removed or completely modified. The malware now has a greater ability to access system resources and data using administrative permission, while working undetected in the background. This can also deny the ability to get security updates from the manufacturer.

It is possible to limit account permission through installing additional administrative tools on the rooted device. But the user should be aware that the device warranty is lost and the phone can become a "brick" if rooting fails. **That is why we do not recommend rooting the device as the solution for all organizations** in order to increase system security. But this can be an exclusive solution for some special case. In the course of this study, the authors have rooted several mobile devices using the Kingo Android Root or Kingo ROOT free software [1]. This software is thought to be among the most trusted rooting solutions. It allows the installation of such apps as SuperSU, Xprivacy, etc. We gained quite interesting results when using the Xprivacy app. The solution follows the actions of each app on a device in a relatively reliable way. It appeared that many free apps that were downloaded from Google Play do unauthorized actions or actions unwittingly authorized by the user with the user's data, and try to access web resources without any notification or disclosure of what is being sent out. They can sometimes begin to perform in this type of offensive manner, not during the first run, but after several runs, or after apps are updated. Using Xprivacy, user can follow the apps' actions, and can partly or fully block them (both actions and apps), but this solution is not for everyone. Such functionality would be very useful to include in the OS built-in toolbox.

One can rebuke a user for allowing an app to needlessly access data, a network, the GPS, etc. But, as already stated above, a user acts in good faith and the desired app just wouldn't even get installed without providing the relevant permission. Some apps ask for additional permission after an update, and users often, without even reading the question, allow these. The apps that show web content within them are very unreliable. These apps are not harmful in themselves, but they can download and execute some harmful scripts afterwards.

Some apps search for specific information within the user and device data, upload it, try to intercept the data stream, change the sensitive data, etc. One also can rebuke a user for installing some cat petting games, but how can one rebuke users of devices where the Chrome browser switches on a microphone which may be used to stream discussions that take place near the phone. This shows that even the current user practice of Android may be unacceptable and hopeless.

It is obvious that there is only one way: forbid BYOD users from installing additional apps on a mobile device which are able to work with company emails and/ or documents. This does not mean that all apps are malware. This means that only verified apps should be installed on an employee's phone and only by the company ICT specialist (network administrator). But then, why should a user bring his own device, if he cannot install apps for his own entertainment? There should also be an option for a browser to erase all data when it is closed.

B. Security Options in an Android OS

It is self-explanatory that authentication mechanisms as well as the encryption of data storage can help to prevent user data from falling into the wrong hands in a case of theft. The following authentication (screen unlock) mechanisms are available in the Android system: a) a combination drawing on the screen; b) entering a numeric PIN code; c) entering a password; d) biometric recognition, e.g. face recognition that is available from the Android 4.0 version (probably, not the best solution). The authors' recommendation is that **the best method is still a complex password that consists of small and capital letters, digits, and special characters and is at least 14 symbols long**.

There are smartphones with double authentication or with repeated authentication if it is not used for some period of time. It is possible to block an Android phone for some period of time when an authorization with the wrong passcode or password takes place. Five incorrect attempts at unlocking a device are possible by default. This blocks the system for 30 seconds. Data storage decryption with a wrong password also blocks the usage of a device for 30 seconds, but 10 wrong attempts are allowed. The recommended policy from PC world would be 5 incorrect login attempts that block a device for 15-30 minutes. But, there are no standard user settings for changing it in Android and we have to look for other solutions.

The next level of Android security allows a configuration of the system to erase all data if the device does not authenticate within a prescribed network within the prescribed period of time [12]. In the latest Android versions, there is a feature which

[1] - www.kingoapp.com/android-root.htm

searches for the lost device, wipes out all of the data and makes the device reset to the factory settings remotely. It is possible to perform a wipe and reset from an Internet portal or by SMS from the registered number if the feature is activated.

One can't overcome the fact that information can be intercepted. It is important that intercepted data cannot be practically decrypted, if the encryption algorithm is unknown, and the key is sufficiently long. That is why one of the most important mobile communication environment criteria is that data decryption is undertaken only by the data receiver. Otherwise, the encryption key is known by a third party which increases the probability of a data leak. If a decryption is made on some proxy server or by some other device at any other transmission step, then it is possible that the decrypted information can be intercepted. This is very important when data is transmitted through different carriers. If there are several different carriers, then the data security control is lost. The same applies to non-encrypted data transmission or when the encryption level is insufficient, i.e. when data fragments can be changed (e.g. a bank account number). The parties that are involved in the communication should be notified that the received message was not changed during the transmission.

Usage of public wireless networks should be minimized. Only new encryption types should be used, i.e. WPAv2 and 802.1x, while WPAv1 and WEP should not. The usage of Bluetooth outside trusted location should be minimized. Only trusted Bluetooth counterparties should be allowed, while others should be denied.

Only the current versions of all software, with the latest security updates, should always be used.

It should be understood that encryption of data storage is not a panacea, but is highly recommended. It will secure data only in cases where the switched off or pass-locked device is stolen. When a user has already entered an encryption key upon starting a device, then all of the activities occur in a non-encrypted way and do not prevent apps from stealing data.

The following issues were discovered whilst testing built-in data encryption tools on Android 4.2 and 4.4 which were installed on a Samsung Galaxy Tab 2 Mini: it wasn't possible to encrypt a device with another password after the previous encryption was removed. The UI was in the Polish language despite the fact that the device was used in only the Latvian and English languages. This should be checked on other Androids. However, it is clear that these solutions still have to develop.

The following useful security configuration options are available through regular Android settings: encrypt data storage, enable remote device blocking and data wipe, setting the device password, change enabled setting "Make passwords visible" by default, disable apps installations from untrusted sources etc. By installing the AppLock app, it is possible to force prescribed apps to run only after entering a passcode.

From the authors' point of view, it would be nice to have better security configuration options in the Android system, like the ones available in the latest Windows OS [13], [14]. Examples of configuration options could be as follows: the number of incorrect unlock/ decryption attempts, the time for which a device should be blocked after a wrong attempts limit is reached, only allowing one to open prescribed URLs, limiting the types of WiFi/ Bluetooth networks, importing security policy templates, choosing a security policy level, viewing logs, etc.

III ARCHITECTURE

A. Main Principles of the Android OS

The main principles of the Android OS are [15], [16], [17]:

1. Android is a processor independent operating system. But, it uses some device specific security features, like ARM v6 eXecute-Never that ensure the separation of user data from processor instructions inside the device's memory.

2. Android has been developed based on the Linux kernel. All device functions, e.g. the camera, GPS, Bluetooth, voice and data transfer is performed using the operating system, not the firmware.

3. Android apps have been developed mostly in the Java programming language and run inside the Dalvik or ART (starting from version 5.0) virtual machine. But many apps, including Android core services, use core libraries. Both, virtual machines and native apps run within the same secure environments – app sandbox that isolates app data and its code execution from other apps. An app gets the prescribed part of a file system to store its data. However, if an app has the appropriate permission then it can access all the device memory (including the SD card). The latest Android versions have an additional layer called SEAndroid that checks all the installed apps on the kernel level [18].

It is possible to create hybrid apps in Android using the WebView component that supports TML, CSS, and JavaScript technologies. These apps are similar to native apps, but they work as web sites with additional options to use the device's camera, accelerometer, etc. Unfortunately, this solution is potentially the most risky, which is why it is not advisable to install such apps from untrusted developers. But it could be a convenient way to develop internal enterprise apps.

B. The Layers of Android Architecture

Android architecture consists of several layers that work one on top of another. The lower layers provide the services for the top levels [15]:

1. The Linux kernel is the basis of the whole system. The Linux kernel enables Android to manage the hardware, memory, security settings, network protocols, and other low level functions. Users and developers do not access this layer directly. This is the layer where the hardware drivers are executed. The basic separation between the apps is also executed at this layer.
2. The core libraries ensure that basic services are available for apps. These libraries are written in C/ C++ language and vary depending on the device hardware. These libraries run as processes inside the Linux kernel.
3. The applications framework layer ensures that the core libraries and virtual machine interfaces are available to the apps.
4. User works at the highest level are called the applications level.

Android apps are divided into two parts:

a) the apps installed by the user;
b) the apps installed by the manufacturer.

It is worth mentioning that the open source Android operating system contains a code from at least three different sources: a) the Android open source project (Android version by Google); b) modifications from mobile device manufacturers; c) third party apps on the market.

C. The Main Security Features of the Android OS

Android has the following main security features:

1. The mandatory sandboxing of every app. Android uses the Linux Mandatory Access Control (MAC) mechanism (it will be described in detail separately) to force apps to work in a sandbox mode that is a part of the SEAndroid solution. A user unique identifier (UID) is assigned to each app during the installation. All executed app processes are attached to this UID. This allows a system to control access to low level resources. The private data storage inside the internal memory is assigned to each app according to this mechanism.
2. Access permissions. Permission labels are assigned to each app. They are displayed during the app installation and a user must accept them. These labels are checked at the application level when an app tries to use the security critical API. Developers can define new access permission in order to secure the interface of their apps in addition to the standard Android access permission.
3. In order to ensure the integrity and authenticity of apps, they are signed a X.509 certificate.
4. There are now enhanced web browsing security options available since the Android 5.0. It enables TLSv1.2 and TLSv1.1. Some enhancements are made in HTTPS and SSL protocols; Smart Lock has been introduced.
5. Android 5.0. also includes an enhanced FORTIFY_SOURCE feature that should provide security from buffer overflow attacks more efficiently, i.e. when an app tries to overflow the device memory in order to get sensitive data.

In the same way as for other operating systems, the latest stable version should be chosen in the Android case. At the moment of writing, the latest version is 5.0.2., which means that a device with a 5.0 version installed could be bought and updated.

D. SEAndroid

Android is a very rapidly evolving OS. The latest versions (including those, modified by the manufacturer) contain new promising solutions. Several Linux security enhancements were introduced starting from the Android 4.3. The Discretionary Access Control (DAC) was changed to Mandatory Access Control (MAC). MAC implements security control over all processes, objects, and operations. According to the developers, MAC can usually restrict erroneous and malicious access even for apps working with root privileges. This was not possible with DAC. The SE (Security Enhanced) Android was developed by US NSA [18]. SEAndroid should solve the following vulnerabilities: a) the malicious usage of administrative privileges, i.e. root exploits; b) the vulnerability of apps when they want to access or modify data without user authentication. It is worth mentioning that SEAndroid only began working in a permissive enforcing mode from version 4.4 and in full enforcing mode from version 5.0.

There also several threats that are not solved by SE Android:

a) It is not possible to forbid things which are allowed by the security policy. It means that the development of a good security policy is a critical task for SEAndroid to be efficient.
b) SEAndroid stops some core vulnerabilities from restricting the vulnerable code from untrusted apps, or making the impact of vulnerability negligible. But SEAndroid cannot stop all core vulnerabilities. That is why additional core security mechanisms should be used together with a SEAndroid solution.

SEAndroid cannot prevent threats that arise from other platform components. Particularly from components that have direct access to system resources, e.g. the RAM or network card.

E. File System

The Android file system is called the YAFFS (”Yet Another Flash File System”). It is built for Flash memory cards that are used as data storage for mobile devices. The classic limitation of the apps’ and users’ permission on folders and the file level is not the most

efficient in the Android system, because there is typically only one mobile device user [19]. The isolation of the apps, even when running them with different UDID, is a partial solution, because they are still executed by the same physical user and apps can ask for and get extensive permission during an install or update.

F. User Accounts

Starting from version 4.3 for tablets and from version 5.0 for smartphones, Android has some built-in user management capabilities. There are the following User Account types available:

The "Owner" user can add, remove, and configure user, guest, and profile accounts, i.e. can do almost everything (this account should be given to the company ICT administrator);

"User" accounts provide full access to apps and services on a device, while all changes made to system settings (like adding a Wi-Fi network) or updating the apps are applied to all user accounts on the device (!). An "Owner" user can restrict whether a "User" can use the phone for calls and SMS (this account should not be used for BYOD due to extensive permission availability).

A "Restricted profile" account (which is currently available only on tablets) can be restricted to run only allowed apps. Currently there is no built-in ability to fully restrict changes made to system settings. Only some additional restrictions, like disallowing location services when using the profile are possible (this account should be used for a user in a BYOD case).

A "Guest" account is a temporary "User" account. The system asks to reset the account or to continue the previous guest session each time the "Guest" account is used (could be useful when a device is given to another trusted user for some period of time).

G. The Fragmentation of Android

Android has the typical Linux issue: the fragmentation of software, dirty code, missing support for older versions (while almost half of all Android devices still use version 2.x). Furthermore, mobile device manufacturers often don't use the latest updates. This is one of the reasons why there are so much malware in Android.

According to F-Secure, mobile device manufacturers are guilty for Android security breaches in most cases, because devices cannot follow the development progress of apps. Mobile device security should be considered in general, because there are many threats that multiply with new apps, holes in internet browsers, messages (SMS), etc. [20], [21].

IV BYOD

A. What is BYOD?

There are different options available for allowing users to access business content and enterprise IT services. One of the options is when an organization distributes devices to employees with a strict security policy enabled. Another option is the so-called Bring Your Own Device (BYOD) when an employee brings their own private mobile device and it is up to the organization as to how to enforce security in such a case. A compromise between usability and security should be found.

Before selecting Android as an option for BYOD, the stakeholders should be aware that Android currently has an insufficient level of security. Currently, the situation with security and integrity is better on iOS and the Windows Mobile systems [22].

B. BYOD Threats with Android or How a Mobile Affects Enterprise Security

Quite a few guidelines and suggestions have already been prepared for somehow improving the situation in the field. Here is a list of the typical, but still insufficient suggestions identified for Android BYOD users [23]:

a) Set a device password (Settings / Location & Security / Set up screen lock),
b) Disable Unknown Source to install apps from (Settings / Applications / Unknown sources),
c) Install Anti-Virus protection,
d) Review application permissions,
e) Check for system updates,
f) Turn off wireless features (GPS, Bluetooth, Wi-Fi and Portable Hotspot) when not in use,
g) Do not Root the device,
h) Be aware of Web Security,
i) Back-up data on the device,
j) Turn off Google location.

Despite all these good suggestions, they are not enough to rescue a situation and serve as myths about adequate security. The problem lies within the inadequacies of Android architecture [18] and the low quality control in the Google Play store and the even lower quality in alternative Android stores.

In general, malware is distributed more on personal equipment, due to the lower security policy which is applied. That is why some enterprise network administrators do not allow work from home in order to protect documents from being infected with malware. Some organizations only allow connection to an enterprise IT system through VPN from specially configured workstations. All of this best practice collapses when the same mobile device is used for private and enterprise needs. Mobile phones and tablets are more vulnerable than enterprise workstations.

The PC may have old, outdated antivirus software installed which does not work or has been damaged by malware. That is why it should be checked occasionally using an antivirus CD. The equivalent check should also be done for mobile devices and memory cards [24]. Unfortunately, this is done quite rarely, but a mobile device or MicroSD card can

already contain viruses at the time of purchase. Not all antivirus software is effective enough. There are common enterprise security measures known to each corporate ICT specialist [25]:

1) using the enterprise cloud with integrated mobile secure synchronization capabilities, does not allow non-encrypted data to appear outside the secure enterprise virtual premises;
2) deploying an enterprise container with a special security level into a mobile device should decrease the risk of enterprise data leaks;
3) enabling mobile access to the enterprise document management system (e.g. SharePoint, Alfresco etc.), this decreases the desire for keeping data in some personal cloud;
4) blocking apps like Dropbox and Google Drive from accessing enterprise data;
5) the user's personal data can still be synchronized with Dropbox, Google Drive, etc. if the encrypted enterprise data containerization solutions are in place.

It is still important for an enterprise IT system to be accessed using a secure channel. That is why, when connecting to enterprise data through mobile internet or through an insecure wireless network, the creation of a VPN (Virtual Private Network) connection to one's own workplace or to the trusted internet provider service is recommended. The channel should also be secured using certificates or even more advanced solutions.

The introduction of a BYOD policy highlights new factors when assessing the security risks of an enterprise's IT infrastructure:

1) A user's full access to administer their personal mobile device is in conflict with the enterprise's general policy and their mobile device in the specific security policy. It increases the risk of data leaks and data vulnerability.
2) A user is free to choose any device model. But, there are many devices with a lower security level than desired (about 40 % of Android devices use old versions or OS versions which are not updated) [8]. This makes enterprise security management quite complex. It also becomes more difficult to track all the vulnerabilities and security updates available for the different device models.

The following is important when a mobile device is lost or stolen, especially for a BYOD case [5]:

1) the possibility of erasing the sensitive data that is kept on the device (user credentials, documents, GPS history, etc.) remotely;
2) the possibility of blocking device usage while pretending to be a device user, i.e. the messages receiving/ sending or accessing the network resources;
3) the sensitive information that is kept on the device must be encrypted;
4) the possibility of automatically blocking the device when it appears inside the untrusted Bluetooth or NFC zone.

It could be that the lost or stolen device cannot be accessed via the network and there is no possibility of initiating a remote wipe of the data. That is why encryption of the information on the device is a mission critical exercise.

A BYOD policy influences the costs of not only infrastructure and software, but also corporate risks and the level of client service. It makes the overall costs smaller, but less predictable. The following typical security mechanisms are needed when a BYOD strategy is in place [5]: 1) authentication and authorization; 2) network access control (NAC); 3) mobile device management (MDM); 4) mobile apps management; 5) encryption of both calls and internet data as well as data storage security.

Typical smartphones can ensure connection and data transfer using the following: mobile telecommunication networks (for calls, SMS, internet, and GPS support), WiFi, Bluetooth, USB cable, Micro SD card. All of them bring additional risks. Apparently, in a BYOD case, best practice is to disable all these options except for calls and SMS, and not allowing them to be enabled automatically. A user can enable them knowingly in a manual way, when needed, and disable them again when they are not being used anymore. But this solution does not cover everything, including a user's level of social responsibility.

C. Options to Restrict Android Users from Enhancing Security

There are different ways in which to forbid a user from installing or changing the configuration of an Android OS:

1) The easy one – install the AppLock free app or similar, set the password for settings, for Google Play, for apps installation and for running specific apps, etc. This option can be used by everyone.
2) There is a built-in multiple users feature in the latest Android versions (starting from 4.3 for tablets, and starting from 5.0 for smartphones). However, the feature should be extended with more configurable options per user account.
3) The complex option – to root the device and to create multiple user accounts there. Restricting the installation of new apps is possible, changing the settings, and forcing the usage of predefined networks.
4) A wholesome option would be to connect the device to the enterprise IT system, domain, or special server, e.g. Google for Work or Windows Intune with the Microsoft System Center Configuration Manager. This will be the place where user rights on the mobile device can be administered in a centralized way. It

will also ensure secure access to documents, and email.

5) Another option is to choose a device with already OEM enabled OS extra features. It is possible to limit such Android versions to only install the apps, for example, from the Nokia or Samsung stores. The manufacturers say that these stores contain apps which have been verified more thoroughly.

There are different options available for ensuring the security on an Android device depending on the selected administration method. These relations are shown in Table 1.

TABLE I

THE WAYS OF ADMINISTRATING AN ANDROID DEVICE FOR A BYOD CASE

Nr.	Administration Object	Original Android Settings	Rooted Android	Connected to the company mail server	Connected to the company workplace management server
1	Data storage encryption	+	+	+	+
2	Login password	+	+	+	+
3	Multiple user support	+ (from Android 4.3 tablets and 5.0 phones)	+	+	+
4	The ability to work from a restricted user account (forbidding the installations)	- (+ if AppLock)	+	+	+
5	Monitoring the apps' activity		+	+	+
6	Requirements for password strength and change frequency		+	+	+
7	Connect to email only through the secure channel			+	+
8	Connect to documents only through the secure channel			+/-	+
9	Centralized administration of user permission				+
10	Centralized administration of apps to run				+
11	Centralized remote apps installation				+
12	Centralized remote device update, antivirus check, backup				+

Apparently, company administrators have the broadest variety of administrative options when a mobile device is connected to a company workplace management server with mobile device management (MDM). For example, Google Android for Work allows for the administering and restricting of user account settings. Placing a restriction on a user to use and install only the allowed apps can be set within this solution as well. Windows Intune together with the Microsoft System Center Configuration Manager has similar capabilities.

The Samsung KNOX solution allows users to use a device for both personal and enterprise needs through separating these two environments. KNOX is a special Android version that is a part of "Samsung for Enterprise" (SAFE). An employee can only use the predefined and monitored apps in the enterprise environment of the device. This environment is administered by the enterprise IT department. An employee can also switch to the personal environment where he can access personal photos, the calendar, games, etc. This data is not available to the enterprise IT department. The IT department can wipe all the data from the enterprise environment when needed, while the personal environment is not affected. If the device is infected by malware, it cannot access the enterprise data and apps. [26], [27].

There are some similar services available from other vendors like Airwatch from Vmware, Blackberry Enterprise Server, Citrix, MaaS360 from IBM, MobileIron, SAP, SOTI, Motorola AME 2000, Huawei AnyOffice Mobile Security Solution, LG Electronics Enterprise Mobility Solution, etc.

The workstations within an organization are typically administrated in a centralized way by adding them to the organization domain. **The centralized management of mobile devices from the workplace management server (using MDM in particular) is potentially one of the most effective solutions for administering a mobile device in BYOD and other business cases too**. The further development and extension of such solutions and their alternatives is one of the tasks for IT in the near future.

V ARCHITECTURAL PROPOSALS FOR BETTER BYOD SUPPORT

Proposal 1 – Android as the Web Server concept

The authors propose viewing the apps on the Android platform as isolated websites on a web

server. There is much more experience accumulated on web server security than there is on Android's one. A single physical web server can host hundreds and thousands of websites from different authors. Each website can be isolated through, for example, Apache <virtualhosts> directive, allowing it (website) to exist only in the prescribed folder. It is bad practice to run a web server with one default user www-data afterwards. In such a case, the creator of any hosted website can still access other folders.

Good practice on the Linux Web server is to run the exclusive Apache instance per each website as a different guest user. Access to the website's folder should be given only to this guest user. There is no exceptional opportunity to get access to browse other folders when the system is configured in such a way. In the proposed solution, a small source of possible vulnerability could be the common RAM.

In Android, this approach has been only partially implemented – each app is executed with a different system user but apps can ask for and get extensive permission during the installation, and access other data that does not belong to them. A user is also often asked for additional permission from apps during the app update process and provides this without even reading and thinking about the content of the dialogue, which creates a messy situation.

This means that Android with default installation is not currently suitable for a BYOD, as apps isolation is insufficient. If it is not possible to isolate apps, then the only option is to forbid the user to install, update, and configure. Furthermore, the new Android feature "User Accounts" is not the final solution, because the app is updated for all user accounts at once which means that the app is the same for all users. The only difference is the profile data per each user inside the app. This means that installing or updating an infected app in the private account can also harm the data inside the user account that is meant for work.

The authors propose conducting a thorough analysis of all web server experience to look for solutions that could be carried over on Android for BYOD.

Proposal 2 - Absolute Virtual Machine Isolation per each User

Each Android app runs on its own virtual machine (VM) process, but apps isolation is not absolute, because Android uses the process VM, not the system VM. The last is more secure, so we are proposing some hybrid models. **Even though VM on Android has not been created as the security solution, it is possible to convert it to be suitable for this purpose.** The ideal conditions for a BYOD case would be if each VM used the isolated storage, isolated part of RAM, and used its own separated processor core. This could be useful for a BYOD case. The performance of the latest mobile devices with 4 and more cores and 1 GB or more RAM allows the authors to predict that it could be implemented.

There are two potential solutions available: A) a user will use the same account for private and work needs; B) a better option is when a user uses two accounts with two different profiles.

If the first option with one account is selected, then there should be a possibility for isolating potentially harmful apps from good ones and from the user data inside them. The current approach of running the VM instance per app is not suggested for the isolation described above. However, to build a totally separated VM (storage, RAM, data) for each app would take up too many system resources.

We propose running the Android VMs in two absolutely isolated processes with two different accounts where one VM is for private less trusted apps, but the second one – for work apps. The existing app separation mechanism should be kept inside these VMs. There should be no chance for private apps to access data outside their VM.

Private apps can also be divided into two more VMs, based on the trust level. Then there would be 3 VMs in total, and each could run on its own processor core. Android VMs were not originally planned as a security solution, but could become so, if apps permissions were managed not on the OS, but on a VM level. **The apps of a VM for work must be managed by the organization MDM server**, see Fig. 1.

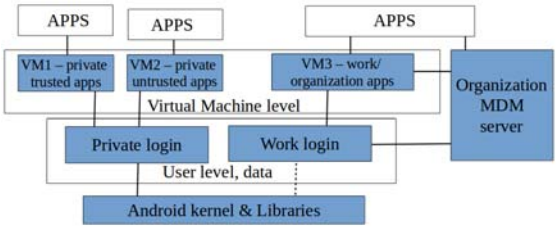

Fig. 1. The proposed *Android* architecture security model for a BYOD.

If there are two different accounts created for the BYOD case, then it is important to separate not only the user data by isolating the data folders, but also to really separate the apps themselves. It means that the same app for two different accounts should be installed twice.

Currently the security issues for enterprises allowing BYOD are very important especially considering the alarming facts published in Veracode analytics in March, 2015 – about 2,400 unsafe mobile apps are installed on employee devices in an average global enterprise [28].

VI CONCLUSIONS AND IMPLICATIONS FOR FURTHER WORK

The Android project has succeeded in some respects. But, today it is not yet ready for BYOD by

default – for secure work and private entertainment with the same device. It is not possible to force all people in an organization to follow the security rules they have signed without some technical restrictive administration tools.

Apps that are indented to reach the masses, like governmental ones, should be developed for at least three platforms: Android, iOS, and Windows Phone. On the other hand, it could be possible to agree on only one platform for a company's internal needs. But it is worth remembering that the Android operating system is not superior to its competitors with regards to security.

When buying a smartphone or tablet, it should be a recent model: its SIM should support 4G/ LTE voice encryption; there should be a possibility to limit voice calls to only use this network type (VoIP is not the preferred encryption solution); it should be updated to the latest OS version; it should have a at least 4 cores processor and at least 1 GB RAM; there should be a possibility of encrypting data storage and setting the device login password; there should be the possibility to administer the device in a centralized way (using the MDM server) in order to minimize the organization's security risks.

The same device should not be allowed for work use and private needs if the device is not specially prepared for this. Connecting a mobile device to the company email system is not enough, because it does not provide the management of all system settings on an Android device. We suggest that any organization should ensure the use of the centralized mobile device management (MDM) server both in a BYOD case or when the organization distributes the device to employees. The following are examples of MDM systems: Google Android for Work, Windows Intune with Microsoft System Center Configuration Manager, Samsung's KNOX within Samsung Approved For Enterprise (SAFE), etc.

Looking into the future, it is desirable to continue to improve the Android architecture, taking over the positive experience of Linux Apache web servers and the experience of virtual machines for cloud services systems.

If it is decided to use the same user account both for work and for private needs on a mobile device, then an organization needs to ensure that potentially harmful apps absolutely never face the good ones and their data. We propose the running of the Android virtual machines as at least two absolutely isolated processes where one is for good (work) apps, while the other is for less trusted (private) apps. Private apps could even be separated between two more VMs. It should not be possible for private apps to access the data outside their VM.

If the decision is to use two different user accounts (which is more preferable), then it is important to isolate not only the user data from each account, but also the apps themselves. In order to achieve the isolation, a VM with its own part of RAM and its own processor core per each user should be run.

The set of apps per each user should also never face the apps and data of another user. It means that the same app from two different accounts should be installed twice. The existing app separation mechanism should be kept inside these VMs.

The proposed solutions can be helpful both for a BYOD case and for the situation where an organization distributes the devices itself. The implementation of these suggestions could make the IT world a bit safer.

VII ACKNOWLEDGEMENTS

This research is part of a project „Competence Centre of Information and Communication Technologies" run by IT competence centre, contract Nr. L-KC-11-0003, activity Nr.1.22, co-financed by European Regional Development Fund.

VIII REFERENCES

[1] D. Kerr. Android dominates 81 percent of world smartphone market. [Online]. Available: http://news.cnet.com/8301-1035_3-57612057-94/android-dominates-81-percent-of-world-smartphone-market/ [Accessed: Dec. 11, 2013].

[2] P. Beuth, W. Merkels. Handy abgehört werden konnte. [Online]. Available: http://www.zeit.de/digital/datenschutz/2014-12/umts-verschluesselung-umgehen-angela-merkel-handy [Accessed: Dec. 25, 2014].

[3] K. Nohl. Attacking phone privacy. Berlin: Security Research Labs. 2010.

[4] OpenSignal. Android Fragmentation Visualized. [Online]. Available: http://opensignal.com/reports/2014/android-fragmentation/ [Accessed: Dec. 20, 2014].

[5] Samsung Electronics Co. Ltd., featuring Gartner. Strategies to Solve Challenges of BYOD in Enterprise. 2013.

[6] A. Goodloe, S. Person. NASA Formal Methods: *4th International Symposium, NFM 2012, Norfolk, VA, USA, April 3-5, 2012*, Proceedings, Springer, 465 lpp.

[7] C. Osborne. Android app malware rates jump 40 percent. [Online]. Available: http://www.zdnet.com/android-app-malware-rates-jump-40-percent-7000019093/ [Accessed: Aug. 7, 2013].

[8] Z. Whittaker. Millions of Android users vulnerable to security threats, say feds. [Online]. Available: http://www.zdnet.com/millions-of-android-users-vulnerable-to-security-threats-say-feds-7000019845/ [Accessed: Aug. 26, 2013].

[9] N. A. Staff. How to check the legitimacy of Android apps. [Online]. Available: http://networksasia.net/article/how-check-legitimacy-android-apps-1324343340, 11.10.2013.

[10] P. Ducklin. Naked security. [Online]. Available: http://nakedsecurity.sophos.com/2013/05/31/android-malware-in-pictures-a-blow-by-blow-account-of-mobile-scareware/ [Accessed: May 31, 2013].

[11] C. Castillo. McAfee Blog Central. [Online]. Available: - http://blogs.mcafee.com/mcafee-labs/phishing-attack-replaces-android-banking-apps-with-malware [Accessed: Jun. 3, 2013].

[12] Android Smartphone Security. [Online]. Available: http://latestandroids.wordpress.com/2013/06/10/android-smartphone-security/ [Accessed: Jun. 10, 2013].

[13] Technet. Security Options. [Online]. Available: http://technet.microsoft.com/en-us/library/jj852268.aspx [Accessed: Dec. 25, 2014].

[14] Technet. Windows Server Security. [Online]. Available: http://technet.microsoft.com/en-us/windowsserver/windows-server-security.aspx [Accessed: Jan. 12, 2015].
[15] Android Community. Android Security Overview. [Online]. Available: http://source.android.com/devices/tech/security/ [Accessed: Jan. 8, 2015].
[16] Android Community. Dashboards. [Online]. Available: http://developer.android.com/about/dashboards/index.html [Accessed: Nov. 12, 2014].
[17] Android Community. Session Initiation Protocol. [Online]. Available: http://developer.android.com/guide/topics/connectivity/sip.htm [Accessed: Nov. 12, 2014].
[18] L. Chanhee, K. Jonghwa, C. Seong-je, C. Jongmoo, P. Yeongung. Unified security enhancement framework for the Android operating system. *Supercomput 67:738–756, DOI 10.1007/s11227-013-0991-y, Springer Science+Business Media New York 2013*, Published online: 6 August 2013 [Accessed: Dec. 29, 2014].
[19] YAFFS. Overview. [Online]. Available: http://www.yaffs.net/yaffs-overview [Accessed: Dec. 19, 2014].
[20] F-Secure Labs. Whitepappers. Mobile Threat Report. [Online]. Available: http://www.f-secure.com/static/doc/labs_global/Research/Mobile_Threat_Report_Q3_2013.pdf. [Accessed: Nov. 12, 2014].
[21] F-Secure. (2014). Mobile Threat Report Q1 2014. [Online]. Available: https://www.f-secure.com/documents/996508/1030743/Mobile_Threat_Report_Q1_2014.pdf. [Accessed: Feb. 8, 2015].
[22] Microsoft Corporation. Low cost devices in government and education- Windows vs. Android. [Online]. Available: http://blogs.technet.com/cfs-file.ashx/__key/communityserver-components-postattachments/00-03-59-75-35/Low-Cost-Devices-in-Government-and-Education-_2D00_-Windows-vs-Android.pdf [Accessed: Jan. 11, 2015].
[23] T. Oh, B. Stackpole, E. Cummins, C. Gonzalez, R. Ramachandran. Best Security Practices for Android, BlackBerry, and iOS. *The First IEEE Workshop on Enabling Technologies for Smartphone and Internet of Things (ETSIoT)*, 2012.
[24] CERT.lv. Datorvīrusu ierobežošana. [Online]. Available: https://cert.lv/uploads/uploads/Seminari/Datorvirusu_ierobezosana.pdf [Accessed: Dec. 11, 2014].
[25] B. Rossi. Mobile content management and BYOD: the Dropbox catch-22. [Online]. Available: http://www.information-age.com/technology/mobile-and-networking/123457826/mobile-content-management-and-byod-dropbox-catch-22 [Accessed: Mar. 20, 2014].
[26] B. X. Chen, I. Austen. Samsung Armors Android to Take On BlackBerry. The New York Times. 2013.
[27] Samsung Electronics Co., Ltd., Enterprise Mobility Solutions. White Paper: An Overview of Samsung KNOX 2013.
[28] Average Large Enterprise Has More Than 2,000 Unsafe Mobile Apps Installed on Employee Devices. Veracode Press Release, March 15, 2015. [Online]. Available: http://www.veracode.com/average-large-enterprise-has-more-2000-unsafe-mobile-apps-installed-employee-devices [Accessed: Mar. 18, 2015].

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 65-70

Actualities of balance diagnostics system model development for persons with disabilities

Aleksandrs Gorbunovs, Bruno Zuga, Janis Kapenieks (sn.), Atis Kapenieks, Rudolfs Gulbis, Zanis Timsans
Distance Education Study Centre, Riga Technical University.
Address: Kronvalda Blvd. 1, Riga, LV-1001, Latvia

Abstract. **In Latvia so far there have not been made comprehensive studies to provide balance function diagnostics for persons with disabilities. It is also recognized that there is a lack of dedicated efficient and widely accessible for patients systems, equipment and tools to make balance function and dysfunction diagnostic, as well an improvement of the patients' movement capabilities. This position paper gives an overview about the latest technical engineering solution trends and developments in the world, which could be suited and developed to create the conformable new, rather simple, accessible and cheap systems. Such systems will have to ensure the diagnostics of balance function for persons with disabilities and encourage them the usage of mentioned equipment to particular balance and movement functionality.**

Keywords: **balance disorder, controller, disability, sensor.**

I INTRODUCTION

At the moment, according to the United Nations and World Bank data, approximately 15 per cent of the whole world population are living with some kind of disability [1], and about 3 per cent of them have significant disabilities [2]. Besides, population aging tendency makes these proportions even worse. The most unfavourable situation is in the group of the world's poorest countries and among indigent population where one fifth of residents are disabled [3]. One of the eight persons in European Union is either with disabilities or chronically ill [4]. And what is more unacceptably, 75 per cent of people with severe disabilities are outside European labour market [5]. During a year more than 3 per cent of the workers in the European Union and European Economic Area get injuries at the workplace [6].

Economic and Social Council of the United Nations (UN) in its Resolution has drawn the attention to the necessity of removing any barriers to the thorough participation of persons with disabilities in society activities and their employment ensuring equal rights and opportunities [7]. Integration of persons with disabilities as much as possible in society activities is declared by the UN "Convention on the Rights of Persons with Disabilities" [8]. This is the UN strategic course of action [9]. In November 2010 European Commission launched similar strategy, so-called Disability Strategy 2010-2020 [10].

Noted UN statement challenges countries and municipalities getting back to work persons with disabilities. It includes wide range of activities and necessity to work out policy documents at the regional level. Various formal and informal reintegration tools should be used to ensure rehabilitation and retention of persons with chronic disease, injuries and disabilities [6].

However, still now persons with disabilities have not sufficient possibilities to return back to work, acquire new professional competences, het a real labour, social and leisure equity. Rehabilitation and retention processes are not going quick enough. Provided service still needs to be improved; it does not satisfy quantitative and qualitative needs of persons with disabilities [11]. Appropriate continual medical and physiotherapeutic actions and care can hasten rehabilitation process.

Variety of patients' supporting tools and systems are also aimed to help for persons with disabilities. However, abovementioned United Nations and European Commission strategies urge researchers to develop new innovative technical solutions. The aim of this position paper is to direct attention to the necessity of creating low-cost helpful efficacious system which would ensure balance diagnostics and balance functions improvement for people with disabilities.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.185

II TOOLS AND METHODS

A. Theoretical background of human balance characteristic parameters and control methods

Human imbalances can be described by the inverted pendulum model. It is used to determine the gravity and acceleration perturbations. If they are identified, it would be possible to determine motoric mechanisms which could allow protecting against abovementioned perturbations [12].

The measurement of the position of the body is carried out by quantifying changes of the centre of gravity in direction from the front to rear and middle-lateral deflection. In the inverted pendulum model the balance dynamic comprises deflections both from the front to rear and from the middle to lateral. The model includes the acceleration controlled by the weight, and perturbations caused by gravitation [13].

Human balance mechanism controls the position of the separate centre of gravity of each foot in respective room area. If this mechanism detects the position deflection of the centre of gravity from admissible value, corresponding muscle group receives a signal which ensures a return to the equilibrium state [13].

If balance control mechanism is not able to control deflections of the centre of gravity effectively enough, balance disorders will appear. Then appropriate technological solutions might help people and allow them controlling wittingly the deflection from admissible norms, and as a result, maintaining the equilibrium state. Accordingly, automatic feedback, which ensures keeping the balance, would be substituted with the conscious balance control. And that could be supported by technological solutions. Besides, over time the conscious balance control will become as a habit of balance ensuring.

The balance ensuring mechanism controls decrease and increase of the load onto a straight line which connects human centre of gravity with the spots under each foot. If the angles between the lines are correct, the lines of the force increasing-decreasing lines are situated in the related tandem positions, and the ankle muscles control them independently. In the interim state the lines exist but the ankle muscles control processes are not orthogonal ones; usually they work at approximately 60 degree angle. The direction and force increasing-decreasing lines control of the ankles allows fixing the muscle groups' activity adding points which are responsible for the ankles position and balance ensuring mechanism in the rest state in different positions [13].

Human balance capabilities might be improved by rather simple user's centre of pressure measurement tool which is known as the centre of pressure and complexity monitoring system (CPCMS). It ensures both data collection and analysis [14, 15].

The coordinates of the centre of pressure (COP) are calculated, taking into account generated reaction force which is an outcome from four pressure sensors F1, F2, F3, F4, using the formula: W = F1 + F2 + F3 + F4, where W is the reaction force, and COP = (X, Y) where X and Y – respective directions. Further, if a distance between the sensors is L, then the COP location could be calculated using formulas as follows [15]:

$$X = \frac{[(F4 + F2) - (F1 + F3)] \times L}{W} \quad (1)$$

$$Y = \frac{[(F3 + F4) - (F1 + F2)] \times L}{W} \quad (2)$$

B. Balance testing systems and tools

Existing balance systems can be divided into two main groups: deep muscle training equipment (spheres, discs, soft platforms, etc.) and digital balance testing systems (Biosway, DBA (Digital Balance Analyser), ICS Balance Platform, etc.). Training equipment strengthens musculature which is responsible for keeping of balance but do not provide diagnostic information. Besides, there is also not possible to detect and register ongoing change dynamics with training equipment. Digital systems are designed for very precise balance function measurement. These data can be used in scientific research. Unfortunately, such systems are not widely available due to the high price. Therefore, the necessity of creating user-friendly, cheap and data assessable system is activated. Planned balance training system prototype might be developed also taking into account widely available nowadays mobile devices and built-in sensors.

The United States company Biodex introduces powerful and efficacious equipment Balance System™ SD which can satisfy needs of wide range of its users. It is not limited for medical purposes only, and can help in balance diagnostics and improvement both for athletes, persons with disabilities, and everybody who looks for the improvement of balance capabilities and development of muscle tone. It is also aimed to help people with injures to improve their kinaesthetic abilities. Five test options and six training activities are provided [16, 17].

Balance diagnostics and exercises are implemented on the special surface platform which can be used to measure both static and dynamic balance capabilities. To ensure users safe on-off movement, it is equipped with locking mechanism. For safety purposes adjustable support handles are also installed. The data, received from the placed in the platform set of sensors, reach high resolution colour touch screen LCD display which allows patients monitoring their exercise progress. All diagnostics results can be recorded and sent to print [17].

However, Balance System™ SD ought to be recognized mainly as the stationary system. Regardless of small wheels, attached to the platform, which make system movement process much easier, it remains still rather uncomfortable for relocation due to its weight and dimensions.

Another Biodex development in the form of Biosway system with lightweight platform can solve relocation problem. It is produced as portable equipment and can be transported in a couple bags. However, even this advantage does not allow patients with disabilities moving on their own [18, 19].

There are also attempts to create more compact and suitable for daily use tools. Researchers have developed some balance testing systems which put to use the sensors already embedded in mobile phones. In practice this works so that the phone is secured on the body with a longer splint and the range of movement anomalies are summarised in a time period.

Balance problem solving application software for mobile phones and tablets might be divided into two groups: a) for performing measurements and b) training. To illustrate both practical applications further in this sub-chapter we will give some examples.

For example, Visual Vertical application software supports balance measurement [20]. It could be installed into iPhone or iPad. These devices ought to be inserted and fixed at the bottom of the opaque bucket. The aim of the use of the basket is to limit patient's view angle and prevent observation of surroundings during the test. Patient's task is, looking at the red line, to amend the bucket so that the line is vertical. The patient must perform this task initially with both eyes and then with each eye separately. The measurement is displayed in degrees for the deviation from vertical. As the measurements might not be precise enough, it is suggested to perform this test several times; the average value is calculated by noted application. The application is cheap (USD 14,99). Pursuant to that, despite some inaccuracies the tool may tempt customers to choose it.

Another balance measurement system iBest is designed on the Android mobile operating system (OS) base and is for mobile devices to be used in this OS. IBest consists of three main components: mobile device, data storage cloud server and Web technology interface. Mobile device offers information about an exercise, detects and evaluates balance and sway abnormality by using of sensors embedded into smart phone, and gives feedback in audio form. Storage server performs trend analysis from the user's profile saved data. Web interface ensures therapists' access to the patients' data and statistics. The system employs three sensors: acceleration, gyroscope, and magnetometer ones; the combination of all these sensors allows getting a higher accuracy [21]:

- Acceleration sensor provides speed and position change tracking. Multiple axes accelerometer also can be used as the orientation sensor in the same plane,
- Gyroscope sensor allows measuring orientation or rotation speed changes,
- Magnetometer sensor like compass provides absolute orientation in space.

The primary function of the iBest device is to help those patients who want to continue the practice at home. Unfortunately, due to small size of the pilot group the researchers did not compared achieved results with accepted solutions in treatment.

Interactive Medical Productions, LLC has launched the game iGotBalance involving users into interactive process which improves their movement patterns, i.e. balance and stability [22]. Interactive game starts with a template where the user can choose different positions to check his (or her) balance: standing on one leg, with opened or closed eyes, as well set up other parameters. The task is, stabilising a body, to keep a ball in the centre of the circle displayed on the mobile device's screen. This application costs just USD 0,99 [23].

Australian and Singapore researchers present their development in a form of portable, relatively cheap and widely available equipment for the Wii game console – the Wii balance board (WBB) to assess of standing balance [24]. Special software was created for this board to be used onto computers. As one of the shortcomings might be the platform's failure to assess the force in horizontal direction which is required for defining a formula of the balance centre. Researchers concluded that, in comparison with the treatment of recognized devices, this solution can provide a satisfactory quality of the measurement.

J.A.Patterson et al. [25] made the measurement of balance functions by using of tri-axial accelerometer sensors installed in smart phones and later on compared them against measurement results gotten from previously validated Biodex Balance System™ SD. The scores were similar. It was concluded that their approach can give positive effect to measure and monitor users balance functions. However, while the balance was assessed only in anterior/posterior direction, the further findings are required [25].

Scientists from the University of Sydney developed special iPhone, iPad and iPod compatible educational application to test patients' angular vestibulo-ocular reflex (aVOR) and detect VOR disorders. This tool provides not only testing, but also great teaching and training opportunities in rehabilitation process for patients with vestibular dysfunctions. The application shows to patients what the spring-up of vestibular disorder is and how it ought to be treated [26].

Advances in Latvia and Lithuania

Pursuant to a big number of patients who do not receive appropriate treatment, care and rehabilitation,

Rezekne University of Applied Sciences and Siauliai University organized rehabilitation treatment for patients with lower body injuries. Research as well patients' rehabilitation was implemented within MODPART project from January 2012 till December 2013. The project had to encourage persons with disabilities to come back in the labour market. Individual psychophysical, vocational rehabilitation and other health facilitation activities were offered to the patients [27].

During project activity period Biosway system and its features were applied in practice. Researchers developed appropriate methods and technology for the rehabilitation and health improvement. Integrated studies were taken on social and medical rehabilitation to mitigate social risks and facilitate reintegration into the labour market [28, 29, 30].

III JUSTIFICATION, PROPOSAL AND DISCUSSIONS

Justification

Many and many mobile communication devices nowadays are already equipped with different types of tools which are used to determine user's position, deflections, and many other requests. Embedded in mobile phones sensors could be used as testing, training, gaming and entertaining instruments. Companies are started development of modern controllers, mini data processing mother-boards, which allow engineers creating a huge amount and different sorts of interactive devices and systems.

While Raspberry Pi controller (i.e. an analogue of computer's motherboard) is great in the interacting with computers, in gaming and entertainment environment, such as music or Internet browsing, although it is good for other applications, for example, used in weather stations or observation posts, etc., and BeagleBone Black, similarly to Raspberry Pi has a great potential for the hobbyists to construct innovative products [31, 32], Arduino microcontroller is aimed for the purposes which differ from them. While Raspberry Pi, for instance, is a microprocessor based single-board computer, Arduino is a microcontroller based physical computing platform [33]. Arduino costs less than USD 30 which is very affordable price, cheaper than Raspberry Pi or BeagleBone Black. But main differences which speak for Arduino pick are as follows. Raspberry Pi and BeagleBone Black are better when a developer plans work on software applications; on the other hand, Arduino is by far better in case of working on hardware projects. Arduino has a real-time capability but Raspberry Pi does not. Arduino has analogue capability that Raspberry Pi does not. So, Arduino is more flexible than other controllers. This gives to Arduino notable advantage in working with any sort of sensors [34]. All that differ Arduino approach from others. From this point of view Arduino features meet better our expectations.

Besides, Arduino microcontrollers attract our attention also thanking to their diverse small sizes and forms which might be used in different applications and conditions [32, 35], in particular, where important role is given to sensors, robotics, interactive objects and so on. This, and noted before advantages makes Arduino microcontroller more attractive to be chosen from the set to reach our goal – creating human balance diagnostic system which must be low-cost, affordable, available, user-friendly, reliable, rather sensitive and precise.

Proposal

The idea is to incorporate "Arduino" microcontroller, acceleration sensors, multi-screen digital devices and server into one joint system (Fig. 1).

Acceleration sensors will be attached to patient's body (it could be done either on the head, chest, wrists, shoulders, or other parts of body including lower body ones). Sensors will be wired connected to Arduino microcontroller. This will allow transmitting of the signals from the sensors to controller; as well provide power supply to ensure an operation of these sensors. As an alternative, the wires could be substituted by wireless connections; although, such solution might cause additional questions.

To ensure wireless data transmission between Arduino and digital data processing devices, Bluetooth module will be used and connected to Arduino. To do this, Bluetooth module could be programmed by AT-commands through Bluetooth module's UART interface.

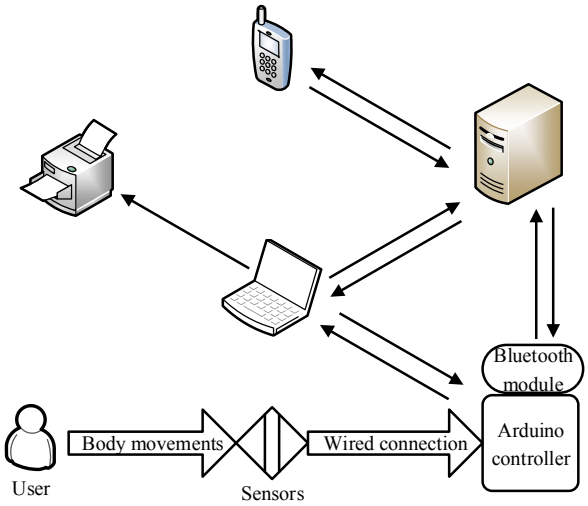

Fig. 1. Human balance diagnostic system proposed model

The data in a form of sensor signals about patient's body movements will be sent through Arduino microcontroller to the server which, depending on patient's needs, will distribute obtained data to the

computers, laptops, mobile phones and tablets. Patient's movement data, e.g. body deflections, will be stored in the server's database. Computers and laptops might be used as a replacer of the server. In this case all the data will be stored onto hard discs of computers (laptops).

Discussions

An important thing, which we ought to keep in mind, is the power supply and data transfer. This both issues are closely related. In our case, when acceleration sensors are attached to the body or clothing, sensors power supply issue from energy sources becomes challenging. We can discuss about patient's comfort. On the one hand, the wires really could not be considered as a comfortable dress supplement. On the other hand, independent energy source in a form of battery, powered by DC voltage, could be more comfortable but due to intensive exploitation may request rather frequent battery replacement; besides, it would provoke another problem of sensors data transmitting, but this is not that what we want.

Next challenge might be balance disorder data precision and interpretation. Patient's body starting position and positions during the test, deflections, moving speed, and other parameters are crucial. It is necessary to work out appropriate algorithms to solve this problem.

Developed system expenses are expected to be low. Raw calculation of such costs indicates that we will be able to be in the limits of fifty USD (i.e. Arduino microprocessor – less than USD 30, Bluetooth module – less than USD 10, sensors – less than USD 10, plus wires or batteries and other fittings). Such price could be considered as a very affordable for almost everybody.

IV CONCLUSIONS

At the moment various social and technical solutions are represented in the world to help persons with disabilities, among them – lower body injuries, vestibulo-ocular reflex disorders, other balance dysfunctions. Unfortunately, these solutions in their majority can be used by the narrow spectrum of users because of much orientation of the system developers and sellers on the definite operating systems and devices (first of all, mobile ones); another problem is that technical solutions in a form of equipment, tools and systems, usually offered in the market, are either sensitive and precise but expensive and not affordable, or cheap but not precise enough.

The solution, proposed in this position paper, may solve abovementioned problem. It might be possible to achieve the aim of research – to create a cheap, affordable for everybody and widely available, balance testing and rehabilitation exercises ensuring, progress monitoring, data transmission and processing, easy to use and rather precise tool.

Data measurement precision and interpretation issue might become an integral part of the model's development. As a next step in new system approval would be the comparison of obtained data from new developed system prototype with the data from existing approved body balance testing systems, e.g. Biosway.

V ACKNOWLEDGMENTS

This study was partly supported by the National Research Program "Innovative solutions in social rehabilitation in Latvian schools in the context of inclusive education" („Inovatīvi risinājumi sociālajā telerehabilitācijā Latvijas skolās iekļaujošās izglītības kontekstā – VPP INOSOCTEREHI"); project No.2 „Development, approbation and implementation of new prototypes and innovative methodology (approaches, methods, techniques, ways) in social rehabilitation for the rendering of new services" („Jaunu prototipu, inovatīvas metodikas (pieeju, metožu, tehniku, paņēmienu) sociālajā rehabilitācijā izstrāde, aprobācija un ieviešana jaunu pakalpojumu sniegšanai"). Contract No. 10-4/VPP-8-8 (RTU PVS ID 1804).

VI REFERENCES

[1] The United Nations. Factsheet on Persons with Disabilities, 2015. [Online]. Available: http://www.un.org/disabilities/default.asp?id=18 [Accessed May 10, 2015].

[2] The World Bank. Disability, Apr. 06, 2015. [Online]. Available: http://www.worldbank.org/en/topic/disability [Accessed May 10, 2015].

[3] Disabled World. Disability Statistics: Facts & Statistics on Disabilities & Disability Issues, 2015. [Online]. Available: http://www.disabled-world.com/disability/statistics/ [Accessed May 10, 2015].

[4] M.Živetere, Cilvēku ar īpašām vajadzībām integrācija darba vidē: vadlīnijas, metodiskie materiāli, [Integration of people with disabilities into working environment: guidelines, methodical materials], Riga: ISMA, 2008.

[5] European Disability Forum, "Nothing about us without us", 2013. [Online]. Available: http://www.edf-feph.org/Page_Generale.asp?DocID=12534 [Accessed Mar. 20, 2015].

[6] W.Oortwijn, E.Nelissen, S.Adamini, S. van den Heuvel, G.Geuskens, L.Burdof, Social determinants state of the art reviews - Health of people of working age - Summary Report, European Commission Directorate General for Health and Consumers, Luxembourg, p. 40, 2011.

[7] UN Economic and Social Council, "Promoting the rights of persons with disabilities and mainstreaming disability in the post-2015 development agenda", United Nations 23rd plenary meeting, Resolution E/RES/2014/6, p. 4, 12 June 2014.

[8] The United Nations. Convention on the Rights of Persons with Disabilities, 2006. [Online]. Available: http://www.un.org/disabilities/convention/conventionfull.shtml / [Accessed May 10, 2015].

[9] UN General Assembly, "Realization of the Millennium Development Goals and internationally agreed development goals for persons with disabilities: a disability - inclusive development agenda towards 2015 and beyond", Report of the Secretary - General A/67/211, p. 19, 30 July 2012.

[10] European Commission, "EU launches Disability Strategy 2010-2020", 2010. [Online]. Available: http://ec.europa.eu/

social/main.jsp?catId=89&langId=en&newsId=933&furtherNews=yes. [Accessed Mar. 20, 2015].

[11] I.Skestere, G.Anca, L.Embrekte, M.Gravis, I.Neimane, A.Pastore, A.Silina, B.Silina, T.Vahlina, L.Velka, K.Alksne-Alksnite, O.Zubkovs, "Koncepcija „Dzīve sabiedrībā"", ["Conception "Life in Society""], p. 43, 2008. [Online]. Available: http://www.lkndz.lv/lv/box/files/filelists/1290686941-koncepcijabrosura.pdf. [Accessed Mar.10, 2015].

[12] D.Winter, "Human balance and posture control during standing and walking," Gait & Posture, vol.3, no.4, pp. 193-214, 1995.

[13] D.Winter, J.Frank, C.Powell, K.Zabjek, "Unified theory regarding A/P and M/L balance in quiet stance," Journal of neoropsychology, vol.75, no.6, pp. 2334-2343, 1996.

[14] A.Enders, and Z.Brandt, "Using Geographic Information System Technology to Improve Emergency Management and Disaster Response for People With Disabilities," Journal of Disability Policy Studies, vol.17, no.4, pp. 223-229, 2007.

[15] C.Huang, P.Sue, M.Abbod, B.Jiang, J.Shieh, "Measuring center of pressure signals to quantify human balance using multivariate multiscale entropy by designing a force platform," Sensors, vol. 13, no. 8, pp. 10151–10166, 2013.

[16] Biodex, "Balance System SD", 2015. [Online]. Available: http://www.biodex.com/physical-medicine/products/balance/balance-system-sd. [Accessed Mar. 12, 2015].

[17] Biodex, "Operation / Service Manual", Biodex Medical Systems Inc., pp.123, 2014. [Online]. Available: http://www.biodex.com/sites/default/files/950440man_10205revc.pdf. [Accessed Mar. 12, 2015].

[18] Biodex, "BioSway", 2015. [Online]. Available: http://www.biodex.com/physical-medicine/products/balance/biosway-portable. [Accessed Mar. 12, 2015].

[19] Biodex, "Balance System and Biosway: Software Manual – User's Guide", pp.123, 2014. [Online]. Available: http://www.biodex.com/sites/default/files/950440man_add_14074.pdf. [Accessed Mar. 12, 2015].

[20] Clear Health Media, "The Easy, Accurate and Low-Cost way to measure subjective visual vertical in your patients", 2015. [Online]. Available: http://www.clearhealthmedia.com/measure-subjective-visual-vertical/ [Accessed Jan. 30, 2015].

[21] A.A.Wai, P.D.Duc, C.Syin, H.Zhang, iBEST: Intelligent balance assessment and stability training system using smartphone: Engineering in Medicine and Biology Society (EMBC), the 36th Annual International Conference of the IEEE, pp. 3683 – 3686, Aug. 26-30, 2014, Chicago, USA.

[22] Interactive Medical Productions, LLC, "The iGotBalance App", 2014. [Online]. Available: http://www.igotbalance.com/ [Accessed Jan. 15, 2015].

[23] iTunes, "iGotBalance", 2014. [Online]. Available: https://itunes.apple.com/us/app/ustabilize-lite/id435404380?mt=8 [Accessed Jan. 15, 2015].

[24] R.A.Clark, A.L.Bryant, Y.Pua, P.McCrory, K.Bennell, M.Hunt, "Validity and reliability of the Nintendo Wii Balance Board for assessment of standing balance", Gait & Posture, vol. 31, pp. 307-310, 2010.

[25] J.A.Patterson, R.Z.Amick, T.Thummar, M.E.Rogers, "Valdation of measures from the smartphone sway balance application: a pilot study", Internation Journal of Sports Physical Therapy (IJSPT), vol. 9, no.2, pp. 135-139, April 2014. [Online]. Available: US National Library of Medicine, National Institutes of Health, http://il.proquest.com. [Accessed Jan. 28, 2015].

[26] H.MacDougall, S.Rogers, "aVOR – angular vestibulo-ocular reflex", 2014. [Online]. Available: https://itunes.apple.com/us/app/avor/id497245573?mt=8 [Accessed Feb. 26, 2015].

[27] Rezekne University of Applied Sciences and Siauliai University, "LAT-LIT project "Designing a Model Geared towards Participation of People at Social Risk Groups in the Labour Market" (MODPART)", Project index LLIV-223", pp. 3, 2013. [Online]. Available: http://www.ru.lv/en/research/scientific_institutes/psri/psri_projects/ [Accessed Nov. 24, 2014].

[28] V.Lubkina, A.Kaupuzs, S.Usca, S., L.Rizakova, A.Ciukmacis, "The testing technology of neuromuscular system, the methodology of its introducting", pp. 164, 2013. [Online]. Available: http://www.ru.lv/ckfinder/userfiles/RAweb/Saturs/zinatne/zinatniskie_instituti/personas_socializacijas_petijumu_instituts/izdevumi/2014/monografiskie_izdevumi/THE%20TESTING%20TECHNOLOGY%20OF%20NEUROMUSCULULAR%20SYSTEM.pdf [Accessed Nov. 24, 2014].

[29] V.Lubkina, A.Kaupuzs, S.Usca, S., L.Rizakova, A.Ciukmacis, "The technology of balance testing and training, the methodology of its introducing", pp. 136, 2013. [Online]. Available: http://www.ru.lv/ckfinder/userfiles/RAweb/Saturs/zinatne/zinatniskie_instituti/personas_socializacijas_petijumu_instituts/izdevumi/2014/monografiskie_izdevumi/THE%20TECHNOLOGY%20OF%20BALANCE%20TESTING%20AND%20TRAINING.pdf [Accessed Nov. 25, 2014].

[30] V.Lubkina, A.Kaupuzs, S.Usca, S., L.Rizakova, A.Ciukmacis, "Methodology of Vibromassage technology and its implementation", pp. 133, 2013. [Online]. Available: http://www.ru.lv/ckfinder/userfiles/RAweb/Saturs/zinatne/zinatniskie_instituti/personas_socializacijas_petijumu_instituts/izdevumi/2014/monografiskie_izdevumi/METHODOLOGY%20OF%20VIBROMASSAGE%20TECHNOLOGY.pdf [Accessed Nov. 26, 2014].

[31] Raspberry PI, "What is a Raspberry Pi?", 2015. [Online]. Available: http://www.raspberrypi.org/help/what-is-a-raspberry-pi/ [Accessed Feb. 27, 2014].

[32] D.Norris, The Internet of Things: Do-It-Yourself at Home Projects for Arduino, Raspberry Pi, and BeagleBone Black, San Francisco: McGrow-Hill Education, 2015.

[33] R.Adams, "What are the differences between Arduino and Raspberry Pi?", 2015. [Online]. Available: http://www.quora.com/ What-are-the-differences-between-Arduino-and-Raspberry-Pi [Accessed Mar. 18, 2015].

[34] L.Orsini, "So you want to get into DIY hacking. Which device should you buy?", 2015. [Online]. Available: http://readwrite.com/2014/05/07/arduino-vs-raspberry-pi-projects-diy-platform [Accessed Mar. 18, 2015].

[35] T.Olsson, Arduino Wearables, Berkley, CA: APress, 2012.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 71-75

Storing an OWL 2 Ontology in a Relational Database Structure

Henrihs Gorskis, Arkady Borisov
Riga Technical University,
1 Kalku Str., Riga LV-1658, Latvia

Abstract. **This paper examines the possibility of storing OWL 2 based ontology information in a classical relational database and reviews some existing methods for ontology databases. In most cases a database is a fitting solution for storing and sharing information among systems, clients or agents. Similarly, in order to make domain ontology information more accessible to systems, in a comparable way, it can be stored and provided in a database form. As of today, there is no consensus on a specific ontology database structure. The main focus of this paper is specifically on OWL 2 as a basis for the description of ontology centric information in a database. The Web Ontology Language OWL 2 is a language for describing ontology information for the Semantic Web. As such it consists of a list of reserved words and grammatical rules for defining many parts of ontology knowledge. Based on this language specification this paper examines the possibility of storing information in a relational database for the description of domain ontology information. By creating a database structure based on OWL2 it is feasible to obtain an approach to storing information about the domain ontology in an utilizable way, by using its descriptive abilities. Nowadays multiple approaches to storing ontology information and OWL in databases exist; most of them are based on storing RDF data or provide persistence for specific OWL software libraries. The examination of the existing approaches provided in this paper, shows how they differ from the goal of obtaining a general, more easily usable and less software library specific database for domain ontology centric information. This paper describes a version of a simple relational database capable of holding and providing ontology knowledge on demand, which can be implemented on a database management system of choice.**

Keywords: **ontology, OWL2, relational database.**

I INTRODUCTION

Ontology knowledge is a powerful tool to share, describe and classify information about a given domain. The ontology describes concepts important to a domain. It does it by naming classes, individuals and relations and describing how these ontology elements interact with each other. By using reasoning on this information, new relationships between concepts and individuals can emerge, and individuals can be classified by reasoning about their attributes. The ontology knowledge can be provided in many different forms and using many different languages or notations to describe the information. In its most basic form, ontology knowledge is stored in a file on a computer. In order to use the ontology more easily, it would be desirable to access it with the same ease as a database. There are many approaches to storing ontology or similar information in a database. Some databases are structured around the information it contains and how the information is used; other approaches store the ontology in its most basic form in RDF triplets, and still other methods include storing API specific data structures in a database for persistence of these variable objects. This paper shows that having a natural and common structured relational database for storing ontology information can be useful for many applications. A general and common database is accessible to many different software applications or agents and requires only the understanding of the database structure by these agents. This makes it possible for the agents or software applications to be developed separately instead of requiring them to use the same API or be written in the same programming language. In order to store ontology knowledge, it is necessary to understand the kind of information stored in modern ontology modelling approaches.

II ONTOLOGY KNOWLEDGE

At its core any ontology describes concepts, individuals and properties and uses structures in order to convey information about elements of the domain. Ontology information can be described using many different notations and approaches. In order to be able to use the most common and full approach, the OWL 2 standard will be used as an ideal. OWL 2 is the web ontology language version 2 for the Semantic Web with formally defined meaning [1]. It provides many useful keywords to frame and describe knowledge. Using these keywords and structures makes it possible

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.168

to describe a doming ontology and the knowledge contained in it. First it provides a way to assert the association of certain pieces of information to classes, individuals or properties. This is done using the "Declaration" operator followed by "Named Individual", "Class", "Object Property", "Data Property" or "Data type" and the name of the entity being associated with one of these types. Next, the operator "Class Assertion" is used to classify an individual, by stating the individual's name and the class it belongs to. The operators "Sub Class Of" and "Sub Object Property Of" are used to create a hierarchy of classes and properties to form taxonomy. Ontology building requires the expert to follow the rules of how an ontology is structured. In order to create a correct taxonomy, any higher level classes must be more abstract than their lower level classes. Individuals belonging to a class must be the most distinct and unique element of the ontology. Further OWL 2 provides a way to define expressly equal or distinct classes and individuals using the operators "Equivalent Classes", "Disjoint Classes", "Different Individuals" and "Same Individual". On the lower level individuals can be described by using their relationship with other individuals. This is done using the "Object Property Assertion" operator and stating the name of the object property and the second individual in this relationship. The operator "Negative Object Property Assertion" can be used to express a distinct lack of a specific relationship between two individuals. For implying a class to members of a relationship, the operators "Object Property Domain" and "Object Property Range" are used. By defining these attributes of the object property it can be reasoned that any individual with this property is of the specified class. OWL 2 has many more operators for defining ontology knowledge. There are very similar operators for defining data properties between individuals and data types, for example "Data Property Assertion" and "Negative Data Property Assertion". Other operators are used for defining complex classes. The operators "Object Intersection Of", "Object Union Of" and "Object Complement Of" are used to define unnamed classes which arise from the interaction of other classes. A whole array of operators is used in order to define the specific attributes of a property; these include: "Symmetric Object Property"," Asymmetric Object Property" ," Reflexive Object Property", "Irreflexive Object Property", "Functional Object Property" , "Inverse Functional Object Property" and "Transitive Object Property".

All the above and other operators are used to meticulously define every single detail regarding the knowledge contained in the ontology about each of the important domain concepts. Using the logic associated with these operators even more information about the concepts can be derived from the specifically expressed definitions. However, not every application using ontology knowledge is required to know every detail about the domain, and not every agent will use reasoning on the ontology. Some software agents are completely content using ontology concepts simply as a dictionary. This makes it reasonable to store the ontology separately from the software applications which end up using it.

III EXISTING APPROACHES

There already exist several approaches to storing and recalling ontology information. Protégé is a very popular tool for creating ontology models. However, it does not provide a solution for accessing the ontology externally. It mainly provides the means of saving the ontology in a file or source so that another program can use the ontology file. Historically, in previous versions of the software many attempts have been made to create a solution for accessing and storing ontology models created in Protégé in a database [2] - [4].

Another popular tool for working with ontology is Apache Jena. It is a programming library for JAVA. Besides many other functions it offers two ways of storing ontology data in databases. Jena comes packaged with server software called "Fuseki". Fuseki is a SPARQL server. It stores ontology information in its own internal data structure and provides access to the data by sending SPARQL queries to it. The second solution Jena offers is called TDB (), which is a native high performance triple store [5]. This approach creates a triplet-based database table in a database of the user's chaise. Apache Jena does provide means to work with ontology structures, however, underneath it relies mostly on RDF data, with the ontology being higher level abstraction of it.

It is worth mentioning another software tool for ontology persistence called OWLDB [6]. OWLDB is a database backend for the OWL API. It provides persistence for OWL API data structures. This means that by using this tool objects created by the OWL API can be stored and recalled from a database. This, however, means that only application written by using both these tools can use this function. The reason for creating this backend, as stated by the authors, was to make the use of ontology information simpler, based directly on OWL and not reliant on previous RDF structures. The approach presented in this paper is in agreement with this sentiment. In order to work with the capabilities of the ontology itself it is not necessary to implement a backwards compatibility with pre-existing approaches. Unfortunately, this project is not being continued anymore.

There also exist many different database types, all of which have their own advantages [7]. The reason a classical relational database was chosen over more specialized ones, was to make the resulting database more accessible to a wider variety of possible uses and software applications.

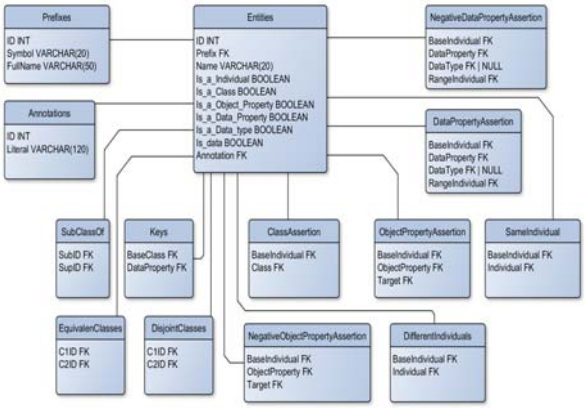

Fig. 2. Core tables

prefix, there will be only one prefix in the prefix table

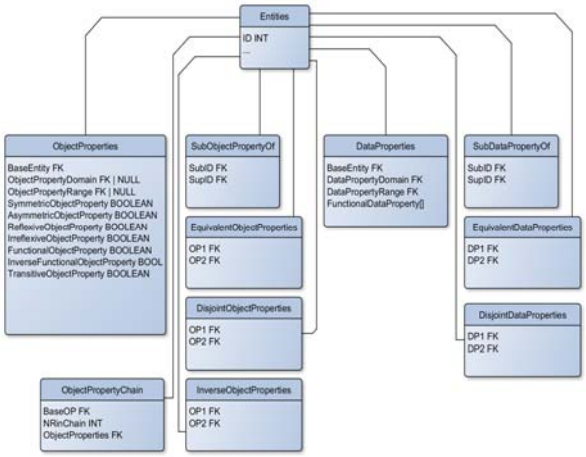

Fig. 1. Property tables

IV OWL 2 BASED DATABASE

This paper proposes a database structure based directly on OWL 2 and uses simple relations to describe ontology knowledge. At the centre of the proposed database architecture a main table is located which holds every element in the ontology (Fig. 1). In this context an element is any piece of information which describes some idea in the ontology. Without additional information such an element can potentially be a class, individual, property, literal, data type and so on. Additionally to the main table many other tables exist, each named after operators in the OWL2 specification. The main table shall hold a unique identifier, the elements name and a list of true or false Boolean variables to describe the element further. The unique identifier will be used in other tables to reference the unique element to provide additional information to it or to use it for the description of other elements of the ontology. This allows the database to reuse the named individual as many times as needed. The list of true or false operators provides hints to the type of element in order to make it easier to find additional information about the element. For example, if the element is hinted to be an individual or a class it is reasonable to search the "class assertion" table for more information about the element as to which classes the element belongs to, or which elements are individuals of this class element, depending on the type of the element. Having a list of hints requires it to be updated in addition when new entries are being made into other tables. However, at the same time, the list of hints simplifies searching for additional information immensely. Without this list of hints, it would be necessary to search the entire database and every table in it to obtain the full picture about every element. Additionally the entity table holds a reference to a prefix entry in the prefix table. By separating prefixes from the entity they can be reused. If every entity in the ontology has the same used by every entity. Finally, the entity holds a reference to an annotation in case it exists. The annotation provides additional information to a human user. Every other table in the database references one or more entities. For some tables the order of referencing entities is important, for other it is not. For example, the table "Sub Class Of" contains information about the hierarchy of classes. The column "Sub ID" holds references to the class which is the sub class in the hierarchy, while the column "Sup ID" holds the references to the class above it. So the order is important, and the naming of the columns reflects this fact. In contrast, the table "Equivalent Classes" has the columns "C1 ID" and "C2 ID". The order of the references is not important since the table describes equivalency. This must be taken into account during searches on the tables. If, for example, one wishes to find the equivalent classes to an entity, one must search for the entity's identifier in both columns, for it can have been placed in the first as well as in the second position during the creation of the ontology. Some tables reference many more entities. The table "Data Property Assertion" holds references to the entity of type individual, which has been given this attribute, the reference to the entity which describes the data property, the reference to a data type entity and the entity representing the data.

Some possible tables have been omitted from this database structure since the data they would have contained are more useful and tied directly to an entity. Besides the class, individual and other assertions being represented directly in the entity table, assertions about object property characteristics have also been added directly to the "Object property" table (Fig. 2). This table references a base entity representing this object property. In addition it can reference entities representing classes, to provide information about the domain and range of this object property. All of the possible characteristics of an

object property are provided as Boolean variables in the table.

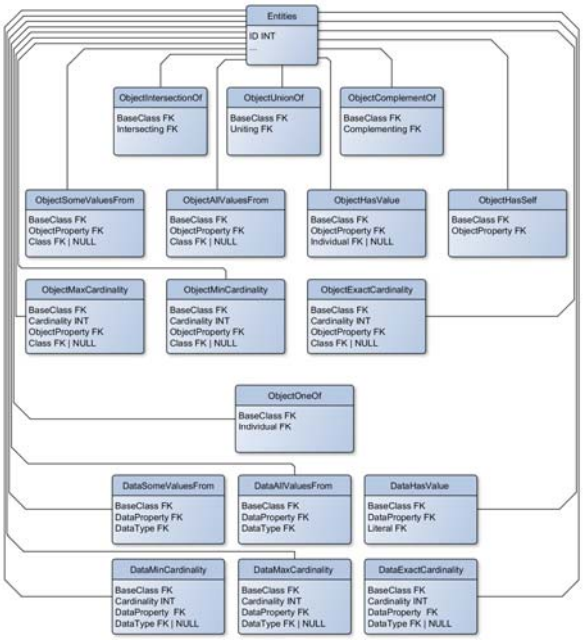

Fig. 3. Complex class tables

All the aforementioned tables provided pieces of information about an entity in a disconnected, but self-describing way. However, there are tables which hold more complex information and require all pieces of it to be obtained, before a conclusion can be reached. For example, the table "Object Intersection Of" (Fig. 3) provides parts of a description of a complex class. A complex class arises from the interaction of multiple other classes. In the case of intersection, a complex class is created, when the combination of other classes creates a new conceptual class. Since an unspecified number of classes can be involved in this interaction, a static table containing all required references cannot be created. Therefore the table "Object Intersection Of" holds only one reference to a participating class at a time. The reference to the base class is a reference to an entity describing the complex class itself, and the column "Intersecting" holds the reference to one of the intersecting classes. This means, all table entries concerning the complex class (having the same base class id) must be obtained, before it is known which classes are involved in the intersection. The same rule applies to all tables describing complex classes.

Finally, the database also holds a table capable of describing various datatypes for data properties (Fig. 4). Datatypes behave like classes with the difference that they do not have individuals, but instead govern literal data. Literals, datatypes and the data themselves are also entities in this database. Similar to complex object classes, complex data types also require multiple entries into tables. Just like the "Object Intersection Of" database table, the "Data Intersection Of" table provides information about all data types whose intersection form a new complex data type.

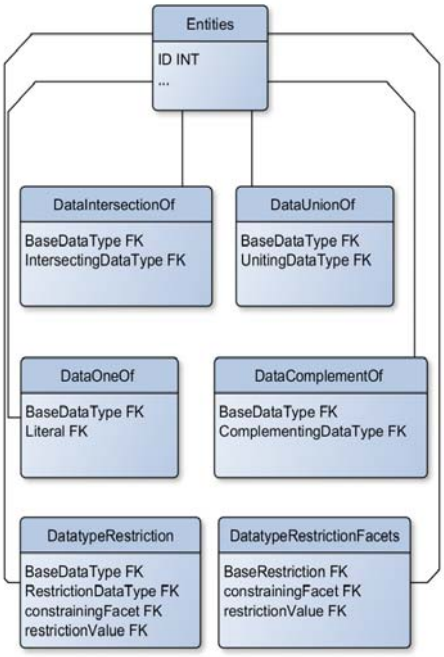

Fig. 4. Datatype tables

Access to the knowledge in the database is very simple. In the case, when a user or software agent is looking for a specific concept, the main entity table is searched for it by name. There can be two entities with the same name in the table. In such a case the related prefix can be consulted. If a prefix was specified within the search parameters, the unique entity can be found. Once the entity is found, it can be provided to the user. Some systems may be satisfied at this stage with the obtained information. Other systems may choose to obtain further information about the found entity. Based on the entity's characteristics further tables can be polled for additional information based on the identification of the entity. In most cases the result of searching other tables will be a list of identifications referencing other related entities. Again, depending on the nature of the obtained information some systems may choose what connected information must be researched further. To do this, the main table is searched again based on the identifications and other entities and their name and characteristics are obtained. This process is repeated until the user or system has obtained all the required information about the original and related entities.

This database structure can be implemented in any standard database system. For example, the creation script of the main entity table for MYSQL looks as follows:

```
DROP TABLE IF EXISTS
      'owl2db'.'entities' ;
CREATE TABLE IF NOT EXISTS
      'owl2db'.'entities' (
  'Id' INT(11) NOT NULL
      AUTO_INCREMENT ,
  'Prefix_fk' INT(11) NOT NULL,
  'Name' CHAR (50),
  'Is_a_Individual' Boolean NOT
      NULL,
  'Is_a_Class' Boolean NOT NULL,
  'Is_a_Object_Property' Boolean NOT
      NULL,
  'Is_a_Data_Property' Boolean NOT
      NULL,
  'Is_a_Data_type' Boolean NOT NULL,
  'Is_data' Boolean NOT NULL,
  'Annotation_fk' INT(11) NOT NULL,
  PRIMARY KEY ('Id'),
  CONSTRAINT FOREIGN KEY
      ('Prefix_fk') REFERENCES
      'prefixes' ('Id') ON DELETE
      CASCADE ON UPDATE CASCADE,
  CONSTRAINT FOREIGN KEY
      ('Annotation_fk') REFERENCES
      'annotations' ('Id') ON DELETE
      CASCADE ON UPDATE CASCADE
)
ENGINE = InnoDB
  AUTO_INCREMENT = 1;
```

V CONCLUSION

This paper described a novel approach to storing specifically ontologies based on OWL 2 in a simple and directly accessible database. Since the structure was based on the OWL 2 language, its capabilities for defining and describing ontology knowledge must be comparable. However, there are some potential downsides resulted from using such a database structure. The main entity table can become very large in size. This can slow down access to the ontology knowledge since every other quarry is using this table to determine the name and characteristics of an entity based on its identifier. This is amplified by storing not only named, but also unnamed entities in this table. Every complex class which does not necessarily have been given a name must still have an entity object in order to define every attribute of the complex class. Existing datatypes like "xsd:integer" also must have entity object in order to maintain the integrity and consistency of the database structure. All these factors contribute to a very large list of entities.

The proposed database does not guarantee or verify the reasonability of the knowledge described in it. There is no mechanism to prevent any entity from being a class, individual and property and other at the same time. It is the ontology expert's responsibility for the ontology knowledge to make sense. However, in some cases it can be of use to have elastic and multi-purpose element of the ontology. A concept being several things at the same time is not automatically a logical fallacy, as long as the resulting ontology is usable for its stated purpose. Another downside to storing ontology knowledge in a database is the lack of a reasoning mechanics. Any such functionality must be provided by an additional software solution connecting to the database, obtaining its contents and adding any new conclusions to the database. However, some basic reasoning functionality like an automatic classification of entities can be added directly to the database using triggers or scripting languages. This is dependent on the database management software capability.

At this point it is unclear how this database will or should handle imports from other ontologies. The prefix table provides some functionality to describe the origin of an entity. If an entity has a prefix from an outside source, further quarrying to other knowledge sources may be required.

Because of the large number of tables, obtaining all information can be difficult. At this point, knowledge obtaining must be performed using a dialog approach, which includes obtaining information piece by piece. This has positive as well as negative aspects to it. On the one hand, this allows the user of the database to obtain only those pieces of information which are important. By having a dialog, the user has a choice at every step. On the other hand, as a negative aspect, this means that in cases where all the related information has to be obtained, the process is slowed down immensely. As future work, an approach to obtaining the information more easily and preferably handled on the server side can be researched as well as an implementation of a quarrying language like SPARQL.

VI REFERENCES

[1] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider and S. Rudolph, *OWL 2 web ontology language primer. W3C recommendation*. 2009 [Online]. Available: http://www.w3.org/TR/owl2-primer/ [Accessed March 15, 2015]

[2] T. Redmond, "An Open Source Database Backend for the OWL API and Protege 4". In *OWLED*, Vol. 614, 2010.

[3] M. Horridge and S. Bechhofer, "The owl api: A java api for owl ontologies". *Semantic Web*, 2(1), 11-21, 2011.

[4] P. Cudré-Mauroux, I. Enchev, S. Fundatureanu, P. Groth, A. Haque, A. Harth and M. Wylot, "NoSQL databases for RDF: an empirical evaluation", in *The Semantic Web–ISWC* , Springer Berlin Heidelberg, 2013, pp. 310-325.

[5] M. Voigt, A. Mitschick and J. Schulz, "Yet Another Triple Store Benchmark? Practical Experiences with Real-World Data", in SDA 2012, September, pp. 85-94.

[6] J. Henss, J.Kleb, S. Grimm and J.Bock, *A Database Backend for OWL. 5th Int. Workshop on OWL: Experiences and Directions (OWLED 2009)*, Chantilly, Virginia, USA,October 2009.

[7] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y.Chen and D. Wilkins, (2010, April). *A comparison of a graph database and a relational database: a data provenance perspective. In Proceedings of the 48th Annual Southeast Regional Conference* , p. 42. ACM.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 76-81

The Influence of Hidden Neurons Factor on Neural Network Training Quality Assurance

Peter Grabusts, Aleksejs Zorins
Rezekne Higher Educational Institution, Faculty of Engineering,
Atbrivosanas alley 115, Rezekne, LV-4601, Latvia

Abstract. **The work shows the role of hidden neurons in the multilayer feed-forward neural networks. The numeric expression of hidden neurons is usually determined in each case empirically. The methodology for determining the number of hidden neurons are described. The neural network based approach is analyzed using a multilayer feed-forward network with backpropagation learning algorithm. We have presented neural network implementation possibility in bankruptcy prediction (the experiments have been performed in the Matlab environment). On the base of bankruptcy data analysis the effect of hidden neurons to specific neural network training quality is shown. The conformity of theoretical hidden neurons to practical solutions was carried out.**

Keywords: **bankruptcy prediction, financial ratio, hidden neurons, neural networks, backpropagation.**

I INTRODUCTION

In literature on artificial neural networks there is often found the concept - "hidden neuron" or "hidden layer". It is generally said that for the network training with complex input data the hidden neurons are required. Indeed, the role of hidden neurons is often not clear. Also, in special literature, their role and functions are described minimally. For example, such types of definitions are given: "Between these layers there is one or more interlayers (hidden layers), which perform the knowledge accumulator functions" [2].

If formally the neural network is treated as consisting of the input layer and output layer, then in the hidden layer a particular training algorithm operation is actually realized.

The work gives an attempt to explain the role of hidden neurons, explain the situation when there is a need for the hidden neurons, the methods are described that allow to select the optimal number of hidden neurons, and also to know whether dealing with specific input data - in the training process the hidden neurons will be needed or not. In practical part on the basis of bankruptcy data example the selection of the number of hidden neurons for training quality assurance is described experimentally.

II THE EMPIRICAL DETERMINATION OF THE NUMBER OF HIDDEN NEURONS

Neural network, in general, has the following structure:

- The network consists of many interconnected neurons;
- Neurons are arranged in levels or layers;
- The first layer introduces the data into the neural network;
- The output determines the response of the network to the input data;
- Between the first and the last layer there may be one or more hidden layers which performance the accumulator functions in the network.

In the work [13] the author J. Kinser describes the methodology that seems interesting as it offers to determine the need for the hidden neurons before training by analyzing the input data. It is called *hidden neuron* theory and describes the procedure which extracts all the necessary information to determine the network architecture before the training process from the input data. The number of possible hidden neurons, its specific role and position for all training sets of data pairs can be determined. The method allows finding network architecture with a minimal number of hidden neurons.

The idea is as follows. The number of hidden neurons depends directly on the internal structure of training data and its essence is to resolve conflicts in the input data - if the input data contains conflicts, then the neural network needs hidden neurons.

It can be explained with XOR example. In XOR problem there are 4 training elements. As the activation function the threshold function is used:

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.213

$$y(a) = \begin{cases} 1, & a > \gamma \\ 0, & otherwise \end{cases} \quad (1)$$

where $0 < \gamma < 0$.

For training data pairs (x:y) the following inequalities must be:

$$\sum_i w_{ji} x_i \begin{cases} > \gamma, & y = 1 \\ < \gamma, & y = 0 \end{cases} \quad (2)$$

Thus, XOR problem is converted with a number of inequalities:

$$\begin{gathered} 0 + 0 < \gamma \\ 0 + w_{12} \geq \gamma \\ w_{11} + 0 \gg \gamma \\ w_{11} + w_{12} < \gamma \end{gathered} \quad (3)$$

It is clearly seen that, if $W_{12} \geq \gamma$ and $W_{11} \geq \gamma$ is fulfilled, then inequality $W_{11} + W_{12} < \gamma$ can not be fulfilled. It is said that inequalities are in a conflict. The solution of the conflict is introduction of hidden neurons, as it is shown in Figure 1.

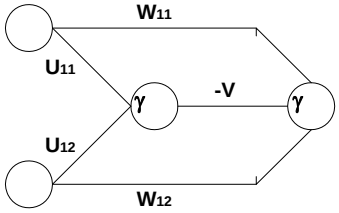

Fig.1. Example solution of XOR problem

The corresponding inequality systems will be as follows:

$$\begin{gathered} 0 + 0 < \gamma \\ 0 + U_{12} < \gamma \\ U_{11} + 0 < \gamma \\ U_{11} + U_{12} \geq \gamma \end{gathered} \quad => \quad \begin{gathered} 0 + 0 + 0 < \gamma \\ 0 + W_{12} + 0 \geq \gamma \\ W_{11} + 0 + 0 \gg \gamma \\ W_{11} + W_{12} + V < \gamma \end{gathered} \quad (4)$$

This means that the conflict has been resolved through the introduction of additional neurons. Searching the conflicts in large training sets the weak point of this method, unfortunately, is in solving the inequality systems.

III PRACTICAL RECOMMENDATIONS FOR THE SELECTION OF THE NUMBER OF HIDDEN NEURONS

Different authors have given different evaluations of the hidden neurons. The most common evaluation is as follows: if there are N input elements, then the total number of input vectors is 2^N. The maximum number of hidden neurons is 2^{N-1}.

The authors of this work [9] offer the following evaluation (known as Baum – Haussler rule):

$$N_{hidden} \leq \frac{N_{train} E_{tolerance}}{N_{pts} + N_{output}} \quad (5)$$

where:

N_{hidden} – number of hidden neurons;

N_{train} – number of training elements;

N_{pts} – number of learning elements;

N_{output} – number of output neurons;

$E_{tolerance}$ – error.

Several authors, among them [1], [11], give the following, seemingly simple, practical recommendations for the feed-forward network implementation:

- Use one hidden layer;
- Use fewer hidden neurons, if possible.

How many hidden layers?

There is no reason to use more than one hidden layer – otherwise the network training can take a long time. Of course, there is always a possibility that several layers perfectly solve the complex problem. Therefore, in the feed-forward model practical realization for the first time it is recommended to use only one hidden layer. If a large number of hidden neurons in the first layer do not offer a good solution to the problem, it is worth trying to use a second hidden layer, reducing the total number of hidden neurons.

How many hidden neurons?

The need to choose the right number of hidden neurons is essential. Using too little – the network lacks the resources for training algorithm implementation. Using too much – increases the training time, even it can be impossible to be adequately trained within a reasonable period of time. Therefore it is necessary to use an absolute minimum of hidden neurons that solves the problem.

For most of the tasks the *geometric pyramid rule* can be used. It states that the number of neurons should be reduced from the input to the output, as it is seen in Figure 2.

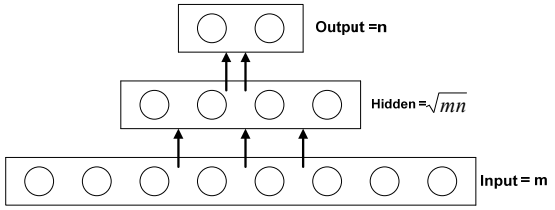

Fig.2. Typical three-layer network 8-4-2

Similarly, the four-layer network could be set up, such as 8-4-2-1. The number of hidden neurons in layers HID_1 and HID_2 should be selected according to the following relationship:

$$HID_1 = nr^2 \quad (6)$$

$$HID_2 = nr \quad (7)$$

where $r = \sqrt[3]{\frac{m}{n}}$.

The above written formulas give only a rough evaluation. In practice, for each specific task the required number of hidden neurons is determined experimentally.

IV EXPERIMENTAL PART – ANALYSING BANKRUPTCY DATA

A. Bankruptcy financial ratios

Bankruptcy diagnostics is a directed financial analysis system which area is a crisis situation control at the enterprises. For this purpose, in the analysis of the general financial situation of the company a separate group of financial ratios is made, using which it is possible to reason about the threat of bankruptcy. In general case there is no theoretical background as to which financial ratios might be used in different bankruptcy studying models. A lot of researchers performing an analysis of the bankrupt enterprise proceed in this way: they calculate several ratios and then select potentially most significant of them.

Beaver [10], Altman [5], and Ohlson [14] are considered the pioneers of the empirical approach, Beaver was one of the first who has applied balance sheet data in bankruptcy research. His analysis was comparatively simple and was based on studying one financial ratio and comparing it with other ratios. He has concluded that ratio R4: Cash flow / Total debt is a very essential indicator which has to be accounted in bankruptcy analysis. Beaver's works became a beginning of multicriteria analysis application, which was later developed by Altman et al.

Altman employed classical multivariate discriminant analysis (MDA) in his research, Altman's Z-scores [6] are widely used as input data in neural network algorithms. It was noted that Z-score has an essential drawback – actually it can only be related to large companies whose stocks are quoted in the stock exchange.

B. Neural network approaches

Neural network approach to bankruptcy prediction [7], [8] became extremely popular in the nineties of the last century. Starting from 1990 research on neural network application in bankrupt analysis was performed periodically: Odom and Sharda [4], Tam and Kiang [17] et al.

Odom and Sharda were one of those who first employed neural network techniques in bankruptcy analysis. In the input of the network, Altman's Z-scores about 128 companies were used. It was shown that neural network approach yields better results than MDA.

Tam and Kiang have compared different techniques applied in bankruptcy diagnostics (MDA, single layer network and multilayer network) and have shown that in the „one-year-ahead" data the multilayer network was most effective whereas in the "two-year-ahead" data the other method turned to be most effective.

Analysing the neural network application in bankruptcy analysis, these conclusions were made:

- Neural networks ensure approximately 90% accuracy compared to the 80-85% accuracy of other methods;
- Bankruptcy can be predicted several years before it happens, the accuracy of prediction being practically the same for the "one-year-ahead" data and for the "two-year-ahead".

Formally speaking, bankruptcy prediction task practically does not differ from that of classical pattern recognition. Financial ratios are forwarded to the network input and network learning is accomplished. Use of hidden neurons is made. In the network output there are only two states: bankrupt or non-bankrupt. General scheme of the network that will be used in the experiments presented here, is shown in Figure 3.

For a detailed description of this well known type of network see [3] and [16].

C. Experimental data

The data on firm bankruptcy were taken from [15], [18]. For the purpose of experiments, balance sheet data of 63 companies were used - 46 bankruptcy and 17 not bankruptcy (for other records from [15] information was incomplete for calculating the necessary financial ratios). It was decided to calculate the following financial ratios on the basis of the data available and further use them in all the experiments:

- R2: Cash Flow / Current Liabilities;
- R3: Cash Flow / Total Assets;
- R7: Current Assets / Current Liabilities;
- R9: Current Assets / Total Assets;
- R31: Working capital / Total assets.

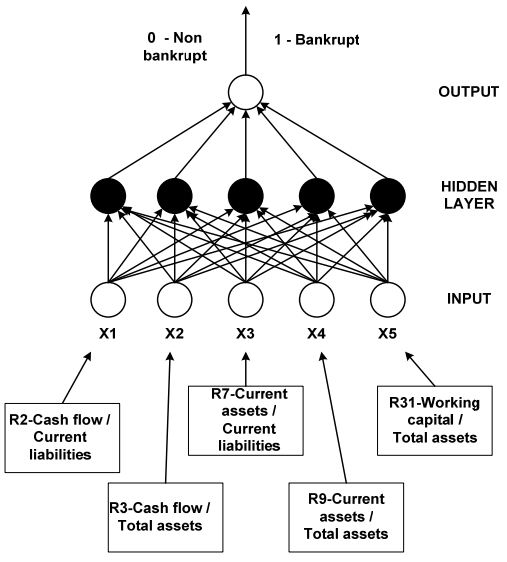

Fig.3. Multilayer feed-forward neural network architecture

The multilayer feed-forward neural network has been chosen as a neural network model. A series of experiments have been performed in order to choose the number of hidden neurons suitable for the purpose of learning. From now on, all the experiments are based on the 5-5-1 model. The parameters of the chosen network model were as follows:

- Input nodes - 5 neurons;
- Hidden layer - 5 neurons;
- Output nodes - 1 neuron (1- bankrupt, 0 - not bankrupt);
- Learning rate - 0,25;
- Stopping condition - the training is stopped if MSE=0,5;
- Momentum rate - α;
- Slope of the tanh activation function - β.

The input nodes represent the financial ratios that the model will use to predict bankruptcy. The learning and momentum rates define the rate and accuracy with which the backpropagation algorithm converges on the error minimization solution. Larger numbers for momentum are used to achieve faster convergence, but run the risk of skipping over the optimal solution. The final parameters were obtained after extensive experiments.

In the first part of the experiment, the learning quality dependence on α value was checked. The given experimental model has shown that at the α values within 0,8 the following occurs: the algorithm does not converge or else correct network learning does not happen. Valid learning results were obtained at α=0,8 and α=0,9 [12].

In the second part of the experiment the learning quality dependence on the slope of the *tanh* activation function - $\beta \in [0,1; 1]$ was checked. As a result of experiments, it was found that acceptable learning quality was achieved at specific β values only. Table 1 shows the most significant results and misclassified cases [12].

TABLE 1.

EXPERIMENTAL RESULTS (PARAMETERS α, β AND ITS MISCLASSIFIED CASES)

PARAMETER α	PARAMETER β	EPOCHS	NO OF MISCLASSIFIED CASES	CASES
α = 0,8	β = 0,8	41	9	14,26,36,37,41,58,59, 60,62
	β = 0,9	889	6	37,50,58,59,60,62
α = 0,9	β = 0,8	46	9	14,26,35,36,37,41,58, 59,62
	β = 0,9	1489	7	37,50,58,59,60,62,63

In Figure 4 error graph for α=0,8 and β = 0,9 are shown.

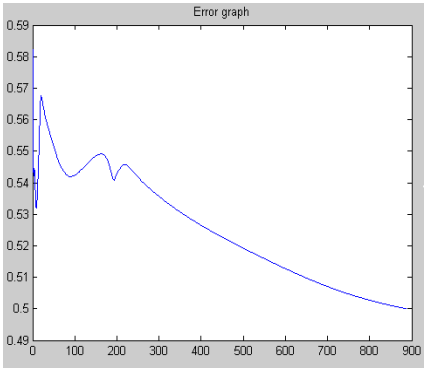

Fig.4. Error graph for α=0,8 and β = 0,9

All the results on the misclassified cases of the methods employed are summarised in Table 2.

From the data of Table 2 it can be seen that for the specific bankruptcy data sample, all the methods are unable to classify data vectors 37, 38, 59 and 62. Calculating in absolute numbers, we obtain that NN-2 correctly classified 88,9% cases, the potential function method and NN-4 - 88% cases, NN-1 and NN-3 - 85,7%. It can be concluded that for the given data sample the neural network method performs bankruptcy data classification more effectively, which actually corresponds to the conclusions about the results achieved by Tam and Kiang [17].

TABLE 2.

SUMMARY TABLE ABOUT USED METHODS AND MISCLASSIFIED CASES

METHOD	NN-1	NN-2	NN-3	NN-4
PARAMETERS	α = 0,8 β = 0,8	α = 0,8 β = 0,9	α = 0,9 β = 0,8	α = 0,9 β = 0,9
MISCLASSIFIED CASES	14		14	
	26		26	
			35	
	36		36	
	37	37	37	37
	41		42	
		50		50
	58	58	58	58
	59	59	59	59
	60	60		60
	62	62	62	62
				63

D. The effect of hidden neurons on experimental data

Since the network best option was with α=0,8 and β=0,9, an experimental research was carried out – how the selecting the number of hidden neurons affects the overall quality of the network and whether it corresponds to theoretical principles of selecting the number of hidden neurons. In Table 3 the performance

of neuron network at different hidden neurons values is shown (see Table 3 and Figure 5).

TABLE 3.

NUMBER OF HIDDEN NEURONS AND QUALITY EVALUATION

METHOD	NN-1	NN-2	NN-3	NN-4	HIDDEN NEURONS
PARAMETERS	α = 0,8 β = 0,8	α = 0,8 β = 0,9	α = 0,9 β = 0,8	α = 0,9 β = 0,9	
QUALITY (%)	88,9	88,9	87,3	90,5	1
	88,9	88,9	87,3	88,9	2
	88,9	88,9	87,3	88,9	3
	85,7	88,9	84,1	88,9	4
	82,5	88,9	84,1	88,9	5
	82,5	88,9	84,1	88,9	6
	85,7	88,9	82,5	88,9	7
	82,5	88,9	85,7	88,9	8
	82,5	88,9	85,7	88,9	9
	85,7	88,9	84,1	88,9	10
	87,3	88,9	87,3	88,9	11
	87,3	88,9	82,5	88,9	12

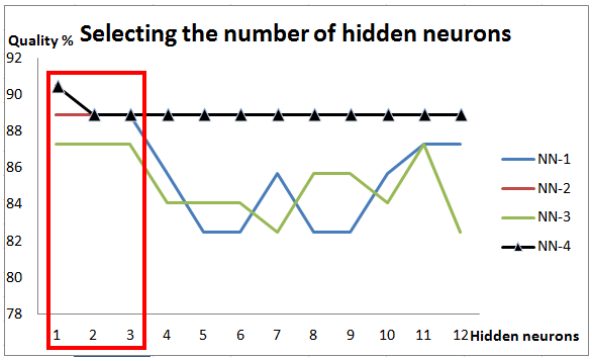

Fig.5. Selecting the number of hidden neurons for bankruptcy data analysis

It can be concluded, that for high-quality network training in this particular case it is enough to use as minimum 1 neuron in the hidden layer, in order to ensure an acceptable quality of training that also corresponds to the theoretical evaluation: according to the formula (6) we have:

$$HID_1 = nr^2 = \left(\sqrt[3]{\frac{m}{n}}\right)^2 = 1{,}7^2 = 2{,}9 \approx 3$$

that is maximum 3 neurons in the hidden layer, which correspond to the theoretical evaluation. Further increase of the hidden neurons, as it is shown in Table 3, is not needed.

V CONCLUSIONS

Working with feed-forward network architecture models and using the back propagation learning algorithm, we face the problem to determine the number of hidden neurons. This paper presents some methods of determining number of hidden neurons. The last chapter gives practical recommendations for testing specific data samples which analyze neural network input data and test the situations where hidden neurons are necessary. This could make it possible to identify the number of hidden neurons before the network training to build up the optimal network. The method is suitable only for feed-forward neural network model. In the experimental part, the application of bankruptcy data leads to the conclusion that using 3 neurons in the hidden layer in a specific case is entirely sufficient for providing the neural network training quality. It allows saving time in network training and testing.

VI REFERENCES

[1] T. Masters, *Practical Neural Network recipes in C++*. Academic Press, 1993.
[2] I. Alexander and H. Morton, *An Introduction to Neural Computing*. Chapman & Hall, London, 1991.
[3] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the theory of neural computation*. Addison Wesley, 1991.
[4] M. Odom and R. Sharda R, A neural network model for bankruptcy prediction, In Proc, Int. Joint Conf, Neural Networks, San Diego, CA, 1990.
[5] E. Altman E, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," Journal of Finance, vol. 13, pp.589-609, 1968.
[6] E. Altman, R. Haldeman and P. Narayanan, "ZETA analysis. A new model to identify bankruptcy risk of corporations," Journal of Banking and Finance 1, pp. 29-54, 1977.
[7] A. Atiya A, "Bankruptcy prediction for credit risk using neural networks: A survey and new results, " IEEE Transactions on Neural Networks, Vol. 12, No. 4, pp. 929-935, 2001.
[8] B. Back, T. Laitinen and K. Sere, "Neural networks and bankruptcy prediction: funds flow, accrual ratios and accounting data, " Advances in Accounting 14, pp. 23-37, 1996.
[9] E.B. Baum and D. Haussler, "What size net gives valid generalization," Neural Computation, 1, pp. 151-160, 1988.
[10] R. Beaver, "Financial ratios as predictors of failure. Empirical Research in Accounting: Selected Studies," Accounting Research, vol. 4, pp.71-111, 1966.
[11] K. Gnana Sheela and S.N. Deepa, "Review on Methods to Fix Number of Hidden Neurons in Neural Networks," Mathematical Problems in Engineering, Volume 2013.
[12] P.Grabusts, "Analysing Bankruptcy Data with Neural Networks," 10th International Conference on Soft Computing, Mendel2004, Brno, Czech Republic, June 16.-18., pp.111-117, 2004.
[13] J.M. Kinser J.M., „The determination of Hidden Neurons," Optical Memories and Neural Networks, 5(4), 245-262, 1996.
[14] J. Ohlson J, "Financial ratios and the probabilistic prediction of bankruptcy," Accounting Res, vol. 18, pp. 109-131, 1980.
[15] G. Rudorfer G, "Early bankruptcy detecting using neural networks," APL Quote Quad, ACM New York, vol. 25, N. 4, pp. 171-178, 1995.

[16] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning representations by backpropagating errors," Nature, 323(9):533-536, 1986.
[17] K. Tam and M. Kiang, "Managerial applications of the neural networks: The case of bank failure predictions," Management Science, vol. 38, pp. 416-430, 1992.
"Bankruptcy data," [Online]. Available: http://godefroy.sdf-eu.org/apl95/ratios95.zip [Accessed: March 30, 2015].

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 82-86

On one Mathematical Model for Dynamics of Propagation and Retention of Heat over New Fibre Insulation Coating

Armands Grickus[1, a], Sharif E. Guseynov[1, 2, b]
[1]Institute of Mathematical Sciences and Information Technologies, Liepaja University, 14 Liela Street, Liepaja LV-3401, Latvia; [2]Institute of Aeronautics, Faculty of Transport and Mechanical Engineering, Riga Technical University, 1A Lomonosov Street, Building-1, Riga LV-1019, Latvia [a]Armands.Grickus@liepu.lv, [b]Sh.E.Guseinov@inbox.lv

Abstract. **In circumstances, when it is important to replace insulation materials with high content of emissions during production it is necessary to create new heat and sound insulation material, which eliminates CO_2 emissions, develop its production techniques and technological machinery – raw material chopper, pulp mixer, termopress, dryer chamber, formatting knifes, determine technical control parameters and control equipment, develop mathematical model of the material and calculation methods for design works. It is necessary to design, manufacture and experimentally test the respective technological equipment for insulation production pilot plant. To get exact physical parameters it is necessary design, manufacture and test unique laboratory equipment for determining the properties of insulation material. The mathematical model describing the dynamics of propagation and retention of heat over fibre insulation coating by taking "inner" specificities (graininess and porosity of layered structure of the considered fibre insulation) of heat insulator into account is proposed in the present paper.**

Keywords: **Insulation material, mathematical model, thermoelastic deformation and thermal movement, temperature distribution.**

I INTRODUCTION

As it is known (see, for instance, [1], [2] and respective references given in these), in the capacity of fibre material could be used wide range of materials, both organic and inorganic, for instance, cellulose, fabrics, wool, cotton, glass, rockwool, basalt fibre, etc. As the insulator could be used chopped polystyrene, polyurethane, cork, peat, bark, etc.

Fig. 1. Foamed polystyrene particles are bonded with cellulose fibers.

Binding together of the insulation particles, forming self-supporting layer of insulation material, useful both for thermal and for acoustic insulation (see, for instance, [3]). Remarkable positive property among others is ability of the material to accept and release water vapour – "breathe" like most of the natural materials. Other – it is stable against setting – opposite to pure cellulose wool insulation.

New insulation has been developed by the Liepaja University scientists. It is based on mix of insulation material particles enclosed in fibrous mass, having insulation properties (as it contains trapped air micro pockets) in the same time it.

The brief discussion on heat retention and energy conservation of the fibre insulation coating: fibre insulation coating can provide energy savings of 20-40% depending on ambient temperature, contents, weather conditions and application thickness. According to test data and results from applications (see, for instance, [4]-[8] and respective references given in these) efficiency is higher in conditions with exposure to convection-based cooling compared to uninsulated surfaces. Efficiency in convection-based cooling conditions is roughly comparable to conventional natural materials, but the difference is

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.504

that coating-based insulated equipment will experience a slow heat drop and faster recovery compared to a slow heat drop and slow recovery in a conventionally insulated structure.

Fibre insulation coating relies primarily on radiant heat blocking while conventional material slows down heat transfer with dense mass. The coating's high saturation of insulation particles provides very low emissivity and transmittance characteristics that block radiative heat transfer. While there is some conductive heat transfer blocking, that's a smaller contribution compared to radiative blocking. A good analogy for low emissivity is "Low E" window coatings that also work by blocking solar radiation wavelengths' heat transfer. These two heat transfer characteristics are very important in understanding how the thin coating blocks transfer of radiant heat energy.

Conventional insulation material's heat transfer rating test is conducted at 0% humidity and 20°C. However, in the real world material becomes saturated with moisture at least up to the volume of relative humidity of surrounding air. Not a problem in very dry conditions, however, in most of the world that translates to 40-60% relative humidity and simply a matter of time before almost all mass-based insulations are saturated to the local relative humidity level. At only 30% moisture content, chopped polystyrene's heat conductivity is reduced almost to that of water at 60°C and very close to window glass at 80°C (see, for instance, [6], [9]). The advantage of coating is that it will not accept moisture and retains low heat transfer characteristics indefinitely.

And just as conventional insulation has a point of diminishing return when it comes to thickness; fibre insulation coating's effectiveness also becomes marginal beyond certain thicknesses at varying temperatures. Because radiant barriers react with thermal energy, they're useful when you have a heat source and do not store the heat energy to maintain a thermal battery as do mass-based conductive materials. This means that fibre insulation coating is very effective on heated equipment and much less so on unheated equipment applications such as preventing cold water lines from freezing. Insulation coating is effective in cold spaces with condensing surfaces based on it's low transmittance characteristics that essentially "take the chill off" and raise them above the dew point ([6], [7], [9]). Overall, insulation coating has many proven application in the field and can also be combined with conventional materials for solutions that leverage the best qualities of both.

II MATHEMATICAL MODEL FOR DYNAMICS OF PROPAGATION AND RETENTION OF HEAT OVER FIBRE INSULATION COATING

In this section we propose the mathematical model describing the dynamics of propagation and retention of heat over fibre insulation coating by taking "inner" specificities (graininess and porosity of layered structure of the considered fibre insulation; see, for instance, [5], [8], [10] as well as [11]) of heat insulator into account. It should be noted that the proposed model has its limitations: it describes only "internal" physical processes includes:

- heat propagation in the insulation material;
- mechanical process, related to tensions in material structure and differences in elasticity of said material under the influence of uneven heat spreading in the insulation material, which has been regarded as non-homogeneous layered structure.

Thus, the proposed mathematical model has the following statement:

1. Four equations concerning sought-for functions $T = T(x, y, z; t)$ and $u(x, y)$

$$\rho \cdot c \cdot \frac{\partial T}{\partial t} + \nabla_{x,y,z} \left\{ k \cdot \nabla_{x,y,z} T \right\} = f, \quad (x, y, z; t) \in \operatorname{int} D \times (0, t_{end}]; \tag{1}$$

$$2 \cdot \frac{\partial^2 u_{xx}}{\partial x^2} + \frac{\partial^2 u_{xx}}{\partial y^2} + \frac{\partial^2 u_{yy}}{\partial x^2} + \frac{\partial^2 u_{yy}}{\partial x^2} + E_l \cdot \alpha_l \cdot \left\{ \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} - \frac{\partial^2 T_0}{\partial x^2} - \frac{\partial^2 T_0}{\partial y^2} \right\} = 0, \quad (x, y, z; t) \in \operatorname{int} D \times (0, t_{end}]; \tag{2}$$

$$\frac{\partial u_{xx}}{\partial x} + \frac{\partial u_{xy}}{\partial y} = 0, \ (x, y, z) \in \operatorname{int} D; \tag{3}$$

$$\frac{\partial u_{xy}}{\partial x} + \frac{\partial u_{yy}}{\partial y} = 0, \ (x, y, z) \in \operatorname{int} D; \tag{4}$$

2. Initial condition

$$T\big|_{t=0+0} = T_0(x, y, z), \ (x, y, z) \in D; \tag{5}$$

3. Eighteen boundary conditions concerning both the thermal field $T = T(x, y, z; t)$ (six boundary conditions) and the thermoelasticity $u(x, y)$ (twelve boundary conditions)

$$T\big|_{x=0+0} = T_{0x}(y, z; t), \quad (y, z; t) \in \left(D \Big/ [0, L_x] \right) \times [0, t_{end}]; \tag{6}$$

$$T\big|_{x=L_x-0} = T_{Lx}(y, z; t), \quad (y, z; t) \in \left(D \Big/ [0, L_x] \right) \times [0, t_{end}]; \tag{7}$$

$$T\big|_{y=0+0} = T_{0y}(x, z; t), \quad (x, z; t) \in \left(D \Big/ [0, L_y] \right) \times [0, t_{end}]; \tag{8}$$

$$T\big|_{y=L_y-0} = T_{Ly}(x, z; t), \quad (x, z; t) \in \left(D \Big/ [0, L_y] \right) \times [0, t_{end}]; \tag{9}$$

$$T\big|_{z=0+0} = T_{0z}(x,y;t), \quad (x,y;t) \in \left(D/[0,L_z]\right)\times[0,t_{end}]; \tag{10}$$

$$\left.\frac{\partial T}{\partial z}\right|_{z=L_z-0} = T_{Lz}(x,y;t), \quad (x,y;t) \in \left(D/[0,L_z]\right)\times[0,t_{end}]; \tag{11}$$

$$u_{xx}\big|_{x=0+0} = u_{xx}^{0x}(y),\ y \in [0,L_y]; \tag{12}$$

$$u_{xx}\big|_{x=L_x-0} = u_{xx}^{Lx}(y),\ y \in [0,L_y];$$

$$u_{xx}\big|_{y=0+0} = u_{xx}^{0y}(x),\ x \in [0,L_x]; \tag{13}$$

$$u_{xx}\big|_{y=L_y-0} = u_{xx}^{Ly}(x),\ x \in [0,L_x]; \tag{14}$$

$$u_{xy}\big|_{x=0+0} = u_{xy}^{0x}(y),\ y \in [0,L_y]; \tag{15}$$

$$u_{xy}\big|_{x=L_x-0} = u_{xy}^{Lx}(y),\ y \in [0,L_y]; \tag{16}$$

$$u_{xy}\big|_{y=0+0} = u_{xy}^{0y}(x),\ x \in [0,L_x]; \tag{17}$$

$$u_{xy}\big|_{y=L_y-0} = u_{xy}^{Ly}(x),\ x \in [0,L_x]; \tag{18}$$

$$u_{yy}\big|_{x=0+0} = u_{yy}^{0x}(y),\ y \in [0,L_y]; \tag{19}$$

$$u_{yy}\big|_{x=L_x-0} = u_{yy}^{Lx}(y),\ y \in [0,L_y]; \tag{20}$$

$$u_{yy}\big|_{y=0+0} = u_{yy}^{0y}(x),\ x \in [0,L_x]; \tag{21}$$

$$u_{yy}\big|_{y=L_y-0} = u_{yy}^{Ly}(x),\ x \in [0,L_x]. \tag{22}$$

In the proposed model (1)-(22) there are the following notations and assumptions:

- $D \overset{def}{\equiv} \left\{(x,y,z): x \in [0,L_x], y \in [0,L_y], z \in [0,L_z]\right\}$ is geometric configuration (the closed 3D domain) of the insulation material that has a rectangular shape with the thickness / depth L_z;
- L_x and L_y are the length and the width of the rectangular insulation material, respectively;
- $\mathrm{int}\,D$ is an open domain that signifies interior of the domain D: $\mathrm{int}\,D \overset{def}{\equiv} D/\partial D$, ∂D contains the frontier points of the D;
- t_{end} is the time within a period of which we investigate the thermal processes occurring interior of the insulation material;
- the sought-for function $T = T(x,y,z;t)$ is the temperature (or rather the thermal field) in the considering point (x,y,z) of the insulation material at the time moment t;
- the prescribed function $k = k(x,y,z) > 0$ is the heat conductivity coefficient in the considering point (x,y,z) of the insulation material;
- the prescribed function $\rho = \rho(x,y,z) > 0$ is the density of the insulation material in the considering point (x,y,z) of the insulation material;
- the prescribed function $c = c(x,y,z) > 0$ is the specific thermal capacity of the insulation material in the considering point (x,y,z) of the insulation material;
- the prescribed function $f = f(x,y,z;t)$ is the power density of external heat sources applied to the considering point (x,y,z) of the heat insulator at the time moment t;
- the sought-for functions $u_{xx} = u_{xx}(x,y)$, $u_{xy} = u_{xy}(x,y)$ and $u_{yy} = u_{yy}(x,y)$ are the components of mechanical stress and thermoelasticity $u(x,y)$ under assumption of ignoring any changes / influences in the direction of the axis OZ;
- $\nabla_{x,y,z} T \overset{def}{\equiv} \frac{\partial T}{\partial x}\cdot\vec{e}_1 + \frac{\partial T}{\partial y}\cdot\vec{e}_2 + \frac{\partial T}{\partial z}\cdot\vec{e}_3$ is the gradient of thermal-vector field, where $\vec{e}_i$ $\left(i = \overline{1,3}\right)$ are unit vectors located on the coordinate axes (OX, OY, OZ), respectively;
- the prescribed function $\alpha_l = \alpha_l(x,y,z) \overset{def}{\equiv} \frac{\Delta l}{l\cdot\Delta T}$ is the linear thermal expansion coefficient (see, for instance, [12]);
- the prescribed function $E_l = E_l(x,y,z) \overset{def}{\equiv} \frac{F\cdot l}{S\cdot\Delta l}$ is the modulus of elongation (so-called "Young modulus"; see, for instance, [13], [14]), which characterizes the deformation taking place in the considering point (x,y,z) of the heat insulator surface S under the impact of temperature voltage both sides relative to the heat insulator (i.e. from the outside and on the inside of premises): this Young modulus characterizes also the properties of the heat insulator to make resistance to tension at the elastic deformation under the impact of the temperature voltage;
- the prescribed function $T_0 = T(x,y,z;t)\big|_{t=0+0}$ is the initial temperature in the considering point (x,y,z) of the insulation material;

- the boundary functions $T_{0x}(y,z;t)$, $T_{Lx}(y,z;t)$, $T_{0y}(x,z;t)$, $T_{Ly}(x,z;t)$, $T_{0z}(x,y;t)$, $T_{Lz}(x,y;t)$, $u_{xx}^{0x}(y)$, $u_{xx}^{Lx}(y)$, $u_{xy}^{0y}(x)$, $u_{xy}^{Ly}(x)$, $u_{yy}^{0y}(x)$, $u_{yy}^{Ly}(x)$, $u_{xy}^{0x}(y)$, $u_{xy}^{Lx}(y)$, $u_{yy}^{0x}(y)$ and $u_{yy}^{Lx}(y)$ are assumed as known functions in their applicable / definitional domains.

In order to make sure that the system of four equations (1)-(4), which describes the interrelated processes generating temperature field $T(x,y,z;t)$ and thermoelasticity $u(x,y)$, had a physical determinacy (i.e. physical meaning), it is necessary to have some initial information of quantitative and qualitative patterns (see, for instance, [4], [5], [15]-[17] and respective references given in these). The initial condition (5) and the boundary conditions (6)-(22) form the required quantitative information. As regards the required information of qualitative pattern, its forming mostly depends on the chosen methods of analysis and solving the constructed model. Obviously, the less constrains are imposed on the model, the wider the range of application of this model becomes. Therefore, it makes sense to pose a question on finding the optimal set of constrains of qualitative pattern. However, in view of the fact that we cannot solve the proposed model (1)-(22) in present paper, the formulated below constrains of qualitative pattern are conditioned only by mathematical correctness of the equations (1)-(4):

- $T(x,y,z;t) \in C^{1,2}\left\{\text{int}\, D \cup \left[0, t_{end}\right]\right\}$;
- $u(x,y) \in C^{2}\left\{\text{int}\, D / \left(0, L_z\right)\right\}$;
- $T_0(x,y,z) \in C^{2}\left\{D\right\}$.

Thus, the proposed model (1)-(22) is the complete statement of the initial-boundary-value problem for investigation of the dynamics of propagation and retention of heat over fibre insulation coating by taking "inner" specificities of heat insulator into account. The analytical and / or numerical solution of the proposed model (1)-(22) will allow finding the sought-for functions $T(x,y,z;t)$ in the domain D and $u(x,y) \overset{def}{\equiv} \left(u_{xx}(x,y), u_{xy}(x,y), u_{yy}(x,y)\right)$ in the domain $D / \left[0, L_z\right]$, and consequently, in the time-interval $\left[0, t_{end}\right]$, during which the thermal processes occurring interior of the insulation material are investigated, we can completely determine the thermal field and the thermoelasticity of the considered insulation material having a "parallelepiped" shape with spatial measurements $L_x \times L_y \times L_z$.

III CONCLUSIONS

In this paper we formulated boundary conditions, among which only one is the second kind boundary condition (so-called the Neumann condition), while others are the Dirichlet boundary conditions. This is due to the practical point of view: as a rule an experimental method, which allows to realize boundary functions of the first kind, is an easier technique (however, such approach is not always expedient!). In addition, in present paper we formulated three conditions / constraints of qualitative pattern (one of many possible variants for constructing the qualitative constraints), the implementation of which together with the formulated boundary conditions and the initial condition unambiguously ensures mathematical correctness of the proposed model.

Qualitative analysis of the proposed model and / or its solving can favour for profound investigation and understanding of the thermal and thermomechanical processes occurring in the insulation materials, and thereby can improve the functionality and reliability of heat insulators.

To conclude with, let us note that authors of this paper are intended to continue the further investigation taking the benefit of both qualitative and quantitative studies for the proposed model (1)-(22) as well as to develop the stable analytical and numerical methods for their solution ensuring the corresponding computer-based implementation.

IV ACKNOWLEDGMENTS

For the second co-author the present article was executed within the framework of the following two Research Projects: 1) The State Research Programme "NextIT": *Next generation Information and Communication Technologies*; 2) The European Regional Development Fund Project No. 2014 / 0029 / 2DP / 2.1.1.1 / 14 / APIA / VIAA / 088: *"Development of an experimental long flight distance unmanned aerial vehicle prototype for multi-purpose environmental monitoring"*.

V REFERENCES

[1] S. R. Reid, and G. Zhou, *Impact Behaviour of Fibre-Reinforced Composite Materials and Structures*. Cambridge, UK: Woodhead Publishing Limited, 2000.

[2] K. L. Pickering. *Properties and Performance of Natural-Fibre Composites*. Cambridge, UK: Woodhead Publishing Limited, 2008.

[3] K. Ohno, K. Esfarjani, and Y. Kawazoe, *Computational Materials Science: From Ab Initio to Monte Carlo Methods*. Berlin: Springer-Verlag, 1999.

[4] R. W. Serth, *Process Heat Transfer. Principles and Application*. Oxford, UK: Elsevier Science & Technology Books, 2007.

[5] R. W. Serth, and Th. Lestina, *Process Heat Transfer: Principles, Applications and Rules of Thumb*. Academic Press, 2014.

[6] W. S. Janna, *Engineering Heat Transfer*. Boca Raton, Florida: CRC Press, 2000.

[7] J.E.Hesselgreaves, *Compact Heat Exchangers: Selection, Design, and Operation*. Oxford, UK: Pergamon Press, 2001.
[8] A. D. Kraus, A. Aziz, and J. Welty, *Extended Surface Heat Transfer*. New York: John Wiley & Sons, Inc., 2001.
[9] "District Heating and Cooling in the United States: Prospects and Issues." *Committee on District Heating and Cooling, National Research Council*, [Online] Available: http://www.nap.edu/catalog/263.html
[10] D. R. Gaskell, *An Introduction to Transport Phenomena in Materials Engineering*. New Jersey: Momentum Press, 2013.
[11] I. G. Staroverov, and Yu. I. Shiller, *Handbook of developer: Heating, Plumbing, Sewerage*. Moscow: Stroyizdat, 1990. (In Russian)
[12] V. Blazy, *Handbook of developer: Physics of Civil Engineering*. Moscow: Technosphera, 2005. (In Russian)
[13] L. D. Landau, and E. M. Lifshits, *Course of Theoretical Physics. Vol. 7: Theory of Elasticity*. New York: Pergamon Press, 2005.
[14] A. S. Nowick, and B. S. Berry, *Relaxation Phenomena in Crystals*. Moscow: Atomizdat, 1975. (In Russian)
[15] A. N. Tikhonov, and A. A. Samarsky, *Equations of Mathematical Physics*. Moscow: Lomonosov Moscow State University Press, 1999. (In Russian)
[16] L. Arkeryd, and C. Cercignani, "A global existence theorem for the initial-boundary-value problem for the Boltzmann equation when the boundaries are not isothermal." Archive for Rational Mechanics and Analysis, Vol. 125, Issue 3, pp. 271-287, 1993.
[17] C. M. Dafermos, and E. Feireisl, *Handbook of Differential Equations. Vol. 1: Evolutionary Equations*. Amsterdam: Elsevier B.V.2004.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 87-91

Modern approaches to reduce webpage load times

Aleksejs Grocevs, Nataly Prokofyeva
Riga Technical University, Faculty of Computer Science and Information Technology.
Address: 1/4 Meza str., Riga, LV-1048, Latvia

Abstract. **Nowadays, many modern websites offer a variety of information services, providing a dynamically generated content for the end user. However, the more users are trying to obtain such content, the slower becomes its loading time in their browsers. This article explains the mentioned problem in detail, as well as examines the roots of this problem.**

Keywords: **Loading time, performance, web site.**

I INTRODUCTION

Nowadays, applications deal with many different tasks. With the appearance of the Internet, it became possible to solve these tasks remotely, quickly transferring large amounts of information. End users consider it a favorable way to work with their data, regardless of their geographical location and available forms of communication. These and many other factors led the software development industry to look for a different set of content transfer opportunities via the Internet, using a traditional web user-interface. Today, websites are filled with user-friendly interfaces, animations and other appealing objects that end users may desire. But before all this appears in the user's eyes, it has to go through the server to the user's browser, and whilst the amount of data increases, the time necessary to exit the path, is growing [1]. This paper examines causes of the increase of this time and offers reduction options.

II INFORMATION GATHERING

Using Firebug extension for Mozilla Firefox browser, developers can view the division of time consumed to process a page request - until the moment of full loading and being displayed in the browser. By exploring the load time of a number of popular Web-sites with this tool, it is possible to see that the load time is mostly spent on data transfer from the server to the browser; for example, in the case of facebook.com all elements of the downloading queue, according to Fig. 1, even after 12 minutes of download, haven't loaded completely.

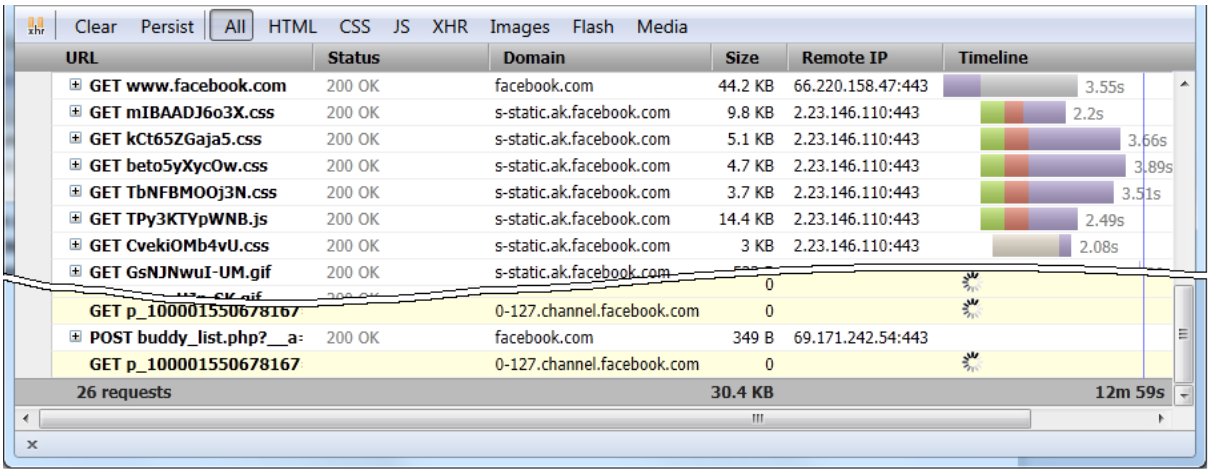

Fig. 1. Facebook.com page element downloading time graph

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.506

The main page load time distribution is shown in the Fig. 2 separately. It can be observed that most part of the page's loading time is taken up by data transmission. This data from the server can be divided into static HTML text with images and dynamic content, such as JavaScript (JSON) and Flash. Basing on the file list load time, a conclusion can be made, that most part of the data comes from different domains, which is done to increase the count of different data sources. Furthermore, it can be noticed, that even after a long wait some files are not fully loaded, which suggests possible problems with external data providers.

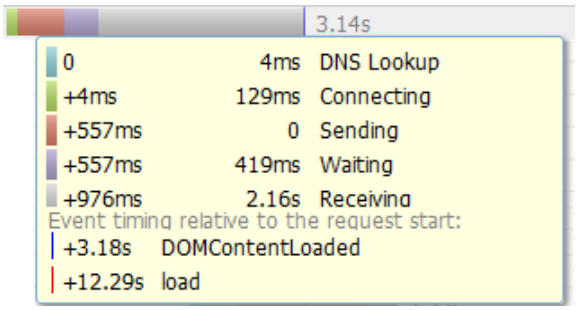

Fig. 2. Facebook.com page loading time distribution

III MULTIPLE FILE CREATION

One of the factors affecting the load time is multiple CSS file creation, although this content may be transferred as one joined file. Similarly, it permits to process JS-format files with an existing JavaScript code [2]. To impose additional complexity during the processing of an abnormally large file, it is possible to merge these files in the deployment phase, when it is deployed on a web server, for example, using Ant [3] or Maven [4] scenarios. This approach not only speeds up data transfer via the Internet, but also reduces the hard drive usage frequency, without requiring frequent reference to the hard disk file from operating system to read next.

Although usually frequently requested files are already in memory, and the increase of server performance will not be large, the main advantage of this approach is the required acceleration of information transfer, which results not only in the files merging and them being read faster from the media, but it also eliminates unnecessary resources overhead by reducing HTTP header transfer amount. Excessive amount of headers is shown in Fig. 3. The transferred header size (1051 bytes) is twice the size of the data received.

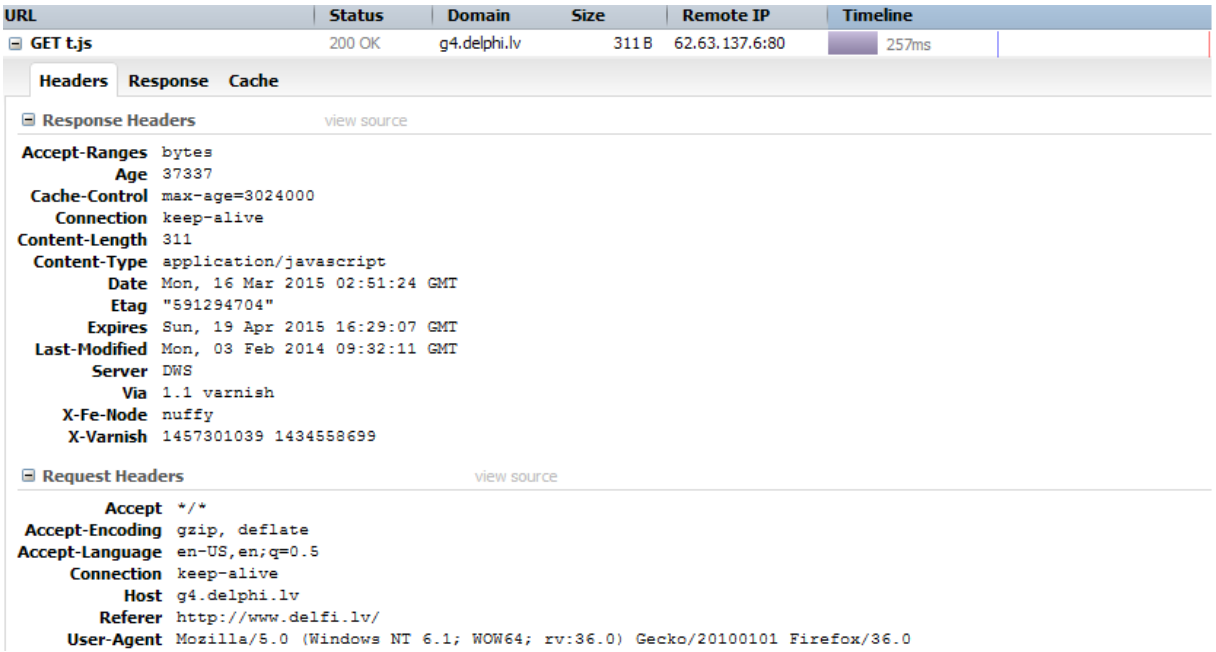

Fig. 3. News site request and response header size comparison

IV CHOOSING INCORRECT PROTOCOL

In addition to static type, modern websites also provide dynamic content, which varies according to the requests and also to the current status of the users or, similarly, according to the instantaneous requested data modifications. Although the majority of sites send new browser data in plain-text format to the client (or as ready-to-insert HTML-code), the usage of specialized protocols, designed for better use of Internet channels, can reduce the total data transfer time. For example, these protocols could be:

A. AMF protocol

AMF is created by Adobe messaging protocol, structure of which was published in 2007. It is intended for binary data transfer between applications, which are operated on the *Flash* platform and written

on the ActionScript base, all run on a server that can be either Flash Media Server (or any other media-server) or server-side application, written in PHP, Java, Ruby or other language. Before transmission, all data is serialized and all lines of text are encoded in Unicode.

B. RTMP

Real-Time Messaging Protocol – also developed by *Adobe*, allows to send not only video and audio data, but also other information, that makes it possible to build applications without using additional unnecessary resources to maintain service.

C. XML

eXtensible Markup Language, - a simplified SGML (Standard Generalized Markup Language) subset, which is designed for data description and a form of storage, that can be understandable to both computers (easy to read data objects) and humans.

D. JSON

JavaScript Object Notation – JavaScript language object, which can be recognized by the browser's JavaScript engine and treated as a ready-to-use object, without further action. The Fig. 4 shows the received code, which was converted into JavaScript object with key-value pairs.

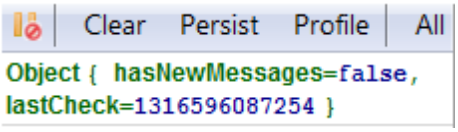

Fig. 4. JSON data evaluation as JavaScript object

E. HTTP2

Second version of HTTP protocol draft has been finalized in the beginning 2015 year and currently is awaiting RFC allocation and standardization finalization [5]. It is too early to estimate implementation and usage benefits, since only latest developer nightly builds of modern browsers supports this protocol, however it must significantly improve data transmission speed, since it is binary protocol as well as AMF and RTMP, and is built upon SPDY protocol extension, including stateful header compression (in HPACK) [6].

V IMPACT OF NETWORK PERFORMANCE

Although the homepage had been already prepared for viewing in ~3 seconds after making the request, its component loading could affect the parallel connection limit. In the file list it can be seen, that the CSS files are transmitted simultaneously. However, all browsers have a fixed maximum of connections to a particular domain by default and in accordance with W3C recommendations [7], so that *Internet Explorer 9* can be used for four simultaneous connections, *Chrome 12* - for six; *Firefox 7* - for fifteen and *Opera 11* - for sixteen.

This means you can simultaneously download from four to six files, depending on the browser used. The creation and closing of the connection takes a long time, that is why all modern browsers (including web-servers) can keep continuous or Keep-Alive connections, which makes it possible to use a single connection to multiple file transfers. The only requirement is that the server must indicate how many seconds it will allow for maintaining the connection without the actual data transfer. This is done to avoid an overflowing connection pool on a web-server, which is typically around 100 sockets providing client connection. If the first 100 clients connect to the server and continue to do so, it will be impossible to serve other clients because of the pending data requests from the user-server.

As one part of the solution we can suggest to use analysis, which aims to define the average and maximum loading times per element page. Using crawl results, it will be possible to determine the required time needed to apply the server-side Keep-Alive interval [8]. The client connects to the site, obtains all required files and then the server disconnects the client to make room in the pool for the next connection, while the client processes information and displays it to the end user, who will spend some time viewing the obtained information and will most likely not make more requests for some time.

A different approach may give the user an ability to increase the competitor-requested number of files by using other (sub) domains, as shown in Fig. 1 - the JavaScript and CSS files are loaded from a separate *s-static.ak.facebook.com* domain, i.e., they are assigned in addition to the number of connections. Usually, such way of division is implemented to separate the static content and use of the web servers with high static file handling capabilities, such as Lighttpd and Nginx, designed especially to serve the heavy duty sites and to provide minimal configuration/ extensibility options.

VI UNIQUE DATA USAGE

In order to enable updated data transmission in the background when making a request for information that the user's browser already had, it is possible to use built-in data storage in user's computers. Studies have shown that Adobe Flash plug-in is available on 99% [9] of all computers that have Internet access. It enables data transfer processing to use options provided by technology, one of them being Local Shared Objects [10].

Local Shared Objects (100KB in size by default) is a flash cookies file with a .sol extension, which in itself contains an array that uses the string as a key and any ActionScript language object as a value. The

user can specify a maximum size of Shared Objects, which can be 0kb, 10KB, 100KB, 1MB, 10MB or unlimited. The available amount of developers is relatively high. They keep a constant part of the site, occasionally checking their integrity and, if necessary, replacing any necessary parts.

Of course, this approach has its drawbacks: it is not possible to access another domain's Shared Object, even by using sub-domain's request because it contradicts with Adobe Flash Sandbox Restrictions rules, and, by default, if the Shared Object approaches the size of 100KB, the user will receive a message and a demand for additional space needed for granting permission, which may confuse users who are not accustomed to it. Although this technology allows creating a local caching, it is generally intended for small data storage configurations, rather than actual file-keeping.

Like Adobe Flash, Microsoft Silverlight supports their storage facility known as ISOSTORAGE (isolated local storage), which is separate from the browser's temporary directory and is sized 1MB. The user, just as in Flash, may change its size in steps of 100KB, 1MB, 5MB and 10MB.

Unlike the Adobe Flash technology, Microsoft Silverlight provides access to stored resources directly from JavaScript code without creating additional difficulties existing in Flash <–> JavaScript interaction. Another opportunity that developers may use - ISOSTORAGE format remembers the file system, i.e. it is possible create files and folders, copy, delete, rename and perform other activities that are available with a conventional file system.

VII WEAK CACHING IMPLEMENTATION

All the files in the example are sent from the Apache Web server that does not use external caching capabilities, which negatively affects the load time. The majority of web sites update their information no more than a few times a day, making it possible to use reverse proxy caching options to speed up data transfer. Conventional proxy server acts as a mediator and usually does not belong to a web page owner. Their task is to maintain some of the traffic that passes through until the next time another user will request the same data to issue from their storage. These proxies tend to be regarded as direct (forward).

Contrariwise, reverse proxy is located close (both in physical and network routing sense) to the web server and mainly serves users through the web server, and not directly.

Reverse proxy can exempt web server from the data compression, and during transition proxy <–> web server does not require small amounts of data transfer. Almost all modern browsers support data compression - they are responsible for sending the Accept-Encoding header in the request. Its value can be GZIP, which means that the browser supports and requests web server capacity-encoded page; or DEFLATE, which means that the browser is not built, or has disabled compression processing. From these two values and the sequence from your options/settings of your web server may be concluded what type of content to transfer in response to the request.

If the compression is available and settings are enabled, the server, before sending the final result, undergoes the GZIP filter that compresses data, particularly the text. This can be observed in data being compressed up to 100 times.

The fact that compressed data is received in headers, symbolized by Content-Encoding and valued as GZIP, which compresses the received requests and responses, can slightly reduce the overall load time.

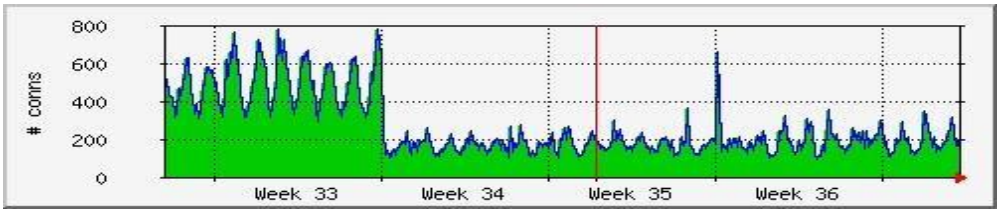

Fig. 5. Reverse proxy implementation performance boost

Another objective that can be attained by a proxy server is the load time balancing that distributes requests among multiple available web servers, depending on their workloads and opportunities to serve new demand [11]. Similarly, the mission can be considered a static page generated in storage and transmission from the cache, rather than a request to the web server where the page will be generated by a script, slowing down the transfer process. In Fig. 3 we can see that in the file, transferred from server, two additional headers appear, which indicates that the server uses reverse proxy usage, Varnish, - it is part of Via, a value of 1.1, used in the HTTP protocol version and in the name of the proxy software X-Varnish. (All HTTP headers that begin with X are not defined in HTTP 1.0 and 1.1 specifications, and are freely available for server-specific information). Its value is the request identifier, by which problems can be found in logs. Another significant reverse proxy server, which can also serve as a direct server, is Squid. It is

used in the Wikipedia Foundation, and the web-server/proxy ratio is 3:1 [12]. Reverse proxy usage enables reducing the requests to the main web server, in certain cases up to four times, as shown in Fig 5.

VIII CONCLUSIONS

In this article web site load time problem has been examined, including additional hints on how to solve occurring problems. Multiple similar file transmission and extra-header addition were described as server-side influencing factors. Browser's inability to persist already received data and attempts to run unprepared JavaScript code are considered as the main client-side problem. Many CSS and JavaScript file-merging processes that were described significantly affect not only the file transfer speed of the network, but also increases the load- and parse time in the user's browser. As a solution, we offered to combine similar content files into one, just like both the CSS and JavaScript files in the support type, and this combination does not affect the final result.

Furthermore, sending and receiving large headers was also discussed. Since servers usually do not need headers to identify users of the static content files, it is better to remove additional unnecessary headers from all static content either using server side – the same web server software, or the use of programming languages - or JavaScript for removing unneeded request headers.

Another element on which this article is focused: choosing the most appropriate protocol. Many websites in AJAX queries return the JavaScript code, which overfills the browser - these responses take time and increase the total amount of data. It is possible to use JSON or JSONP protocols only for transferring data and information to determine which one from pre-loaded functions to call.

From the set of already implemented improvements, we described the distribution of static files across multiple servers/domains using additional servers and other domain names in order to circumvent the browser restrictions. In addition, the reverse proxy was studied, which was already encountered in examples and also reduced the overall load time.

IX REFERENCES

[1] S. Souders, High Performance Web Sites: Essential Knowledge for Front-End Engineers, O'Reilly Media, Inc, 2007.

[2] R. Page, Website Optimization: An Hour a Day, John Wiley & Sons, 2012

[3] J. Fuller, *Apache Ant Recipes for Web Developers*, FastPencil, Inc, 2010.

[4] Srirangan, *Apache Maven 3 Cookbook*, Packt Publishing, 2011

[5] *Hypertext Transfer Protocol version 2* [Online]. Available: http://www.rfc-editor.org/internet-drafts/draft-ietf-httpbis-http2-17.txt [Accessed: Mar 15, 2015]

[6] *HPACK - Header Compression for HTTP/2* [Online]. Available: http://http2.github.io/http2-spec/compression.html [Accessed: Mar 15, 2015]

[7] W3C, HTTP/1.1 Connections, [Online]. Available: http://www.w3.org/Protocols/rfc2616/rfc2616-sec8.html [Accessed: Mar 15, 2015]

[8] S. Shivakumar, Architecting High Performing, Scalable and Available Enterprise Web Applications, Morgan Kaufmann, 2014

[9] Adobe Inc, *Adobe - Flash Player Statistics*, [Online]. Available: http://www.adobe.com/products/player_census/flashplayer/ [Accessed: Mar 15, 2015]

[10] Adobe Inc, *Adobe - Adobe Flash Player : What Is a Local Shared Object*, [Online]. Available: http://www.adobe.com/products/flashplayer/articles/lso/ [Accessed: Mar 15, 2015]

[11] H. ElAarag, S. Romano et al, *Web Proxy Cache Replacement Strategies: Simulation, Implementation, and Performance Evaluation*, Springer Science & Business Media, 2012

[12] Possibility Outpost, *High Scalability - High Scalability - Wikimedia architecture* [Online]. Available: http://highscalability.com/wikimedia-architecture [Accessed: Mar 15, 2015]

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 92-99

Mathematical Modelling of Aquatic Ecosystem

Sharif E. Guseynov[1, 2, a], Jekaterina V. Aleksejeva[1, b]
[1]Institute of Mathematical Sciences and Information Technologies, Liepaja University, 14 Liela Street, Liepaja LV-3401, Latvia; [2]Institute of Aeronautics, Faculty of Transport and Mechanical Engineering, Riga Technical University, 1A Lomonosov Street, Building-1, Riga LV-1019, Latvia [a]Sh.E.Guseinov@inbox.lv, [b]Jekaterina.V.Aleksejeva@gmail.com

Abstract. **In present paper we consider the complete statements of initial-boundary problems for the modelling of various aspects of aqueous systems in Latvia. All the proposed models are the evolutionary models: all they are nonstationary and continuous qualitative models having the dynamic parameters and aimed at analysis, evaluation and forecast of aquatic environmental quality (i.e. quality of reservoirs, lakes and seas). In constructing these mathematical models as research tools classic apparatus of differential equations (both ODE and PDE) as well as apparatus of mathematical physics were used.**

Keywords: **Aqueous ecosystem, environmental pollution, mathematical modelling, multicomponent dynamic model.**

I INTRODUCTION

Since the second half of the XX century the increasing growth of anthropogenic impact on the environment has led to exacerbation of many ecological problems. Prospects for reducing (full solution is not even discussed) the influence of such ecological problems on human life-sustaining activities are connected with implementation of "sustainable development" concept (see, for instance, [1]-[5] and respective references given in these), the main point of which is to ensure the stable coexistence between mankind and nature, which in its turn requires at least to save and reproduce the resource potential; improving the structure of ecosystem exploitation on the basis of scientific approach and performing an analysis for identifying the objective characteristic of ecological situation in order to take a reasonable (even non-optimal) solution; improving the level of human health. All these necessary conditions for implementation of "sustainable development" concept require: development of models and algorithms for unbiased and adequate evaluation of ecosystems' stability; finding out the relationships for dynamics of environmental problems; improvement of both models and methods of ecological and economic forecasts; studying of the combined effect of the anthropogenic loads influencing the ecosystem at large, etc.

The main toolkit in all these studies is mathematical modelling (see, for instance, [1], [2], [5], [6]-[11]) and / or simulation (see, for instance, [7], [12]-[15]), as well as analytical and numerical methods for solving the constructed models. Among the mathematical models the special place is taken by qualitative models, which are one of the most efficient tools of system analysis – the main approach to the study of ecological systems at various levels. System analysis still remains the only tool, in case if carrying out of large-scale field studies and experiments is encumbered because of technical and economic reasons, or is just impossible.

In the construction of mathematical models (both qualitative and quantitative) it is necessary to take into account existing peculiarities of ecological problems. The reason of the specific character of ecological problems is rooted, firstly, in the utmost uncertainty of problems formulation, and, secondly, in the complexity of problems, when it is required to take into account diverse factors starting from geology factors up to the pollution level of atmospheric air, seas and oceans.

As far as engineering practice is concerned, it is known that the error of quantitative assessments and calculations could be as large as hundreds percent. The results obtained in [6] in the study of two-boxed model for the investigation of metabolic processes in the Baltic Sea clearly show the hypersensitivity of water unit even to insignificant discharge amount of human activity wastes (the results of [3] show a similar hypersensitivity in the Black Sea). Eventually, the goal of using formal methods in environmental design, mathematical modelling and forecasting is a more accurate calculation for the scope of ecological damage, forthcoming financial costs as well as

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.192

effectiveness evaluation for the applied countermeasures.

Latvia is one of the most successful European countries in ecological problems solving. In accordance with the Environmental Performance Index 2014 (see [16]) Latvia takes 40th place among 178 countries in the world by the purity of the environment. The Environmental Performance Index, which produced by a team of environmental experts at the Yale University and the Columbia University, is constructed through the calculation and aggregation of 20 indicators reflecting national-level environmental data. These indicators are combined into nine issue categories (Environmental Health: 1. Health Impacts, 2. Air Quality, 3. Water and Sanitation; Ecosystem Vitality: 4. Water Resources, 5. Agriculture, 6. Forests, 7. Fisheries, 8. Biodiversity and Habitat, 9. Climate and Energy), each of which fit under one of two overarching objectives.

However, Latvia has room for improving the ecological problems. First of all, for improving the problem of Baltic Sea health. Latvia has more than 500 kilometers of beach along the Baltic Sea, which is among the most polluted seas in the world. According to the assessment presented in [17], none of the open basins of the Baltic Sea had an acceptable status. Only a very few coastal areas along the Gulf of Bothnia can be considered healthy. Implementation of special international agreements and cross-border programs as well as various studies and research is the way towards a healthy and sustainably-used Baltic Sea.

The second opportunity of improving the problem of the environment of Latvia concerns air pollution, which is particularly heavy during windless, cloudy weather. The main air pollutants are sulfur dioxide, formaldehyde, phenols, ammonia, and nitrogen oxides. The transport and energy sectors are the main sources of air pollutant emissions in Latvia. On average more than 80% of atmosphere pollution in Europe cities comes from motor transport as result of gas burn out. The chemical content of exhaust depends on gas type and quality, processing technology, burn out technique inside of motor, its technical condition and following filtration of exhaust (see, for instance, [12], [19] and respective references given in these). Improving air quality is one of the main priorities of Latvia's environmental policy [18]).

II MODELS OF AQUATIC ECOSYSTEM

A. The qualitative model for determination of pollutant concentration dynamics in the "layered in respect to depth" Baltic Sea at a known velocity of pollutant transfer

In this subsection a mathematical model is proposed for determination of pollutant concentration dynamics in the Baltic Sea. For constructing of this model the Baltic Sea was divided into parallel narrow horizontal layers, which are perpendicular to the internal normal in respect to the surface of the sea. Moreover, it was supposed that the velocity of pollutant transfer in each layer is priory known. The given assumption, obviously, is a quite strong requirement that reduces the value of the offered model to some extent, since determination of pollutant transfer velocity for any non-homogenous environment, including the Baltic Sea, is considered to be independent and quite tough problem. However, despite the limiting role of the given assumption, the offered mathematical model can be used as primary model for deeper monitoring of the eco-state of the Baltic Sea. Indeed, if roughly assuming that in each layer for some small period of time the velocity of pollutant transfer is constant (this could be achieved by executing the following procedure: significantly increasing the quantity of divisions into sea layers and at the same time significantly reducing the time for study of the pollutant distribution process), then, by using the offered model, after finding its analytical solution, it is possible to determine the quantity of the transferred pollutant amount in each layer with quite high accuracy. Then there appears a question: what is the limitation of the model if after the abovementioned procedure it becomes possible to determine the dynamics of pollutant concentration with a fairly high degree of accuracy? The answer is simple:

- the ideal accuracy of the solution found is only possible with an infinite increase of the layers number and with an infinite reduction of the time period at which the dynamics of pollutant concentration is investigated;
- sufficient accuracy of the solution found is possible at a very large increase of the layers number and at essential reduction of research time period that demands huge computing resources – on the edge of possibilities of modern supercomputers relevant to the possible maximum volume of RAM, CPU throughput, and maximal number of servers joined into cluster having the outmost parallel computing performance, etc.;
- the model is unstable, i.e. rather quite small changes in initial data can lead to a rather large discrepancy between the solution found and the ideal exact solution, which exists, is unique and unknown and which is intended to be reached. Therefore, it is necessary not just to solve the offered model, but to develop a stable method of its solution, having a regularization (by Tikhonov) property. It should be noted that this issue is relatively easily solved.

So, by denoting as D the studied three dimensional domain of the Baltic Sea, and as $u(x,y,z;t)$ the desirable concentration of harmful substance in this domain D at the moment of time t, the proposed model implies the following: it is necessary to

determine the function $u(\xi,t)$, where $u(\xi,t),(\xi,t)\in D\times[0,T]$, $D\subset\mathbb{R}^3$, which satisfies the equation

$$\frac{\partial u(\xi,t)}{\partial t}=\nabla_{\xi}\left(a(\xi)\cdot\nabla_{\xi}u(\xi,t)\right)+\left\langle\alpha,\nabla_{\xi}u(\xi,t)\right\rangle+\beta\cdot u(\xi,t)+f(\xi,t),\quad \xi\in int\{D\},\ t\in(0,T], \tag{1}$$

where $D\overset{def}{\equiv}\{\xi=(x,y,z)\in[0,l_1]\times[0,l_2]\times[l_3,0],\ l_j>0\ (j=1,2),\ l_3<0\}$ is n-layered (in respect to axis OZ) 3D parallelepiped domain; $z\in(z_{i-1},z_i),\ i=\overline{1,n};\quad z_0=l_3,\ z_n=0;$ $a(\xi)\equiv a_i=const>0$ if $z\in(z_{i-1},z_i),\ i=\overline{1,n};$ n is the number of layered media parallel to the plane XOY in respect to the vertical axis OZ; $\alpha=(\alpha_1,\alpha_2,\alpha_3)$, $\alpha_j=const\ (j=\overline{1,3});$ $\beta=const;$ $\langle q,r\rangle$ denotes a scalar product of vectors $q\in\mathbb{R}^3$ and $r\in\mathbb{R}^3$; a symbol $\nabla_{\eta}Q(\eta,\bullet)$ denotes gradient as to the variable η of the scalar function $Q(\eta,\bullet);$ $int\{D\}$ and ∂D correspondingly denote internal and boundary points of the restricted three-dimensional domain $D\subset\mathbb{R}^3$; the initial condition

$$u(\xi,t)\big|_{t=0}=u_0(\xi),\ \xi\in D; \tag{2}$$

the boundary conditions

$$\gamma_{j,i-1}^{\{i-1\}}\cdot u(\xi,t)\big|_{z=z_{i+j-2}+\varepsilon_j}+\gamma_{j,i}^{\{i-1\}}\cdot\frac{\partial u(\xi,t)}{\partial z}\bigg|_{z=z_{i+j-2}+\varepsilon_j}=u_{j,i-1}(x,y,t);\ i=\overline{1,(n+1)},\ j=1,2, \tag{3}$$

where $(x,y,t)\in\{D/[l_3,0]\}\times[0,T];$ $\varepsilon_j=+0$ if $j=1,$ $\varepsilon_j=-0$ if $j=2;$ $\gamma_{j,i}^{\{k\}}=\forall const<0$ if $(j=1)\wedge\left(i=k=\overline{0,n}\right),$ $\gamma_{j,i}^{\{k\}}=\forall const>0$ if otherwise;

$$\theta_{j,i-1}^{\{i-1\}}\cdot u(\xi,t)\big|_{\omega=\omega_{ji}}+\theta_{j,i}^{\{i-1\}}\cdot\frac{\partial u(\xi,t)}{\partial\omega}\bigg|_{\omega=\omega_{ji}}=\vartheta_{j,i}(\xi/\omega,t);\ i=1,2;\ j=1,2, \tag{4}$$

where $(y,z,t)\in\{D/[0,l_1]\}\times[0,T];$ $(x,z,t)\in\{D/[0,l_2]\}\times[0,T];$ $\omega=x$ if $i=1,$ $\omega=y$ if $j=2;$ $\theta_{j,i}^{\{k\}}=\forall const<0$ if $(j=1)\wedge(i=k\in\{0;1\}),$ $\theta_{j,i}^{\{k\}}=\forall const>0$ if otherwise; $\{\omega_{ji}\}_{j=1,2}^{i=1,2}=\begin{pmatrix}0+0 & 0+0\\ l_1-0 & l_2-0\end{pmatrix};$ the consistency conditions

$$\gamma_{j,i-1}^{\{i-1\}}\cdot u_0(\xi,t)\big|_{z=z_{i+j-2}+\varepsilon_j}+\gamma_{j,i}^{\{i-1\}}\cdot\frac{\partial u_0(\xi,t)}{\partial z}\bigg|_{z=z_{i+j-2}+\varepsilon_j}=u_{j,i-1}(x,y,t)\big|_{t=0+0};\ i=\overline{1,(n+1)},\ j=1,2, \tag{5}$$

where $(x,y,t)\in\{D/[l_3,0]\}\times[0,T];$ the value and meaning of constants $\varepsilon_j\ (i=1,2),$ $\gamma_{j,i}^{\{k\}}\ \left(j=1,2;\ i=\overline{0,(n+1)};\ k=\overline{0,n}\right)$ are the same as in case of boundary conditions (3);

$$\theta_{j,i-1}^{\{i-1\}}\cdot u_0(\xi,t)\big|_{\omega=\omega_{ji}}+\theta_{j,i}^{\{i-1\}}\cdot\frac{\partial u_0(\xi,t)}{\partial\omega}\bigg|_{\omega=\omega_{ji}}=\vartheta_{j,i}(\xi/\omega,t)\big|_{t=0+0};\ i,j=1,2, \tag{6}$$

where $(y,z,t)\in\{D/[0,l_1]\}\times[0,T];$ $(x,z,t)\in\{D/[0,l_2]\}\times[0,T];$ the variable $\omega\in\{x,y\},$ and constants $\theta_{j,i}^{\{k\}}\ \left(j=1,2;\ i=\overline{0,2};\ k=0,1\right),$ $\{\omega_{ji}\}_{j,i\in\{1,2\}}$ have the same values and meaning as for the boundary conditions (4); the conjugating conditions

$$\begin{cases}u(\xi,t)\big|_{z=z_i+0}=u(\xi,t)\big|_{z=z_i-0},\\ (x,y,t)\in\{D/[l_3,0]\}\times[0,T];\\ a_i\cdot\dfrac{\partial u(\xi,t)}{\partial z}\bigg|_{z=z_i+0}=a_{i+1}\cdot\dfrac{\partial u(\xi,t)}{\partial z}\bigg|_{z=z_i-0},\\ (x,y,t)\in\{D/[l_3,0]\}\times[0,T]\end{cases} \tag{7}$$

for $\forall i=\overline{1,(n-1)}.$

In the formulas (3)-(6) the parameters ω and $\varepsilon_i\ (i=1,2)$ are introduced only for a compact record of these conditions.

In the mathematical model (1)-(7) the initial data are represented by a piecewise-constant function $a(\xi);$ velocity vector $\alpha;$ constant $\beta\in\mathbb{R}^1;$ constants $l_i\ (i=\overline{1,3}),$ $T>0;$ functions $f(\xi,t)\in C\{D\times[0,T]\};$ $u_{j,i}(x,y,t)\in C^{1,1,2}\{\{D/[l_3,0]\}\times[0,T]\}(i=\overline{0,n};\ j=1,2);$ $\vartheta_{j,1}(y,z,t)\in C^{1,1,2}\{\{D/[0,l_1]\}\times[0,T]\},(j=1,2);$ $\vartheta_{j,2}(y,z,t)\in C^{1,1,2}\{\{D/[0,l_2]\}\times[0,T]\},(j=1,2);$ constants $\gamma_{j,i}^{\{k\}}\left(j=1,2;\ i=\overline{0,(n+1)};k=\overline{0,n}\right);$ $\theta_{j,i}^{\{k\}}\left(j=1,2;\ i=\overline{0,2};\ k=0,1\right).$

B. Multi-component dynamic model for determination of the main characteristic for the substance circulation process (taking circulation of nitrogen as an example) in the near-surface layers of the natural aquatic environments (natural reservoirs, large basins, lakes and seas)

In this subsection a mathematical model is proposed describing circulation of nitrogen in natural aquatic environments for determination of major characteristics of this process. The constructed model is parametric and is described by a system of partial differential equations. It allows decreasing significantly the uncertainty of parameters of the sophisticated circulation process in the non-homogeneous media, the example of which is the natural aquatic environment in the presence of powerful sources and sinks. Reducing the uncertainty of parameters is carried out in the way of (a) identification of relationships imposed on the parameters, (b) finding of non-trivial special points of

the system-model and taking into account the stability preservation conditions in respect to these special points for several structures of the circulation process.

The proposed model is constructed under the following assumptions: the average temperature of the studied natural aquatic environment for the whole period of the circulation process study is assumed to be constant; the total amount of nitrogen in the studied aquatic environment can be represented in the form of organic and inorganic nitrogen, where the organic nitrogen in its turn can be represented also in two forms – as animated and unanimated nitrogen; the reserve of nitrogen in the studied aquatic environment is replenished (a) by means of molecular nitrogen adoption of the air by some free nitrogen fixing agents and (b) by some types of seaweed; processes of nitrogen fixing by phytoplankton and bacterial plankton have linear interdependence (with a known coefficient of linearity); mineral nitrogen in the aquatic environment is used only by some types of seaweed and planktons; protein substances of plants and animal species formed in the studied aquatic environment, after dying off of the organisms are subjected to mineralization with the assistance of bacteria, moreover, decomposition of such protein substances evolves nitrogen, generally in the form of ammonia; in the studied process of circulation there occurs direct denitrification (denitrification is a decomposition of nitrogenous compounds with generation of free nitrogen) as well as circumstantial denitrification (circumstantial denitrification – this is a chemical effect of NO_2^-, ammonium salts and amino acids with generation of free nitrogen); the studied circulation process also includes deammonification (decomposition by microorganisms of nitrogen-containing organic compounds with generation of free nitrogen).

So, the offered parametric model anticipates determination of the vector $u \overset{def}{\equiv} \left(u_{ph.pl.}, u_{bac.pl.}, u_{zoo.pl.}, u_{inor.nit.}, u_{or.nit.}\right)(\xi,t) \in C^{0,1}\left\{D \cup [0,T]\right\}$ of dynamic variables/functions $u_{ph.pl.}(\xi,t)$, $u_{bac.pl.}(\xi,t)$, $u_{zoo.pl.}(\xi,t)$, $u_{inor.nit.}(\xi,t)$ and $u_{or.nit.}(\xi,t)$, where $D \overset{def}{\equiv} \left\{\xi = (x_1, x_2, x_3):\ l_{x_3} \le x_3 \le 0;\ 0 \le x_i \le l_{x_i}\ (i = 1,2)\right\}$, from the non-linear system

$$
\begin{cases}
\dfrac{1}{u_{ph.pl.}(\xi,t)} \cdot \dfrac{\partial u_{ph.pl.}(\xi,t)}{\partial t} = k_1(\xi) + k_2(\xi) \cdot u_{or.nit.}(\xi,t) \\
+k_3(\xi) \cdot u_{inor.nit.}(\xi,t) - k_4(\xi) \cdot u_{zoo.pl.}(\xi,t) - k_5(\xi) - u_{flowing}(t), \\
(\xi,t) \in \operatorname{int}\{D\} \cup (0,T],\ \operatorname{int}\{D\} \overset{def}{\equiv} D / \partial D, \\
\dfrac{1}{u_{bac.pl.}(\xi,t)} \cdot \dfrac{u_{bac.pl.}(\xi,t)}{\partial t} = k_6(\xi) - k_7(\xi) - k_8(\xi) - k_9(\xi) \\
-u_{flowing}(t) + k_{10}(\xi) \cdot u_{or.nit.}(\xi,t) - k_{11}(\xi) \cdot u_{inor.nit.}(\xi,t) \\
-k_{12}(\xi) \cdot u_{zoo.pl.}(\xi,t),\ (\xi,t) \in \operatorname{int}\{D\} \cup (0,T], \\
\dfrac{1}{u_{zoo.pl.}(\xi,t)} \cdot \dfrac{u_{zoo.pl.}(\xi,t)}{\partial t} = k_4(\xi) \cdot u_{ph.pl.}(\xi,t) + k_{12}(\xi) \cdot u_{ph.pl.}(\xi,t) \\
+k_{13}(\xi) \cdot u_{or.nit.}(\xi,t) - k_{14}(\xi) - k_{15}(\xi) - k_{16}(\xi) \cdot u_{flowing}(t), \\
(\xi,t) \in \operatorname{int}\{D\} \cup (0,T], \\
\dfrac{1}{u_{inor.nit.}(\xi,t)} \cdot \dfrac{u_{inor.nit.}(\xi,t)}{\partial t} = \dfrac{c_1(\xi) + c_2(\xi)}{u_{inor.nit.}(\xi,t)} + k_{14}(\xi) \cdot \dfrac{u_{zoo.pl.}(t)}{u_{inor.nit.}(\xi,t)} \\
+k_8(\xi) \cdot \dfrac{u_{bac.pl.}(t)}{u_{inor.nit.}(\xi,t)} - k_{11}(\xi) \cdot u_{bac.pl.}(\xi,t) - k_3(\xi) \cdot u_{ph.pl.}(\xi,t) \\
-u_{flowing}(t),\ (\xi,t) \in \operatorname{int}\{D\} \cup (0,T], \\
\dfrac{1}{u_{or.nit.}(\xi,t)} \cdot \dfrac{u_{or.nit.}(\xi,t)}{\partial t} = \dfrac{c_3(\xi) + c_4(\xi)}{u_{or.nit.}(\xi,t)} + k_5(\xi) \cdot \dfrac{u_{ph.pl.}(\xi,t)}{u_{or.nit.}(\xi,t)} \\
+k_7(\xi) \cdot \dfrac{u_{bac.pl.}(\xi,t)}{u_{or.nit.}(\xi,t)} + k_{15}(\xi) \cdot \dfrac{u_{zoo.pl.}(\xi,t)}{u_{or.nit.}(\xi,t)} - k_{10}(\xi) \cdot u_{bac.pl.}(\xi,t) \\
-k_2(\xi) \cdot u_{ph.pl.}(\xi,t) - k_{13}(\xi) \cdot u_{zoo.pl.}(\xi,t) - k_{17}(\xi) - u_{flowing}(t), \\
(\xi,t) \in \operatorname{int}\{D\} \cup (0,T],
\end{cases}
\tag{8}
$$

with initial conditions

$$
\begin{cases}
u_{ph.pl.}(t)\big|_{t=0} = u^0_{ph.pl.}(\xi);\ u_{bac.pl.}(t)\big|_{t=0} = u^0_{bac.pl.}(\xi); \\
u_{zoo.pl.}(t)\big|_{t=0} = u^0_{zoo.pl.}(\xi),\ \ \xi \in D, \\
u_{inor.nit.}(t)\big|_{t=0} = u^0_{inor.nit.}(\xi);\ u_{or.nit.}(t)\big|_{t=0} = u^0_{or.nit.}(\xi),\ \xi \in D,
\end{cases}
\tag{9}
$$

where functions $u_{ph.pl.}(\xi,t)$, $u_{bac.pl.}(\xi,t)$, $u_{zoo.pl.}(\xi,t)$, $u_{inor.nit.}(\xi,t)$ and $u_{or.nit.}(\xi,t)$ denote respectfully the desired amount of phytoplankton, bacterial plankton, zooplankton, inorganic nitrogen and organic nitrogen at the time moment $t \in [0,T]$ at the point $\xi \in D$; T is the given finite moment of time, during which the process of substance circulation in the near-surface layers of natural aquatic environments is studied; $l_{x_i}\ (i = \overline{1,3})$: $l_{x_1} > 0, l_{x_2} > 0, l_{x_3} < 0$ are the known boundaries of the considered "parallelepiped" domain of natural aquatic environments; the initial functions $u^0_{ph.pl.}(\xi)$, $u^0_{bac.pl.}(\xi)$, $u^0_{zoo.pl.}(\xi)$, $u^0_{inor.nit.}(\xi)$, $u^0_{or.nit.}(\xi)$, $\xi \in D$ are a priori known functions, which are continuous in the closed domain D; the functions $k_i(\xi)\ (i = \overline{1,17})$, $\xi \in \operatorname{int}\{D\}$, $c_i(\xi)\ (i = \overline{1,4})$, $\xi \in \operatorname{int}\{D\}$ and $u_{flowing}(t)$, $t \in (0,T]$ have the following meaning: $k_1(\xi)$ is a coefficient of nitrification of phytoplankton; $k_2(\xi)$ is a coefficient of organic nitrogen absorption by phytoplankton; $k_3(\xi)$ is a coefficient of mineral nitrogen absorption by phytoplankton; $k_4(\xi)$ is a

coefficient of phytoplankton consumption by zooplankton; $k_5(\xi)$ is a coefficient of emission of organic nitrogen by phytoplankton; $k_6(\xi)$ is a coefficient of nitrification of bacterial plankton; $k_7(\xi)$ is a coefficient of emission of organic nitrogen by bacterial plankton; $k_8(\xi)$ is a coefficient of emission of mineral nitrogen by bacterial plankton; $k_9(\xi)$ is a coefficient of denitrification: denitrification – is decomposition of nitrogenous compounds with free nitrogen generation; $k_{10}(\xi)$ is a coefficient of organic nitrogen processing by bacteria; $k_{11}(\xi)$ is a coefficient of mineral nitrogen processing by bacteria; $k_{12}(\xi)$ is a coefficient of bacterial plankton consumption by zooplankton; $k_{13}(\xi)$ is a coefficient of organic nitrogen consumption by zooplankton; $k_{14}(\xi)$ is a coefficient of mineral nitrogen emission by zooplankton; $k_{15}(\xi)$ is a coefficient of organic nitrogen emission by zooplankton; $k_{16}(\xi)$ is a coefficient of zooplankton discharge by higher level predaceous organisms; $k_{17}(\xi)$ is a coefficient of organic nitrogen supply to the lower layer of the considered domain D of natural aquatic environments – the domain of near-surface layers of natural aquatic environments; $c_1(t)$ is the inflow of inorganic nitrogen that is equal to the sum of (a) main feeding river sinks, (b) rain surface sinks, (c) inner drainage of mineral fertilizers into the soil, (d) industrial and household discharges, (e) atmospheric precipitations on the water area: water area – is a region of water surface within the set boundaries; $c_2(t)$ is the inflow of mineral nitrogen from lower layer silt; $c_3(t)$ is the inflow of organic nitrogen from lower layer silt; $c_4(t)$ is the inflow of organic nitrogen, equal to the sum of (a) main feeding river flows, (b) atmospheric precipitations on the water area, (c) decaying biomass of phytoplankton and higher level of marine vegetation, (d) emissions of living phytoplankton; $u_{flowing}(t)$ the power of flow (water circulation; evaporation; etc.).

Remark 1. In the model (8), (9) modeling units – are tons of nitrogen per time unit (for instance, per year).

Remark 2. Model (8), (9) having any fixed value of spatial variable $\xi = \xi_{fix} \in D$ becomes a Cauchy problem for the system of ordinary non-linear differential equations, and in this case it is an adequate model for determination of nitrogen circulation in large basins, reservoirs, lakes, ponds, etc.

Remark 3. In the proposed model (8), (9) to describe the functional significance of the coefficient $k_9(\xi)$, which denoted coefficient of denitrification, it was discussed that the denitrification is a decomposition of nitrogenous compounds with generation of free nitrogen. This definition requires a more precise specification. The transition of nitrogen into the gaseous form of nitrates $\left(NO_3^- \to NO_2^- \to N_2\right)$ is called denitrification. Generally speaking, this process in the presence of denitrifying bacteria occurs fairly rapidly only when there are nitrates present, the amount of easily assimilable organic substance is sufficient, as well as anaerobic conditions fulfilled. If the medium together with NO_2^- contains ammonium salts (or amino acids), the free nitrogen can be generated due to their chemical interaction, namely, $2KNO_2 + (NH_4)SO_4 \to 2NH_4NO_2 + K_2SO_4$; $NH_4NO_2 \to N_2 + 2H_2O$. This process is called (see, for instance, [20]) circumstantial denitrification. Nitrogen losses are also possible in the result of deammonification $(NH_4 \to N_2)$, and also when generating fugitive nitrogen oxides with the participation of certain amino acids and a variety of unstable nitrogen containing compounds ([11]).

As it was mentioned at the beginning of this subsection, in the mathematical model (8), (9) we supposed that all the above mentioned processes occur to a greater or lesser degree, however, in [21], it is shown that the most frequently denitrification is observed in soil, and not in the layers of water.

In the model (8), (9) for its simplification all the above mentioned processes are joined into one process called denitrification with one functional coefficient $k_9(\xi)$, covering the other possible processes associated with the loss of nitrogen in natural aquatic environments.

C. Model for determination of oxygen condition sensitivity for the Baltic Sea

In this subsection a mathematical model for determination sensitivity of the oxygen condition in the Baltic Sea is proposed. As it is shown in the work [5], the models sensitivity study anticipates identifying of special features of ecosystem dynamics within the domains bounded by the bifurcation surfaces.

Methods of the sensitivity theory can be used to solve the two types of problems that arise when constructing dynamic multi-component models, including simulation models: (a) at the stage of establishment of the model structure and for binding parameters of the models, which provide concordance during some observations time period; (b) at the stage of studying the model with already established structure for finding irregular special points (that is, "weak places") and stationary points (that is, anaerobic places) of the aquatic ecosystem, as well as for determination of critical values of parameters when modelling toxic effect and anthropogenic eutrophication (eutrophication – is the process of total productivity growth of the water reservoir ecosystem,

which includes water masses, bottom deposits and microorganisms living there; eutrophication – is the increase of the level of water primary production owing to the increase in the concentration of biogenic elements there, that is mainly nitrogen and phosphorus).

The model proposed below allows determination of peculiarities of oxygen state dynamics, which cannot be determined by simple enumeration of parameters and/or when using purely statistical methods of parameters estimation.

Moreover, the proposed model allows determining the qualitative properties of the oxygen state dynamics, which significantly reduces the uncertainty in sophisticated interrelations between parameters of dynamics and the environment, allows making adequate high precision forecasts in respect to anthropogenic impact onto the ecosystem and, therefore, reveals trends in the ecosystem state of the sea. Finally, the proposed model can be used as a primary model to determine the condition that leads to anaerobic zones growth strengthening in the Baltic Sea.

As a major restriction of the proposed model there can be noted the impossibility of its usage for the study of the seasonal Baltic Sea ecosystem dynamics in the conditions of aerobic and anaerobic states being below the halocline level.

Remark 4. This model is essentially based on the results of works [7], [12] taking into account the modifying changes from works [22]-[25], namely, in the Sjoberg anaerobic model non-autonomous system of ordinary differential equations, where in the direct form there are present functions of vertical turbulent mixing, function of lighting and function of temperature, is substituted by autonomous system of equations (such substitution is approved by Floquet and Lyapunov's theorem: the information on the theory of Floquet and Lyapunov, namely, the importance and benefits of Floquet-Lyapunov theorem can be found in book [26] and fundamental monograph [27]; but the widest applications of Floquet-Lyapunov theorem are given in the book [28]).

It is required to determine a vector $u(\xi,t) \overset{def}{\equiv} \left(\overline{u}_{inor.phosph.}, \underline{u}_{inor.phosph.}, \overline{u}_{or.phosph.}, \underline{u}_{or.phosph.}, \overline{u}_{ph.pl.}, \underline{u}_{oxyg.}\right)(t)$ of dynamic variables/functions $\overline{u}_{inor.phosph.}(t)$, $\underline{u}_{inor.phosph.}(t)$, $\overline{u}_{or.phosph.}(t)$, $\underline{u}_{or.phosph.}(t)$, $\overline{u}_{ph.pl.}(t)$ and $\underline{u}_{oxyg.}(t)$ from the system of ordinary non-linear differential equations

$$
\begin{cases}
\dfrac{\dot{\overline{u}}_{inor.phosph.}(t)}{\overline{u}_{inor.phosph.}(t)} = \dfrac{F_{oxyg.}^{inor.phosph.} + F_{ph.pl.}^{inor.phosph.}}{\overline{u}_{inor.phosph.}(t)} + (F_0 + H)\cdot c_1 \cdot \dfrac{\underline{u}_{inor.phosph.}(t)}{\overline{u}_{inor.phosph.}(t)} + \\
+k_1 \cdot T^{\circ} \cdot \dfrac{\overline{u}_{or.phosph.}(t)}{\overline{u}_{inor.phosph.}(t)} - \left(H + F_{oxyg.} + F_0\right)\cdot c_2 - k_2 - k_3 \cdot I \cdot T^{\circ} \cdot \overline{u}_{ph.pl.}(t), \\
\dfrac{\dot{\underline{u}}_{inor.phosph.}(t)}{\underline{u}_{inor.phosph.}(t)} = \dfrac{F_0 \cdot q_0^{inor.phosph.}}{\underline{u}_{inor.phosph.}(t)} + H \cdot c_2 \cdot \dfrac{\overline{u}_{inor.phosph.}(t)}{\underline{u}_{inor.phosph.}(t)} \\
+k_4 \cdot \dfrac{\underline{u}_{oxyg.}(t)\cdot \underline{u}_{or.phosph.}(t)}{\underline{u}_{inor.phosph.}(t)} + k_5 \cdot \dfrac{\underline{u}_{or.phosph.}(t)}{\underline{u}_{inor.phosph.}(t)} \\
-(F_0 + H)\cdot c_1 - k_6 \cdot \underline{u}_{oxyg.}(t), \\
\dfrac{\dot{\overline{u}}_{or.phosph.}(t)}{\overline{u}_{or.phosph.}(t)} = (F_0 + H)\cdot c_1 \cdot \dfrac{\underline{u}_{or.phosph.}(t)}{\overline{u}_{or.phosph.}(t)} + k_7 \cdot \dfrac{\overline{u}_{ph.pl.}(t)}{\overline{u}_{or.phosph.}(t)} \\
-\left(H + F_{oxyg.} + F_0\right)\cdot c_2 - k_8 - k_9 - k_1 \cdot T^{\circ}, \\
\dfrac{\dot{\underline{u}}_{or.phosph.}(t)}{\underline{u}_{or.phosph.}(t)} = \dfrac{F_0 \cdot q_0^{or.phosph.}}{\underline{u}_{or.phosph.}(t)} + H \cdot c_2 \cdot \dfrac{\overline{u}_{or.phosph.}(t)}{\underline{u}_{or.phosph.}(t)} + H \cdot c_2 \cdot \dfrac{\overline{u}_{ph.pl.}(t)}{\underline{u}_{or.phosph.}(t)} \\
+k_9 \cdot \dfrac{\overline{u}_{or.phosph.}(t)}{\underline{u}_{or.phosph.}(t)} - (F_0 + H)\cdot c_1 - k_{10} - k_5 - k_4 \cdot \underline{u}_{oxyg.}(t), \\
\dfrac{\dot{\overline{u}}_{ph.pl.}(t)}{\overline{u}_{ph.pl.}(t)} = k_3 \cdot I \cdot T^{\circ} \cdot \overline{u}_{inor.phosph.}(t) - \left(H + F_0 + F_{oxyg.}\right)\cdot c_2 - k_7, \\
\dfrac{\dot{\underline{u}}_{oxyg.}(t)}{\underline{u}_{oxyg.}(t)} = \dfrac{F_0 \cdot q_0^{oxyg.}}{\underline{u}_{oxyg.}(t)} + \dfrac{H \cdot q_1^{oxyg.}}{\underline{u}_{oxyg.}(t)} - (F_0 + H)\cdot c_1 - \dfrac{k_{11}}{\underline{u}_{oxyg.}(t)} \\
-k_{12} \cdot k_4 \cdot \underline{u}_{or.phosph.}(t) - k_5 \cdot \dfrac{\underline{u}_{or.phosph.}(t)}{\underline{u}_{oxyg.}(t)}
\end{cases}
\tag{10}
$$

and initial conditions

$$
\begin{cases}
\overline{u}_{inor.phosph.}(t)\big|_{t=0} = \overline{u}^0_{inor.phosph.};\ \underline{u}_{inor.phosph.}(t)\big|_{t=0} = \underline{u}^0_{inor.phosph.}; \\
\overline{u}_{or.phosph.}(t)\big|_{t=0} = \overline{u}^0_{or.phosph.};\ \underline{u}_{or.phosph.}(t)\big|_{t=0} = \underline{u}^0_{or.phosph.}; \\
\overline{u}_{ph.pl.}(t)\big|_{t=0} = \overline{u}^0_{ph.pl.};\ \underline{u}_{oxyg.}(t)\big|_{t=0} = \underline{u}^0_{oxyg.}.
\end{cases}
\tag{11}
$$

In the model (10), (11) such designations and assumptions take place: the range of definition for each equation forming the system (10) is determined by the semi interval $(0,T]$; functions $\overline{u}_{inor.phosph.}(t)$, $\underline{u}_{inor.phosph.}(t)$, $\overline{u}_{or.phosph.}(t)$, $\underline{u}_{or.phosph.}(t)$, $\overline{u}_{ph.pl.}(t)$ and $\underline{u}_{oxyg.}(t)$ denote accordingly the desired amount of inorganic phosphorus above the layer of Baltic Sea, where the salinity gradient is maximal (i.e. above the layer, where sea water salinity maximally changes – such layer is called halocline: halocline is a vertical zone in the oceanic water column in which salinity changes rapidly with depth, located below the well-mixed: for instance, see http://www.britannica.com/EBchecked/topic/252980/halocline), inorganic phosphorus below the halocline, organic phosphorus above the halocline, organic phosphorus below the halocline, phytoplankton above the halocline and oxygen below the halocline at the moment of time $t \in [0,T]$: T is the given time period, during which the dynamics of oxygen condition of the Baltic Sea is studied; constants $\overline{u}^0_{inor.phosph.}$, $\underline{u}^0_{inor.phosph.}$, $\overline{u}^0_{or.phosph.}$, $\underline{u}^0_{or.phosph.}$, $\overline{u}^0_{ph.pl.}$ and $\underline{u}^0_{oxyg.}$ from the initial conditions (11) are supposed to be known priory; values of constants (let's call them inner model

constants) I, T°, H and parameters values (let's call them inner parameters of the model) $k_i\ \left(i=\overline{1,12}\right)$, $c_i\ \left(i=1;2\right)$, F_0, $F_{oxyg.}^{inor.phosph.}$, $F_{ph.pl.}^{inor.phosph.}$, $F_{oxyg.}$, $q_0^{inor.phosph.}$, $q_0^{or.phosph.}$, $q_0^{oxyg.}$, $q_1^{oxyg.}$ of the differential equation system (10) are given below in accordance with research results [12]: $I=8$ hours; $T^{\circ}=8$ ºC; $H=2.125$ km/(twenty-four hours); $k_1=0.004$ 1/ºC per twenty-four hours; $k_2=0.0000693$ 1/(twenty-four hours); $k_3=0.00000004$ 1/t Phosph. hour ºC·(twenty-four hours); $k_4=0.000537$ 1/km^3 O_2·(twenty-four hours); $k_5=0.001558$; $k_6=0.00000217$ 1/km^3 O_2·(twenty-four hours); $k_7=0.065$ 1/(twenty-four hours); $k_8=0.00001846$ 1/(twenty-four hours); $k_9=0.00022$ 1/(twenty-four hours); $k_{10}=0.00022$ 1/(twenty-four hours); $k_{11}=0.0004206$ 1/km^3 O_2·(twenty-four hours); $k_{12}=0.000199$ 1/km^3 O_2·(twenty-four hours); $c_1=0.00002765$ 1/km^3; $c_2=0.0001053$ 1/km^3; $F_0=1.308$ km^3/(twenty-four hours); $F_{oxyg.}^{inor.phosph.}+F_{ph.pl.}^{inor.phosph.}=28.7$ t Phosph./ (twenty-four hours); $F_{oxyg.}=1.308$ km^3/(twenty-four hours); $q_0^{inor.phosph.}=0.001$ km^3 Phosph./km^3; $q_0^{or.phosph.}=0.003$ km^3 Phosph./km^3; $q_0^{oxyg.}=0.01$ km^3 O_2/ km^3; $q_1^{oxyg.}=0.01$ km^3 O_2/km^3.

Remark 5. Due to the fact that almost 30 years have passed from the moment of publishing the work [12] and in this period significant geopolitical changes occurred in Europe, in particular, in the Baltic Sea Countries (breakdown of the USSR; acquisition of sovereignty of three Baltic states; decrease of industrial power of some of the Baltic Sea Countries (Latvia, Lithuania, Estonia, Poland, Russia) and, as consequence, change of anthropogenic impact on the Baltic Sea by those countries; increase of industrialization of developed Baltic Sea Countries – Germany, Sweden, Finland, Denmark and other factors), it is evident that the values of model (10), (11) inner constants and parameters, which were calculated in the work [12] and borrowed by us without any changes, obviously have to be specified more precisely. As it could be seen from the offered model (10)-(11), the corresponding changes/specifications of such inner constants and parameters of the model do not change the essence of the proposed model, except the coordinates of positive stationary points (one of the positive stationary points, within the small vicinity of which the model is stable will be the desired condition of the Baltic Sea!) of the system (10) will change: in [7], [12] the Sjoberg model had two positive stationary points, one of them was stable and, therefore, it described the real condition of oxygen in the Baltic Sea in the 70's of the XX-th century.

Remark 6. In the model (10), (11) units of modelling – tons of phosphorus (t Phosph.) per twenty-four hours and cubic kilometer of oxygen (km3), namely, measuring units of functions $\overline{u}_{inor.phosph.}(t)$, $\underline{u}_{inor.phosph.}(t)$, $\overline{u}_{or.phosph.}(t)$, $\underline{u}_{or.phosph.}(t)$, $\overline{u}_{ph.pl.}(t)$ are t Phosph/(twenty-four hours), but measuring units of function $\underline{u}_{oxyg.}^{0}$ are km^3 O_2.

III CONCLUSIONS

In the present paper there is proposed a complex of continuous qualitative models having dynamic parameters for analysis, evaluation and forecast of aquatic (water reservoirs, lakes and the Baltic Sea) ecosystem status in Latvia. All the constructed models are described in terms of differential equations and mathematical physics and are considered to be evolutional models. Below it is given a short description of the models proposed. The first model is aimed at discovering the pollutant concentration dynamics in the Baltic Sea. This model can be used as the primary model for the better monitoring of the Baltic Sea. The second model is designed to determine the basic characteristics of nitrogen circulation in the natural aquatic environments. This model allows to reduce substantially (in comparison with the other models constructed to this date and known to the authors of this paper) the uncertainty in the parameters of the sophisticated process of circulation in the inhomogenous media, which is typical for natural aquatic environment with powerful sources and sinks. Let us note that it is possible to consider minimizing the number of uncertain model parameters issue by (a) identifying relationships imposed on the parameters, (b) finding non-trivial special points of the system-model, and taking into account those points found in a consequent way for the several structures of the circulation processes following the conditions of stability maintenance. In further studies authors plan to carry out the abovementioned minimization for the uncertain model parameters number. The third is a mathematical model that allows to determine the sensitivity of oxygen condition in the Baltic Sea. This model provides a possibility to (a) identify those features of oxygen condition dynamics, which are not revealed during the regular enumeration and/or using pure statistical methods of parameters evaluation; (b) to identify qualitative properties of the oxygen condition dynamics, which reduces significantly the uncertainty in the complex interrelations between dynamics parameters and the environment, making it possible to forecast the anthropogenic impact on the ecosystem with a high degree of accuracy, and, therefore, to identify existing trends in the ecosystem of the sea. However, it should be noted that this model cannot be used for the study of the seasonal Baltic Sea ecosystem dynamics in the conditions of aerobic and anaerobic states being below the halocline level, and this is regarded to be a limitation of the model. For this reason, third model can be used only as a primary

model for identification of condition that leads to anaerobic zones growth strengthening in the Baltic Sea.

To conclude with, let us note that authors of this paper are intended to continue the further investigation taking the benefit of both qualitative and quantitative studies for all three models as well as to develop the stable analytical and numerical methods for their solution ensuring the corresponding computer-based implementation.

IV ACKNOWLEDGMENTS

The present article was executed within the framework of the following two Research Projects: 1) The State Research Programme "NextIT": *Next generation Information and Communication Technologies* (for both authors); 2) The European Regional Development Fund Project No. 2014 / 0029 / 2DP / 2.1.1.1 / 14 / APIA / VIAA / 088: *"Development of an experimental long flight distance unmanned aerial vehicle prototype for multi-purpose environmental monitoring"* (for the first author).

V REFERENCES

[1] G. I. Marchuk, *Mathematical modelling in environmental problem*. Moscow: Science, 1982.

[2] J. Jennifers, *Introduction to system analysis: application for ecology*. Moscow: World, 1981.

[3] Yu. P. Zaitsev, and G. G. Polikarpov, "Ecological processes in the critical areas of the Black Sea: synthesis of the results of two research dimensions from the middle of the XX to the beginning of XXI centuries." Marine Ecological Journal, Vol. 1, Issue 1, pp. 33-55. 2002.

[4] J. Hofbauer, and K. Sigmund, *The theory of Evolution and dynamical systems*. London: Cambridge University Press, 1988.

[5] L. M. Nedostup, "Sensitivity of the water ecosystem models that are affected by the anthropogenic factor." in *Natural waters preservation, protection and quality improvement problems*. Moscow: Science, 1982, pp. 139-155.

[6] B. Bolin, "Model studies of the Baltic Sea." University of Stockholm, Institute of Meteorology, Ambio Special Report (Report GH-4), No. 1, 1972. pp. 115-119.

[7] S. Sjoberg, "A mathematical and conceptual framework for models of the pelagic ecosystems of the Baltic Sea." University of Stockholm, Asko Laboratory, Report No. 1, 1980. 206 p.

[8] Yu. M. Barabasheva, L. I. Brodsky, and G. I. Devyatkova, "About evaluation of parameters in the point model of aquatic ecosystem: Theoretical ecology." Moscow: Lomonosov Moscow State University publishing, 1987, pp. 105-110.

[9] Sh. E.Guseynov, E. A. Kopytov, and O. V. Schiptsov, *Mathematical Models of an exhaust concentration dynamics in urban atmosphere*. Riga: Transport and Telecommunication Institute Press, 2010.

[10] J. S.Rimshans, I. N. Esau, S. S. Zilitenkevich, and Sh. E. Guseynov, "Analytical-Numerical Solution for the One Dimensional PBL Turbulence Model." Proceedings of the 18th Symposium on Boundary Layers and Turbulence under the aegis of the American Meteorological Society, Stockholm, Sweden June 09-13, 2008. http://ams.confex.com/ams/pdfpapers/139877.pdf.

[11] Yu. N. Sergeyev, "Problem of mathematical modeling of multicomponent physico-biological marine system." Herald of the Leningrad State University, Issue: 24, pp. 114-125, 1972.

[12] S. Sjoberg, F. Wulff and P. Wahllstrom, "Computer Simulations of Hydrochemical and Biological processes in the Baltic." University of Stockholm, Asko Laboratory, Report No 1, 1972. 180 p.

[13] L. F. Serdiutskaya, "Study of mathematical models of environmental systems using multivariate factor analysis." International Journal on Engineering Simulation, Vol. 17, pp. 417-428, 2000.

[14] L. F. Serdyutskaya, "About some aspects of factorial analysis application for the problems of environmental simulation." in *Modelling and diagnostics of the sophisticated processes and systems*, Kiev: Scientific thaught, 1997, pp. 35-40.

[15] V. I. Belyaev, Ye. D. Korenyuk, and V. K. Hrusch, *Computer modelling of the subsurface water circulation and pollution dynamics*. Dnepropetrovsk: Science and Education, 2001.

[16] The Environmental Performance Index 2014. Yale Center for Environmental Law & Policy, Yale University 2015. [Online]. Available: http://epi.yale.edu/epi/country-profile/latvia [Accessed: May 10, 2015].

[17] HELCOM Initial Holistic Assessment. Ecosystem Health of the Baltic Sea in 2003-2007. Baltic Sea Environment Proceedings, No. 122, 2008, 66 p.

[18] Latvian Ministry of Environment. Environmental Policy Strategy in 2009–2015. Informative section. Riga, 2009, 53 p.

[19] Sh. E.Guseynov, J. S. Rimshans, and E. A. Kopytov, "Solution of the Model of Exhaust Concentration Dynamics in Urban Atmosphere under Unknown Turbulent Air Flow Velocity." International Journal of Procedia Environmental Sciences, Series: Urban Environmental Pollution, 2011, No 4, pp. 35-42, 2011.

[20] S. I. Kuznetsov, *Micro flora of the lakes and its geochemical activity*. Moscow: Science, 1970.

[21] W. Eppley, E. U. Renger, and E. L. Uenrick, "A study of plankton dynamics and nutrient cycling in the central gyre of the North Pacific Ocean." Journal on Limnology and Oceanography, Vol. 18, Issue 4, pp. 534-555, 1973.

[22] Baltic Marine Environment Protection Commission: Helsinki Commission. Baltic Sea Environment Proceedings, No. 16, 1986, 176 p.

[23] Baltic Marine Environment Protection Commission: Helsinki Commission. First Periodic assessment of the State of the Marine Environment of the Baltic Sea area, General Conclusions. Baltic Sea Environment Proceedings, No. 17A, 1986, 56 p.

[24] V. B. Georgievsky, "Identification and verification of the water ecosystem models." in *Natural waters preservation, protection and quality improvement problems*. Moscow: Science, 1982, pp. 156-163.

[25] V. B. Georgievsky, "Identification and mathematical modeling of the eutrophication processes of the sea ecosystems." Ambio Special Report, Vol. 5, pp. 165-176, 1983.

[26] J.Hofbauer, and K. Sigmund, *The theory of Evolution and dynamical systems*. London: Cambridge University Press, 1988.

[27] V. I. Arnold, *Additional chapters of the ordinary differential equations theory*. Moscow: Science, 1978.

V. A. Yakubovich, and V. M. Starshinsky, *Linear differential equations with periodic coefficients and their applications*. Moscow: Science, 1972.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 100-108

On one Approach for Stable Estimate of Technical System Efficiency

Sharif E. Guseynov[1, 2, a], Alexander I. Urbah[2, b], Sergey A. Andreyev[2, c]
[1]Institute of Mathematical Sciences and Information Technologies, Liepaja University,
14 Liela Street, Liepaja LV-3401, Latvia;
[2]Institute of Aeronautics, Faculty of Transport and Mechanical Engineering,
Riga Technical University,
*1A Lomonosov Street, Building-1, Riga LV-1019, Latvi*a
[a]Sh.E.Guseinov@inbox.lv, [b]Aleksandrs.Urbahs@rtu.lv, [c]S.A.Andreyev@gmail.com

Abstract. **In present paper the problem of efficiency evaluation of technical system by measurable structural design parameters is investigated. To accomplish the purpose of considered problem the mathematical model is constructed in the form of a finite-dimensional operator equation, where desired elements are both influence weights of the calculated structural design parameters and technical effectiveness indicator of the system. First, the constructed model is reduced to the normal system, and then the apparatus of the ill-posed inverse problem theory is used for the reduced problem: a regularizing operator is constructed and an algorithm for finding the regularization parameter is developed.**

Keywords: **Efficiency evaluation of technical systems, ill-posed and inverse problems, mathematical model, regularizing algorithm.**

I INTRODUCTION

The technical system according to one of numerous definitions (see, for instance, [1], [2]) is a set of elements of the system (mechanisms, assemblies, components and similar components of the system) that interact with each other during the process of performing the specified functions. In the named definition it is worth to emphasize that the concepts "the technical system" and "the system components" can be mutually expressed one by the other, and depending on the purpose of the research, the required accuracy, the information level about the reliability and the similar factors and characteristics, the considered / introduced concrete concept "the technical system" for the certain problem-target can be converted into "the system element" for another problem-target. However, in contrast to the above given definition of the technical system, in this paper we distinct the "system" and "elements" concepts and we try to identify, first, the level of influence (in other words: the importance / significance / weight / weight of influence) of possible defects in each element of the technical system on the operability of all other elements (with or without possible defects in these elements) of the same technical system; second, the resulting / cumulative impact of possible defects in all elements of the given technical system on the operation of the whole investigated technical system. Here, anticipating things, it is necessary to note the following:

- if such mutual influence of possible defects in the elements of the technical system are missing or negligible then the investigation results of the constructed mathematical model, which describes the considering problem of the defects mutual influence, should reflect this fact;
- the technical system possible defects in all elements resulting / cumulative influence on the operating of the whole studied technical system is not a simple sum of the effects of these defects, and, moreover, may not be a linear combination (with constant coefficients) of the effects of these defects.

The aforementioned considerations lead us to the following two natural questions:

1. What is the measure of influence of each element possible defects of the given technical system on each of all the other technical system elements performance? Is it possible to define these individual measures of defects mutual influence in cases when there is only given a set of calculated / controllable parameters measurement values of the system?
2. If the answer to the first question, which has been formulated above, is positive, then is it

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.191

possible to stably (in the sense of Tikhonov: see, for instance, [3]) find out the effectiveness of the technical system under investigation without any additional assumptions and information?

Before attempting to answer these two questions, let us briefly discuss the general scheme of the technical systems effectiveness quantitative evaluation. It should be emphasized that speaking about the technical system effectiveness usually is meant the effectiveness of its use as an active mean for typical operations, for which purposes the system is intended. If the considered technical system is a multi-purpose, and if such a multipurpose technical system for each objective has a high efficiency, it means that such a system has wide functionality, i.e. it has qualities that reflect the ability of the system to meet the challenges associated with achieving each goal. In applied researches, the effectiveness of technical systems usually has to solve the following two problems: (1) the problem of choice, which essence is to evaluate operations effectiveness with the use of technical system; (2) the problem of choice, which essence lies in the choice of a rational strategy for the use of technology in the operation. The problem of effectiveness is closely related to the reliability problem in technical systems see, for instance, [4], [5]). The complexity growth of technical systems reduces their reliability and, therefore, reduces their effectiveness. Unreliable technical system cannot be an effective way of achieving the goal. The operation effectiveness analysis provides an approach to determine the required parameters of the technical system reliability that are used as active means in the operation. This is one of the important aspects of the reliability and effectiveness problems interconnection. Another aspect of the relationship of these problems is to determine the rational ways of use for the available resources to improve the reliability of the developing technical systems.

The essence of this work – to give an unequivocal answer for the formulated above two important issues without requiring additional information about the technical state of system elements. As it will be shown the obtaining of such a clear answer is done by transformation of the technical system efficiency indicator finding task to the inverse problem, which has the form of the first kind finite-dimensional operator equation, and a subsequent development of a regularized method for its sustainable solutions.

II GENERAL SCHEME OF EFFICIENCY EVALUATION OF TECHNICAL SYSTEMS AND METHODS OF WEIGHTING COEFFICIENTS

There are many methods for the quantitative evaluation of the effectiveness of technical systems (see, for instance, [1], [4], [5]), the majority of which are operation research methods: the criterion of efficiency and constraints (significant and sign constraints) are made; the output effect and costs (or the result of the operation of the technical system) are determined, wherein, the costs can be determined at the production or design stages of the considered technical system, as well as during operation of the system. However, all these methods have a common scheme, which can be represented as the functional

$$E[y;t] \overset{def}{\equiv} F(y_1,...,y_m;C)(t), \tag{1}$$

where $m \in \mathbb{N}$ is a number of elements in the technical system; $y_i = f_i(x_1,...,x_n;t)$ is a technical indicator of the system i-th $\left(i=\overline{1,m}\right)$ element, which is also called the i-th partial quality indicator; $n \in \mathbb{N}$ is a number of controllable / calculated (mainly structural-constructive) parameters of the technical system; $\{x_j\}_{j=\overline{1,n}}$ are controllable / calculated parameters of the system; t is time; C is total expenses (in other words, the total cost) for the development, design and operation of the system.

As is clear from (1), the variation in methods for finding the efficiency $E[y;t]$ is caused by specific types of functions $f_i(x_1,...,x_n;t)$ $\left(i=\overline{1,m}\right)$ and F. Namely, constructing the function $f_i(x_1,...,x_n;t)$ $\left(i=\overline{1,m}\right)$ and F in different ways, we obtain different formulas for calculating E, at the same time, if designed function $f_i(x_1,...,x_n;t)$ $\left(i=\overline{1,m}\right)$ and F will have an analytical form, then the calculation of $E[y;t]$ is not difficult, since the controllable / calculated parameters of the investigated technical system are known for each case / state of the system; if the explicit form of the functions is not known, then in order to calculate $E[y;t]$ are mainly used statistical methods or sometimes only evaluation of $E[y;t]$ according to one of the most important quality partial indicators $y_{imp.}$ $\left(imp \in \{1;...;m\}\right)$, and over the other partial quality indicators are placed constraints so that they do not exceed certain limits:

$$\begin{cases} E[y;t] = y_{imp.}, \\ \underline{y_i} \le y_i \le \overline{y_i} \quad \left(i=\overline{1,m} / \{imp\}\right), \end{cases} \tag{2}$$

where $\underline{y_i}$ and $\overline{y_i}$ are respectively the lower and upper limits of i-th $\left(i=\overline{1,m}\right)$ partial quality indicator.

The effectiveness evaluation of the technical systems according to the rule (2) has the disadvantage – the solution of the corresponding optimization problem (i.e., the same problem (2) with the criterion $E[y;t] = y_{imp.} \to \sup$ or the problem of implementation choice for practical optimization technical system version will be ambiguous, since the

criterion choice is ambiguous: you can get a lot of options for systems with the same or almost the same main partial quality indicator $y_{imp.}$ with significantly different other partial indicators that satisfy the constraints; therefore, it cannot be guaranteed that the determined version of the system will be closer to the optimum.

One of the widely used and well-studied methods for the effectiveness estimation of technical systems is the method of weighting coefficients, in which the function F is represented as a linear combination of functions $y_i \ \left(i=\overline{1,m}\right)$:

$$\begin{cases} E\left[y;t\right] \overset{def}{\equiv} \sum\limits_{i=1}^{m} \lambda_i \cdot y_i, \\ \lambda_i \in \Lambda \overset{def}{\equiv} \left\{ \lambda_i \in \mathbb{R}_+^1 : \sum\limits_{i=1}^{m} \lambda_i = 1 \right\} \left(i=\overline{1,m}\right), \\ \underline{y_i} \le y_i \le \overline{y_i} \ \ \left(i=\overline{1,n}\right), \\ \underline{y_i} \overset{def}{\equiv} \inf\limits_{x;t} f_i\left(x_1,...,x_n;t\right), \\ \overline{y_i} \overset{def}{\equiv} \sup\limits_{x;t} f_i\left(x_1,...,x_n;t\right) \ \ \left(i=\overline{1,m}\right). \end{cases} \tag{3}$$

In (3) weighting coefficient $\lambda_i \in \Lambda$ is called the coefficient importance / significance of the i -th $\left(i=\overline{1,m}\right)$ partial quality indicator and generally speaking, is the desired number; set Λ is called the set of importance or preference. Without going into details, we list the well-known, the main advantages and disadvantages of the method of weighting coefficients (3), the final destination of which is the choice of the best option among the alternatives in the development, design or operation of complex technical systems. As the merits may be listed the following: ease of formalization; the possibility of taking into account all the main partial quality indicators; opportunity to take into account the preferences of the decision maker about the problem during the weighting coefficients setup; clear physical / technical / economic sense. The main drawbacks the following may be indicated: subjectivity in weighting coefficients; the accountability lack of dependence on the weighting coefficients values of partial quality indicators.

In conclusion of this section we note that the independent objectives are both determination of the weighting coefficients of partial quality indicators of the technical system and the problem of finding / evaluation of these partial system indicators. Currently, there are a lot of analytical, analytical and numerical, logical, statistical, sets and graphs based, expert, etc. methods, which are successfully used, for solving these independent problems. Exhaustive information about the basic and the most common these methods, as well as a fairly complete overview of them with reasonable instructions of strengths, weaknesses and areas of application can be found in [5], [6]. Exhaustive information on the design of experiments in technics / technology and processing of the obtaining experimental data can be found in [7]-[9]. Therefore, in this paper we will not deal with any problem of finding controllable / calculated structural design parameters $\left\{x_j\right\}_{j=\overline{1,n}}$ of technical system, or partial quality indicators estimation problem $\left\{y_i\right\}_{i=\overline{1,m}}$ of technical elements of the system, and in the following sections of this paper, we assume that all of them are known to us aposteriori.

III MATHEMATICAL FORMULATION OF THE PROBLEM

We consider the functional

$$E\left[y;t\right] \overset{def}{\equiv} \sum_{i=1}^{n} \lambda_i \cdot y_i, \tag{4}$$

appearing in the method of weighting coefficients (3). Obviously, the formula (4) to calculate the efficiency of technical systems from the known information about the partial quality indicators $y_i \ \left(i=\overline{1,n}\right)$ of system elements is unable to determine which of the controllable/calculated structural design parameters of technical system in a given period of time has a significant impact on partial indicators of the elements of the technical system and, consequently, the common efficiency indicator of the system. Furthermore, according to this formula it is impossible to determine how different is the influence of the same structural design parameters to different parts of the system, which partial quality indicators (consequently, their reliability) significantly differ from each other or, conversely, very similar. In other words, the formula (4) does not distinguish the influence measure of each controlled parameter of the system on common efficiency indicator of the system, both in time and in the given set of controlled technical parameters for each element of the system. Therefore, the partition of the elements of technical system under consideration by their influence both on the partial quality indicators of elements, and the common efficiency of the whole system is impossible. Obviously, the disadvantage of the formula (4) is removed if for each element of the system we will take into account each controllable technical parameter with its "individual influence" – the weight, which will depend on both the time t and on each controlled/calculated indicator (index j) of each element (index i) of considered technical system. In other words, instead of the method (3) we propose to consider the method (let us call it the method of individual weighting coefficients), in which

the functional of efficiency indicator of technical system has the form

$$E[i;t] \overset{def}{\equiv} C_1 \cdot \sum_{j=1}^{n} w_{ij}(t) \cdot \hat{x}_{ij}(t) + C_2, \qquad (5)$$

where $n \in \mathbb{N}$ is a number of controllable/calculated structural design parameters of technical system; $m \in \mathbb{N}$ is the elements number of the technical system; for each fixed $i \left(i = \overline{1,m}\right)$ and $t \in [T_{start}, T_{end}]$ the parameters $\left\{\hat{x}_{ij}(t),\ j = \overline{1,n}\right\}$ is a result of standardization by means of any method applied to initial controllable structurally design parameters of technical system, for example, by the formula (for instance, see the fundamental books [10], [11]) $\hat{x}_{ij}(t) = \left(x_{ij}(t) - m_j\right)/\sigma_j$, through m_j is denoted the average deviation of the j-th indicator and by σ_j is denoted the standard deviation of the j-th indicator; $[T_{start}, T_{end}]$ is time interval during which the technical system effectiveness is studied and the structural design parameters values of technical system are fixed (locked); T_{start} and T_{end} are respectively the initial and final moments of the time interval; constants C_i $(i = 1;2)$ are chosen in such a way that they are, respectively, the centers of two clusters of technical system elements with the worst and the best values of structural and design parameters, wherein, if the clusters overlap, it is possible to apply FCM (the fuzzy c-means clustering algorithm) or PCM (the possibilistic c-means algorithm) algorithm (see, for instance, [12]); for each i-th $\left(i = \overline{1,m}\right)$ element of technical system the desired weight of the j-th $\left(j = \overline{1,n}\right)$ controllable / calculated structural design parameter into the common effectiveness indicator of the entire technical system at every fixed time $t \in [T_{start}, T_{end}]$ denoted as $\left\{w_{ij}\right\}_{i=\overline{1,m}}^{j=\overline{1,n}}$. Coefficients $\left\{w_{ij}\right\}_{i=\overline{1,m}}^{j=\overline{1,n}}$ we call the weights of individual influence.

Thus, in the proposed formula (5) the numbers $n \in \mathbb{N}$, $m \in \mathbb{N}$, $T_{start} \in \mathbb{R}^1_+$, $T_{end} \in \mathbb{R}^1_{++}$, $\left\{\hat{x}_{ij}(t) \in \mathbb{R}^1_+\right\}_{i=\overline{1,m}}^{j=\overline{1,n}}$ and $C_i \in \mathbb{R}^1$ $(i = 1;2)$ are known data; the weight of individual influence $\left\{w_{ij}\right\}_{i=\overline{1,m}}^{j=\overline{1,n}}$ and the common effectiveness indicator $E[i;t]$ of technical system are sought. It is obvious that in formula (5) there are exactly $m \cdot (n+1) \cdot T_{s \to e}$ (here $T_{s \to e}$ means the length of the time interval $[T_{start}, T_{end}]$ measured in an integer number of conventional units) unknown $w_{ij}(t)$ and $E(i,t)$, which should be uniquely defined via $\left(m \cdot n \cdot T_{s \to e} + 2\right)$ initial data $x_{ij}(t)$, *A, B,* connected by exactly $m \cdot T_{s \to e}$ linear algebraic equations.

Indeed, the individual influence coefficient $\left\{w_{ij}\right\}_{i=\overline{1,m}}^{j=\overline{1,n}}$ are time-dependent, i.e. $w_{ij} = w_{ij}(t)$ for $\forall i \in \{1,2,...,m\}$ and $\forall j \in \{1,2,...,n\}$, since the influence measure of the same structural design parameter of the considered technical system on the partial component of the same element of the technical system at different time may vary. In other words, generally speaking, $w_{i_{fix} j_{fix}}(t_1) \neq w_{i_{fix} j_{fix}}(t_2)$ for $t_1 \neq t_2$. However, taking in formula (5) the stationary weights in the individual influence, we are thus considering the case when the weights are different only relatively n controlled / calculated structural design parameters of technical system, and are constants in time $t \in [T_{start}, T_{end}]$ and relatively elements $i = \overline{1,m}$. Therefore, the introduced functional (5) under these assumptions is equivalent to functional $E[i;t] = C_1 \cdot \sum_{j=1}^{n} w_j \cdot \hat{x}_{ij}(t) + C_2$. Further, for the time interval $[T_{start}, T_{end}]$ constructing a discrete mesh in the form of $t_k \in [T_{start}, T_{end}]$: $t_k = T_{start} + (k-1) \cdot \tau$, $k = \overline{1,K}$; $\tau = (T_{end} - T_{start})/(K-1)$; $K \in \mathbb{N} / \{1\}$, and then introducing the notation $\hat{x}_{ikj} \equiv \hat{x}_{ij}(t_k)$ and $E_{ik} \equiv E[i;t_k]$ for $i = \overline{1,m}$; $j = \overline{1,n}$; $k = \overline{1,K}$; we obtain the following final analytical formula for the functional in the proposed method of the individual weighting coefficients:

$$E_{ik} = C_1 \cdot \sum_{j=1}^{n} w_j \cdot \hat{x}_{ikj} + C_2. \qquad (6)$$

So, summarizing the abovementioned and abstracting from the subject area of the considered problem, we can formulate the following mathematical problem: it is required to find a stable solution of a linear algebraic equations system (6) relatively $(m \cdot K + n)$ unknown $\left\{E_{ik}\right\}_{i=\overline{1,m}}^{k=\overline{1,K}}$ and $\left\{w_j\right\}_{j=\overline{1,n}}$ for $(n \cdot m \cdot K + 5)$ known initial data $\left\{\hat{x}_{ikj} \in \mathbb{R}^1_+\right\}$ $\left(i = \overline{1,m};\ k = \overline{1,K};\ j = \overline{1,n}\right)$; $n \in \mathbb{N}$; $m \in \mathbb{N}$; $K \in \mathbb{N} / \{1\}$; $C_i \in \mathbb{R}^1$ $(i = 1;2)$. Obviously, the system (6) is not a normal system, and therefore, the concept of its "solution" should be clarified.

IV REDUCTION OF THE PROBLEM TO THE INVERSE PROBLEM

Let us introduce an augmented matrix $\tilde{X}$ with dimensions $(m\cdot K)\times(m\cdot K+n)$, which of elements $\{\tilde{x}_{ikj}\}$ are calculated by the following rule:

$$\tilde{x}_{ikj} \overset{def}{\equiv} \begin{cases} -1 \ if \ (j>n)\wedge(K\cdot i+k=j+(K-n)); \\ 0 \ if \ (j>n)\wedge(K\cdot i+k\neq j+(K-n)); \\ C_1\cdot\hat{x}_{ikj} \ if \ j\le n. \end{cases} \quad (7)$$

Then the system (6) has the form:

$$\sum_{j=1}^{m\cdot K+n}\tilde{x}_{ikj}\cdot\tilde{w}_j=-C_2 \ \ \forall\left(i=\overline{1,m};\ k=\overline{1,K}\right), \quad (8)$$

where

$$\tilde{w}_j \overset{def}{\equiv} \begin{cases} w_j \ if \ j\le n; \\ E_{ik} \ if \ j=n+(i-1)\cdot K+k, i=\overline{1,m},\ k=\overline{1,K}. \end{cases}$$

Introducing the notation $\tilde{W}\equiv\{\tilde{w}_j\}_{j=\overline{1,(m\cdot K+n)}}$; $\tilde{C}\equiv\left(\underbrace{-C_2,...,-C_2}_{m\cdot K}\right)^T$, the system (8) becomes

$$\tilde{X}\tilde{W}=\tilde{C}. \quad (9)$$

Thus, the original problem (6) is reduced to the problem (9), where we want to define the $(m\cdot K+n)\times 1$ size vector $\tilde{W}$ under the known $(m\cdot K)\times(m\cdot K+n)$ size matrix $\tilde{X}$ and the $(m\cdot K)\times 1$ size vector $\tilde{C}$. The system (9) consists of $m\cdot K$ equations with $(m\cdot K+n)$ unknown $\tilde{w}_j$, i.e. number of unknown is larger exactly by n values than the amount of equations. In other words, the system (9) is an underdetermined system of linear algebraic equations. Hence, the question arises: what should we mean by a solution of the system (9)? We introduce the following concept.

Definition 1. Let us call the element $\tilde{W}$ as a generalized solution of underdetermined system (9) if

$$\tilde{W}_{g.s}=\arg\min_{\tilde{W}\in\mathbb{R}^{m\cdot K+n}}\left\|\tilde{X}\tilde{W}-\tilde{C}\right\|^2_{\mathbb{R}^{m\cdot K}}, \quad (10)$$

where the norm in a Euclidean space $R^{m\cdot K}$ is understood as a Schur norm: $\|y\|_{\mathbb{R}^{m\cdot K}}=\sqrt{\sum_{j=1}^{m\cdot K}y_{ij}^2}$.

It is obvious that in order to find the generalized solutions of the system (9), according to the definition 1, it is required to solve the unconditional optimization problem $\left\|\tilde{X}\tilde{W}-\tilde{C}\right\|^2_{\mathbb{R}^{m\cdot K}}\to\min_{\tilde{W}\in\mathbb{R}^{m\cdot K+n}}$. For this purpose, we construct the Euler equation (for instance, see [3]) for the functional $V(\tilde{W})\overset{def}{\equiv}\left\|\tilde{X}\tilde{W}-\tilde{C}\right\|^2_{\mathbb{R}^{m\cdot K}}$ and then we equate it to zero: $\delta V\overset{def}{=}2\cdot\left\langle\Delta W,\tilde{X}^T\tilde{X}\tilde{W}+\frac{1}{2}\cdot\tilde{X}^T\tilde{X}\Delta W-\tilde{X}^T\tilde{C}\right\rangle=0.$ Hence, due to the arbitrariness of the increment ΔW, we will have

$$\tilde{X}^T\tilde{X}\tilde{W}=\tilde{X}^T\tilde{C}, \quad (11)$$

The finite-dimensional operator equation (11) is the Euler equation for the unconditional extremum problem $\left\|\tilde{X}\tilde{W}-\tilde{C}\right\|^2_{\mathbb{R}^{m\cdot K}}\to\min_{\tilde{W}\in\mathbb{R}^{m\cdot K+n}}$. The operator equation (11), in contrast to the system (9), is a normal system: there are exactly $(m\cdot K+n)$ linear algebraic equations and the same number of unknowns $\{\tilde{w}_j\}_{j=\overline{1,(m\cdot K+n)}}$. So, we see that the generalized solution $\tilde{W}_{g.s}$ of the system (9) is a classical solution of the normal system of linear algebraic equations (11), and vice versa. Since the matrix $\tilde{X}$, elements of which are calculated by the formula (7), is strongly sparse matrix, the determinant of the principal matrix $\tilde{X}^T\tilde{X}$ can be arbitrarily small, i.e. $\det(\tilde{X}^T\tilde{X})\ll 1$ or even $\det(\tilde{X}^T\tilde{X})=0$. Consequently, it is wrong to solve the system (11) by any direct methods, i.e. the desired solution $\tilde{W}$ of the system (11) cannot be expressed by the formula $\tilde{W}=(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{C}$, from which according to the well-known Cauchy inequality $\left\|\tilde{W}\right\|\le\left\|(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\right\|\cdot\left\|\tilde{C}\right\|$ implies stability of the problem (9) solution by $\tilde{C}$ and $\tilde{X}$ in the sense of a generalized solution. In addition to the abovementioned problem of ill-conditioned system (11) (this causes the instability of its solution), violation of the condition $\det(\tilde{X}^T\tilde{X})\neq 0$ also violates the Hadamard correctness (see, for instance, [3], [13], [14]) of the system (9) in the sense of a generalized solution, namely, the generalized solution, as the solution of the normal system (11), may be non-unique, and then is determined up to elements of the kernel $\ker\tilde{X}$ of main matrix $\tilde{X}$. Therefore, in this case if $\tilde{W}^{(1)}$ is a generalized solution of the system (9), then each column-vector $\tilde{W}=\tilde{W}^{(1)}+\tilde{W}^{(2)}$, where $\tilde{W}^{(2)}$ is an element of the kernel $\ker\tilde{X}$, i.e. $\tilde{W}^{(2)}\in\ker\tilde{X}$, will also be a generalized solution of the system (9), moreover, all these solutions can arbitrarily differ from each other. Therefore, there should be formulated solution selection criterion. The following concept gives the criterion:

Definition 2. The generalized solution with minimal Schur norm we call the normal generalized solution. In addition, the pseudosolution of the system $Az=u\ \left(z\in\mathbb{R}^n;\ u\in\mathbb{R}^m;\ A=\{a_{ij}\}_{i=\overline{1,m}}^{j=\overline{1,n}}\right)$ is called (the

pseudosolution concept first introduced in the works [15], [16]; a vector $z^{(\text{pseudo})} \in \mathbb{R}^n$, which minimizes the residual $\|Az-u\|$ on the entire space $\mathbb{R}^n$. Since the system $Az=u$ may have more than one pseudo-solution, then we can talk about the set F_A of pseudosolutions $Az=u$. The normal solution of the system $Az=u$ respecting to the vector $z^{(1)} \in \mathbb{R}^n$ is called (for instance, see [3]) a pseudosolution $z^{(0)} \in \mathbb{R}^n$ with minimal norm $\|z-z^{(1)}\|$, i.e. such that $\|z^{(0)}-z^{(1)}\| = \inf\limits_{z\in F_A} \|z-z^{(1)}\|$. For simplicity we can always assume that $z^{(1)} \equiv 0$, and then the normal solution with respect to vector $z^{(1)} \in \mathbb{R}^n$ can be simply called the normal solution. It is important to emphasize that for any system $Az=u$ the normal solution always exists and is unique. It is appropriate to note that the normal solution of the system $Az=u$ can also be defined as a pseudo-solution that minimizes a given positive definite quadratic form in the coordinates of the vector $z-z^{(1)}$. Since the set of generalized solutions $\left\{\tilde{W}_{g.s} \in \mathbb{R}^{m\cdot K+n}: \ \left\|\tilde{X}\tilde{W}_{g.s}-\tilde{C}\right\|^2_{\mathbb{R}^{m\cdot K}} = \min\limits_{W\in\mathbb{R}^{m\cdot K}} \left\|\tilde{X}\tilde{W}-\tilde{C}\right\|^2_{\mathbb{R}^{m\cdot K}}\right\}$ is a compact set (i.e. convex and closed set), then by the Weierstrass theorem, there exists a unique element $\tilde{W}_{g.s}$ of this compact set, which has the least distance from zero, i.e. there exists a unique element $\tilde{W}_{g.s}$ with the smallest norm. In other words, the normal generalized solution is unique, and it is uniquely determined. It is known, such a solution of the system (9) can be found using the pseudoinverse Moore-Penrose matrix: $\tilde{W}_{g.s} = \tilde{X}^{+}\tilde{C}$, where through $\tilde{X}^{+}$ is denoted the pseudoinverse Moore-Penrose matrix with the dimensions $(m\cdot K+n)\times(m\cdot K)$, and is given by

$$\tilde{X}\tilde{X}^{+}\tilde{X} = \tilde{X}, \tag{12}$$

or by theoretical asymptotic formula

$$\tilde{X}^{+} = \lim_{\varepsilon\to 0}\left(\tilde{X}^T\tilde{X}+\varepsilon\cdot I\right)^{-1}\tilde{X}^T, \tag{13}$$

where through I is denoted an identity matrix.

It should be emphasized that both the relation (12) and the asymptotic formula (13) are inconvenient for practical finding of $\tilde{X}^{+}$. Relatively easy formula for finding practical pseudoinverse matrix $\tilde{X}^{+}$ is the following formula:

$$\tilde{X}^{+} = \tilde{X}_1^{+}\tilde{X}_2^{+} = \tilde{X}_1^T\left(\tilde{X}_1\tilde{X}_1^T\right)^{-1}\left(\tilde{X}_2\tilde{X}_2^T\right)^{-1}\tilde{X}_2^T, \tag{14}$$

where $\tilde{X} = \tilde{X}_1\tilde{X}_2$ is an ambiguous skeleton decomposition of the matrix $\tilde{X}$, at that $\tilde{X}_2$ has the dimension $(m\cdot K)\times r$, but $\tilde{X}_1$ has the dimension $r\times(m\cdot K+n)$, where $r = rang\tilde{X} = rang\tilde{X}_1 = rang\tilde{X}_2$. At different skeletal decompositions we always obtain the same normal generalized solution for pseudoinverse matrix $\tilde{X}^{+}$.

Now let us ask whether the normal generalized solution is stable? Since the normal generalized solutions $\tilde{W}_{g.s}$ satisfies the inequality $\left\|\tilde{W}_{g.s}\right\| \le \left\|\tilde{X}^{+}\right\|\cdot\left\|\tilde{C}\right\|$, then it is obvious, the normal generalized solution is stable for variations of right side of the system (9). However, the construction of pseudoinverse matrix $\tilde{X}^{+}$ by formulas (12)-(14) implies that the solution (9) in the sense of the normal generalized solution does not become stable for arbitrary variations of the main matrix $\tilde{X}$ of system (9). Therefore, if the elements of matrix $\tilde{X}$ initially and/or during the computer calculations may start perturbing (for studied in this paper problem such disturbances are quite typical, because the original data $\tilde{x}_{ikj}\ \left(i=\overline{1,m};\ j=\overline{1,n};\ k=\overline{1,K}\right)$, calculated according to the formula (7), are the results of experimental measurements of controlled/calculated structural design parameters of considered technical system, and the measurement results always have errors), then the method of Moore-Penrose does not give a true solution of the system (9). Consequently, the problem of finding a stable solution of a finite dimensional operator equation (11), and thus, a stable solution of the system (9), is still open, and it is required to construct the regularized algorithms for the normal generalized solution finding. Note that if the equation (11) is solvable in a classical way, then its normal generalized solution is called a normal solution.

Thus, the obtained problem (11) refers to the theory of inverse problems. As already mentioned above, we cannot take every found element $\tilde{W}$ from (11) as an exact or approximate solution. It is intuitively clear that there is needed the possible solutions selection principle. It usually requires the use any available additional information about the solution. Such information may have qualitative or quantitative nature, and the desire to use quantitative information allows select a compact set $M \subset F$, where initially Hadamard ill-posed problem becomes well-posed problem (according to Tikhonov). In other words, the use of additional quantitative information eliminates the instability of the inverse problem (11) solution.

V THE CONSTRUCTION OF THE REGULARIZED SOLUTION ALGORITHM

Thus we consider the problem

$$\tilde{X}^T\tilde{X}\tilde{W} = \tilde{X}^T\tilde{C}, \ \ \tilde{W}\in F;\ \tilde{X}^T\tilde{C}\in U, \tag{15}$$

where the inverse operator $\left(\tilde{X}^T\tilde{X}\right)^{-1}$ does not exist, or exists, but, generally speaking, is not a continuous

operator on the image of the set F when mapping it using the operator $\tilde{X}^T\tilde{X}$, i.e. on the set of $\tilde{X}^T\tilde{X}F$. Furthermore, the problem (15) is essentially incorrect, i.e. in the equation (15) small change of the right-hand side $\tilde{X}^T\tilde{C} \in U$, which is related to its approximate nature, may lead outside the set $\tilde{X}^T\tilde{X}F$.

Thus, in (15) the set F of possible solutions of the equation (15) is not compact. Suppose that it is a priori known that the function $\tilde{W}_{ex}$ is a normal generalized solution of equation (15) at the precisely given right-hand side part $\tilde{X}^T\tilde{C} = \left(\tilde{X}^T\tilde{C}\right)_{ex}$, i.e. $\tilde{X}^T\tilde{X}\tilde{W}_{ex} = \left(\tilde{X}^T\tilde{C}\right)_{ex}$. If in (15) instead of the precisely given right-hand side part $\left(\tilde{X}^T\tilde{C}\right)_{ex}$ we know its approximation $\left(\tilde{X}^T\tilde{C}\right)_{\delta}$, as well as the value δ, that characterizes the closeness degree of $\left(\tilde{X}^T\tilde{C}\right)_{\delta}$ to $\left(\tilde{X}^T\tilde{C}\right)_{ex}$, the problem is to find an approximate normal generalized solution $\tilde{W}_{\delta}$ to the exact solution $\tilde{W}_{ex}$, having the stability towards small changes $\left(\tilde{X}^T\tilde{C}\right)_{\delta}$ property by the known initial data $\left\{\left(\tilde{X}^T\tilde{C}\right)_{\delta}; \delta\right\}$. As it has been already mentioned in the previous section of this work, for an approximation of the normal generalized solution $\tilde{W}_{\delta}$ of equation (15) we cannot use the exact solution of the approximate equation $\tilde{X}^T\tilde{X}\tilde{W}_{\delta} = \left(\tilde{X}^T\tilde{C}\right)_{\delta}$. In other words, the vector $\tilde{W}_{\delta} = \left(\tilde{X}^T\tilde{X}\right)^{-1}\left(\tilde{X}^T\tilde{C}\right)_{\delta}$ is not an approximate normal generalized solution of equation (15) with the approximate right-hand side $\tilde{X}^T\tilde{C} = \left(\tilde{X}^T\tilde{C}\right)_{\delta}$: $\rho_U\left(\left(\tilde{X}^T\tilde{C}\right)_{ex}, \left(\tilde{X}^T\tilde{C}\right)_{\delta}\right) \le \delta$, since the vector $\tilde{W}_{\delta}$ exists for not any arbitrary element $\tilde{X}^T\tilde{C} \in U$. Furthermore, as it was shown in the previous section of this work, the normal generalized solution lacks stability property to small changes in the right hand side $\tilde{X}^T\tilde{C}$ of equation (15). The value $\delta \ge \rho_U\left(\left(\tilde{X}^T\tilde{C}\right)_{ex}, \left(\tilde{X}^T\tilde{C}\right)_{\delta}\right)$ characterizes the error of the equation right hand side part of (15). Therefore, obviously, it is necessary to determine an approximate normal generalized solution $\tilde{W}_{\delta}$ of equation (15) using an operator (action/technique), which depends on the numerical parameter $\delta \ge \rho_U\left(\left(\tilde{X}^T\tilde{C}\right)_{ex}, \left(\tilde{X}^T\tilde{C}\right)_{\delta}\right)$. And the values of this operator must be taken according to the error δ, namely, the consistency must be such that the approaching in the metric space U of the right hand side part $\left(\tilde{X}^T\tilde{C}\right)_{\delta}$ of equation (15) to the exact value $\left(\tilde{X}^T\tilde{C}\right)_{ex}$, the found (by means of an operator, which depends on a parameter δ) approximated normal generalized solution $\tilde{W}_{\delta}$ approaches to the desired exact normal generalized solution $\tilde{W}_{ex}$ of the equation $\tilde{X}^T\tilde{X}\tilde{W}_{ex} = \left(\tilde{X}^T\tilde{C}\right)_{ex}$ in the Schur metric F. Thus, the main task – the problem of the equation approximate normal generalized solution finding (15), resistant to minor changes in the right-hand side part, – reduces to the solution of the following two "side" problems: the first "side" problem is to find a regularizing operator for the equation (15) relatively to the element $\left(\tilde{X}^T\tilde{C}\right)_{ex} = \tilde{X}^T\tilde{X}\tilde{W}_{ex}$; the second "side" problem is to determine the regularization parameter α by additional information about the problem, for example, the error value, which is given for the right-hand side $\left(\tilde{X}^T\tilde{C}\right)_{\delta}$ (in this case $\alpha = \alpha(\delta)$).

Let us discuss now these "sides" problem. First of all we assume that a regularizing operator exists: we denote it by $R = \{r_{km}\}$. Then, as a solution to the original problem (15) the element $\tilde{W} = R\tilde{X}^T\tilde{C}$ may be chosen. Let us construct a regularizing operator R. For this purpose the algorithm, enables finding the inverse $\left(\tilde{X}^T\tilde{X}\right)^{-1}$ matrix, is based on the regularization method idea by academician A.Tikhonov ([3]). Another regularization algorithm, which has proved itself to be suitable enough while solving various both linear finite-dimensional operator equations of a first kind, and linear infinite-dimensional operator equations of a first kind is presented below. The algorithm consists of the following steps:

Step 0. The initial data are given $n \in \mathbb{N}$; $m \in \mathbb{N}$; $\left\{\hat{x}_{ikj} \in \mathbb{R}^1_+\right\}$ $\left(i = \overline{1,m};\ k = \overline{1,K};\ j = \overline{1,n}\right)$; $C_i \in \mathbb{R}^1$ $(i = 1;2)$ in equation system (6).

Step 1. The origin system (6) is reduced to a system (15).

Step 2. The sequence of parameters, for example, $\alpha_l = 2^{-l}$ $(l = 1,2,3,...)$, is taken and for the two (eg., the first two) adjacent parameters $\alpha = \alpha_{l_0}$ and $\alpha = \alpha_{l_0+1}$ of this sequence is solved the following system equations relatively to $(m \cdot K + n)^2$ unknowns r_{li}^{α}:

$$\alpha \cdot r_{li}^{\alpha} + \sum_{j=1}^{m\cdot K+n} r_{lk}^{\alpha} \cdot \beta_{ji} = \left(\tilde{X}^T\tilde{X}\right)_{il} \quad \forall\left(i, l = \overline{1,(m\cdot K+n)}\right),$$

where $\beta_{ji} \overset{def}{\equiv} \sum_{p=1}^{m\cdot K+n} \left(\tilde{X}^T\tilde{X}\right)_{jp} \cdot \left(\tilde{X}^T\tilde{X}\right)_{ip}$.

Step 3. For each of the parameters $\alpha = \alpha_{l_0}$ and $\alpha = \alpha_{l_0+1}$ the coordinates $\left\{\tilde{w}_l^{\alpha}\right\}_{l=\overline{1,(m\cdot K+n)}}$ of the column-vector $\tilde{W}^{\alpha}$ are found by formula

$$\tilde{w}_l^{\alpha} = \sum_{i=1}^{m\cdot K+n} r_{li}^{\alpha} \cdot \left(\tilde{X}^T\tilde{B}\right)_i \quad \left(l = \overline{1,(m\cdot K+n)}\right).$$

Step 4. The condition $\left\|\tilde{W}^{\alpha_{l_0}} - \tilde{W}^{\alpha_{l_0+1}}\right\|_{\mathbb{R}^{m\cdot K+n}} \le \varepsilon_0$ is verified, where ε_0 is the desired accuracy (for example, $\varepsilon_0 = 10^{-3}$) of the approximate normal generalized solution regarding the exact solution (which is unknown). If the answer is positive, then we finish the algorithm: the approximate normal generalized solution $\tilde{W}^{\alpha_1} \approx \tilde{W}^{\alpha_2}$ is found and, therefore, the desired $w_j \left(j = \overline{1,n}\right)$ and $\left\{E_{ik}\right\}_{i=\overline{1,m}}^{k=\overline{1,K}}$ in the initial equation (6) are found, wherein we have $w_j = \tilde{w}_j$ for $j \le n$, and for $\forall\left(i = \overline{1,m};\ k = \overline{1,K}\right)$ such that $j = n + (i-1)\cdot K + k$, we have $E_{ik} = \tilde{w}_j$ for $j = \overline{(n+1),(m\cdot K+n)}$. If $\left\|\tilde{W}^{\alpha_{l_0}} - \tilde{W}^{\alpha_{l_0+1}}\right\|_{\mathbb{R}^{m\cdot K+n}} > \varepsilon_0$, then we move to the step 2, where we assume $\alpha_{l_0} = \alpha_{l_0+1}$ and $\alpha_{l_0+1} = \alpha_{l_0+2}$. The given theoretical reasoning of the papers [17]-[19] guarantees the algorithm is finite, and the result of the algorithm will be a normal generalized solution of the system (6).

Remark. The desired optimal regularization parameter $\alpha = \alpha_{opt}$ can be found by means of qualitatively different method – generalized residual method (see, for instance, see [20]), which is more versatile and accurate method, but also requires more complex calculations. A qualitatively new approach to find quasi-optimal regularization parameter, which combines simplicity and accuracy, is in detail described (with many appendices) in the monographs [13], [14]).

VI CONCLUSIONS

In this paper we investigate the problem of the effectiveness estimation of the technical systems from the measurable structural design parameters. A fundamentally new approach, where, in contrast to the traditional approaches, is no need to determine in advance the kind of partial quality indicators of the technical system and to find the weights of these partial indicators (the weights determination of particular indicators is subjective and ambiguous: hence, it cannot be guaranteed to identify which option of the destination choice weights provides optimal effectiveness of the studied technical system). The proposed in this paper approach is based on the apparatus of the inverse and ill-posed problem theory, namely, first, a mathematical model is built, which is the underdetermined finite-dimensional operator equation with respect to unknown influence weights of the calculated structural design parameters, as well as the desired effectiveness indicator of the technical system; the built model is reduced in a special way to the normal system of algebraic equations with approximate initial data; the parametric regularizing operator is built for a stable solving of the obtained normal system – this operator, in case of an appropriate parameter choice (which is called the Tikhonov regularization parameter), guarantees the stability of the approximate solution found for the normal system; the fairly simple to implement and efficient algorithm of finding the regularization parameter is offered for the constructed regularizing operator; a closed formula determines the calculated structural design parameters weights of the technical system and, thus, the effectiveness rate of the technical system under study.

It should also be noted that the proposed approach eliminates one of the main drawbacks of a well-known and frequently used method of weighting coefficients (this method is the main method for multicriterion linear programming problems solving) – the accountability lack of the weighting coefficients dependence on the partial quality indicators values of the considered technical system.

VII ACKNOWLEDGMENTS

The present article was executed within the framework of the following two Research Projects: 1) The State Research Programme "NextIT": *Next generation Information and Communication Technologies* (for the first co-author); 2) The European Regional Development Fund Project No. 2014 / 0029 / 2DP / 2.1.1.1 / 14 / APIA / VIAA / 088: *"Development of an experimental long flight distance unmanned aerial vehicle prototype for multi-purpose environmental monitoring"* (for the first and the second co-authors).

VIII REFERENCES

[1] M. A. Yastrebenetskii, and G. M. Ivanova, *Reliability of Automated Control Systems for Engineering Processes*. Moscow: Energoatomizdat, 1989.

[2] V. S. Avduyevsky, V. I. Kuznetsov, N. D. Kuznetsov, V. A. Melnikov, V. P. Mishin, V. F. Utkin, K. V. Frolov, B. V. Gnedenko, I. N. Kovalenko, and B. F. Lomov, *Reliability and Efficiency: Methodology. Organization. Terminology*. Moscow: Engineering Industry, 1986.

[3] A. N. Tikhonov, and V. Ya. Arsenin, *Solutions of Ill-Posed Problems*. Moscow: Science, 1986.

[4] B. V. Gnedenko, Yu. K. Belyayev, and A. D. Solovyov. *Mathematical Methods of Reliability Theory. Principal Characteristics of Reliability and Statistical Analysis*. Moscow: URSS, 2013.

[5] I. A. Ushakov, Yu. K. Belyaev, V. A. Bogatirov, and V. V. Bolotin, *Reliability of Thechnical Systems*. Moscow: Radio and Communication, 1985.

[6] V. G. Totsenko, *Methods and Systems for Decision-Making Support*. Kiyiv: Naukovo Dumka, 2002.

[7] V. V. Rykov, and V. Yu. Itkin. *Mathematical Statistics and Planning an Experiment*. – Moscow: Press of The Gubkin Russian State University of Oil and Gas, 2009.

[8] J. Antony, *Design of Experiments for Engineers and Scientists*. London: Elsevier Science & Technology Books, 2003.

[9] R. L. Mason, R. F. Gunts, and J. L. Hess, *Statistical Design and Analysis of Experiments with Applications to Engineering and Science*. New Jersey, USA: WILEY-Interscience, 2003.

[10] S. Ch. Albright, W. Winston, and Ch. Zappe, *Data Analysis and Decision Making*. Mason, USA: South-Western Cengage Learning, 2010.

[11] P. P. Biemer, and L. E. Lyberg, *Introduction to Survey Quality*. New Jersey, USA: Wiley-Interscience, 2003.

[12] C. Hwang, and F. C-.H. Rhee, "Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means." IEEE Transaction on Fuzzy Systems, Vol. 15, Issue 1, pp. 107-120, 2007.

[13] S. M. Yunusov, V. P. Labendik, and Sh. E. Guseynov, *Monitoring and Diagnostics of Aircraft Gas Turbine Engines: Improvement of Models and Methods for Diagnosis of Gas Path of Gas Turbine Engines*. Saarbrücken, Germany: Lambert Academic Publishing, 2014.

[14] S. A. Andreyev, and Sh. E. Guseynov, *Regularizing Algorithms for Diagnosing: Applied to Gas Turbine Engines in Operation*. Saarbrücken, Germany: Lambert Academic Publishing, 2013.

[15] V. K. Ivanov, "About linear ill-posed problems." Herald of The USSR Academy of Sciences, Vol. 145, Issue 2, pp. 270-272, 1962.

[16] V. K. Ivanov, "About ill-posed assigned problems." Mathematical Collections, Vol. 61, Issue 2, pp. 211-213, 1963.

[17] Sh. E. Guseynov and V. I. Dmitriev, "Investigation of resolving power and solution detailedness of inverse problems of magnetotelluric sounding." Herald of the Lomonosov Moscow State University, Series 15: "Computational Mathematics and Cybernetics", Vol. 1, pp. 17-25, 1995.

[18] Sh. E. Guseynov and M. Okruzhnova, "Choice of a quasi-optimal regularization parameter for the first kind operator equations." Transport and Telecommunication, Vol. 6, Issue 3, pp. 471-486, 1995.

[19] Sh. E. Guseynov and S. M. Yunusov, "New regularizing approach to determining the influence coefficient matrix for gas-turbine engines." in *Dynamical Systems, Differential Equations and Applications*. Vol. I, American Institute of Mathematical Sciences (AIMS), 2011, pp. 614-623.

[20] V. A. Morozov, *Regularization Methods for Ill-Posed Problems*. Florida, USA: CRC Press, 1993.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 109-113

Modeling of time dependent thermal process in sliding electrical microcontact

Alexander Ilyin, Igor Plokhov, Igor Savraev, Oksana Kozyreva
Pskov State University, Electromechanical Faculty,
Department of Electric Drive and Automation Systems.
Address: Lenina 8, Pskov, 180000, Russia

Abstract. **Temperature has great influence on mechanical, electrical and chemical processes that occur in transition layer of sliding contact. The aim of the research is creating a three-dimensional finite-element model for calculating time dependent thermal process in sliding electrical microcontact.**

The article starts with the description of the contact element representing a discrete microcontact and physics in it. Then the authors suggest use modern simulation software COMSOL Multyphisics for modeling.

The developed 3D-model uses modules of electric currents, heat transfer in solids, and electromagnetic heat source for computing. For each module the assumptions, the initial and boundary conditions are made.

The outcomes of modeling are the transient processes of average overheat in the elements of contact-details surface layers. The transient processes depend on geometric size of the microcontact (size of contact element, height of surface elements, thickness of oxide films, and overlap), contact-details physical properties (density, electrical conductance, thermal conductivity, and heat capacity), external influences (electrical current and friction heat), and temperatures of the neighbor elements.

The results of the research will be used in the numerical simulation model of sliding electrical contact.

Keywords: **sliding electrical contact, simulation model, time dependent thermal process.**

I INTRODUCTION

Electrical contacts can be found at a variety of practical application. Good performance of the brush slip rings has particular importance for large-power electro-generators. The research is concerned with the modeling and the simulation of processes in the sliding electrical contact. Basic concepts of the processes in sliding electrical contacts and algorithms of the simulation model are given in [1, 2] that was based on researches in [3–13].

There are different modeling methods. Computer simulation is an important feature in engineering systems or any system that involves many processes. It can be used to estimate the performance of systems too complex for analytical solution. The scale of events being simulated by computer simulations has far exceeded anything possible using traditional mathematical modeling.

Computer simulation is the imitation of the operation of a real-world process or system over time using a single computer or a network of computers. The act of simulating requires that a model representing the key characteristics and behaviors of the system be developed. The model represents the system itself, whereas the simulation represents the operation of the system over time.

The developing simulation model of sliding electrical contact is:

- stochastic, as it uses random number generators to model random events;
- stencil code, because stores data in regular grids and require only next-neighbor access.

For some past time several algorithms of the simulation model were improved: microscale surfaces generation, computation of the contact surfaces overlapping, evaluation of the constriction resistance of the contact spots cluster, and other [14, 15]. The physicochemical properties of the contact materials depend on temperature varying within wide range. Therefore thermal processes have significant influence on other physical processes: electrical, mechanical, and chemical. Thus we have to propose a new improved model for calculating time dependent thermal process in sliding electrical contact. In this work we will develop a computational model of heat conduction in a microcontact, run test computing experiments, and discuss how to use the results in the simulation model.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.194

II MATERIALS AND METHODS

Model scheme

To solve transient heat conduction in the microcontact we designed the model of the contact element (Fig. 1). The contact element is composed of two interacting surface elements. Each surface element is a hemisphere of radius $dx/2$ based on a dx square face of a rectangular parallelepiped with the height $h - dx/2$, where dx is the size of the contact element and h is the height of the surface element. The hemispheres overlap on the distance δ depending on heights of the surface elements and the applied load. In the overlapping volume is located a heat source P_{fr} equal to the power of friction in the sliding contact. An electric current I flows through the contact element. The hemisphere consists of two domains with different values of electric conductivity. In the domain $r \le dx/2 - f$ the conductivity is equal to the contact material conductivity. In the domain $dx/2 - f < r \le dx/2$ the conductivity is equal to the material of oxide film. There f is the thickness of the film. In figures and expressions we will use index *br* (brush) for one contact and index *r* (rotor) for the other.

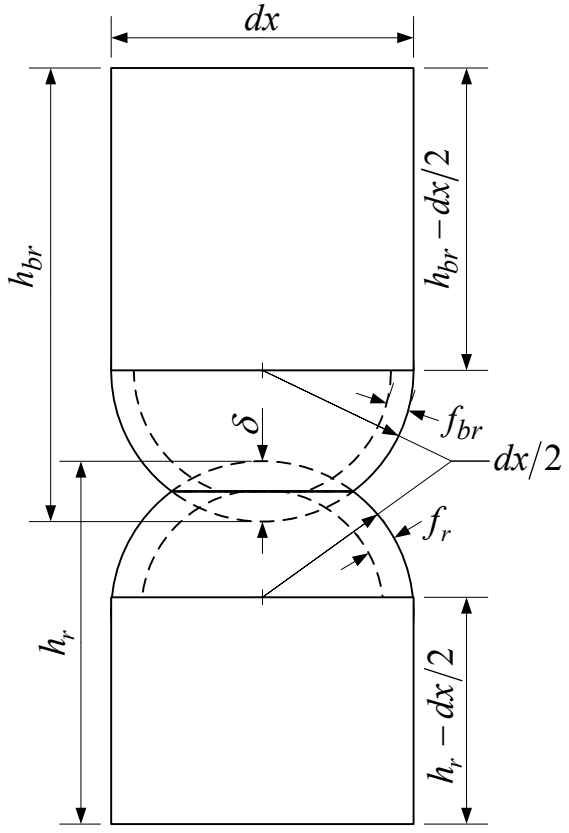

Fig. 1. Scheme of the contact element

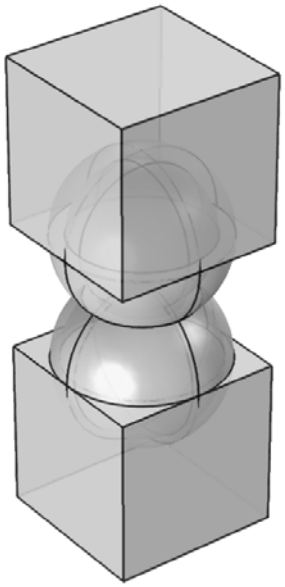

Fig. 2. Three-dimensional view of the contact element

It is necessary to determine: the time dependence of the average temperature of the surface elements, heat fluxes between the surface elements, heat fluxes to the neighbor surface elements, and heat fluxes to the contact bodies. View of the contact element in three-dimensions is shown on Fig. 2.

Using the model of contact element we developed computational model in COMSOL Multiphysics (Fig. 3). The computational model includes the neighbor and body elements for each surface element. The model uses modules: Electric Currents, Heat Transfer in Solids, and Electromagnetic Heat Source.

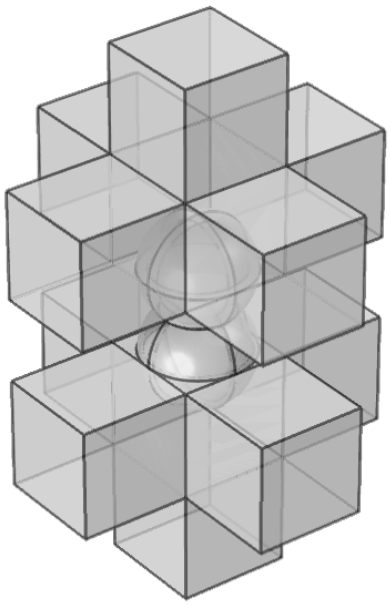

Fig. 3. COMSOL Multiphysics model

Initials and boundaries

The model represents closed system and the external surfaces are thermally insulated from the environment.

The surface element of the brush has the initial overheat θ_{br}. We will use the term overheat for the difference between the temperature of an element and the environment temperature. The initial overheats of the four neighbor elements are θ_{br1}, θ_{br2}, θ_{br3}, θ_{br4}. The most top cubic element represents the body of the brash and has the initial overheat θ_{br0}. For the rotor surface element the initial overheats in the same way are θ_r, θ_{r0}, θ_{r1}, θ_{r2}, θ_{r3}, θ_{r4}.

In the overlapping domain of the rotor and brush hemispheres is situated the heat source with total power P_{fr} that depends on friction ratio, sliding velocity, and applied load.

The physics Heat Transfer in Solids is applied to all of the model domains. The physics uses for calculation equation

$$\rho C_p \frac{\partial T}{\partial t} + \rho C_p u \cdot \nabla T = \nabla \cdot (k \nabla T) + Q, \qquad (1)$$

where ρ is density, C_p is heat capacity at constant pressure, and k is thermal conductivity for an element material; T is temperature field (K), u is velocity field (m/s), Q is heat source (W/m^3) [16].

The module Electric Currents uses physics Current Conservation for calculation of all domains with equations

$$\begin{aligned} &\nabla \cdot J = Q_j, \\ &J = \left(\sigma + \varepsilon_0 \varepsilon_r \frac{\partial}{\partial t}\right) E + J_e, \\ &E = -\nabla V, \end{aligned} \qquad (2)$$

where σ is electrical conductivity of a material; J is current density (A/m^2), Q_j is boundary current source (A/m^3), E is the electric field strength (V/m), J_e is external current density (A/m^2), V is the electric potential (V).

The initial electric potential value of the elements is 0 V. The boundary electric potential of the most bottom side of the rotor body cubic element also is 0 V. For the most top surface of the brush body cubic element is specified the normal current density J defined as

$$J = \frac{I}{dx^2}. \qquad (3)$$

All of the other external surfaces are electrically isolated from the environment.

The multiphysics module Electromagnetic Heat Source uses Joule heating as addition to Q in the equation (1).

Assumptions and properties of materials

In this section we will describe assumptions made in the model that may affect on the accuracy of the model.

All of the neighbour elements have the same height as the brush surface element and the rotor surface element.

In reality the size of the brush and the rotor is about three orders more than the contact element size dx. Therefore the properties of materials for the cubic elements of the contact bodies are specified in ratio due to units

$$\rho' = \rho\left[\frac{\text{kg}}{(10^{-3}\,\text{m})^3}\right] = 10^9 \rho\left[\frac{\text{kg}}{\text{m}^3}\right], \qquad (4)$$

$$C_p{}' = C_p\left[\frac{\text{J}}{\text{kg}\cdot\text{K}}\right], \qquad (5)$$

$$k' = k\left[\frac{\text{W}}{10^{-3}\,\text{m}\cdot\text{K}}\right] = 10^3 k\left[\frac{\text{W}}{\text{m}\cdot\text{K}}\right], \qquad (6)$$

$$\sigma' = \sigma\left[\frac{\text{S}}{10^{-3}\,\text{m}}\right] = 10^3 \sigma\left[\frac{\text{S}}{\text{m}}\right]. \qquad (7)$$

For the oxide films domains the electrical conductivity σ is equal to zero, since the electrical conductivity of metal oxides is several orders less than for metals and it can be neglected.

Experimental procedure

For the test computational experiment the parameters provided in Table 1 are used.

TABLE 1.
PARAMETERS OF THE TEST EXPERIMENT

#	Parameter	Value
1	dx	8 μm
2	h_{br}	11 μm
3	h_r	13 μm
4	f_{br}	1.5 μm
5	f_r	0.8 μm
6	δ	1.9 μm
7	ρ_{br}	2500 kg/m^3
8	C_{pbr}	500 J/(kg·K)
9	k_{br}	40 W/(m·K)
10	σ_{br}	0.02 S/m
11	ρ_r	6500 kg/m^3

12	C_{pr}	250 J/(kg·K)
13	k_r	75 W/(m·K)
14	σ_r	33.333 S/m
15	P_{fr}	0.8 mW
16	I	10 mA
17	θ_{br}	0 K
18	θ_{br0}	1 K
19	θ_{br1}	0 K
20	θ_{br2}	0 K
21	θ_{br3}	3 K
22	θ_{br4}	0 K
17	θ_r	0 K
18	θ_{r0}	–1 K
19	θ_{r1}	0 K
20	θ_{r2}	0 K
21	θ_{r3}	0 K
22	θ_{r4}	2 K

The modelling time t is 10 μs with 100 steps each of the interval $dt = 10^{-7}$ s.

III RESULTS AND DISCUSSION

The next figures show some results of the test experiment.

Fig. 4 presents the plot of the temperature distribution in the cut yz-plane for x-coordinate is 0 on the last step of modeling $t = 10$ μs. As we can see the brush element has the temperature spot in the contact zone caused by the friction and Joule heating. The brush element overheat more than in the rotor element mostly due to lesser value of the thermal conductivity.

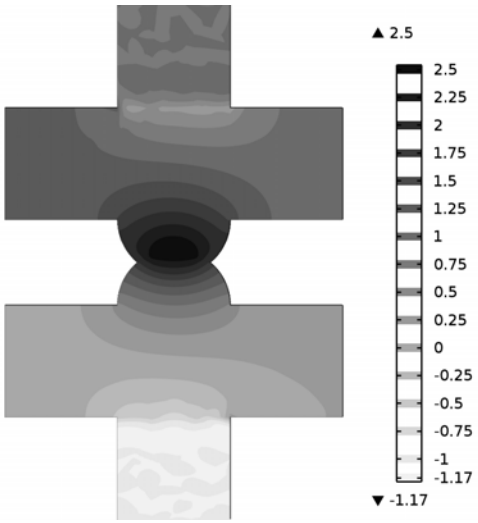

Fig. 4. Temperature distribution in yz-slice

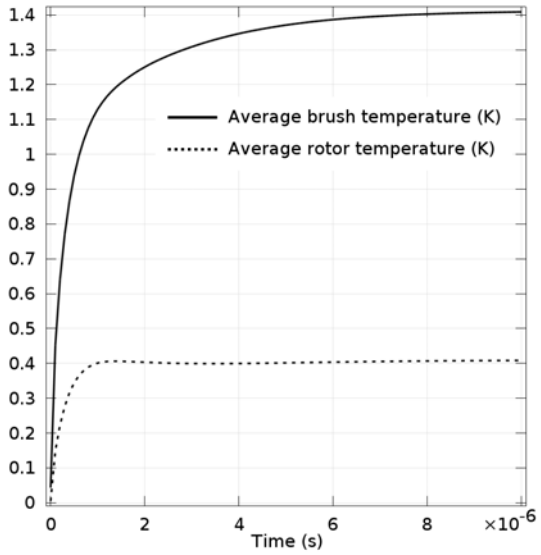

Fig. 5. The transient processes of the average overheats of the brush and the rotor elements

Fig. 5 shows the transient processes of the average temperatures in the brush and rotor elements.

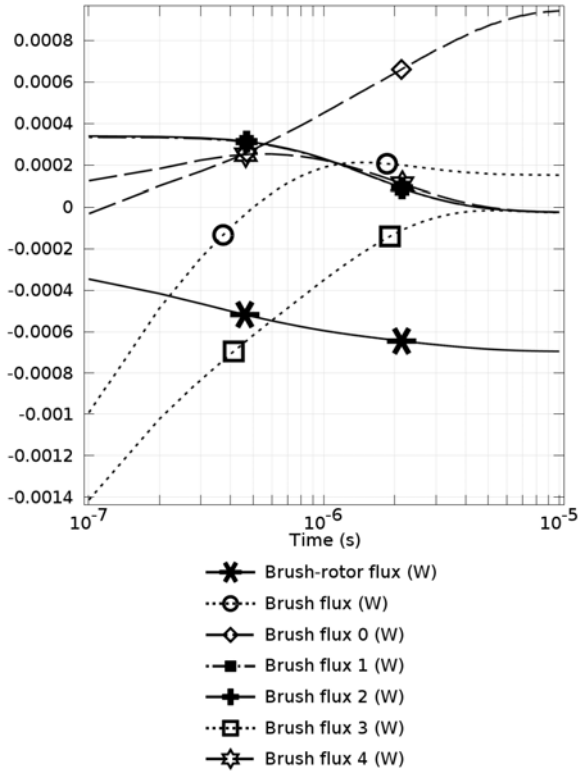

Fig. 6. The transient processes of the heat fluxes from the brush element to the neighbor elements

Fig. 6 illustrates the transient processes of the heat flux from the brush element to the neighbor elements. The brush flux (dotted line with circles) is the total flux from the brush element to the others. For better

understanding the trends on Fig. 6 are shown on a logarithmic scale.

The results of this and other test experiments show that all of the transient processes may be approximated using second-order aperiodic elements.

IV CONCLUSION

This paper describes a three-dimensional computational model in COMOSL Multiphysics for calculating non-steady-state thermal processes in microcontact depending on the parameters of the contact element and the material properties of contacts. The outcomes of the model will be used in the simulation model of sliding electrical contact. Since the transient processes depend on many parameters (see Table 1) so we have to use experimental design theory for planning set of experiments.

V REFERENCES

[1] И. В. Плохов, *Комплексная диагностика и прогнозирование технического состояния узлов скользящего токосъёма турбогенераторов.* Диссертация доктора технических наук. СПб: СПбГПУ, 2001.

[2] A. Ilyin, I. Plokhov, A. Isakov, "The simulation model of a sliding contact", *ENVIRONMENT. TECHNOLOGY. RESOURCES: Proceedings of the 9th International Scientific and Practical Conference* June 20–22, 2013. Volume II. Rezekne, Latvia, 2013, pp. 111–115.

[3] Р. Хольм, *Электрические контакты.* – М.: Иностранная литература, 1961.

[4] J. A. Greenwood, "Constriction resistance and the real area of contact", *British Journal of appl. Physics.* 1966. V.17. P.1621–1631

[5] Milenko Braunovic, Nikolai K. Myshkin, Valery V. Konchits. *Electrical Contacts: Fundamentals, Applications and Technology.* – CRC Press, 2010. – 672 p.

[6] В. В. Кончиц, В. В. Мешков, В. В. Мышкин, *Триботехника электрических контактов.* Минск: Наука и техника. 1986.

[7] А. В. Чичинадзе, Э. Д. Браун, Н. А. Буше и др. *Основы трибологии (трение, износ, смазка)*: Учебник для технических вузов. 2-е изд. переработ. и доп. / Под общ. ред. А. В. Чичинадзе. – М.: Машиностроение, 2001. – 664 с.

[8] Е. Федер, *Фракталы*: Пер. с англ. – М.: Мир, 1991. – 254 с.

[9] Д. Б. Горохов, *Контактное взаимодействие фрактальных шероховатых поверхностей деталей машин*: Диссертация на соискание ученой степени кандидата технических наук. — Братск, 2005.

[10] К. Н. Войнов, В. А. Ходаковский, М. А. Шварц, «Математическое моделирование шероховатых поверхностей». *Трение, износ, смазка*. – 2009. – №41.

[11] В. Т. Омельченко В. Т, *Теория процессов на контактах.* – Харьков: Вища школа. Изд-во при Харьк. ун-те, 1979. – 128 с.

[12] Ф. Крейт, У. Блэк, *Основы теплопередачи* / Пер. с англ. М.: Мир. 1983.

[13] M. F. Gomez, *Characterisation and modelling of brush contacts.* – Humburg, 2005.

[14] А. В. Ильин, И. В. Плохов, О. И. Козырева, «Моделирование процессов электрофрикционного взаимодействия в узлах скользящего токосъема» *Научно-технический вестник Поволжья.* – Казань: Научно-технический вестник Поволжья, 2013. – №4. – С. 166–173. ISSN 2079-5920

[15] А. В. Ильин, И. В. Плохов, О. И. Козырева, «Моделирование микрорельефа поверхностей контактирующих деталей» *Научно-технический вестник Поволжья.* – Казань: Научно-технический вестник Поволжья, 2013. – №5. – С. 180–183. ISSN 2079-5920

[16] *COMSOL Multiphysics Reference Manual*, Version 4.4, November 2013.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 114-120

Multi factor authentication as a necessary solution in the fight with information technology security threats

Anita Jansone, Kaspars Lauris, Ivars Šaudinis
Liepaja University, Faculty of Science and Engineering,
Address: Liela iela 14, Liepaja, LV-3401, Latvia

Abstract. **In the publication multi factor authentication solutions are offered as a necessary tool for decrease of information technology security risks. The work includes description of authentication process from the viewpoint of information technology security aspect, as well as authentication factors are described, which can be used in authentication process modules. Some recommendations for decrease of security risks are given using multi factor authentication solutions. In the work a multi-factor authentication security testing experiment is described, which involves use of Linux remote console - SSH service. Analysis of data of unauthorised access efforts obtained during tests is described: it is determined from which countries or regions, as well as in which days there is the highest threat to information technology security.**

Keywords: **authentication, authentication security testing, authentication factor, authentication security risk, identification, password.**

I INTRODUCTION

Nowadays in the world of information technologies static authentication data - passwords or the same in combination with other authentication factors are less and less used in newer authentication solutions. Use of static passwords in authentication process involves high security risk. Even if a user of a system composes a "secure" password, there are no guarantees that the same user will not use this password in another system, where the security level is low, thus putting other systems to risk. As Jonathan Klein, the president of a well-known company *Usher*, which deals with mobile identification solutions for enterprise needs, says *„I think the password is going the way of the dinozaur"* [11]. He believes that use of passwords will come to the end one day because there are more modern and secure authentication technologies being able to substitute the use of passwords. Multi factor authentication solutions are offered to decrease authentication security risks. In case of multi factor authentication, even if the lawbreaker obtains your passwords it is impossible to access your data in the system, because he does not have the code card assigned to you by the bank or a code calculator or your mobile phone, where an application generates a dynamic password, which is an additional authentication factor.

There have been lots of publications, where IT security experts make qualitative statements about the fact that multi factor authentication solutions significantly decrease security risk. But there are not any publication providing quantitative data for proving those statements. Therefore, it was decided to perform multi factor authentication security tests and obtain quantitative data about multi factor authentication security risk, providing and proving the hypothesis - *if the process of authentication is organized in several stages, and at each stage different authentication factors are used, then it is possible to decrease authentication security risk.*

To determine what authentication solutions it is possible to create, it is important to find out what authentication data or factors it is possible to use in the authentication process.

II AUTHENTICATION SOLUTIONS AND FACTORS

Identification is a quite simple process. A person should submit own data to the system and the system can launch authentication and authorisation processes. To submit data to the system it is enough to enter only a user name or to scan a fingerprint etc. Without identification the system cannot associate authentication factors with the person. During authentication the person's identity is checked by comparing one or several factors to the information about this person, kept in the data base. Authentication data used to check person's identity is the information of restricted access. Ability of the person and the

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.183

system to keep authentication data confidential reflects the security level of a certain system.

Identification and authentication processes always happen as a single two-step process [2]. User identification is always done at the first step followed by authentication at the second step. Without having completed both processes the person cannot get access to the system.

From the viewpoint of information technology security authentication processes can have one-stage, two-stage and multi-stage authentication mechanisms [3]. At each stage the authentication process can be static, dynamic or biometric depending on the authentication data used in the authentication process.

The number of stages and factors used in the certain authentication process depends on the authentication solution. In its turn, the authentication solution depends on how important or sensitive are the data in the system, where user authentication needs to be performed. It is important to use different authentication data or factors at each stage of the multi-stage authentication process.

The authentication data used at each stage can be put into several categories [4]:

1. something you **know** – password, personal identification number - PIN;
2. something you **have** – code card, code calculator, smart card, mobile device etc.;
3. something you **are** – user biometric data, for example, voice, fingerprints, etc.

*A. „Something you **know**" authentication factor*

As it was mentioned before, there are three factor groups. The first of them corresponds to the condition „something you **know**". This authentication factor group is formed by passwords, PIN, phrase or cognitive passwords. However, this authentication factor group contains only static authentication data. Insecure passwords create the main authentication risk, but if the person composes a safe password, it is a powerful authentication factor, which decreases the security risk [4].

Companies usually elaborate information system security policy due to this is defined by legal standards. Such document is elaborated by an information system security expert, and should be certified and approved by the head of institution. Usually such documents also include password composition and application policies. The password composition policy determines the frequency of password changes, the length of passwords and the complexity of password composition algorithm. Furthermore, it is also important to keep in memory all the composed passwords and to implement prohibition in the system for users to use only two passwords all the time, by changing them periodically. Information system users usually choose phrase-type passwords. Such passwords are easier to remember, but are more difficult to guess by using *„brute-force"* methods. They perform the same functions as traditional passwords, and usually are easier to remember because users put some meaning into them. Phrase-type passwords are simple sentences with some modifications. For example – the phrase „Neviens nav ideāls", where „s" is substituted by „$" and „a", or „ā" by „@", and as the result a phrase-type password is obtained – „Nevien$N@vIde@l$".

Another interesting password composition mechanism is used by cognitive passwords [2]. Usually cognitive passwords are formed as a number of questions, which can be answered only by a certain person. For example - "What is your date of birth?", "What is your mother's name?", "What is your pet's name?" etc. The most effective way would be answering a series of such questions during authentication, however, that would significantly delay the authentication process. Therefore cognitive passwords are usually used in case users forgot their passwords. In this case, by providing the correct cognitive password a user can renew the usual password.

*B. „Something you **have**" authentication factor*

The second authentication factor group corresponds to the condition "something you have". This means that the user has a device, which either contains identified authentication data or is able to generate them. Such devices are called "smart-cards" or "tokens", or just security system devices. However, such devices are prone to risk - they can be stolen, lost or duplicated.

Smart-cards have integrated microprocessors and memory, where one or several certificates are stored. The certificates contain subject identification and/or authentication data, which a person can use for identification and/or authentication. The certificates are generated using asymmetric cryptography, such as encryption or digital e-signature. Smart-cards represent a secure authentication factor, they are easy to carry and use complex encryption keys in the identification or authentication process. When a smart-card user wants to perform authentication process, the smart-card is inserted into the reader. Afterwards, the user is usually asked to enter the PIN code or the password, which is another authentication factor. Smart-cards are able to provide both identification and authentication processes. However, it is to admit, that smart-cards are not effective identifiers, because they can be easily given to other persons, changed or stolen [2], therefore, they should be always used together with other authentication factors.

A security system device is a password generating tool. Security system devices are equipped with informative displays, they generate passwords, which

have limited usage time interval, for example, 30 seconds. Security system devices use several elements - a unique security system device identifier, which is different for each device, time and encryption key. Security system devices usually are not the only one authentication factor in the authentication process. When using the security system device usually a multi-stage authentication process is created, where at the first authentication stage static authentication data are used - passwords or phrase-type passwords, but at the second authentication stage only a single-use password is used, which is generated by the security system device. But even such authentication devices are prone to security risk. Such devices can be lost, compromised, the battery can expire or the device can malfunction and as the result the device will not be suitable for authentication process. The security system devices are usually activated by entering the *PIN* code, which decreases the authentication security risk. If the device is lost, it will not be easy for the finder to activate and use it.

Two of the most popular security system device types are synchronous and asynchronous dynamic password generators. However, in the authentication process static security system devices are used as well. Synchronous and asynchronous security system devices work as single-use password generators. Single-use passwords are dynamic authentication data, which are changed after the certain time interval or they can be used in the authentication process. Synchronous dynamic password generators generate passwords within the known time period, for example, within 30 second interval, and one of the password generation elements is time. That means that there should be time synchronisation between the security system device and the authentication system. In order for the person to be identified, it is necessary to enter the password generated by the security system device into the authentication system interface. Moreover, the security system device itself shall be activated by using the *PIN* code or the password, which is the second authentication factor. The generated single-use password can provide user identification and/or authentication, but the *PIN* code or the password provides only authentication. Asynchronous dynamic password generators do not use time as the password generating element. In this case, the security system device generates a password only when the person has entered the code generated by the authentication system. By using asynchronous dynamic password generators the authentication process includes also the*„challenge-response"* process. For example, when a person wants to be authenticated in the system, at the first stage it is necessary to enter a user name and a password. After the authentication system has checked the entered credentials, it generates a request code by using the security system device identifier, which is then displayed in the interface. The request code is unique for each authentication. The person enters this request code into the own security system device, which then generates a single-use password that has to be entered by the person into the authentication system in order to finish the authentication process.

Static security system devices can be magnetic cards, smart-cards, *RFID* cards and tags, diskettes, USB devices [2] etc. Various static security system devices contain encryption keys, such as electronic signature, private encryption key *„private key"* or encrypted authentication data. In order to provide authentication process, static security system devices usually require an additional authentication factor - a static or a biometric authentication factor. However, there are cases, when the private encryption key serves as the only authentication factor. Such authentication type is widely used in Linux/Unix system administration in order to perform remote console connection from one resource computer to another.

*C. „Something you **are**" authentication factor*

The third authentication factor group corresponds to the condition "something you are". These are called biometric authentication factors. Biometric authentication factors can be used both for person's identification and authentication. Biometric authentication data can be obtained from person's biological parameters or behaviour features [5]. As to behaviour features, it is possible to say that they correspond to the authentication factor group "somehow you do it". The most popular biologic parameters used during authentication process in companies are fingerprints, venous structure of palms and fingers, eye cornea features, voice or face structure. The last three parameters do not anticipate physical contact with the biometric sensor, thus becoming less reliable. Behaviour features, used in authentication process, relate to keyboard use dynamics. To identify the person, the interval between pressing keys is measured when a word or a phrase is entered. Previously use of biometric authentication in companies was hard to implement due to high expenses, it was necessary to purchase biometric sensors and to perform complex implementations in authentication modules of information systems. However, nowadays smart-phones are very popular, which have such equipment that can be used as biometric sensors, thus making biometric authentication available for the enterprise information systems. Smart-phones are equipped with digital cameras able to perform face or eye cornea recognition, microphones for voice recognition and a keyboards for determining the rhythm of key usage.

Biometric authentication data cannot be easily lost, stolen, broken, guessed, copied or shared. Comparing with passwords, smart-cards or security system devices, biometric authentication factors are much

more resistant to social engineering attacks, because it is necessary for a person to be present during authentication process. Google Intelligence security analyst Alan Goode believes that *„Biometric systems can be much more convenient than tokens and other systems, and are useful to augment existing security methods like passwords. For added security they are also sometimes used as a third factor."*

The main shortage of biometric systems is inability to provide 100% precision in its functioning [5]. To use a biometric system, it is firstly necessary for a person to submit one or several biologic data samples to the system for configuration, for example, fingerprints. When the person tries to perform authentication, the fingerprint is compared with the saved sample and, if it is similar enough to the sample stored in the authentication system, the person's authentication is performed. Measuring precision of a biometric system is usually described by two error rates – *„False Non Match Rate"* and *„False Match Rate"* [5]. The first one "False Non Match Rate" describes the number of authentication efforts in the system containing person's biometric data samples, which were not approved by error. The second one "False Match Rate" describes the number of successful authentication efforts in the system, not containing the person's biometric data samples. The majority of biometric systems can be adjusted to decrease one of these rates, however, this happens at the expense of the other rate. Mark Diodati, the analyst of the company Gartner, said: *„It's important to understand that when a user supplies a password or a number from an OTP (one time password) token, it is either correct or it isn't. With biometrics you never get a definitive yes or no"*. Various biometric systems provide different security levels, determined by error rates. A good fingerprint sensor offers low error rates, thus providing a better security level in comparison to the non-contact biometric sensors, for example, microphone or digital camera for voice or face recognition. Fingerprint readers do not work properly in the environment, where users may have dirty fingers. Thus, voice recognition sensors are not suitable for the environment with high noise level. Sensors with low error rates can be used as the only authentication factor. But if error rates are high, then biometric authentication factors should be used together with other authentication factors, organizing a multi-stage authentication process, which significantly decreases the security risk.

III MULTI FACTOR AUTHENTICATION SECURITY TESTING

In order to start testing of authentication security, a testing server was created with the name Omega:

- the chosen equipment *AMD Athlon II X3 440* three core processor, *DDR3 4GB* RAM, *SATAII 500Gb* hard disk, network card *1 Gbit*;
- the installed operation system *Gentoo GNU/Linux* with minimal configuration;
- *Omega* testing server was assigned a real *IP* address;
- *ssh* service installed for server console connection;
- installed *Google authenticator* for a two-stage authentication process implementation [6];
- configured so that in case of successful *ssh* service authentication, the work author receives an e-mail about unauthorised connection and the server shuts down. In order to implement this, *Mutt* e-mail client *GNU* software and *Postfix* e-mail server *GNU* software was installed that will forward a message sent by the *Mutt* client to the given e-mail;
- for the privileged user *root* an insecure password is set *„password"*, users *admin*, *test* are created with passwords *„admin"* and *„test"*.

The aim of the test is to determine whether a multi-factor authentication solution decreases the security risk, by proving the hypothesis of this publication. The research is done in two parts:

a) In the first part of authentication security testing, for *ssh service* authentication a single-factor authentication is chosen with static authentication data - passwords, which were composed as "insecure";
b) In the second part of authentication security testing, for *ssh* service authentication a two-stage authentication is activated, where in the first stage static authentication data are used - passwords, and in the second stage - a single-use password is generated by *Google authenticator*.

After all the installation and configuration works are finished, all the functions of the test server *Omega* are checked prior to testing of the authentication security in the internet environment. Various password guessing tools are considered, such as*Ncrack, Medusa, Hydra*, in order to be able to perform password guessing using the remote *ssh* service console [7]. For Omega test server password guessing a tool *Hydra* was chosen, because the newest version of it was available. *Hydra* software was downloaded and installed on another test server. From the web-site *http://xato.net/* 10000 of popular passwords were downloaded and saved in the file *passwords1.txt* [8]. To perform the test, the superuser *root* was assigned a password *„rumbarumba"*, which was added to the file *passwords1.txt*. In the console of Omega server the corresponding command was entered to start the password guessing.

Example:

#hydra –l root –P password1.txt 84.237.XXX.X ssh

At the time of password guessing simulation the *Omega* test server was configured for the first part of authentication security testing, a single stage was used during authentication with static passwords without dynamic password generation by means of *Google authenticator*.

```
[root@Samba:/usr/src/hydra]# hydra -l root -P password1.txt 84.237.      ssh
Hydra v6.3 (c) 2011 by van Hauser / THC and David Maciejak - use allowed only for legal purposes.
Hydra (http://www.thc.org/thc-hydra) starting at 2014-11-21 15:55:03
[DATA] 16 tasks, 1 servers, 10000 login tries (l:1/p:10000), ~625 tries per task
[DATA] attacking service ssh on port 22
[STATUS] 256.00 tries/min, 256 tries in 00:01h, 9744 todo in 00:39h
[22][ssh] host: 84.237.       login: root   password: rumbarumba
[STATUS] attack finished for 84.237.      (waiting for children to finish)
Hydra (http://www.thc.org/thc-hydra) finished at 2014-11-21 15:56:16
[root@Samba:/usr/src/hydra]#
```

Fig.1. Hydra password guessing tool in action

After a known period of time the *Omega* server was compromised, and after sending the e-mail message the *Omega* test server shut down. In the Fig. 1 it is possible to see the *Hydra* password guessing tool in action. In the Fig. 1 it is seen that within a comparatively short period of time - 73 seconds, for the *Omega* server superuser *root 9744* passwords were tested. In order to start authentication security testing of the *Omega* test server in the internet environment, the superuser *root* was assigned an insecure password *„password"* and in the firewall of the network router access from the internet was activated for the port 22 of the *Omega* server. Such password was not chosen randomly. According to the researches this is the most popular password in the year 2014 chosen by computer users [9].

The first part of authentication security testing started on 21.11.2014 and.16:34 and on 23.11.2014 at 1:47 a hacker with *IP* address beginning with *176.102.3XX.XXX* registered in Ukraine, guessed the user name *admin* and the password *„admin"*. The duration of the test was 33 hours and 14 minutes until the moment the *Omega* test server shut down, sending an e-mail message about the compromised system.

As the privileged user password was not guessed, the authentication security test was retaken. The repeated test started on 24.11.2014 at 9:00 and on 25.11.2014 at 1:51 a hacker with *IP* address beginning with *222.161.XXX.XXX* registered in the Public Republic of China, guessed the privileged user *root* password *„password"*. The repeated test took 16 hours and 51 minute.

In the second authentication security testing part for *ssh* service the above-mentioned two-stage authentication was activated using *Google authenticator* dynamic passwords at the second stage and no one except the test server administrator could be able to perform authentication. In five months it has been more than 1.7 million unauthorized connection attempts, approximate 10335 in a day. Unauthorized connection attempt statistics has been published in the portal *http://ssh-stats.liepu.lv*. At the site can be seen the statistics of the unauthorized connection attempt geographic and *Logwatch* application ssh service original log files.

IV TESTING RESULT ANALYSIS AND RECOMMENDATIONS

The testing security experiment proves the stated hypothesis, because from the moment the multi-factor authentication was activated, no one was able to compromise the test server. Such authentication security testing proves that dynamic passwords represent a secure authentication tool, which is hard to compromise nowadays. However, it is possible to decrease the security risk to a larger extent if at the first stage of authentication a secure password is used. The company *TrustWave* published a research in IT security field *„Trustwave Global Security Report 2014"* stating that more than 31% of all IT incidents happen due to insecure passwords [1][10]. This indicates that each third password composed by users is *„insecure"*. From this it follows that it is unsafe to use static authentication data - passwords as the only authentication factor. In authentication solutions it is necessary to use the authentication data, which are not created by a person, for example, single-use dynamic passwords. In order to decrease authentication security risk, it is necessary to use multi-factor authentication solutions and static authentication data should always be used together with other authentication factors.

The unauthorised connection effort data obtained during authentication security tests can be analysed. What do we get from it? It is possible to determine regions and countries with the highest threat level. It is possible to estimate whether it is necessary for business to have contacts from such countries or regions and to make decision about blocking access to IT resources or to improve system authentication security by introducing multi-factor authentication.

At the beginning of the research *IP* address origin is analysed with the aim to determine the countries or regions with the highest threat level. The results are shown in Fig. 2. According to the results, the highest threat level is from Hong Kong, which makes 58.42% of the total threat volume. Likewise, high threat level is from the Public Republic of China – 28% and from

Australia – 9.65%, France – 2.46%, Italy, Moldavia, Korea, Kazakhstan, Portugal and Germany. If it is not important for your business to have connections with Hong Kong and China, restrict the access to your IT resources for these countries. If you have collaboration partners in such countries, it is necessary to perform access control and use multi-factor authentication in authentication solutions.

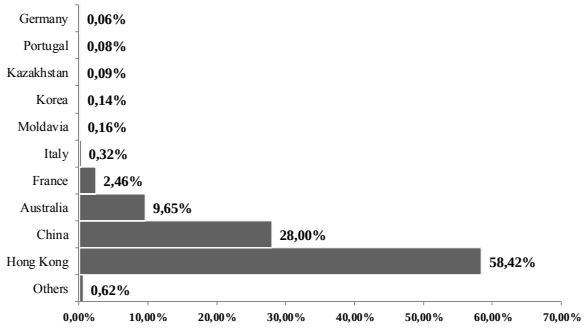

Fig.2. *SSH service* unauthorized connection attempts geographic statistics

Further, the data acquired from the statistics of unauthorised connection efforts to *SSH services* are analysed in order to determine the days of week, when the security threat level is the highest. Analysis of data obtained within 23 weeks was performed. The results are shown in Fig. 3.

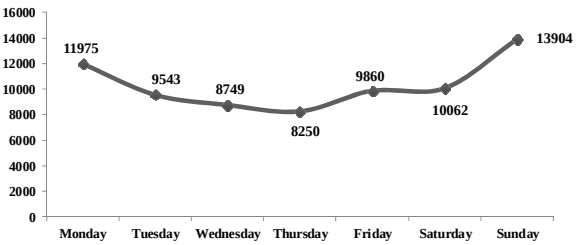

Fig. 3. SSH service average unauthorized connection attempts by days of week

According to the obtained data shown in Fig. 3, the highest IT security threat level is observed on Saturdays, Sundays and Mondays. The reason why exactly in these days IT security threat level is the highest is a good topic for another research. However, authors of this paper can express some qualitative statements regarding this situation. Namely, on Saturdays the threat level can be high, because on weekend the number of employees in corporations decreases thus decreasing the level of IT resource monitoring. A lawbreaker, having obtained the access to the system on Saturday, will be able to perform lots of actions during the weekend without being noticed. In its turn, on Sundays the threat level is high for IT resources because on this day availability and application level of IT resources is the lowest. Therefore, it is recommended to monitor and control IT resources not only during working days, but also in the weekends and on holidays, so that in case of IT incidents the responsible person could perform the necessary actions to prevent the incident.

V CONCLUSION

When applied, various authentication technologies prove that there are more secure solutions rather than password authentication solutions, which use dynamic and biometric authentication factors. This statement is proven also by the authentication security testing experiment described in this publication. The multi-factor authentication process, which uses static and dynamic passwords, provides lower authentication risk even if the static password is considered as unsafe. At the moment in the world passwords are being replaced by other authentication technologies and this is the question of time, when passwords will no longer be used in the authentication process!

There will always be authentication security risks for each authentication solution, and it is impossible to create a 100% secure authentication solution in the nearest future. Therefore, it is necessary to continue the research in the field of authentication solutions and security risks, determining what security risks are new authentication solutions prone to. Using theoretical research methods, performing direct and indirect observations and tests of authentication security, it is necessary to find out, what security risks are authentication security system devices and biometric sensors prone to.

In another research it would be interesting to find out, why such high level of IT security threats comes directly from the Public Republic of China, including the special administrative region thereof - Hong Kong. It would be interesting to find out whether inhabitants of this country are initiators of IT threats or is it the politics of the country, which is favourable for hackers from other countries to perform unlawful actions from IT resources of PRC.

VI REFERENCES

[1] A. Jansone, K. Lauris, "Authentication solutions and security risks", presented at International Joint Conferences on Computer, Information, Systems Sciences & Engineering (CISSE 14), Dec. 9, 2014.

[2] J.M. Stewart, D. Gibson, M. Chapple, "Certified Information Systems Security Professional – Study Guide (Sixth Edition)", Canada, John Wiley & Sons, 2012, pp. 1-47.

[3] L. Lāce, "Divpakāpju autentifikācija", Riga, CERT.LV - the Information Technology Security Incident Response Institution of the Republic of Latvia, Feb. 19, 2013. [Online], Available: https://www.esidross.lv/2013/02/19/divpakapju-autentifikacija/ [Accessed: Feb. 20, 2015].

[4] C.P. Pfleeger, S.L. Pfleeger, "Analyzing Computer Security – A Threat/Vulnerability/Countermeasure Approach", USA, Pearson Education, 2012, pp.38–64.

[5] P. Rubens, "Biometric Authentication: How It Works", Aug. 17, 2012. [Online]. Available: http://www.esecurityplanet.com/trends/biometric-authentication-how-it-works.html [Accessed: Feb. 20, 2015].

[6] "Install Google Authenticator". [Online], Available: https://support.google.com/accounts/answer/1066447?hl=en [Accessed: Feb. 21, 2015].
[7] "Brute Forcing Passwords with ncrack, hydra and medusa". [Online], Available: http://hackertarget.com/brute-forcing-passwords-with-ncrack-hydra-and-medusa/ [Accessed: Feb. 21, 2015].
[8] M. Burnett, "10,000 Top Passwords" Jun. 20, 2011. [Online], Available: https://xato.net/passwords/more-top-worst-passwords/#.VI7ZHnvLLCD [Accessed: Feb. 26, 2015].
[9] A. Kooser, "Worst passwords of 2014 are just as awful as you can imagine", Jan. 20, 2015. [Online], Available: http://www.cnet.com/news/worst-passwords-of-2014-are-just-as-awful-as-you-can-imagine/ [Accessed: Mar. 2, 2015].
[10] D. Kaplan, "2014 Trustwave Global Security Report", May 21, 2014. [Online], Available: https://www.trustwave.com/Resources/Trustwave-Blog/The-2014-Trustwave-Global-Security-Report-Is-Here/ [Accessed: Mar. 4, 2015].
[11] L. Eadicicco, "Passwords Are A Horrible Way To Keep Us Safe — Here Are The Potential Alternatives" May 11, 2014. [Online], Available: http://www.businessinsider.com/password-alternatives-2014-5 [Accessed: Mar. 5, 2015].

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 121-126

Wireless sensor networks lifetime assessment model development

Aleksejs Jurenoks, Leonids Novickis
Riga Technical University, Riga, Latvia
Contactpoint: aleksejs.jurenoks@rtu.lv, lnovickis@gmail.com

Abstract. **In the recent year's low power computing systems have gained popularity. Networks, which use low power computer systems and transmitted data by using wireless connection are called wireless sensor networks, which main task is to get the information from sensors and transmission network. Nowadays, the most topical researches pertaining to wireless sensor networks are grounded on the new optimization of structure of network transmission protocol, the routing optimization in transmission network, optimization of network structure, as a result of which the life circle of wireless network sensors is possible to increase. In the present article the methodology for determining the life circle of network is discussed. The approaches in detection of life circle pertaining to the important network nodes are described.**

Keywords: **Arduino; Sensor Network; life circle; Data processing.**

I INTRODUCTION

More institutions are concerned with the problems of processing real time data every day. The bigger the data flow, the longer the processing time is necessary. According to IDC data in year 2008 information data size of "digital universe" was 1,6 ZB and it will be increased at least twice to year 2015.

In the recent years low power computing systems have gained great popularity. Nowadays, the technologies allow to decrease the existing size of the system creating an inexpensive low power computing systems with the ability to use both physical and wireless networks for transmission of information. The main popularity this type of technology has gained in the field of receiving information from sensors and control systems. Network that uses low power computing and transmitting the data via a wireless connection is referred to as wireless sensor networks, the main task of which is receiving information from the sensors and transmission in the network.

The main drawback of wireless sensor networks (BST) is the low power batteries which significantly limit the lifetime of the network.

Therefore, this article discusses the developed model of sensor network, which allows to assess the life circle of network by using the way of working of each network element. The methodology for the calculation of the parameters of the main model are described in details. The life circle of the separate network nodes is estimated by using the defined parameter values.

II RELATED WORKS

Nowadays, the most topical researches pertaining to wireless sensor networks are grounded on the new optimization of structure of network transmission protocol, routing optimization in the transmission network, optimization of network structure, as a result of which the life circle of wireless network sensors is possible to increase.

In the network exists unequal power consumption; thus, the net nodes become disable at the moment when, despite the fact that a larger part of the network continues to operate, the electricity finishes in some network nodes. [2, 15]

There are a number of methods that allow to solve this problem [11, 13]. These include the choice of the capacity of individual battery, the deployment of the node density, the adjustment of power transmitter, the application of energy-efficient data transfer protocol, positioning of network nodes and other methods that are associated with the introduction of an additional network costs. Relatively recently a new method for balancing energy through the mobility of network node, which provides the opportunity for reconfiguration of dynamic network or change of network topology, has been offered [14].

However, the main drawback of this method of study is absence of mathematical model of the dynamic network.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.508

III NETWORK LIFE EXPECTANCY VALUATION MODEL

The life circle of network depends on how long its elements are working. Taking into account that verbosity is often included in networks, different requirements exist for the quality of the results received from the network.

While working out the network, it is necessary do timely assessment of the approximate working time pertaining to each network element until the moment when the battery change of the element will be necessary. It is important to understand what factors influence the autonomous working time of network elements.

It is well known that the power consumption of individual elements depends on several factors to consider when designing networks:

- The technical parameters of node;
- Frequency of data collection;
- Physical and channel level protocols;
- Network topology;
- Use of routing protocol.

Any sensor network has three types of nodes - terminals, routers, and data collectors. Let us assume that the data collectors do not affect the life circle of the overall system because they are provided with independent power supply or are equipped with a much more powerful autonomous power supply. Currently, there is a variety of creators/indicators of network life circle.

A. The Index Based on the Working Node

The total time of network operation can be marked as T_k^n time, in which at least k of the n units are in operation [1, 3, 4, 5]. However, there is one drawback of this marking- the types of network nodes are not defined. Mostly in the networks there are primary nodes - which provide data retransmission and significantly impact the overall network performance. If one of the "bearing" nodes stops working, it automatically means that the network ceases to exist. That is why in many works [6, 7], analyzing the life circle of system, variable m, which means the number of critical important elements in the network, which must be active at all times, is defined. For example they could be the nodes of network cluster management [8]. For the other nodes, metrics is used:

$$T_k^{n-m}. \quad (1)$$

In the researches it is often used a case when $k = n$. In this case the network is considered as able-bodied as long as all the network nodes are active or:

$$T_n^n = min_{i \in V_n} T_i, \quad (2)$$

where: V_n – the quantity of network nodes;
T_i - the life circle of each network node.

B. The Indicator Based on the Area Coverage

This indicator is related to the use of network sensor- the quantity of information that is necessary to get from the concrete network segment. There are two approaches on the determination of the indicators based in the coverage. On the basis of the first approach the network is considered as able-bodied as long as α percent of the overall network coverage is covered by at least one sensor (α - coverage) [9]. The second approach is based on provision of redundancy and requires that in each network segment should be at least k – number of active sensors (k - coverage) [10].

The biggest disadvantage pertaining to the detection of indicators based coverage is the complex process of algorithmic presentation.

C. Index Based on the Delay of Information Transmission

According to the research, this index is mostly used when working with information gathering sensor network systems. The full definition of the index is provided in the paper [3]. Let us improve this index in order to cover previously mentioned index groups.

Suppose that the environment, which is carried out by the sensor network monitoring system, at each time interval occurs in certain activities which need to be identified. The events can be considered as external events, such as alerts, temperature changes of the system and other technical information, as well as described determinate event of the internal network, such as a regular transmission of information.

Let us introduce the parameters of quality pertaining to working network for each zone at specified time interval t. Suppose that $N_k(t)$ – in the range k there is the total amount of events in the time interval t and $I_k(t)$ - the total amount of the events from the $N_k(t)$ number that are delivered to a data collector at the determined period of time. The total time t is defined for the whole range k. The parameters Δt_k are defined on the basis of the necessary network intensity in the determined network segment. In the result zone at k time in the t interval the indicator of network working quality will be equal to:

$$Q_k(t) = \begin{cases} \frac{I_k(t)}{N_k(t)}, & N_k(t) \neq 0 \\ 1, & N_k(t) = 0 \end{cases}. \quad (3)$$

For acting network there is a threshold c_k that exists in order to indicate the lowest ratio value Q_k below which the network can not be regarded as working age. Thus, for the possible duration of the network can be considered:

$$\forall t < \tau_1, \forall k \in [1..m]: Q_k(t) \geq C_k. \quad (4)$$

IV CALCULATION OF CONSUMED ENERGY PERTAINING TO THE NODES OF SENSOR NETWORK

A. The terminal lifespan assessment method

The terminal node of the sensor network is intended to reading information from the sensors for the further transmission of data in the network.

The main difference in the node from the router is the impossibility of information transmission. This means that the terminal node does not use power and time for reception of information.

The consumed terminal energy in one operating cycle of the system can be defined by the following formula:

$$P_{term} = P_{tx} * t_{tx} + P_s * t_s + P_a * t_a, \quad (5)$$

where P_{tx} – the average conumed energy in the transmission mode (W);

t_{tx} – time that is required for the information transmission in the network (s);

P_s – average power consumed in a sleeping mode (W);

t_s – time in which the system is in the sleeping mode (s);

P_a– consumed power on processing of data (reading the information from sensors) (W);

t_a – time that is intended for processing the information (s).

It is assumed that $t_{tx} + t_a > t_s$, which means that the system has sufficient time to be in the sleeping mode. The value of parameters is determined by the technical data of processor and properties of software algorithm.

For operating wireless network, it is important to deliver the data to the recipient node and, depending on the situation, the coefficient is impossible to define precisely.

There might appear a situation that the system will be unable to deliver the information to the recipient, and then it will be necessary to resent it as long as the information is delivered. As a result, the total sleeping time $\Delta t_{tx} \leq t_s$ will be decreased. Consequently, it means that the system can dedicate time for resending information no more than the amount of time t_s. If the information has not been delivered, it will be considered as lost.

B. The router lifespan assessment method

It is known that in sensor networks oriented on events, which are using asynchronous data access environment, a router is the weakest point regarding the duration of life circle of the system.

It is based on the fact that a node spends very little time or at all does not spend time in the sleeping mode, but consumes power on reception of information from the network and the transmission of information into network.

The total power consumption of the router very much depends on the selected routing protocol that in the end determines the table of service data flow for safety.

A router has four operating conditions:

t_s - sleeping mode of the system;

t_s - The information reception from a network;

t_{tx} - Information sharing in a network;

t_a - Reading the information from the sensors.

Since the performance of the network is highly determined, it can be considered that the router $\Delta t_{rx} + \Delta t_{tx} \leq t_s$. While transmitting the information, an error may appear in the network; consequently, it will take more time to send and receive data; as a result, it is necessary to introduce an additional parameter p_e - percentage denoting error probability.

Using error probability coefficient in the network can be considered that:

$$p_e(\Delta t_{rx} + \Delta t_{tx}) \leq t_s. \quad (6)$$

The amount of time the system spends on the reception of all the information is equal to:

$$r_{rx} = \sum_{i=1}^{n} t_i * I_i, \quad (7)$$

where t_i- the time necessary for the reception of information from one terminal (s);

I_i - i-the intensity of data flow.

The amount of time that the system will spend on sending the information in the network is equal to:

$$r_{tx} = \sum_{i=1}^{n+1} t_i * I_i. \quad (8)$$

There is a situation when the system is not able to transfer all the necessary information that has been received by the node $r_{rx} > r_{tx}$. Before the information is received the system checks whether there will be sufficient time for transferring the information.

When planning the router load, the total life cycle of the router Δr must be taken into account.

$$\Delta r < p_e(r_{rx} + r_{tx)} + r_a, \quad (9)$$

where r_a-the time consumed by the system to receive information from the sensors (s).

Let us look at the situation $r_{rx} + r_{tx} + r_a \leq r_s$, when the router is able to fulfill its task and transfer the information to the network. In this case, the router's power consumption will be equal to:

$$P_m = P_a r_a + P_{tx} r_{tx} + P_{rx} r_{rx} + (r_s - r_{tx} - r_{rs}) P_s, \quad (10)$$

where P_{tx} – the average power of router in the transmission mode, P_{rx} - the average power of router in the receiving mode, P_a - the average power of router receiving information from sensors, P_s - the power of router in the sleeping mode.

C. The Impact of Operating Range on the Work of the Terminal

Each network element has defined range in which it can operate. The amount of the consumed energy on transmission of information is influenced by the distance to which the information has been transmitted. The router of the system and the terminal can operate in several modes that will reduce the amount of consumed energy by rationally deploying elements of the network, as well as determining the location of network elements.

Each transmitter is provided with optimal transmission distance, working in which power consumption is more efficient. The power of transmitter determines the speed of the transmission of information. In the study it is believed that the system is operating in the same environment, and the speed of information transmission is not affected externally.

It is known that the system consumes certain power for the transmission of data to the maximum distance [12]. Using the formula for determination of radio signal amplitude provided by Boris Vvedenskij, it is possible to determine the admissible maximum distance to which the information can be transmitted.

$$E_{np} = \frac{4\pi\sqrt{P*G}}{\lambda*r^2} * h_1 * h_2, \qquad (11)$$

where

E_{np} - wave amplitude in the receiving point;

P - power coefficient of retranslator;

G - coefficient of wave amplifier;

λ - wavelength;

r - distance of radio communications;

h_1, h_2 - height to which radio antenna was lifted.

It is assumed that all network nodes in our system use antennas of the same type and do not use additional intensifiers; thus, the simplified formula can be used:

$$E_{np} = \frac{4\pi\sqrt{P}}{\lambda*r^2}. \qquad (12)$$

Changing the distances of network node, it is important not to lose the quality of signal. It is assumed that in case of increasing the distance the E_{np} value will remain unchanged. In the result the P_k coefficient of consumed power pertaining to router is equal to:

$$P_k = \frac{{E_{np}}^2*\lambda^2*r^4}{(4\pi)^2}, \qquad (13)$$

When planning the operation scheme of sensor network there is a possibility to equalize the consumed power of network element by adapting the power of transmitter and operating range to certain network segment. This approach will allow to equalize the total duration of life circle pertaining to nodes by decreasing the left out quantity of energy in the nodes.

As a result, the consumed power of router while transmitting the information in the network is equal to:

$$P_{tx} = \begin{cases} P_{tx}, & K_r = 1 \\ P_{tx} * \frac{{E_{np}}^2*\lambda^2*{K_r}^4}{(4\pi)^2}, & K_r > 1, \end{cases} \qquad (14)$$

where K_r - the coefficient of working range against standard working range.

D. IEEE 802.15.4 Standard Frame Properties

The maximum size of the load efficiency pertaining to frame is dependent on the length of service field. The standard describes the maximum length of the physical level of the frame$aMaxPHYPacketSize = 127bytes$. If to take into account the minimum size of addressing technical frame comprises 4 bytes, the maximum coefficient of efficiency load will be equal to 112 bytes.

IEEE 802.15.4 standard operates at 2.4 GHz frequency with a permissible transfer rate of channel f = 250 kbit/s. The result time required for transmission of information in the network can be calculated using the following formula:

$$t_{DATA} = \frac{L+O}{f}, \qquad (15)$$

where L – the number of bits per packet,

O - the size of the service field in bits.

V THE PRACTICAL EXAMPLE OF CALCULATION OF POWER CONSUMPTION

A. The Terminal Node power consumption

Let us assume that the frequency of device processor is f_{cpu}: one activity takes C numbers of processor cycles and it is necessary to execute M number of measuring amount in order to secure the preparation of the information to be sent to the network. In case of recalculation, the time necessary for the transition form the sleeping mode t_s must be taken into account.

The result time that is necessary for system to be in the active mode is equal to:

$$t_a = \frac{C*M}{f_{cpu}} + t_s, \qquad (16)$$

Let us assume that the system operates:

- using a single AA battery types – accordingly the initial energy $E_0 = 9\ kJ$;
- Each operating cycle requires two CPU cycles - $C = 18$;
- For the system to perform qualitative measurement 4 measurement cycles are needed - $M = 4$;

- The system uses low power processor $f_{cpu} = 16\,MHz$;
- The size of packet comprises 20 bytes - $L = 20$;
- The time that is required for the transition from the system sleep mode - $t_s = 0.08s$;
- The total time of system cycle - $t_c = 20s$.

Using the technical information of BSN node, let us calculate the consumed energy by one cycle of the terminal mW.

$$P_{term} = P_{tx} * (t_{data} + t_{ACL}) + P_a * t_a + P_s * (t_c - t_{data} - t_{ACL} - t_a). \quad (17)$$

The result is that, using output data, the terminal device for one cycle consumes 136.81 mW. Knowing the *w* - energy consumption of one time per 1 second and the *J* - initial amount of energy let us calculate the total life circle of the system by using the following formula:

$$s = \frac{J}{w}. \quad (18)$$

B. The Router Node power consumption

Let us assume that the frequency of device processor is f_{cpu}: one activity takes *C* numbers of processor cycles and it is necessary to execute *M* number of performances in order to accomplish the task. In case of recalculation, the time necessary for the transition from the sleeping mode must be taken into account - t_s.

The result time that is necessary for system to be in the active mode is equal to:

$$t_a = \frac{C*M}{f_{cpu}} + t_s. \quad (19)$$

The time that is necessary for receiving information from terminals is equal to:

$$t_{rx} = \sum_{i=1}^{n} \frac{L}{f} + t_{DATA}. \quad (20)$$

where t_{DATA} for IEEE 802.15.4 with frame size of 11 byte = 0.352.

The time that is necessary for sending information in the network is equal to:

$$t_{tx} = \sum_{i=1}^{n+1} \frac{L}{f} + t_{DATA}. \quad (21)$$

Assume that the system works:

- using a single AA battery types – accordingly the initial energy $E_0 = 9\,kJ$;
- the amount of processor cycles required for each working cycle: $C = 18$;
- For the system to perform qualitative measurements 4 measurement cycles are needed - $M = 4$;
- The system uses low power processor $f_{cpu} = 16\,MHz$;
- The system uses the 2.4 GHz frequency with the transmission rate $f = 250\,kbit/s$;
- The size of packet is 20 bytes - $L = 20$;
- The time that is required for the transition from the system sleeping mode $t_s = 0.08s$;
- The total time of the cycle pertaining to the system $t_c = 20s$;
- The number of servicing terminals in the chain - $n = 1$.

Using technical information, let us calculate the average power consumption in the router per time unit:

$$P_m = \frac{P_{tx}t_{tx} + P_{rx}t_{rx} + P_a t_a + P_s(t_c - t_{tx} - t_{rx} - t_a)}{t_c}. \quad (22)$$

The result is that, using output data, the terminal device for one cycle consumes 137.54 mW. Knowing the energy consumption of one time per 1 second (w) and the initial amount of energy (J), let us calculate the total life circle of the system by using (17).

The results of the calculations show that on the same terms the energy of the router will expire faster than the energy of terminals. Furthermore, the bigger number is serviced by the router, the greater the energy balance remains in the terminal node.

VI DEFINITION OF THE DURATION OF NETWORK

In the process of operation, the network has the time when on the basis of external factors or lack of energy in various network elements system is unable to carry out its task and requires reconfiguration of the network.

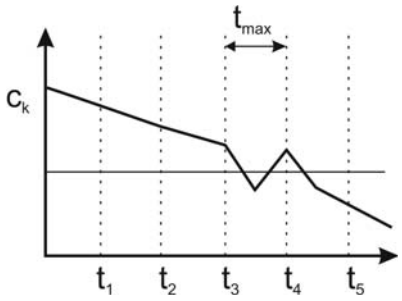

Fig. 1. Network reconfiguration delay time.

Let us define the number of points N, which indicate the moment when the system crosses the border - c_k.

According to the drawing (Fig.1), the quantity is depicted by the points t_1, t_2, .. t_n. Suppose that there is the maximum allowable time t_{max} (s), when the system can be located below the range c_k required for carrying out network reconfiguration. Overall system remains active if the following condition is fulfilled:

$$t_{n+1} - t_n < t_{max}. \quad (23)$$

By controlling the overall network management it is not possible in real time to find out the amount of energy left in nodes because it requires additional steps to transfer information. The only way to control the ability of life pertaining to the node is the delayed control while transmitting the information. If the time T of system reconfiguration is greater than t_{max}, it is possible to verify the exit of an individual node from the rank and file.

This definition can be formalized by using the algorithm pertaining to the collection of information in the nodes (Fig.2.).

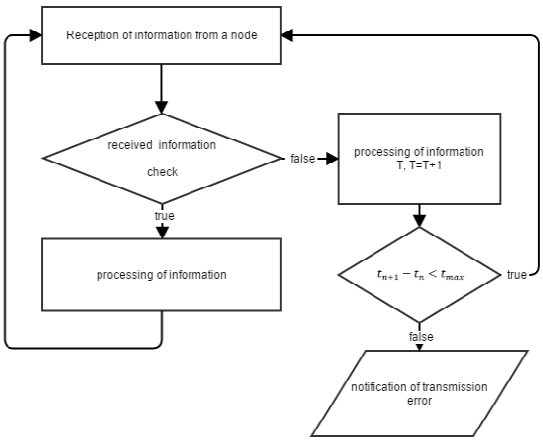

Fig.2. Information collection and processing algorithm.

VII CONCLUSIONS

The article describes the developed methodology for calculation of life circle pertaining to wireless sensor networks, which allows to evaluate the life circle of each network element and to determine the overall life circle of the network using the developed structure of network. Described methodology is design to fit any Wireless sensor network devices that are using IEEE 802.15.4 transmission protocols. Additional protocol changes structure need to be taken to use it with any 802.x technology. This approach differs from the others with as follows:

- description of each element of network by using the quantity coefficient of the consumed power;
- using the coefficient of object-range transmission for calculation of power;
- uses the coefficient of changes of network configuration and determines the denial of time pertaining to the system.

It is known that the total amount of consumed energy in the nodes depends on many factors:

- technical solutions of hardware;
- frequency of operation/performance;
- intensity of data flow;
- algorithms for access to the environment.

In the article the developed methodology allows efficient planning of the structure pertaining to the network by dividing the tasks between the network nodes, thus increasing the overall life circle of network.

In further article we will describe wireless sensor network reconfiguration method that contributes to remaining energy balancing for network nodes.

VIII REFERENCES

[1] Chen B., Jamieson K., Balakrishnan H., Morris R. Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks / ACM Wireless Networks Journal. 2001. P. 85–96.

[2] Chen D., Varshney P. K. QoS Support in Wireless Sensor Networks: A Survey / Proc. of the 2004 International Conference on Wireless Networks (ICWN 2004), Las Vegas, Nevada, USA. 2004. P. 227–233.

[3] Chen Y., Zhao Q. On the lifetime of wireless sensor networks/ Communications Letters, IEEE. Nov. Vol.9, no.11. P.976–978.

[4] Hellman K., Colagrosso M. Investigating a Wireless Sensor Network Optimal Lifetime Solution for Linear Topologies / Journal of Interconnection Networks. 2006. P. 91–100.

[5] Kang I., Poovendran R. Maximizing static network lifetime of wireless broadcast ad hoc networks / Communications, 2003. ICC '03. IEEE International Conference on. Vol. 3. May. P. 2256–2261.

[6] Kumar S., Lai T. H., Balogh J. On k-coverage in a mostly sleeping sensor network / Proceedings of the 10th annual international conference on Mobile computing and networking. MobiCom '04. New York, NY, USA: ACM, 2004. P. 144–158.

[7] Lu G., Krishnamachari B., Raghavendra C. An adaptive energy-efficient and low-latency MAC for data gathering in wireless sensor networks / Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International. 2004. P. 224–231.

[8] Raicu I., Schwiebert L., Fowler S., Gupta E. K. S. Local load balancing for globally efficient routing in wireless sensor networks / International Journal of Distributed Sensor Networks. 2005. Vol. 1. P. 2005.

[9] Soro S., Heinzelman W. B. Prolonging the lifetime of wireless sensor networks via unequal clustering / Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05). IPDPS '05. Washington, DC, USA: IEEE Computer Society, 2005. P. 236–243.

[10] Vass D., Vincze Z., Vida R., Vidacs A. Energy Efficiency in Wireless Sensor Networks Using Mobile Base Station / EUNICE 2005: Networks and Applications Towards a Ubiquitously Connected World / Ed. by C. Kloos, A. Marin, D. Larrabeiti. Springer US, 2006. Vol. 196 of IFIP ternational Federation for Information Processing. P. 173–186.

[11] Vullers R., van Schaijk R., Doms I. et al. Micropower energy harvesting / Solid-State Electronics. 2009. Vol. 53, no. 7. P. 684 – 693.

[12] Wu K., Gao Y., Li F., Xiao Y. Lightweight deployment-aware scheduling for wireless sensor networks / Mob. Netw. Appl. 2005. Vol. 10, no. 6. P. 837–852.

[13] Zhang H., Shen H. Balancing Energy Consumption to Maximize Network Lifetime in Data-Gathering Sensor Networks / IEEE Trans. Parallel Distrib. Syst. 2009. Vol. 20, no. 10. P. 1526–1539.

[14] Ефремов В. В., Маркман Г. З. "Энергосбережение"и "энергоэффективность": уточнение понятий, система сбалансированных показателей энергоэффективности / Известия Томского политехнического университета. 2007. Т. 311, № 4. С. 146–148.

[15] Energy Harvesting Microstrain solutions URL (02.2015): http:/www.microstrain.com/energy-harvesting.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 127-133

Land Cover Recognition using Min-Cut/Max-Flow Segmentation and Orthoimages

Sergejs Kodors
Rezekne Higher Education Institution,
Atbrivoshanas Str. 115, Rezekne, LV-4601

Abstract. **The geospatial information is significant for many socio-technical activities like urban planning, the prediction of natural hazards, the monitoring of land use, weather forecasting, cadastral surveys etc. It is possible to acquire geospatial information from a distance using remote sensing technologies, but remotely sensed images don't have semantics without a previous recognition. The classification of geospatial information is expensive and time consuming process. The paper describes the automatic land cover recognition method, which is based on min-cut/max-flow segmentation. The raw data are othoimages with a high resolution. The proposed method is tested and evaluated by Cohen's kappa coefficient.**

Keywords: **land cover, min-cut/max-flow segmentation, orthoimage, recognition.**

I INTRODUCTION

It is very important to acquire and to analyze the state of an environment before making some decisions. Therefore geospatial information is the significant information, which is used in many socio-technical activities like an urban planning, the prediction of natural hazards, the monitoring of land use, weather forecasting, cadastral surveys etc.

Nowadays it is possible to acquire the geospatial information from a distance using the modern technologies like aerial vehicles, satellites or spacecrafts. The remote sensing is the scientific field how to acquire the geospatial information without a physical contact. All remote sensing techniques are based on the processes of energy emission and reflection from the observed objects of a land surface. These techniques are particularly effective, if a monitoring region is very large and problematically accessed for the ground surveys.

If the geospatial information can be acquired with a single attempt, the interpretation of this information is an expensive and time consuming process.

The paper describes the automatic land cover recognition method, which is based on min-cut/max-flow segmentation. The land cover recognition is the applied science of the artificial intellect. The goal of the land cover recognition is the classification of the geospatial objects using the remote sensing data. The proposed method uses orthoimages as the raw data.

Min-cut/max-flow method separates images into segments using the prior knowledge about the locations of an object and a background. Traditionally the locations of the object and the background, which are called seeds, are manually marked, while the unmarked pixels are labeled in the process of the segmentation. If the manual part is replaced by the automatic method, when the object and the background have the semantic meaning (segments are classes), then min-cut/max-flow segmentation can be used as the recognition method.

The proposed method of the land cover recognition is based on the logical reasoning, which is defined using the color histograms of orthoimages. The logical reasoning is the simplest approach, which can be replaced by the more complicated supervised or unsupervised methods. As a result, the paper provides the framework for the plenty different realizations of the recognition systems, which can be applied for different purposes.

II ORTHOIMAGE AND GEOSPATIAL INFORMATION

An orthoimage is the digital type of an orthophotograph. The orthphotograph is the aerial panorama of the land surface, which is orthorectified (the aerial photograph, which has not distortions caused by a perspective). Traditionally the orthophotograph was produced using orthophotoscope instruments. The usage of computer software systems instead of the hardware to create the digital orthophotograph was initiated in the late 1970s [1]. The orthoimage is produced by the rectification

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.171

algorithms, which analyze the visibility of the geospatial objects using the perspective images and photo orientation parameters [1], [2].

The geographical information systems (GIS) use two representation formats of the geospatial information, which are called raster and vector data formats. The orthoimages belong to the group of the raster format. The raster images are represented as a regular grid, where the pixel is the smallest element. The raster images are widely used in the computer graphics, but the geospatial raster images have two additional features: a geographical location and the proportional geometrical sizes in the real world. The smaller spatial size of the pixel identifies the higher resolution of the image. The medium resolution of the orthoimage is 10 – 100 meters per one pixel [3].

The pixel-based methods of the land cover recognition are less effective than the object-based methods, when the image resolution increases [3], [4]. The pixel-based methods classify each pixel independently, but the object-based methods classify the whole segments. The segments are the sets of the pixels, which have the similar features like a color, a texture and brightness. The modern orthoimages can have the resolution, which is equal to the half of the centimeter (2012) [5]; therefore the segmentation is the incredibly important process to recognize the modern orthoimages. As a result, the land cover recognition method, which is based on min-cut/max-flow segmentation, gives strong promises to classify the land cover.

III IMAGE SEGMENTATION

The image segmentation is the process of the labeling each pixel, where the label is the unique number, which identifies the similar regions.

The recognition is the classification, which labels each pixel too, but the label has the semantic meaning and identifies the category; therefore the segmentation methods can be adopted for the classification problems.

The different methods of the image segmentation were proposed by different authors:

- the edge detection [6-8];
- the watershed algorithm [9];
- the expectation-maximization method [7];
- the genetic algorithms [7], [10];
- Kohenen's self-organizing maps [10];
- min-cut/max-flow segmentation [11]-[14].

The paper describes the land cover recognition method, which is based on min-cut/max-flow segmentation.

IV MIN-CUT/MAX-FLOW SEGMENTATION

Min-cut/max-flow segmentation is based on the next principles:

- The adjacent pixels are connected by some energy, which must be applied to break these connections.
- The pixels of the same segment have the higher energy potential, then the pixels between two segments.
- The goal of the method is to find a solution, when the minimal energy is applied to break the image into two parts: an object and a background.
- The image is transformed into the weighted directed mathematical graph to solve the energy minimization problem.
- The energy minimization problem is solved using the algorithms of the maximal flow.

So, min-cut/max-flow segmentation is organized as an energy minimization problem, when the solution is found as the maximal flow through a network, which is constructed as the transformation of image into the mathematical graph, which describes the connections among the pixels and pixels' similarities.

A. Mathematical Graph

The mathematical graph is the collection of two dependant sets: $G = \{V; E\}$, where V is the vertices, but E - the edges. The vertices are the nodes or the elements, which are connected by the edges. The edges of the graph are written as $(a, b) \in E$, where $\{a, b\} \in V$. If a movement between two nodes is restricted by one direction, the graph is called directed. If the edge has the weight, which represents the cost of the movement, the graph is called weighted. The visual representation of the mathematical graph is depicted in Fig. 1.

The image transformation is executed using the next principles:

- Each pixel is replaced by the node of the graph (see Fig. 2).
- The nodes of the adjacent pixels are connected by two edges: the edge from p to q and the reverse edge from q to p (see Fig. 2). The edges between two nodes of the adjacent pixels are called N-links [13].
- The weights of the N-links are calculated by the function of smooth cost $V_{p,q}(L_p, L_q)$, where L is the labels of adjacent vertices p and q.
- 4-path or 8-path schema can be used to define the adjacent pixels.
- Two terminal nodes are added to the graph: one is called source S, but another - sink T (see Fig. 2).
- The source S is connected with each node of the pixel by the edge, and each node of the pixel is connected with the sink T by the edge (see Fig. 2). The edges, which connect the nodes of the pixels with the terminals, are called T-links [13].

- The weights of *T*-links are calculated by the function of label cost $D_p(L_p)$.

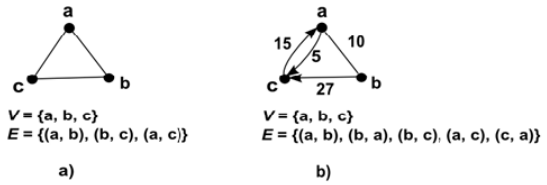

Fig. 1. a) Undirected mathematical graph, b) weighted directed mathematical graph

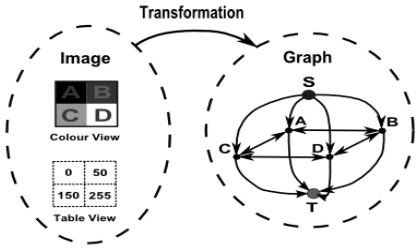

Fig. 2. Principal schema of image transformation into mathematical graph

- The terminal *S* is associated with the segment '0' and the terminal *T* – with the segment '1'.

The weights of the mathematical graph must satisfy the next conditions:

- The weights must be assigned to each edge of the graph [11], [13].
- The weights can be calculated using any non-negative symmetric function [11].
- The weights between the adjacent pixels must be higher, if the pixels belong to the same segment, and they must be lower in the border of two segments [11], [12], and [14].
- The weight of the edge (*p*, *q*) may differ from the weight of the reverse edge (*q*, *p*) [13].

B. Energy minimization

When the image transformation into the mathematical graph is completed, the algorithm starts to search the optimal solution. The optimal solution is the minimal cut, which divides the image into two parts using the minimal energy. The volume of the energy, which must be applied to separate the image, is calculated using the formula (1):

$$E(L) = \sum D_p(L_p) + \sum V_{p,q}(L_p, L_q), \qquad (1)$$

where *p* – a current pixel, *q* – an adjacent pixel, *L* – the set of the labels, which indicate the segments of the pixels, $D_p(L_p)$ – the cost to assign the label L_p to the pixel *p*, $V_{p,q}(L_p, L_q)$ – the energy, which must be applied to break the connection between the pixels *p* and *q* [11], [13]. The component $D_p(L_p)$ is called label cost, but $V_{p,q}(L_p, L_q)$ is called smooth cost [15]. The label cost and the smooth cost are used to calculate the weights of the mathematical graph.

There are the examples of minimal cuts in Fig. 3:

- If the cut is $C = \{(S, A), (S, B)\}$, then the nodes *A* and *B* remain connected with the sink *T*, that means they belong to segment '1' and the minimal energy is $D_A(0) + D_B(0)$ (see Fig. 3a).
- If the cut is $C = \{(A, T), (B, T)\}$, then the nodes *A* and *B* remain connected with the source *S*, that means they belong to segment '0' and the minimal energy is $D_A(1) + D_B(1)$ (see Fig. 3b).
- If the cut is $C = \{(S, A), (A, B), (B, T)\}$, then the node *A* remains connected with the sink *T*, but *B* – with the source *S* and the minimal energy is $D_A(0) + D_B(1) + V_{A,B}(0, 1)$ (see Fig. 3c).

C. Maximal flow

According to the theorem of Ford-Fulkerson, the problem of the minimal cut can be solved by finding the maximal flow, when the edges are associated with the pipelines of water, but the weights of the graph are presented as a flow capacity [11], [13]. According to the principle, that "the pixels of the same segment have the higher potential energy, then the pixels of the border", the boundary pixels create the bottlenecks for the flow from the source *S* to the sink *T* and the bottlenecks are the min-cut of the graph (see Fig. 4), which can be find by the logical expression (2):

$$Edge\ is\ boundary,\ \text{iff}\ \ flow\ capacity = flow\ AND\ flow > 0. \qquad (2)$$

There is the example of the bottleneck in Fig. 4:

- (*A*, *B*): *flow capacity* > *flow*, (*A*, *B*) is not the boundary.
- (*B*, *C*): *flow capacity* = *flow*, (*B*, *C*) is the boundary. *B* belongs to the source, but *C* – to the sink, according to the direction of the flow (*flow* > 0).
- (*C*, *D*): *flow capacity* > *flow*, (*C*, *D*) is not the boundary.

The maximal flow can be found, for example, by Dinic's algorithm, which is based on two principles:

- The shortest paths from *S* to *T* must be filled firstly.
- A residual graph is constructed in all iterations. The residual graph represents the remaining paths and their flow capacity. The flow capacity of *T*-links is calculated by the formula $f(t) = f(t-1) - dx(t)$, where *f* – new flow capacity, *t* – a current iteration and *dx* – a launched flow in the iteration *t*. When the flow capacity of *N*-link is decreased, the flow capacity of the reverse edge must be increased by the formula $f(t) = f(t-1) + dx(t)$.

D. Seeds

The paper describes the seeds-based method. The seeds are the marked pixels, which identify the location of the object and the background. Seeds

organize the input and output points of the water flow. These seeds are converted by the label cost function $D_p(L_p)$: if pixel is marked, *T*-link has the flow capacity equal to infinity, else it is 0.

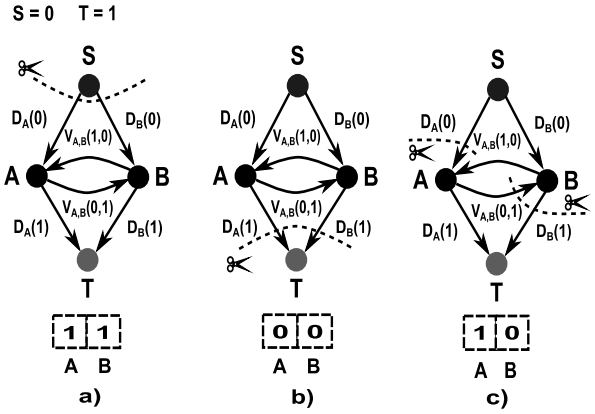

Fig. 3. Different cases of minimal cut

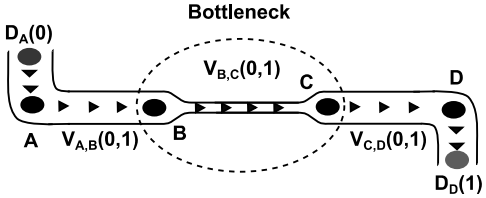

Fig. 4. Bottleneck of water flow

E. Algorithm of Min-Cut/Max-Flow Segmentation

1) select the prior knowledge about the image;
2) define the label cost $D_p(L_p)$;
3) define the smooth cost $V_{p,q}(L_p, L_q)$;
4) transform image into the weighted directed mathematical graph;
5) find the maximal flow;
6) mark the pixels of the source and the sink;
7) mark the pixels with the bottlenecks using the function (2);
8) fill in the segments using the pixels of the source and the sink as start points;
9) mark the pixels of the border as the category of the source.

V LAND COVER RECOGNITION USING MIN-CUT/MAX-FLOW SEGMENTATION

The paper discusses the land cover recognition, which classifies the pixels into three categories: water, field and forest.

The orthoimages with the high resolution have too many classes, which can be sensed: cars, road markings, parasols and other small objects. All these small objects create a noise for the land cover recognition. This noise can be filtered by the small resolution, therefore the resolution of 2.5 m./pix. is selected.

The land cover recognition using min-cut/max-flow segmentation is based on the principle, that the seeds of the object and the background are placed automatically using some recognition method, therefore it is the recognition system, which integrates two modules:

- A seeds marker, which marks the position of the objects using some recognition method.
- A segmenter, which segments the objects using the seeds defined by seeds maker.

As a result, the quality of the recognition depends on the potential of the min-cut/max-flow segmentation, which itself depends on the quality of the seeds maker.

The quality of the land cover recognition is assessed by Cohen's kappa coefficient (3) [16], [17]:

$$k = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}\cdot x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}\cdot x_{+i})}, \qquad (3)$$

where r indicates the number of rows in the error matrix (the headers of columns and rows are categories, the rows indicates the result of classification, but the columns – the waiting result), x_{ii} – the number of observations in the row i and the column i, x_{i+} - the sum of the i-th row observations, x_{+i} - the sum of the i-th column observations, N – the number of the elements.

Cohen's kappa evaluation method needs the classified image and the sample of waiting result, which are compared. The orthoimage is depicted in Fig. 5a. and the waiting result of the classification is depicted in Fig. 5b:

A. Potential of min-cut/max-flow segmentation

If the seeds are manually marked by a human, then it is possible to accept, that the seeds maker has the best possible quality. As a result, it is possible to evaluate the potential of min-cut/max-flow method as the method of the land cover recognition, because the result only depends on the quality of the segmenter in this case. The experiment has showed the result of Cohen's kappa coefficient equal to 0.93 (see Fig. 5c).

B. Automatization using logical reasoning

According to the chapter 4E, the first step is to select the prior knowledge about the image. It is done using the color histograms of three categories (see Fig. 6-8). It is possible to see that the categories have the intersections, therefore the disjoint ranges are selected for the logical reasoning of the seeds maker, but the intersections will be classified by the segmenter. But the min-cut/max-flow algorithm divides the image into two segments, therefore the land cover recognition must be organized as two steps process:

1) water/land segmentation: $D_p($ 'water' $) = \infty$, if $x_p < 28$; and $D_p($ 'surface' $) = \infty$, if $x_p > 39$;

2) forests/fields segmentation: D_p('forest') = ∞, if x_p < 79; and D_p('field') = ∞, if x_p > 111.

The smooth cost can be defined using the function (4) and Manhattan distance (5):

$$V_{p,q}(L_p, L_q) = \begin{cases} \infty, & if\ |x_p - x_q| = 0 \\ \frac{1}{|x_p - x_q|}, & if\ |x_p - x_q| > 0 \end{cases} \quad (4)$$

$$\Delta x = |x_p.red - x_q.red| + |x_p.green - x_q.green| + |x_p.blue - x_q.blue|. \quad (5)$$

The function (4) satisfies all conditions described in the 4A chapter.

The method was evaluated by Cohen's kappa coefficient using the image of Fig.5a. The results of the experiment are depicted in Fig.5d and in Table 1.

TABLE I.

ASSESSMENT OF LAND COVER RECOGNITION

	Land cover recognition	
Cohen's kappa	0.7458	
	User's accuracy	**Producer's accuracy**
Water	86.27%	98.78%
Forests	94.72%	76.04%
Fields	68.04%	79.02%

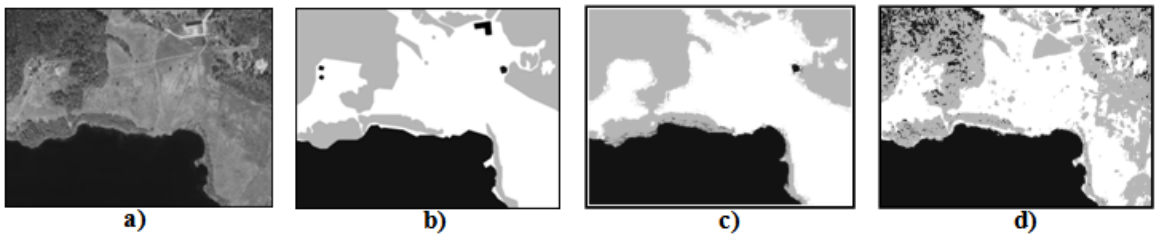

Fig. 5. a) orthoimage; b) waiting result; c) result of segmenter examination; d) result of land cover recognition

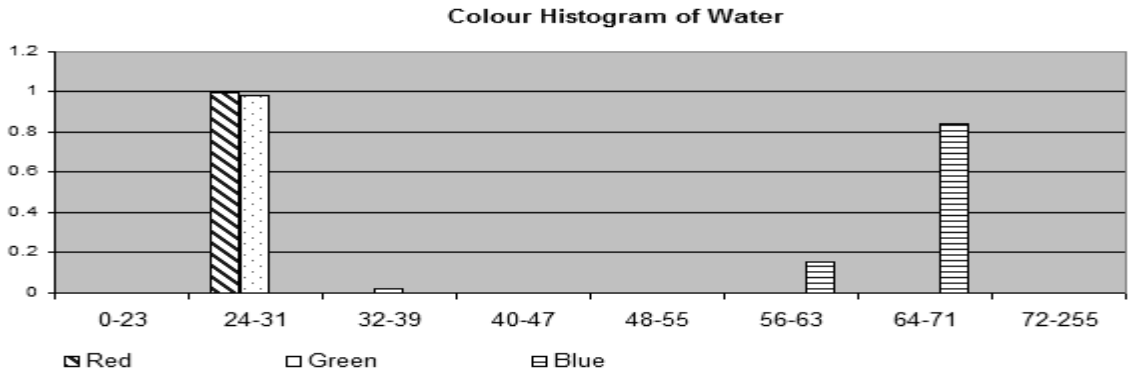

Fig. 6. Colour histogram of water

Fig. 7. Colour histogram of forest

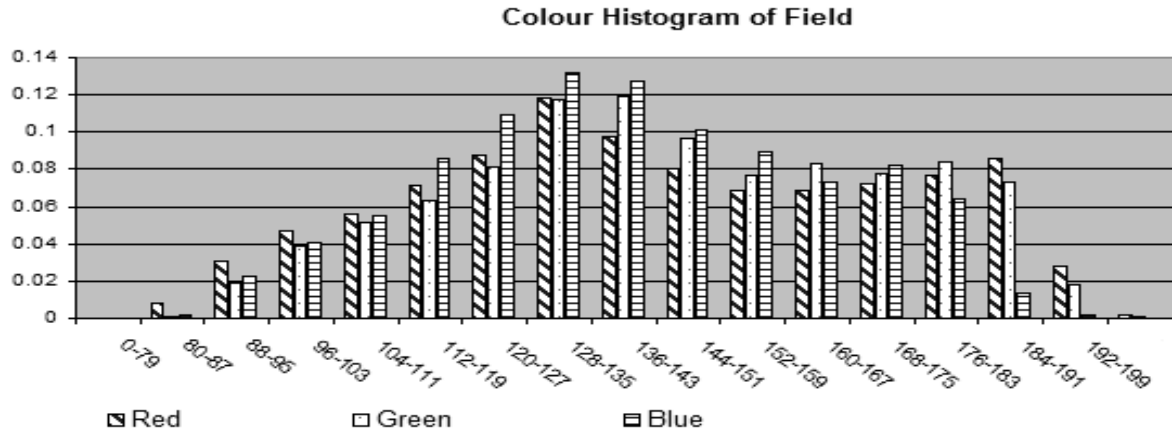

Fig. 8. Colour histogram of field

VI CONCLUSIONS

It was proved that the min-cut/max-flow segmentation can be used for the automatic land cover recognition. The method is organized as two step process: the seeds finding and the segmentation; when min-cut/max-flow algorithm classifies the remaining parts using the energy minimization principle.

The experiment was organized in two steps:

- the assessment of the potential of the proposed method using the conditions of the ideal seeds maker (manually marked seeds);
- the assessment of the method using the seeds maker, which was based on the logical reasoning (fully-automatic recognition).

The results of the recognition were measured by Cohen's kappa coefficient, which uses the image of the waiting result to express the quality of recognition.

Changing the seeds maker's algorithm, it is possible to construct the different recognition systems, which have the different quality of the recognition and which can be applied for different computer vision problems. The paper is connected with the land cover recognition problem. The seeds maker of the described solution was based on the logical reasoning, which had been defined using the color histograms of three categories: water, fields and forests. The assessment of this method has showed the result – Cohen's kappa coefficient equal to 0.746.

According to the fact, that the recognition system consists from two parts, its quality depends on each of them. The experiment has showed that min-cut/max-flow segmentation has the potential, which is equal to Cohen's kappa coefficient 0.93, therefore the seeds maker must be modified to increase the accuracy of the method.

According to the principle, which is used by min-cut/max-flow algorithm, the ideal seeds maker must mark the pixels with 100% precision, when the number of the seeds is not so important, it can be only one point for each simple segment. In this case, the simple segment is the segment, which does not have some subcategories, which can have the greater distances (which can form the smaller bottlenecks) than the distances between the main classes.

VII ACKNOWLEDGEMENT

The author expresses gratitude to the State Land Service of Latvia for providing samples for research purposes.

VIII REFERENCES

[1] Yongwel Sheng, Peng Gong and Gregory S., "Biging.True Orthoimage Production for Forested Areas from Large-Scale Aerial Photographs," American Society for Photogrammetry and Remote Sensing, Photogrammetric Engineering & Remote Sensing, Vol. 69, No. 3, pp. 259-266, March 2003.

[2] Fahmi Amhar, Josef Jansa and Christine Ries., "The Generation of True Orthophotos using a 3D Building Model in Conjunction with a Conventional DTM," IAPRS, Vol. 32, Part 4, pp. 16-22, 1998.

[3] T. Veljanovski, U. Kanjir and K. Ostir., "Object-based image analysis of remote sensing data," Geodetski Vestnik, vol. 55, no. 4, pp. 678-688, 2011.

[4] Y. Babykalpana and K. ThanushKodi., "Classification of land use land cover change detection using remotely sensed data," International Journal on Computer Science and Engineering (IJCSE), Vol. 3, no. 4, pp. 1638-1644, 2011.

[5] Military Memorial Park, Pákozd, "The world's highest spatial resolution aerial orthophoto map produced by aeroplane," *Military Memorial Park, Pákozd*. [Online]. Available: www.aerialrecord.com. [Accessed 05.04.2013].

[6] S. S. Al-amri, N. V. Kalyankar and S. D. Khamitkar, "Image Segmentation by using Edge Detection," International Journal on Computer Science and Engineering, Vol. 2, pp. 804-807, 2010.

[7] Y. Ramadevi, T. Sridevi, B. Poornima and B. Kalyani, "Segmentation and object recognition using edge detection techniques," International Journal of Computer Science & Information Technology (IJCSIT), Vol. 2, no. 6, pp. 153-161, December 2010.

[8] Sergejs Kodors and Imants Zarembo, *Urban Objects Segmentation Using Edge Detection*, Environment. Technology. Resources, Proceedings of the 9th International Scientific and Practical Conference. Vol. 2, pp. 26-29, 2013.

[9] G. Ignatov, H. Nikolov, D. Petkov and G. Georgie, "Segmentation of satellite images by means of morphological and object-oriented approaches," International Society of Photogrammetry and Remote Sensing (ISPRS), Vol. 36, part 7, p. 4, May 2006.

[10] Mohamad Awad, Kacem Chehdi and Ahmad Nasri, "Multicomponent Image Segmentation Using a Genetic Algorithm and Artificial Neural Network," IEEE Geoscience and Remote Sensing Letters, Vol. 4, No. 4, pp. 571-575, October 2007.

[11] Zhenyu Wu and Richard Leahy, "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, pp. 1101-1113, November 1993.

[12] Vladimir Kolmogorov and Ramin Zabih, "What Energy Functions Can Be Minimized via Graph Cuts?," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 2, pp. 147-159, February 2004.

[13] Yuri Boykov and Vladimir Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 9, pp. 1124-1137, September 2004.

[14] Yury Boykov and Gareth Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation," International Journal of Computer Vision, Vol. 70, No. 2, pp. 109–131, 2006.

[15] Andrew Delong, Anton Osokin, Hossam N. Isack and Yuri Boykov, "Fast Approximate Energy Minimization with Label Costs," International Journal of Computer Vision, Vol. 96, no. 1, pp. 1–27, January 2012.

[16] Sergejs Kodors and Imants Zarembo, *Land Cover Recognition with Logical Reasoning using Orthophoto Images*, RCITD 2013, Proceeding of the 1st Research Conference In Technical Disciplines, pp. 116-120, November 2013.

[17] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data", remote sens. environ., Vol. 37, pp. 35-46, 1991.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 134-137

Models of data and their processing for introductory courses of computer science

Mikola Kryshchuk[1], Juris Lavendels[2], Vjaceslavs Sitikovs[2]
[1]Ukrainian National Technical University,
[2]Riga Technical University

Abstract. **Evolution of secondary school course on Informatics during 1990s is discussed. This evolution gave for secondary school graduates rather high level knowledge of applications usage. On the other hand Informatics became purely pragmatic matter within generally academic nature of secondary school curricula. As the result graduates of secondary schools are not prepared enough for mastering of university's course in Computer Science. Authors are not aware of serious studies and methods of prevention of such negative habits caused by evolution of Informatics course. In this article one method applied in university's introduction course of Computer Science is considered. The method provides use of algorithmic system closed to the human mind, and to some extent compensates topics removed from the Informatics course in secondary school.**

Keywords: **Algorithm analysis and design, informatics, model of data structures.**

I INTRODUCTION

Universities, including technical universities, enroll secondary school graduates with good computer end-user skills and with the confidence that their knowledge in the field of ICT is complete and good [1, 2]. School graduates often choose ICT (Computer Science) studio (specialty), although their knowledge regarding specific of the work of Computer Science technician is superficial. Moreover, understanding of specific of such position like programmer is absent at all.

In this situation university's staff involved in Computer Science introductory course teaching is complicated position to work with the very "florid" audience with a distinctly different starting knowledge, interests and motivations. It is necessary to ensure that the course to be interesting and exploitable for whole audiences regardless of prior knowledge of particular student.

II EVOLUTION OF THE INFORMATICS COURSE IN SECONDARY SCHOOL

The course of Informatics in Latvia in the past 50 years experienced a significant evolution [1] and this process is continuing right now. In Latvia, the course of Informatics in secondary schools emerged in 1963/1964 school year in some schools with in-depth learning of mathematics. Since 1985 the course related to Informatics and computational foundations became mandatory in secondary schools. At that time was developed a tendency to teach pupil to think in algorithmically manner, to develop algorithms as well as using schools' computer equipment to start program development. The course had pronounced academic nature and it was in conformance with the society requirements to the education in the ICT field. During this time the secondary school prepare pupils very well to university's engineering courses, including Computer Science. The first year students well understood the basics of Computer Science. Students understood that the computer just executes algorithm created by human and understood that this algorithm should be presented in the form suitable for computer - in some programming language.

At the beginning of the 1990s both in Latvia and Ukraine appear new trends which give the opportunity to pupil in a great extent to choose training courses by themselves. In practice it leads to reduction of exact sciences hours in pupils' personal portfolios.

In parallel with these trends computers appeared in human communications. Using of personal computers evolved very quickly. It was arose public demand for schools to prepare pupils to use personal computers in everyday life - for working with text documents, using of electronic mail etc. Schools included rapidly computers to their infrastructure and mentioned above public demand was implemented in Informatics course without increasing the number of course's hours. Training of computer usage as end-user extrudes from the courses classical academic topics leaving only one of them - "information and communication technology concepts". On the other hand Informatics course in secondary schools now is considered in terms of ECDL certificate program.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.178

Therefore Informatics course was transformed from academic topic to highly pragmatic course [3, 4, 5], which is basically in contradiction with the mission of the secondary school. This solution ensures public demand, however to the large extent undermine students' preparedness for study in university, especially for engineering studies [3].Preparation of students for engineering studies partially is resolved by optional courses considering classic Informatics topics, however pupils' knowledge and preparedness for such a course differ essentially.

III SOME ISSUES OF COMPUTER SCIENCE INTRODUCTORY COURSE IN UNIVERSITIES

In previous time university's introductory course for Computer science begins with practical implementation of already acquired algorithmic skills - understanding of data concept and realization of algorithms on flowchart level. This practical realization was based on some programming language that students are study [6]. At that the emphasis was made on the solution in the form of an algorithm, whereas the programming language was considered just as a tool of algorithm formulation for the computer and for checking of achieved result.

At present time the situation has changed significantly – students have solid experience with the computer, they have not difficulties with the file creation, copying, editing, using of different applications and they desire immediately to work on their computers. Initially it seems that mastering of programming basics could not be difficult for nowadays secondary schools graduates. However in practice students cannot develop algorithm for particular task solution and realization of this algorithm on some programming language because students have not knowledge of the exact and detailed formulation of the solution to another executor. In other word they have no knowledge of algorithmization. On the other hand, it is not usually possible to provide in university's program an introductory course of informatics as it will cause the delay of the course of Computer Science and thereby will lead to the change the whole study programme [7, 8].

To prepare students for practical programming in a short time we use the model of program - algorithmic system which provides accurate language for input data description, actions with the data and language for description of program's results. In parallel with the model there are considered elements of programming language as equivalent for algorithm's description in the form of flowchart.

IV DATA MODEL AND ITS APPLICATION FOR DESIGN OF ALGORITHMS

Our approach to mastering of algorithmization and programming basics is rooted on a use of simplified model of the computerized solution. The model consists of imaginary data storage - data bureau with predefined semantics and of manipulation with data (elements of flowchart with also exactly-defined semantics). The approach provides equivalence of a solution model and elements of programming language. In a such way students can on the very beginning of the course realize elementary algorithms based on graphic image of the algorithm and a on particular operators of programming language that are equivalent to considered flowchart's elements.

Introduced data bureau is useful both the very beginning of the course and considering the placement and relationship, e.g. global and local data in the context of multi-modular program. Data storage is imagining as a bureau with drawers, at that for each type of data is stipulated specific type of drawer. In each drawer is kept just one value (Fig. 1).

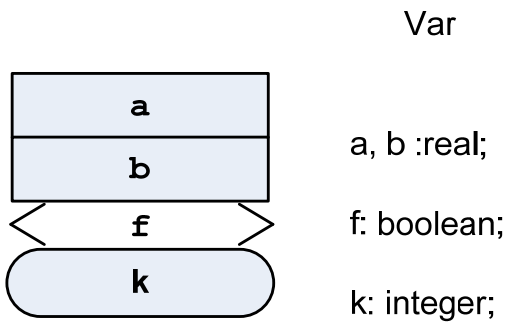

Fig. 1. Data bureau and its equivalence of variables' descriptions in programming language

Students have vague imagination of variables in programming languages [9] because they have no appropriate grounding of computer operations; however they understood immediately semantics of equivalent data bureau:

- Any change of the data value means new data recording to the drawer;
- Inserting data into the drawer the data acquire the type (integer, real, …) that means conversion to the shape of the drawer or refusal in the case when such conversion is not possible;
- It is possible to take the value from the drawer unlimited number of times, but only if it has previously been recorded there;
- If a new value is inserted to the drawer then previous value is lost and is not renewable.

Depending on the nature of the course semantics of data bureau can be given a different granularity, for instance:

- inserting the value to the drawer (without further specification) or
- previous leaflet with written value is removed from the drawer and is destroyed (if there is leaflet in the drawer), the new stored value is

written on the new leaflet, which is inserted to the drawer.

In the similar manner are explained data structures (arrays, records, collections, etc.). Elements of flowchart and equivalent operators of programming language also are explained in relation to the bureau, see examples in Fig. 2 and Fig. 3.

```
S:= p*(p-a)*(p-b)*(p-c);
```

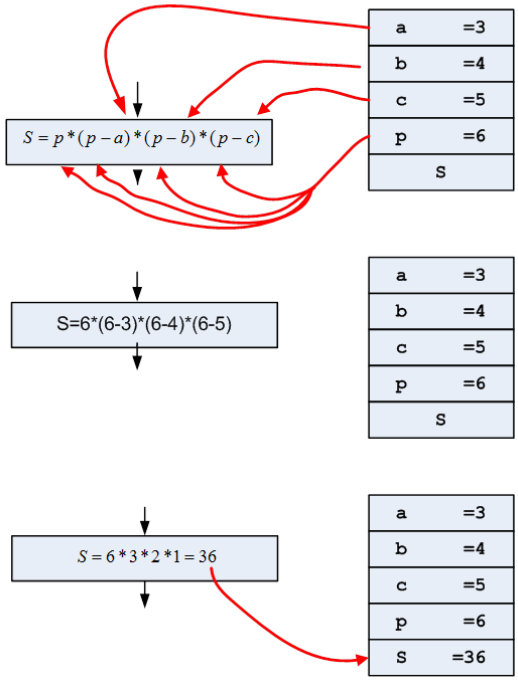

Fig. 2. Semantics of process and assignment

Explaining the teaching material using data bureau, students learn algorithmization in a graphical format proper to human thinking. Semantics of data bureau and elements of flowchart is equivalent to the semantics of programming language operators. This in a great extent facilitates mastering of programming, for instance parameters transfer mechanism and explaining of variables visibility (Fig. 4.).

Described method of algorithmization and methodology of algorithm's realization in program form are practically approved and applied. Semantics of data bureau is practically applied with the different levels of details - at the beginning of the second month we could give up from data bureau on "leaflets" level. During the work data bureau progressively remains unnecessary – it is replaced by data description of a programming language. However considering new programming features, such as internal subprograms and parameters transmission, organization of the external program, visibility o variable etc. it is very useful to introduce new features again on the base of data bureau.

The proposed methodology puts students to work strictly in the algorithmic system with precisely defined semantics. This one hand helps the audience fairly quickly mastering the principles of algorithmization and algorithmic thinking, on the other hand, to some extent compensates the lack of important sections in secondary education informatics course.

This methodology leaves the syntaxes of a programming language in second plan - actually constructions of programming language remain as a solution for presentation of algorithm in the form acceptable for computer. Thus facilities of programming languages are considered just as means of description of a solution in the prescribed form, such as C or PASCAL. This approach provides a simple shift of programming language that is important because the industry is booming. Another advantage of the approach is its direct applicability in the teaching of programming languages built in application software, such as VBA.

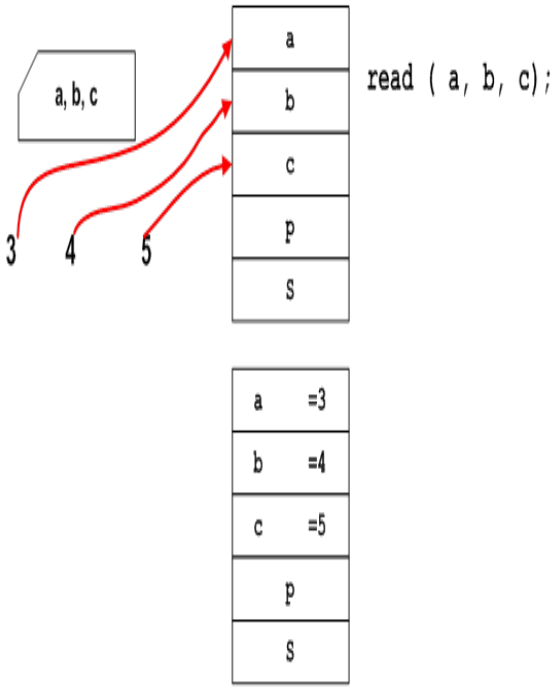

Fig. 3. Semantics of data input.

V CONCLUSION

During 1990s have been made significant changes in Informatics course of secondary school that provided public demand – secondary school graduates are skilled computer users in application level.

Simultaneously with positive aspects of this evolution also appeared negative trends - school graduates are poorly prepared for the university's course of Computer Science. These negative trends are little studied, and we are not aware of the universal method for correcting of these negative characteristics.

In this article is described one method (fig.1, 4) that could be applied in introductory course of Computer Science, which provides applying to teaching algorithmic system model close to the human mind, and to some extent compensates informatics course sections removed from secondary school curricula.

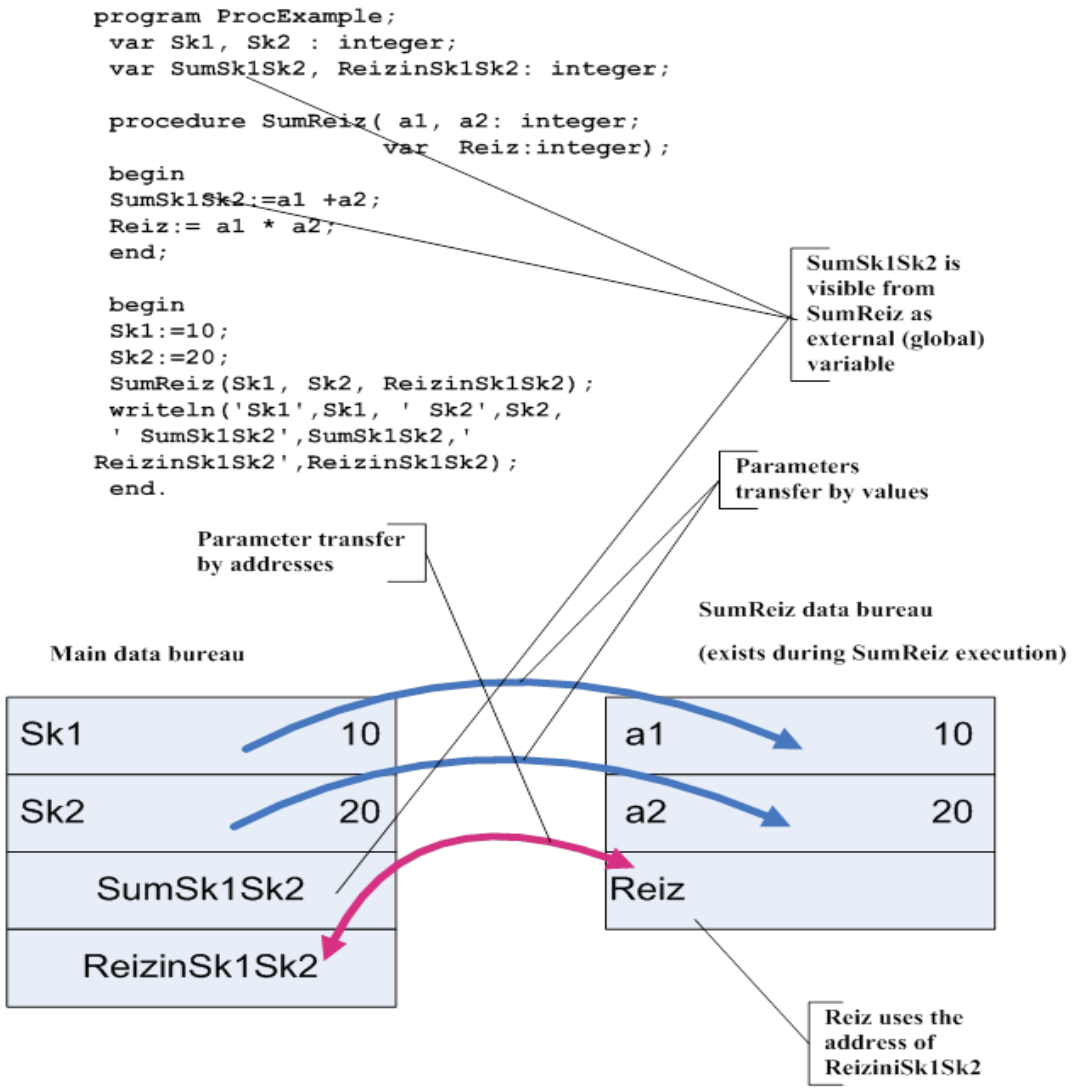

Fig. 4. Visibility of variables and mechanisms of parameter transfer to sub-programs.

VI REFERENCE

[1] V. Vezis, "Informatics at the school," Dr. sc. comp work summary, University of Latvia, Riga, Latvia, 2004., pp. 36.

[2] V. Vezis, "Contents of School Informatics," in *Proceedings of The LatSTE'2004*, LU, Riga, Latvia, 2004, pp. 92–99.

[3] "Latvia general secondary education standard informatics," 2013. [Online]. Available: http://likumi.lv/doc.php?id=257229

[4] "Ukrainen general secondary education standards," 2011. [Online]. Available: http://zakon4.rada.gov.ua/laws/show/1392-2011-%D0%BF

[5] "Ukrainen general secondary education standard informatics," 2014. [Online]. Available: http://refdb.ru/look/2479127-pall.html

[6] P. Zarins, V. Vēzis, "Mathematics and informatics didactic course optimisation problems and their solutions," in Mathematics teaching: history and perspective (3. international Scientific Conference,) LPA, Liepaja, Latvia, 2002 pp. 157–161.

[7] A. Anohina-Naumeca, J. Grundspeņķis, J. Lavendels, V. Šitikovs, V. Vinogradova, "Improvement of the University Programme on the Base of Modularity," in European Integration Studies: Research and Topicalities, 2008, pp. 25-30. ISSN 1822-8402.

[8] A. Anohina-Naumeca, J. Grundspeņķis, J. Lavendels, V. Šitikovs, V. Vinogradova, "Modularity in study programme "Computer systems" development," in The quality of education: business computer systems (The Riga symposium), RTU, Riga, Latvia, 2010 pp 13-22. ISSN 1691-631X.

[9] I. Rivkind, T. Lisenko, L. Chernikova, V. Chahodko, Informatics, (in Ukrainean) Geneza, Kiev, Ukraine, 2010, pp. 302. [Online] . Available: http://4book.org/uchebniki-ukraina/10-klass/483-informatika-10-klas-rivkind.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 138-150

iOS Applications Testing

Ivans Kulesovs
Faculty of Computing, University of Latvia.
Address: 19 Raina Blvd., Riga, LV-1586, Latvia,
Enterprise 2.0 Department, C.T.Co SIA.
Address: 15/25 Jurkalnes Str., Riga, LV-1046, Latvia

Abstract. **Mobile applications conquer the world, but iOS devices hold the major share of tablets market among the corporate workers. This study aims to identify the aspects (i.e. features and/ or limitations) that influence the testing of the native iOS applications. The aspects related to general mobile applications testing are identified through the systematic literature review of academic sources. iOS applications testing aspects are identified through the review of non-academic (multivocal) literature sources. The identified aspects are merged and discussed in detail using the reviewed sources and based on the author's professional experience in iOS applications testing. The references to the credible sources are provided in order to support the professional experience findings. The study eliminates the gap that exists in the academic world in regards to iOS applications testing. The practitioners are also encouraged to fulfill their iOS applications testing strategies with the identified aspects.**

Keywords: **iOS applications, mobile computing, software testing, software quality.**

I INTRODUCTION

According to Clearwater Technology Mobile Computing Sector Report [1], the mobile computing industry is expected to be worth almost US $330 billion by 2015. According to the same study, 67% of corporations allow workers to use the tablets, but 51% of corporations even buy the tablets for them. iOS devices hold 75% of the tablets market share among the corporate workers.

With the growth of platform abilities applications become more complex [2] to satisfy the increasing user needs. The increased complexity means that there are many aspects that should be taken into consideration when testing mobile applications. Mobile workers mostly use native business applications on their devices; otherwise there would not be such a dominant position of the single operating system. That is why iOS native applications are the subject of the main interest for this study.

Despite the fact that the topic being hot, there are only some academic studies [3] – [5] performed that systemize the generic aspects that should be taken into consideration when testing the mobile applications without specifying the platform. Other studies – [6] and [7] that include the clear distinction between the platforms, concentrate on some narrow topic. On the other side, there are different iOS testing checklists, mind maps, blogs etc. available in the internet. This motivates the author to perform the systematic literature review of academic literature in the field of mobile testing and perform the literature review of the available non-academic (or multivocal, as per [8]) sources in the field of iOS testing.

It was decided to concentrate both reviews on aspects of manual testing of such quality characteristics as functional suitability, performance efficiency, compatibility, reliability, maintainability, and portability according to [9]. That is why the test automation, security, and usability testing are out of scope (except parts that are closely related to or are on the border line with the quality characteristics mentioned above).

The following research question was formulated:

RQ: *Which aspects (i.e. features and/ or limitations) influence the testing of functional suitability, performance efficiency, compatibility, reliability, maintainability, and portability of the iOS native business applications?*

The results of both reviews are merged in order to answer the research question. The goal of the study is to eliminate the gap that currently exists between academic and non-academic sources in the field of iOS applications testing, as well as to provide the sufficient details for practitioners to make their iOS applications testing strategy more complete and solid.

The paper is organized as follows: in Section II the research methodology is described; in Section III the merged results gathered through both reviews are presented; in Section IV the details of the identified iOS testing aspects are discussed, and conclusions based on the findings are presented in Section V.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.187

II RESEARCH METHODOLOGY

The systematic literature review (SLR) of the academic sources was performed in order to gain the aspects of the mobile applications testing. The multivocal literature review (MLR) was performed in order to gain the exclusive aspects of iOS applications testing. The idea to perform two types of the review to consolidate the data from different sources was taken from work by Tom et al. [10] Fig. 1 shows the stages of sources selection for the whole review process applied in this paper.

The procedure described by Kitchenham and Charters [11] was followed in order to conduct the systematic literature review. The qualitative review approach was applied in order to include a rigor into the systematic review of multivocal literature as suggested by Ogawa and Malen [8]. They define the multivocal sources as accessible, but non-academic writings on the topic.

A. Systematic Literature Review

The search of academic literature for SLR was performed in two iterations. The first iteration was executed using the databases and search criteria that are described below. 12 papers were selected as relevant to answer RQ. The second iteration was executed based on the references in the papers selected after the first iteration. Some relevant sources were found, but they appeared to be non peer-reviewed. The details of each iteration can be found in Table 1.

1) Databases. The following databases were used to search the keywords described in the Search Keywords section: Springer Links (Springer), IEEE Xplore (IEEE further in the text), and ACM Digital Library (ACM).

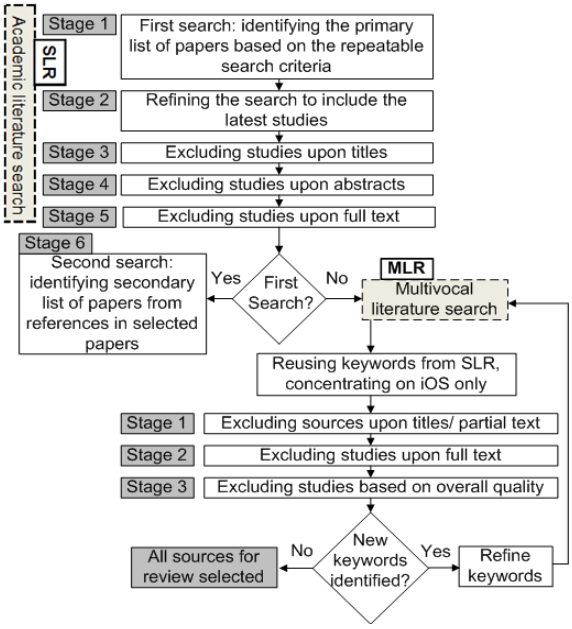

Fig. 1. Process of sources selection for SLR and MLR

2) Search Keywords. Appropriate keywords were searched in metadata. Due to the search engines differences, metadata should be treated as a search within the title, OR abstract, OR keywords for ACM and as a search within the title only for Springer, while IEEE has an option to search within all metadata at once.

TABLE I

NUMBER OF PAPERS LEFT AFTER EXCLUSION/ INCLUSION DURING EACH SLR STAGE

ITERATION	STAGE	NUMBER OF ACADEMIC WORKS AFTER STAGE
1	1. Initial repeatable search (duplicates removed)	946
	2. Refined search to include works from 2014	972
	3. Exclusion upon titles	33
	4. Exclusion upon abstract	18
	5. Exclusion upon full text	12
2	6. Secondary search based on references in selected results	0
	Total:	12

Because preliminary search of keywords “iOS application” and “testing” or “iOS application” and “quality” returned small amount of results, the keyword “iOS” was substituted with “mobile”. It was also given a try to shorten the word “application” to “app”. The following search string was used: (("iOS apps" OR "iOS applications") OR ("iPhone OS apps" OR "iPhone OS applications") OR ("mobile apps" OR "mobile applications")) AND ("quality" OR "testing" OR “verification” OR “validation”).

3) Inclusion/ Exclusion Criteria. Only peer-reviewed papers in English were selected. There was no limitation given on the type of the source (i.e. journals, conference proceedings etc.). Papers starting from the year 2007 were chosen, because it is the year when iOS (iPhone OS at that time) was released. The year 2013 was chosen as the last year of publication for the search results repeatability. The search was also refined by adding papers from the year 2014 in order not to miss the latest available information.

Irrelevant papers were excluded upon title, then upon abstract, and then upon full text. The main credit was given to the papers that offered some categorization or general overview of mobile applications testing. Papers that mention only the specific testing type of mobile applications (i.e. unit

testing, security, usability etc.) or that are related to test automation were excluded from the results after additional acquaintance with abstract because they do not focus on the aspects asked in RQ. Only non-shortened version of papers were included if two versions of the same paper for different occasions (e.g. conference proceedings and magazine) were identified.

4) Data Extraction and Synthesis. The data extraction phase involved the extraction of aspects and categories of aspects related to RQ from the selected studies. The categories of multiple non-overlapping aspects are mentioned in some papers, while the detailed description of aspects from single category is mentioned in others. The data synthesis phase includes the merge of aspects from the different papers that appeared to have the same meaning. In order to make the data more usable the aspects were divided between 4 large clusters: Environment, Application Lifecycle, Inside the Application, and (functional or performance aspects of) UI/ UX.

B. 2.2. Multivocal Literature Review

1) Data Sources and Search Strategy. Sources for MLR were searched in Google (http://www.google.com/). The combination of the same keywords as for SLR, excluding the "mobile applications" OR "mobile apps" part, was used for the first search iteration. The keyword "checklist" was added for the second iteration. The first 50 relevant articles per iteration (see Appendix B) based on the Google ranking algorithm were taken for subsequent analysis.

2) Inclusion/ Exclusion Criteria. The sources were excluded during three stages by evaluating 1) Title/ partial text available in the search results; 2) full text; 3) overall quality. The sources related to iOS testing only were included into the final results, i.e. the sources containing only information about general mobile testing aspects were excluded. The sources on security or unit testing, as well as the sources on testing automation were excluded as well. Duplicates were excluded upon title during the first exclusion stage. Some sources were excluded on the second stage because they directly referred to other sources found. The inclusion/ exclusion progress is depicted in Table 2.

TABLE II

NUMBER OF PAPERS LEFT AFTER EXCLUSION DURING EACH MLR STAGE

ITERATION	INITIAL	STAGE 1	STAGE 2	STAGE 3
Iteration 1	50	20	5	5
Iteration 2	50	20	5	4
Total				9

3) Data Extraction and Synthesis. The data extraction phase involved the extraction of aspects and aspects categories asked in RQ. Some sources already contain categorized lists of aspects while other are materials written in narrative. The data synthesis phase includes the merge of aspects from the selected sources. The identified aspects were divided between the same clusters as done for SLR.

III RESULTS

A. Summary of Reviews

Despite the fact that the search criteria for SLR includes studies starting from 2007, the first selected study was published in 2009 [4], but the most productive years are 2012 (5 studies: [6], [12], [13], [3], and [14]) and 2013 (3 studies: [15], [16], and [17]). 2 studies [7] and [18] were published in 2011, and 1 study [5] was published in 2014. 2 studies [6] and [7] are related to narrow topic of mobile application lifecycle, 1 study [17] is related to user complaints about iOS applications, and other 9 sources [3] – [5], [12] – [16], and [18] are related to general testing of mobile applications.

Between the sources selected through MLR, 7 sources [19] – [25] were published in 2013. 1 source [26] was published in 2012 and 1 source [27] in 2014. 5 sources [20] – [22], [27], and [26] are blog posts, 2 sources [19] and [25] are testing checklists, 1 source [23] is a white paper, and 1 source [24] is a mind map. All the blog posts describe the testing only of one or some aspects, while other sources try to cover the whole iOS testing field.

B. Aspects of iOS Applications Testing

The aspects that influence the testing of iOS applications gathered through SLR and MLR are shown in Table 3. If a source is referred in the table before the details of an aspect, it means that aspect is just mentioned in the source without pointing the details that are related to iOS applications testing.

There are three types of iOS devices: iPad, iPhone, and iPod mentioned in [19], and [23] – [25] that have different screen size, resolution & pixel ratio, processing efficiency, memory, and storage capacity, as per [3] – [5], [13], [14], [16] – [18], and [23] – [25]. It is claimed in [4] that functionalities, usability issues in the interface design, and user behavior "to be tested in emulator", while other sources [13], [16] – [18], and [27] state that almost everything should be tested on the real device to get the reliable test results. There are also different types of the external accessories, both wired and wireless [12], [15], like headphones [12], [15], [25] and keyboard [12], [15] that can be connected to the device.

It is claimed in many sources [3], [5], [13], [16], and [23] – [25] that the variety of operating systems (OS) is an important testing aspect, while OS upgrade is mentioned explicitly only in [16]. It is possible to set the restrictions on the usage of different hardware or OEM software completely or for the specific application within the iOS [19], [22], [24].

Mobile devices have limited power, processing, and memory resource [3], [4], [6], [12], [14], [16], [17]. Thus resources consumption efficiency plays an important role in application success [3], [4], [12], [14], [17], [19]. Applications should also be checked on different networks, i.e. strong WiFi connection, cellular network (LTE, 3G, EDGE), and in Airplane mode [3] – [5], [12], [21], [22], [24], [25].

Different network conditions (e.g. slow connection, packets loss etc.) should be taken into consideration as well [24]. Different regional settings, like data and time formats [24], [25], as well as time zones [25] are also the subject of interest.

iOS application lifecycle consists of several phases, and there are specific conditions that can uniquely influence application's behavior while being in the definite phase. An application can be just installed and launched for the first time [14], [19], [25], work in foreground, stay in background [6], [7], [12], [19], [24], receive memory warnings [6], [7], [12], [18], [24], be interrupted by a call or SMS [15], [19], system alert [15], push notification [19], [24], GPS signal [15], or audio/ video from another application [19], [24], [25]. It can even crash [16], [17], [23], [24]. Or it can also be updated to the next version [16], [24], [25].

[22] warns about the need to check an extended (Chinese) on-screen keyboard, while [4] mentions on-screen keyboard as a generic aspect that should be taken into consideration. According to [19] and [24] data can be shared via email or Bluetooth, or another network between the applications. According to [24] and [25] it is necessary to check application's logging and analytics features. Testing of In-App Purchase component is mentioned in [24]. Testing of Web View component is mentioned both in [5] and [24].

TABLE III

ASPECTS OF IOS APPLICATIONS TESTING

ENVIRONMENT		
HARDWARE		
Devices	iPad, iPhone, iPod [19], [23] – [25].	Screen size, resolution & pixel ratio, processing efficiency, memory, storage capacity; [3] – [5], [13], [14], [16] – [18], [23] – [25] motion activities [24], [25].
Simulator	[3], [4], [13], [16], [18], [27].	
External Accessories	Headphones [12], [15], [25], keyboard [12], [15]; wired/ wireless [12], [15].	
OPERATING SYSTEM		
OS Variety	[3], [5], [13], [16], [23] – [25].	OS upgrade [16].
Restrictions and Privacy Settings	[4], [5], [16], [17]. Safari, Camera, Siri, IAP (in-app purchase), Location Services, Contacts, Calendars, Photos, Social Networking, Microphone, Motion Activities, Cellular Data Use, Background App Refresh [19], [22], [24].	
RESOURCES		
Limitations	Lack of storage, amount of memory, running out of battery, processing capabilities. [3], [4], [6], [12], [14], [16], [17]	
Consumption	Memory consumption, battery consumption. [3], [4], [12], [14], [17], [19]	
CONNECTIVITY		
Network Types	WiFi, Cellular networks; [3] – [5], [12], [21], [25] Bluetooth [3], [4], [12], [25]; Airplane mode [24], [25].	
Network Conditions	[3], [4], [12], [14], [15], [17]. Strong/ no/ poor connection; connection loss [19], [21], [22], [24]. Ask for connection [4].	
INTERNALIZATION		
Region Formats	[14]. Date format, hour format [24], [25].	
Date/ Time Settings	Switching between time zones, system time too fast/ too slow. [24]	
APPLICATION LIFECYCLE		
Installing and Launching	[14], [19], [25].	
Background	[6], [7], [12], [19], [24].	
Crash	[16], [17], [23], [24].	
Low-Memory Warnings	[6], [7], [12], [18], [24].	
Interruptions	[4], [6], [7], [12], [14].	

	Call/ SMS [15], [19]; push notifications [19], [24], system alerts [15]; GPS signal [15]; audio/ video [19], [24], [25].
Application Update	[16], [24], [25].
INSIDE THE APPLICATION	
Keyboard	[4]. Extended keyboard [22].
Data Import/ Export	Email; Bluetooth/ network (peer to peer) [19], [24].
Logging/ Analytics	[24], [25].
In-App Purchases	[24].
Web View	[5], [24].
UI/ UX	
Gestures	[5], [24].
Smooth Animation	[17], [24].
Pull to Refresh	[24].
Orientation	Portrait, landscape. [4], [12], [24], [25]
Half Pixels	[24].
Localization	[16], [25]. Native characters and special symbols [19].
Accessibility	VoiceOver, accessibility zoom etc. [20], [26], [23] – [25]

An application can be manipulated with a variety of gestures [5], [24]. When animated transitions occur, they must run smoothly [17], [24] irrespectively of the task executed in parallel. Testing for half pixels glitches and testing of Pull to Refresh feature are mentioned in [24]. The necessity of checking the application both in portrait and landscape is noticed in [4], [12], [24], and [25]. The importance of localization testing is mentioned in [16] and [25]. [19] identifies the need for testing of native characters and special symbols. It should also be checked that application works as designed when accessibility features of OS are enabled [20], [23] – [26].

IV DISCUSSION AND IMPLICATIONS

Despite the fact that the Results section shows the identified aspects of iOS applications testing gathered through SLR and MLR, the author feels the necessity to discuss the details of identified aspects. There are also some aspects that are known to the author (like iAd, update of Xcode, AirDrop etc.), but they are missing in the reviewed literature. Some of the details are provided in the reviewed sources. Others are added based on the author's more than four years of professional experience of leading more than 20 iOS applications testing projects for several Global Fortune 500 [28] and other multinational companies, giving the references to iOS Developer Library [29] or other credible sources where possible.

A. Hardware

1) Devices. While there are four types of iOS devices, business applications are mostly developed for iPads [30], and sometimes have reduced iPhone versions [31]. iPods and Apple Watch devices generally are out of scope.

iPad 1st generation devices, as well as iPhone 3G, iPhone 3GS, and iPhone 4 are not taken into consideration anymore when new applications for iOS are developed. Only iPhone 4 supports iOS 7 (the latest iOS version at the moment of writing is iOS 8), but three other mentioned devices already are not [32].

iPad 2 [33] and iPad mini [34] both have non retina display (i.e. a display with lower pixel density than the latest iOS devices) and generally the same hardware options. They are the least powerful iPad devices that support the latest iOS version. Special checks that application design fits the small screen of the device and that every UI control can be easily interacted with should be performed on iPad mini.

iPad 4 [35] and iPad 3 [36] both have retina displays, but iPad 4 is more powerful than iPad 3. Generally, it is enough to have only one device of any generation to cover this category of devices.

iPad Air [37] and iPad mini retina [38] both have faster GPU (but still retina display) and M7 64-bit core processor that has built-in hardware for motion activities like *accelerometer, gyroscope,* and *compass.*

iPad Air 2 [39] and iPad mini 3 [40] are equipped with Touch ID [41] technology. iPad Air is also equipped with even faster GPU and M8 64-bit core processor that has a barometer sensor in addition.

iPhone 4 [42] and higher all have retina displays. iPhone 4S [43] has a faster dual core processor in comparison with iPhone 4. iPhone 5 [44] and iPhone 5C [45] are both packed with even faster next generation processor. iPhone 5S [46] is packed with already mentioned M7 64-bit core processor and fingerprint identity sensor.

Despite the fact that iPhone 5th generation devices have larger screen size in comparison with iPhone 4th generation devices, applications designed for iPhone 4th generation devices can still run on iPhone 5th generation devices, but there are black bars above and below application content, unless properly named and to a larger screen accordingly sized launch image is provided [47].

iPhone 6 [48] and iPhone 6 Plus [49] have M8 64-bit core processor, larger and even larger screen size, and already mentioned fingerprint identity sensor, barometer, and other sensors. Separate image resources should be prepared for applications to look smooth on iPhone 6 Plus.

iPhone 6th line supports Near Field Communications-based mobile payment technology Apple Pay [50]. Apple Pay can be used also on iPhone 5th line, but only when paired with Apple Watch. Payments can be made also on iPad Air 2 and iPad mini 3, but only within the applications.

Generally speaking, one device from each generation would be enough to cover the whole set of iPhones, in case of application under test does not rely on the specific function of the device like motion activity or finger print of iPhone 5S and higher or Siri (advanced voice control) that is available only starting from iPhone 4S.

2) Device vs. Simulator. The author's professional experience supports the statement expressed in [13], [16] – [18], and [27] that for achieving good quality of the application, it should be tested on the device rather than on the simulator, because testing results can vary. It also should be taken into consideration that application can behave differently when it is built in debug, not in release mode [51].

3) External Accessories. There are different kinds of accessories, both wired and wireless [12], [15], that can be attached to the device: headphones [12], [15], [25], keyboard [12], [15], stylus etc. It can occur that an application handles the inputs and outputs from/ to external accessories in a different way than it does without them, or it does not handle them at all [52]. External accessories from different manufacturers can behave differently, e.g. styluses from different manufacturers can have different configurations inside the application in order to handle the palm (interaction) rejection etc. [53]

B. Operating System

1) iOS Variety. Release of the new iOS version almost always leads to the major retesting cycle for the non-trivial applications. New Xcode [54] version (that includes new version of iOS SDK and compiler) [55] is shipped together with the new iOS version. Thus, there can be completely different test results when the same code is built by the different Xcode versions.

The following update strategy is followed by the development organizations which the author works or worked for when the new iOS version is released:

1. Current application version built by previous the Xcode is checked on the new iOS version (preliminary checks are done already on Beta or GM versions).
2. Major failures, if any, are fixed, and the application is released with remark that it supports the latest iOS version.
3. More thorough testing cycle follows when the application current version is built by the new Xcode afterwards.

It is possible to leave the application version built by the previous Xcode for some period of time if, for example, active development currently is not planned. But here is a list of situations when developers are forced to rebuild the application with the new version of Xcode:

- New iOS version does not support the methods that were previously deprecated, but still used in the application; new supported methods that substitute the deprecated ones are available with the new Xcode, e.g. detection of UDID [56].
- Apple announces that all the new applications or application updates submitted to the App Store must be optimized for new iOS and built with the latest Xcode [57].
- Application should be redesigned for the marketing purposes, because of the iOS redesign (as it occurred with iOS 7 [58]), but new UI is achieved using the latest Xcode.

It is worth mentioning that devices with the previous iOS version should always be available and handled carefully in case some of the applications developed within the organization still support it. It should not be forgotten that there is no official way to install any previous major iOS version after the release of the latest major iOS version [59]. It should be taken into consideration that not all users update iOS version as soon as it is released [60], but can continue to use the previous one for quite a long period of time. From the author's experience, it is especially applicable for enterprise users – they update iOS version only after the enterprise infrastructure that supports the latest iOS version is ready.

2) Restrictions and Privacy Settings. In iOS a user can set different restrictions, both system and application wise, on the usage of different hardware or OEM software. For example, it is possible to restrict the usage of Safari, Camera, Siri, IAP (In-App Purchase), Location Services, Contacts, Calendars, Photos, Social Networking, Microphone, Motion Activities, Cellular Data Use, Background App Refresh etc. [19], [22], [24] The application should handle cases when it tries to access the restricted item. The user should also be warned about the restriction

and instructed how to remove it [61] or offered to remove the restriction within the application if it is possible.

C. Resources

1) Limitations and Consumption. Due to the fact that a mobile device has more limited storage, memory, power, and processing capabilities than an ordinary PC [3], [4], [6], [12], [14], [16], [17], examination of how the applications handle these limits and operate within these limits are of special interest. The application should check for the free storage availability when the new data is added/ downloaded. Otherwise, from the author's experience, the user will not be able to operate with the data that already is inside the application due to crashes. The application should be checked for efficient battery consumption as well [3], [4], [12], [14], [17], [19]. It can be verified using Xcode Instruments tools [62]. Battery usage logging can also be enabled on the device that is provisioned for the development [63]. Instruments tools can also be used for profiling the efficiency of memory and processor resource utilization.

D. Connectivity

1) Network Types. During Alpha testing the application is mostly checked in the laboratory environment [4]: on the strong WiFi connection and in the Airplane mode. The working on the cellular data (LTE, 3G, EDGE) should be checked as well [3] – [5], [12], [21], [25], especially if the application utilizes a lot of traffic. The user, at least, should be warned when large data synchronization occurs on the cellular network.

2) Network Conditions. There are different network conditions possible [3], [4], [12], [14], [15], [17], [19], [21], [22], [24] (e.g. slow connection, packets loss, etc.). It should be checked if under these conditions:

- The application handles different network conditions on the first launch [24].
- Proper error messages are shown on timeouts and other network errors [24].
- The interaction with UI (i.e. the main thread) is not blocked [24].
- The corrupted data is not stored, or at least can be redownloaded.

For simulating different network conditions Apple Network Link Conditioner can be used [21]. This tool is a part of Xcode Developer Tools [64] and can simulate network conditions on the device if the network connection from Mac is shared. It can also be enabled directly on the device that is provisioned for development.

Sometimes it is also necessary to check a poor connection or a connection loss/ switching in the real world. From the author's experience, the most common cases that should not be simulated, but should be checked in the field are:

- The traffic loss while the device "thinks" that it is still connected to the network (e.g. entering the elevator or walking outside the network coverage).
- Switching from WiFi to the cellular network and vice versa, switching from one WiFi access point to another, switching between different cellular network types.
- Only cellular network conditions (e.g. inbound/ outbound connection speed, packet loss ratio etc.) can be simulated, but the device will still think that it is on WiFi. Thus, the real cellular network should be used to check the cellular network specific functionality of the application.
- The situation when the device is not connected to any network should be checked separately to make sure that this condition is treated the same way as the Airplane mode.

E. Internalization

1) Region Formats. Applications should be tested using different region formats [14] that have different hour format (24 or 12) [24], [25] and different coma separators. For example, German Switzerland and United States regions cover these both differences. From the author's experience, the specific Arabic and Israel region formats should be explicitly tested if the application's functionality is directly related to the calendar and weekend days.

2) Date/ Time Settings. When the application receives updates from backend, and especially when creation/ update timestamps for items are visible (but the same also applies for locally created items), it is necessary to check how the application behaves with different time settings [24]:

- When switching between time zones.
- When the system time is too fast or too slow.

Besides checking that functionality works properly itself, it is necessary to check that relative times are properly calculated [24].

F. Application Lifecycle

1) Installing and Launching. The application should be installed both on the device that already contained some version of the application and on the clean device after the factory reset. The user should be warned through a message or a progress bar in case the access to the application functionality is given in more than 5 seconds after launching [19].

2) Background. The background mode is one of the major cycles of iOS application lifecycle. If the application cannot be sent to the background in approximately 5 seconds, then iOS kills it. The same is applicable when going back to the foreground [65]. That is why it is necessary to check that the application changes the state in sufficient amount of time even with the large amounts of data inside. The application can also perform the refreshes in the

background using the special multitasking feature provided since iOS 7 or if it uses the Location Services, plays audible content in background, etc. It should be checked that all the data is preserved [25], but specific data is updated and is not corrupted after the application is returned to the foreground. All the animations should be restarted as well – it does not occur automatically.

3) Locked Device. Apple warns that improper design or implementation of cryptographic operations can introduce performance or battery life problems. Locking the device with passcode can influence the applications that can operate in background. What is more, the device denies the access to the keychain and files [66]. From the author's experience, the incidents including data loss and crashes can occur if the application needs access to the keychain during the background activity, but the situation when the keychain is not available is not handled properly. It usually takes a long time to isolate the cause of such incidents. It is easy to crash the application just by frequent locking with a passcode and unlocking the device if the file data protection strategy is poorly designed.

4) Crashes. There is an option to use crash reports [24] when the tester cannot reproduce the exact steps that led to the crash. Some crash reports if symbolicated (i.e. converted to the proper stacktrace using debug symbols of the build) [67] can give a hint on the exact scenario that led to crash. Others are not useful if the crash occurred not in the application, but in iOS itself.

5) Low-Memory Warning. When iOS needs more memory, it unloads applications that are currently in the background [7], [12], [24]. Prior to iOS 6, if the application needed even more memory it could unload cashed images (if cashing was performed) and not visible views of the currently running application. In such cases it was possible to see only the placeholders of images or the application could even crash if unloaded data reload was not properly handled during the further navigation activities. Now developers must handle actions to perform when memory warning is received completely on their own [68], [69]. If the application utilizes a lot of memory (usually it means that there are memory leaks in the application) then it can be fully unloaded from the device memory by iOS itself [69]. Low-memory warnings can be simulated by Xcode Instruments (but only for the simulator) [69], [70]. From the author's experience, they can be easily reproduced on the device when many heavy pages are loaded in Safari or when photos or videos are made while the application under test is running on the background. Working with very large data or quick and frequent refreshes of data in UI collections can cause low-memory warnings when the application under test is running on the foreground.

6) Interruptions. The application should preserve its state and should not freeze if it receives an incoming call or SMS [15], [19], system alert [15], or local, or push [19], [24] notification while being in the foreground, especially when activities occur on the main thread.

It is possible to open the application through the push notification if it is received when the application is in the background or closed. Different navigation start points should be checked in case the application also does some navigation actions inside itself on confirming the push notification. The application icon badge update should also be checked including the case when several updates are received in a row [71], [72].

For applications that play audio/video it should be checked that other audio/ video streams are paused on in-application audio stream start. It should be checked if audio continues to play or not when the application is in the background (to play or not - it depends on the requirements) [19], [24], [25]. It is worth mentioning that audio/ video inside the Web Views is handled in a different way than audio/ video played natively [73].

7) Application Update. The migration process of the application from the previous versions should be tested before the new version of the application that will be available to the final user is released [24], [25]. After the application is updated from the previous version it should be checked that:

- The data is not corrupted [74].
- The user preferences stay in place [24].
- Saved credentials are still there [24].
- Previously registered push notifications are still received [24], [74].

The updates should be performed using the different possible paths starting from the very first application release [24]. From the author's experience, in some cases (e.g. due to the requirements change, or the incomplete data model design during the first release) data model changes can be so significant that users are asked to perform the backup of their data and to perform the clean install of the application. Encrypted data migration is also the subject of interest. When there is a backend server and it is updated as well, it is necessary to check the old application versions on the new server version if there is no mechanism that does not allow connecting to the server with the old versions of the application.

G. Inside the Application

1) Keyboard. Editable UI elements should be focused through auto scroll after onscreen keyboard appears. In practice, it is often forgotten to check how they behave in case of split, undocked, extended [22] or external keyboard [12], [15], only docked and merged onscreen keyboard is taken into account. From the author's experience, non standard keyboard appearances often influence the usability of those editable elements that are placed near the screen bottom border.

2) Data Import/ Export. Many applications support different file formats that they can operate with. There are different ways how supported file formats can be imported into or exported from the application. They are:

- Open In from email, web browser, or other applications [75];
- via Air Drop [76];
- via Email [19], [24];
- via iTunes [77];
- via Photos application/ Camera [78];
- via Bluetooth/ network (peer to peer) [19], [24];
- import (download and open) from URL [79].

It should be checked that the application handles (i.e. is registered to open and can open [80]) supported file formats in non case sensitive manner [81]. Naming of the exported data should be verified as well.

Files can be sent via email from the application. In most cases, iOS native email client is used for this purpose. It should be checked that there are default to, subject, and body set on email creation. The application should also properly handle the case when there is no email account configured [24].

From the author's experience, there are not many problems encountered when images are imported from the Photos application using the native view controller. But in case when custom view controller is used, it should be more strictly checked how it is synchronized with Photos application. The robustness of the Camera component usage is also the subject of worries. The Camera component tests should include the device orientation change, rotation lock, background, etc., i.e. the same aspects that should be checked for every mobile application.

3) Logging/ Analytics. Public analytics engines are often used for collecting crash reports, feature usage statistics and other logs for further development activities and testing thoroughness prioritization [24, 25]. Analytics is mostly used for publically available applications without own backend server. Based on the author's experience, if analytics is used then the main points that should be checked are:

- Analytics gathering should handle situations when the data is not available or has another format than expected. It is better to send the wrong one or no statistics than to break the UX.
- The statistics should not be sent via cellular networks. In most cases only WiFi connection should be used.
- The analytics should not gather the data about the user without his/ her permission. The user should be warned about how and where the data will be used [82].
- Collecting the data should not break the UX in any other way.

Enterprise applications can have other, more strict and extensive rules for logging depending on the corporate policy. Own logging protocols are used in such cases.

4) In-App Purchase. In-App Purchase (IAP) is a business model that allows the user to buy virtual or digital consumables, non-consumables, and subscriptions within the application that is distributed via Apple App Store. It should be checked that the purchased items are available on all the devices that are registered for the particular user, and that purchases are restored after the application reinstall, clean install, and iOS update or clean install. [83]

IAP products can be tested using special test users on Apple test environments. It is also possible to test auto-renewable subscriptions on these environments, because they have compressed durations for testing purposes. [84]

IAP password cashing system setting is of the special interest. The password can be saved for 15 minutes or asked each time the user makes any IAP [85]. The application should be checked for handling both options [24].

5) iAd. iAd is Apple's platform that allows to "generate revenue and promote … apps" by showing an advertisement within the applications [86]. Test advertisements, including the erroneous one can be sent "over local networks or USB using iAd Producer, or over the carrier network using Apple's test servers" [87]. There are two types of advertisement available: banner views and full-screen advertisements [88]. Apple suggests checking that the application shows only fully loaded advertisements. The application should pause other activities when the user begins the interaction with a banner and should restart them when the user finishes (or system cancels) the interaction with a banner. Advertisements should appear quickly and response to the device orientation changes [88].

6) Web View. Web View is a part of WebKit. Web Views are used to represent the web content inside the native mobile applications [89]. The native application is called a hybrid when most of the data inside it is represented using Web Views [90]. Web Views are often used in order to open different file formats [91] or to login to the different content providers. It should be checked that the links inside the Web Views are opened in the way they are designed to (they are opened in the same view by default [92], but they often should be opened in a default browser, for example). From the author's experience, unnecessary scroll bars and bouncing effects should be eliminated if any.

H. UI/ UX

1) Gestures. The application can be manipulated with a variety of gestures like tap, double tap, touch and hold, pinch, pan, swipe, etc. It should be checked that gestures bring the same user experience as

suggested in iOS Human Interface Guidelines [93]. The applications made by Apple can be used for reference. Based on the author's practice, if some elements on the definite application screen support non-trivial gestures other than single tap, other screen elements around should be checked for interaction using the same non-trivial gestures. It also should be verified that unexpected interactions with multiple UI elements at once are not allowed, because such actions often lead to crash.

It is worth mentioning that minimal suggested tappable area is 44 x 44 px [94].

2) Smooth Animation. The animation is used to improve UX when the application responds to the user actions or when it provides the user with a feedback about the occurring on the screen. But they should not be "excessive or gratuitous" otherwise they "can obstruct app flow, decrease performance, and distract users from their task" [95]. The animation should be smooth irrespectively of the currently running background tasks, thus the author recommends to test different animated transitions for smoothness while heavy background tasks occur.

3) Pull to Refresh. Pull to Refresh [24] feature is a very common user experience mechanism that is used for performing delta data loads in mobile applications [96]. It should be verified that Pull to Refresh mechanism loads only the new data, not the whole available data set, and that already loaded data is persisted in case of the Pull to Refresh update failed. It also should be checked how it behaves when the current data/ time and/ or data settings (e.g. format and zone) are changed.

4) Orientation. The application should be checked in both orientations if applicable [4], [12], [24], [25], [97]. Based on the author's experience, it can occur that UI elements are wrongly placed on the orientation change, and the application can crash when the user interacts with misplaced elements (it often occurs with popovers) [98]. The application can also crash when the device is rotated during the execution of heavy operations. Executing the actions after the rotation with the rotation lock option enabled is also the subject of interest, because there are several ways how the device orientation can be checked and how the device orientation change is detected by the application [97], [99].

5) Half Pixels. Sometimes there are half-pixels [24] and other unexpected blurs [100] noticed when using the application. They occur when UI elements are scaled or when their size and origin are calculated, but not rounded to the whole pixels. The same applies to the fonts. These UI glitches are more visible on the non-retina displays and are often inspected in practice using the 3-fingers accessibility zoom [101].

6) Localization. The following should be checked in case the application supports localizations:

- Localized text in images [24].
- Localized translated text fits the available area [24].
- The same text is localized in exactly the same way when used in different parts of the application.
- Right-to-left text input and alignment [24] for Arabic and Hebrew languages.
- Native and special characters:
  - persistence in a database or a file;
  - printing and display [19];
  - writing to log;
  - handling both by the client and the server.

7) Accessibility. There are plenty of accessibility features available in iOS [20], [23] – [26], i.e. VoiceOver, accessibility zoom, bold text, invert colors etc. [101] They all change the way how the system and the applications look and respond to the gestures. Thus it should be checked that enabling the accessibility features of the system does not break the application.

V CONCLUSION

The literature review of both academic and multivocal literature was performed. The majority of the sources selected for the review, both academic and multivocal, were published during the last three years period.

The results of SLR are mostly related to general mobile applications testing aspects like limited resource utilization, orientations, localizations etc., while the results of MLR provided the needed details of iOS application testing aspects (like definite restrictions and privacy settings, iOS accessibility features, etc.), as well as identified some new aspects like IAP, date/ time settings etc. The identified aspects were divided between 4 large clusters: *Environment*, *Application Lifecycle*, *Inside the Application*, and (functional or performance aspects of) *UI/ UX*. The details of each aspect were discussed based on the selected sources and the author's professional experience giving the appropriate references to Apple Developers Library**Error! Bookmark not defined.** and other credible sources. Some aspects that were not identified through literature reviews, but are known to the author (iAd, update of Xcode, AirDrop etc.) were discussed as well.

The future work will focus on the details of some of the aspects, as well as iOS applications test automation, stress testing, and security testing.

The author concludes that the study eliminates the gap that existed in the academic world in regards to the identification and detailed description of iOS application testing aspects. These details should also be useful for practitioners who want to make their iOS testing strategy more solid and complete.

VI ACKNOWLEDGMENTS

The author thanks C.T.Co Ltd. (http://www.ctco.lv/) for allowing and stimulating the conduction of the research based on the real software projects, and Darja Smite and Vladislavs Simanovics for their valuable reviews.

VII REFERENCES

[1] "Mobile Computing Sector Report", 2011. [Online]. Available: http://www.clearwatercf.com/documents/library/Mobile_Report_FINAL.pdf [Accessed: May 10, 2014].

[2] Crittercism, "Mobile Experience Benchmark". [Online]. Available: http://pages.crittercism.com/rs/crittercism/images/crittercism-mobile-benchmarks.pdf [Accessed: May 10, 2015].

[3] H. Muccini, F. Di Antonio, and P. Esposito, "Software testing of mobile applications: challenges and future research directions", in *Proc. IEEE 7th Int. Workshop Automation of Softw. Test*, Zurich, AST, 2012, pp. 29–35.

[4] V.L.L. Dantas, F.G. Marinho, A.L. da Costa, R.M.C. Andrade, "Testing requirements for mobile applications", in *Proc. 24th Int. Symp. Comput. and Inform. Sci.*, Guzelyurt, ISCIS, 2009, pp. 555-560.

[5] J. Gao, B. Xiaoying, T. Wei-Tek, T. Uehara, "Mobile application testing: a tutorial", *IEEE Computer*, vol. 47, no. 2, pp. 46-55, Feb. 2014.

[6] D. Franke, S. Kowalewski, C. Weise, and N. Prakobkosol, "Testing conformance of lifecycle-dependent properties of mobile applications", in *Proc. 5th Int. Conf. Softw. Testing, Verification and Validation*, Montreal, ICST, 2012, pp. 241 – 250.

[7] D. Franke, C. Elsemann S. Kowalewski, C. Weise, "Reverse engineering of mobile application lifecycles", in *Proc. 18th Work. Conf. Reverse Eng.*, Limerick, WCRE, 2011, pp. 283 – 292.

[8] R.T. Ogawa, B. Malen, "Towards rigor in reviews of multivocal literatures: applying the exploratory case study method", *Review of Educ. Research*, vol. 61, no. 3, pp. 265–286, Fall 1991.

[9] *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models,* ISO/IEC 25010, 2011.

[10] E. Tom, A. Aurum, R. T. Vidgen, "An exploration of technical debt", *Journal of Systems and Software*, vol. 86, no. 6, pp. 1498-1516, 2013.

[11] B. Kitchenham, S. Charters, „Guidelines for performing Systematic Literature Reviews in Software Engineering", EBSE Tech. Rep., 2007, vers. 2.3.

[12] D. Franke, S. Kowalewski, C. Weise, "A mobile software quality model", in *Proc. 12th Int. Conf. Quality Softw.*, Xi'an, Shaanxi, QSIC, 2012, pp. 154 – 157.

[13] H.-K. Kim, "Mobile applications software testing methodology", *Commun. in Comput. and Inform. Sci.*, vol. 342, 2012, pp. 158-166.

[14] E. H. Marinho, R.F. Resende, "Quality factors in development best practices for mobile applications", in *Proc. Computational Sci. and Its App.*, Salvador de Bahia, Brazil, ICCSA, 2012, pp. 632-645.

[15] D. Amalfitano, A.R. Fasolino, P. Tramontana, and N. Amatucci, "Considering context events in event-based testing of mobile applications", in *Proc. IEEE 6th Int. Conf. Softw. Testing, Verification and Validation Workshops*, Luxembourg, ICSTW, 2013, pp.126-133.

[16] K. Haller, "Mobile testing", *ACM SIGSOFT Softw. Eng. Notes,* vol. 38, no. 6, pp. 1-8, Nov. 2013.

[17] H. Khalid, "On identifying user complaints of iOS apps", in *Proc. 35th Int. Conf. Softw. Eng.*, San Francisco, CA, ICSE, 2013, pp. 1474 – 1476.

[18] D. Franke, C. Weise, "Providing a software quality framework for testing of mobile applications", in *Proc. 4th Int. Conf. Softw. Testing, Verification and Validation*, Berlin, ICST, 2011, pp. 431 - 434.

[19] "Testing Criteria for iOS Apps - App Quality Alliance", Oct. 21, 2013. [Online]. Available: http://www.appqualityalliance.org/files/AQuA_testing_criteria_for_iOS_for_v1.0%20final%2022_oct_2013.pdf [Accessed: May 10, 2015].

[20] "Tips For Accessibility Testing Of iOS Apps | Pat's Tapestry", May 24, 2013. [Online]. Available: http://patstapestry.wordpress.com/2013/05/24/tips-for-accessibility-testing-of-ios-apps/ [Accessed: May 10, 2015].

[21] "Testing iOS Apps for Tough Network Conditions | Nearsoft", Oct 20, 2013. [Online]. Available: http://nearsoft.com/blog/testing-ios-apps-for-tough-network-conditions/ [Accessed: May 10, 2015].

[22] "TestElf Blog — We Find These Common Bugs When Testing iOS Apps", Jul 24, 2013. [Online]. Available: http://blog.testelf.com/post/56341438836/we-find-these-common-bugs-when-testing-ios-apps [Accessed: May 10, 2015].

[23] "The Essential Guide to iPhone & iPad App Testing", Dec 10, 2012. [Online]. Available: http://core.ecu.edu/STRG/materials/uTest_eBook_Mobile_Testing.pdf [Accessed: May 10, 2015].

[24] "iOS Testing mind map 1.2 – Now with more stuff | Neglected Potential", Oct 8, 2013. [Online]. Available: http://www.neglectedpotential.com/2013/10/ios-testing-mind-map-1-2/ [Accessed: May 10, 2015].

[25] "iOS Devices | Dave Addey", Sept 22, 2013. [Online]. Available: http://daveaddey.com/postfiles/AgantReleaseChecklist2013.pdf [Accessed: May 10, 2015].

[26] "iOS Accessibility - A Useful Guide For Testing | Rosie Land", Sep 2, 2012. [Online]. Available: http://www.rosiesherry.com/2012/09/02/ios-accessibility-a-useful-guide-for-testing/ [Accessed: May 10, 2015].

[27] "Testing iOS Applications", Mar 12, 2014. [Online]. Available: http://blog.smartbear.com/mobile/testing-ios-applications/ [Accessed: May 10, 2015].

[28] "Global 500 – Fortune", Mar 18, 2015. [Online]. Available: http://fortune.com/global500/ [Accessed: May 10, 2015].

[29] "Developer Library - Apple Developer". [Online]. Available: https://developer.apple.com/library/ [Accessed: May 10, 2015].

[30] "Apple - iPad in Business". [Online]. Available: http://www.apple.com/ipad/business/ [Accessed: May 10, 2015].

[31] "Apple - iPhone in Business". [Online]. Available: http://www.apple.com/iphone/business/ [Accessed: May 10, 2015].

[32] "iOS 6: Which software features does my iPhone, iPad, or iPod touch support? - Apple Support". [Online]. Available: http://support.apple.com/kb/ht5457 [Accessed: May 10, 2015].

[33] "iPad 2 - Technical Specifications". [Online]. Available: http://support.apple.com/kb/sp622 [Accessed: May 10, 2015].

[34] "iPad mini - Technical Specifications". [Online]. Available: http://support.apple.com/kb/SP661 [Accessed: May 10, 2015].

[35] "iPad (4th generation) - Technical Specifications". [Online]. Available: http://support.apple.com/kb/sp662 [Accessed: May 10, 2015].

[36] "iPad (3rd generation) - Technical Specifications". [Online]. Available: http://support.apple.com/kb/sp647 [Accessed: May 10, 2015].

[37] "iPad Air - Technical Specifications". [Online]. Available: http://support.apple.com/kb/SP692 [Accessed: May 10, 2015].

[38] "iPad mini with Retina display - Technical Specifications". [Online]. Available: http://support.apple.com/kb/SP693 [Accessed: May 10, 2015].

[39] "iPad Air 2 - Technical Specification". [Online]. Available: http://support.apple.com/kb/SP708 [Accessed: May 10, 2015].

[40] "iPad mini 3 - Technical Specification". [Online]. Available: http://support.apple.com/kb/SP709 [Accessed: May 10, 2015].

[41] “Use Touch ID on iPhone and iPad - Apple Support”. [Online]. Available: http://support.apple.com/en-us/HT5883 [Accessed: May 10, 2015].
[42] “iPhone 4 - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/sp587 [Accessed: May 10, 2015].
[43] “iPhone 4S - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/sp643 [Accessed: May 10, 2015].
[44] “iPhone 5 - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/sp655 [Accessed: May 10, 2015].
[45] “iPhone 5c - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/SP684 [Accessed: May 10, 2015].
[46] “iPhone 5s - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/SP685 [Accessed: May 10, 2015].
[47] “iPhone Development 101: Default (Launch) Image Sizes for iPhone & iPad”. [Online]. Available: http://www.idev101.com/code/User_Interface/launchImages.html [Accessed: May 10, 2015].
[48] “iPhone 6 - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/SP705 [Accessed: May 10, 2015].
[49] “iPhone 6 Plus - Technical Specifications”. [Online]. Available: http://support.apple.com/kb/SP706 [Accessed: May 10, 2015].
[50] “Apple - Apple Pay”. [Online]. Available: https://www.apple.com/apple-pay/ [Accessed: May 10, 2015].
[51] “Xcode Overview: Debug Your App”, Oct 20, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/ToolsLanguages/Conceptual/Xcode_Overview/DebugYourApp.html [Accessed: May 10, 2015].
[52] “Monitoring Accessory-Related Events”, Feb 24, 2012. [Online]. Available: https://developer.apple.com/library/ios/featuredarticles/ExternalAccessoryPT/Articles/MonitoringEvents.html [Accessed: May 10, 2015].
[53] “Tablet Stylus Test Lab: Comparison of Pencil, Intuos, Pogo Connect & Jot Script”, May 12, 2013. [Online]. Available: http://www.apartmenttherapy.com/tablet-stylus-test-lab-comparison-of-pencil-intuos-pogo-connect-jot-script-tech-test-lab-reviews-196850 [Accessed: May 10, 2015].
[54] “Xcode - What's New - Apple Developer”. [Online]. Available: https://developer.apple.com/xcode/ [Accessed: May 10, 2015].
[55] “What’s New in Xcode”, Apr 20, 2015. [Online]. Available: https://developer.apple.com/library/mac/documentation/DeveloperTools/Conceptual/WhatsNewXcode/Articles/Introduction.html [Accessed: May 10, 2015].
[56] “iOS 7.0 Release Notes”, Oct 22, 2013. [Online]. Available: https://developer.apple.com/library/ios/releasenotes/General/RN-iOSSDK-7.0/ [Accessed: May 10, 2015].
[57] “Apple to require all new App Store submissions to be “optimized for iOS 7,” built with latest Xcode starting Feb. 1 | 9to5Mac”, Dec 17, 2013. [Online]. Available: http://9to5mac.com/2013/12/17/apple-to-require-all-new-app-store-submissions-to-be-optimized-for-ios-7-built-with-latest-xcode/ [Accessed: May 10, 2015].
[58] “Apple - Press Info - iOS 7 With Completely Redesigned User Interface & Great New Features Available September 18”, Sep 10, 2013. [Online]. Available: http://www.apple.com/pr/library/2013/09/10iOS-7-With-Completely-Redesigned-User-Interface-Great-New-Features-Available-September-18.html [Accessed: May 10, 2015].
[59] “Why Apple won’t allow you to downgrade your iPhone from iOS 7 to iOS 6 | ITProPortal.com”, Sep 29, 2013. [Online]. Available: http://www.itproportal.com/2013/09/29/why-apple-wont-allow-you-to-downgrade-your-iphone-from-ios-7-to-ios-6/ [Accessed: May 10, 2015].
[60] “App Store Distribution - Support - Apple Developer“. [Online]. Available: https://developer.apple.com/support/appstore/ [Accessed: May 10, 2015].
[61] “About Restrictions (parental controls) on iPhone, iPad, and iPod touch - Apple Support”. [Online]. Available: http://support.apple.com/kb/ht4213 [Accessed: May 10, 2015].
[62] “About the Instruments User Reference” [Online]. Available: https://developer.apple.com/library/ios/documentation/AnalysisTools/Reference/Instruments_User_Reference/Introduction/Introduction.html [Accessed: May 10, 2015].
[63] “Logging Energy Usage on an iOS Device”, Oct 16, 2014. [Online]. Available: https://developer.apple.com/library/mac/recipes/Instruments_help_articles/LoggingEnergyUsageinaniOSDevice/LoggingEnergyUsageinaniOSDevice.html [Accessed: May 10, 2015].
[64] “Designing for Real-World Networks” [Online]. Available: https://developer.apple.com/library/ios/documentation/NetworkingInternetWeb/Conceptual/NetworkingOverview/WhyNetworkingIsHard/WhyNetworkingIsHard.html [Accessed: May 10, 2015].
[65] “Background Execution”, Sep 17, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/iPhone/Conceptual/iPhoneOSProgrammingGuide/BackgroundExecution/BackgroundExecution.html [Accessed: May 10, 2015].
[66] “iOS Security”, Oct 2012. [Online]. Available: https://www.apple.com/br/ipad/business/docs/iOS_Security_Oct12.pdf [Accessed: May 10, 2015].
[67] “Analyzing Crash Reports”, Apr 08, 2015. [Online]. Available: https://developer.apple.com/library/ios/documentation/IDEs/Conceptual/AppDistributionGuide/AnalyzingCrashReports/AnalyzingCrashReports.html [Accessed: May 10, 2015].
[68] “Handling Memory Warnings - UIViewController Class Reference”, Sep 17, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/uikit/reference/UIViewController_Class/Reference/Reference.html [Accessed: May 10, 2015].
[69] “Performance Tips”, Sep 17, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/iPhone/Conceptual/iPhoneOSProgrammingGuide/PerformanceTips/PerformanceTips.html [Accessed: May 10, 2015].
[70] “iOS Simulator User Guide: Interacting with iOS Simulator”, Mar 09, 2015. [Online]. Available: https://developer.apple.com/library/mac/documentation/IDEs/Conceptual/iOS_Simulator_Guide/InteractingwiththeiOSSimulator.html [Accessed: May 10, 2015].
[71] “About Local Notifications and Remote Notifications”, Mar 09, 2015. [Online]. Available: https://developer.apple.com/library/ios/documentation/NetworkingInternet/Conceptual/RemoteNotificationsPG/Introduction.html [Accessed: May 10, 2015].
[72] “Registering, Scheduling, and Handling User Notifications”, Mar 09, 2015. [Online]. Available: https://developer.apple.com/library/ios/documentation/NetworkingInternet/Conceptual/RemoteNotificationsPG/Chapters/IPhoneOSClientImp.html [Accessed: May 10, 2015].
[73] “About AV Foundation “, Nov 18, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/AudioVideo/Conceptual/AVFoundationPG/Articles/00_Introduction.html [Accessed: May 10, 2015].
[74] “Technical Note TN2285: Testing iOS App Updates”, Oct 07, 2014. [Online]. Available: https://developer.apple.com/library/ios/technotes/tn2285/_index.html [Accessed: May 10, 2015].
[75] “Technical Q&A QA1587: How do I get my application to show up in the "Open in..." menu on iOS for a specific document type?”, Apr 14, 2014.[Online]. Available: https://developer.apple.com/library/ios/qa/qa1587/_index.html [Accessed: May 10, 2015].
[76] “Share content with AirDrop from your iPhone, iPad, or iPod touch - Apple Support“. [Online]. Available: http://support.apple.com/kb/ht5887 [Accessed: May 10, 2015].
[77] “Apple - iTunes - Everything you need to be entertained“. [Online]. Available: http://www.apple.com/itunes/ [Accessed: May 10, 2015].
[78] “About the Camera and Photo Library”, Jul 17, 2012. [Online]. Available: https://developer.apple.com/library/ios/documentation/AudioVideo/Conceptual/CameraAndPhotoLib_TopicsForIOS/Introduction/Introduction.html [Accessed: May 10, 2015].

[79] "Using NSURLDownload", Oct 22, 2012. [Online]. Available: https://developer.apple.com/library/mac/documentation/Cocoa/Conceptual/URLLoadingSystem/Tasks/UsingNSURLDownload.html [Accessed: May 10, 2015].
[80] "About Document Interaction", Nov 15, 2010. [Online]. Available: https://developer.apple.com/library/ios/documentation/filemanagement/conceptual/documentinteraction_topicsforios/Introduction/Introduction.html [Accessed: May 10, 2015].
[81] "Technical Q&A QA1697: Why doesn't my device load a file that loads fine in the Simulator?", Jun 01, 2010. [Online]. Available: https://developer.apple.com/library/ios/qa/qa1697/_index.html [Accessed: May 10, 2015].
[82] "App Store Review Guidelines - Apple Developer". [Online]. Available: https://developer.apple.com/appstore/resources/approval/guidelines.html [Accessed: May 10, 2015].
[83] "Getting Started with In-App Purchase - In-App-Purchase-Guidelines.pdf", Dec 01, 2014. [Online]. Available: https://developer.apple.com/in-app-purchase/In-App-Purchase-Guidelines.pdf [Accessed: May 10, 2015].
[84] "Testing In-App Purchase Products", Oct 02, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/LanguagesUtilities/Conceptual/iTunesConnectInAppPurchase_Guide/Chapters/TestingInAppPurchases.html [Accessed: May 10, 2015].
[85] "Use Restrictions to prevent purchasing on your iPhone, iPad, or iPod touch - Apple Support". [Online]. Available: http://support.apple.com/kb/ht6088 [Accessed: May 10, 2015].
[86] "About iAd", Sep 17, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/UserExperience/Conceptual/iAd_Guide/Introduction/Introduction.html [Accessed: May 10, 2015].
[87] "About iAd Tester - Apple Support". [Online]. Available: http://support.apple.com/kb/HT5245 [Accessed: May 10, 2015].
[88] "Testing iAd Apps", Sep 17, 2014. [Online]. Available: https://developer.apple.com/library/ios/documentation/userexperience/conceptual/iAd_Guide/TestingiAdApplications/TestingiAdApplications.html [Accessed: May 10, 2015].
[89] "Introduction to WebKit Objective-C Programming Guide", Nov 19, 2012. [Online]. Available: https://developer.apple.com/library/mac/documentation/Cocoa/Conceptual/DisplayWebContent/DisplayWebContent.html [Accessed: May 10, 2015].
[90] "What is a Hybrid Mobile App? -Telerik Developer Network", Mar 25, 2015. [Online]. Available: http://developer.telerik.com/featured/what-is-a-hybrid-mobile-app/ [Accessed: May 10, 2015].
[91] "Technical Q&A QA1630: Using UIWebView to display select document types", Aug 25, 2009. [Online]. Available: https://developer.apple.com/library/IOs/qa/qa1630/_index.html [Accessed: May 10, 2015].
[92] "Simple Browsing", Nov 09, 2012. [Online]. Available: https://developer.apple.com/library/mac/documentation/Cocoa/Conceptual/DisplayWebContent/Tasks/SimpleBrowsing.html [Accessed: May 10, 2015].
[93] "iOS Human Interface Guidelines: Interactivity and Feedback", Apr 08, 2015. [Online]. Available: https://developer.apple.com/library/ios/documentation/userexperience/conceptual/MobileHIG/InteractivityInput.html [Accessed: May 10, 2015].
[94] "iOS Human Interface Guidelines: Adaptivity and Layout", Apr 08, 2015. [Online]. Available: https://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/LayoutandAppearance.html [Accessed: May 10, 2015].
[95] "iOS Human Interface Guidelines: Animation", Apr 08, 2015. [Online]. Available: https://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/Animation.html [Accessed: May 10, 2015].
[96] "United States Patent: 8448084", May 21, 2013. [Online]. Available: http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=8,448,084.PN.&OS=PN/8,448,084&RS=PN/8,448,084 [Accessed: May 10, 2015].
[97] "Supporting Multiple Interface Orientations", Dec 13, 2012. [Online]. Available: https://developer.apple.com/library/ios/featuredarticles/ViewControllerPGforiPhoneOS/RespondingtoDeviceOrientationChanges/RespondingtoDeviceOrientationChanges.html [Accessed: May 10, 2015].
[98] "UIPopoverController Class Reference", Dec 16, 2013. [Online]. Available: https://developer.apple.com/library/ios/documentation/UIKit/Reference/UIPopoverController_class/Reference/Reference.html [Accessed: May 10, 2015].
[99] "Technical Q&A QA1688: Why won't my UIViewController rotate with the device?", Apr 18, 2013. [Online]. Available: https://developer.apple.com/library/ios/qa/qa1688/_index.html [Accessed: May 10, 2015].
[100] "OS X Human Interface Guidelines: Icons and Graphics", Apr 08, 2015. [Online]. Available: https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/OSXHIGuidelines/IconsGraphics.html [Accessed: May 10, 2015].
[101] "Use Accessibility features in iOS - Apple Support". [Online]. Available: http://support.apple.com/kb/HT5018 [Accessed: May 10, 2015].

The full list of the reviewed multivocal sources with indicating the exclusion phase can be found here: https://dspace.lu.lv/dspace/bitstream/handle/7/2739/iOS Applications Testing - Multivocal Sources.pdf

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 151-154

The concept of creating an information technology of process formation of the main educational programs based on professional competencies

Lilia Motaylenko
FSG-fEIHPE "Pskov State University», Faculty of Informatics.
Address: pl. Lenin 2, Pskov, 180000, Russia

Abstract. This paper gives the definition of information technology, describes the concept of creating an Information Technology of process formation of Main Educational Programs (IT MEP) based on professional competences and gives conceptual scheme of creating IT MEP.

Keywords: **information technology, main educational program, professional competence, conceptual scheme, algorithm.**

I INTRODUCTION

Modern education in Russia is going through the stage of reforming, including changes of Federal State Educational Standards (FSES). Any reforming introduces into the system, in this context - the educational system, disturbances that lead to its instability. One of the tasks of creating information technology is a development of a sequence of actions that is operations, leading to the creation of control actions in order to stabilize the system.

Changes of FSES inevitably result in a change in the process of the formation of Basic Professional Educational Programs or Main Educational Programs (MEP) [1], their informative part, as well as a work plan. The following discussion focuses on the Higher Education (HE).

II INFORMATION TECHNOLOGIES. BASIC CONCEPTS AND DEFINITIONS

Let`s systematize basic concepts and definitions.

Before you decide, let`s consider a general definition of technology.

Technology (from the ancient Greek: arts, skill, ability, thought, reason, method, mode of production) is a set of methods, processes and materials used in any field of business, as well as a scientific description of technical methods of production [2].

According to the Dictionary of Philosophy, edited by I. Frolov, "technology is a complex developing system of artifacts, manufacturing operations and processes, resource sources, subsystems of social consequences of information, management, financing and interaction with other technologies" [3].

Standard ISO / IEC 2382 - 1: 1993. Terms and Definitions (the Standard) [4] information (01.01.02) (in information processing) is any fact, concept, or meaning derived from Data (01.01.03) and associated context, or selected from Knowledge (01.01.01).

Data (01.01.03) Representation of Information (01.01.02) is representing information in some formalized manner suitable for communication, interpretation or processing [4].

The Standard does not define data as a source of information. Data are fixed values of different types. Fixing can be carried out both manually (by an observer) and automatically, i.e. by technical means. This concept is read from the definition of knowledge to have been processed. From this point of view information is data processed with manual or automatic methods which lead to their semantic recognition.

The concept of information technology is quite extensive.

The Standard specifies information technology (01.01.22) as an Art and applied sciences that deal with Data (01.01.03) and Information (01.01.02). The example of information technology is given: capture representation, processing, security, transfer, interchange, presentation, management, organization, Storage (12), and retrieval of data and information [4].

At the same time in the Federal Law of the Russian Federation "On Information, Information

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.180

Technologies and Protection of information" [5] the definition is formulated somewhat differently:

"Information technologies are processes, methods for finding, collecting, storing, processing, representing, distributing information and ways to implement these processes and techniques."

The apparent difference is given in the methods mentioned above.

In the theory of information processes and systems basic information processes are listed, which are used as a part of IT [6], including the following operations: information extraction, transfer, processing, storage, presentation and usage.

Information technologies are a basis for creating information systems.

Defined by the Standard information system (1.1.13) is information processing system (01.01.12), together with associated organizational resources, that provides and distributes Information (01.01.02).

The information system is a combination of various resources and information technologies.

III CONCEPTUAL SCHEME

The structure of the system described at the conceptual level allows to identify qualitatively main subsystems, components and connections among them [6, 7].

The conceptual scheme indicates entities that may exist in the area of a concern, i.e. entities that exist, existed or could ever exist. It also points out facts and events that are possible or required for these entities [8, 9].

The main role of the conceptual scheme is the following: to provide a common basis for understanding the general behavior of the total area of a concern, to determine the allowable evolution and the manipulation the information about the subject area, to provide a basis for the interpretation of external and internal syntactic forms, providing the information about the problem domain, to provide a basis for the transformation of external and internal schemes.

In the description of the conceptual scheme of the formation of MEP with the professional competencies the concepts such as class, entity, attribute, relationship are included.

Let`s consider the conceptual scheme of the formation of MEP based on professional competencies (Fig. 1).

The classes of the conceptual scheme are experts, criteria, educational process.

Entities of the class Experts are outer experts that is inspectors who are representatives of higher institutions, employers who are potential employers for graduates, as well as inner experts: Staff and Students.

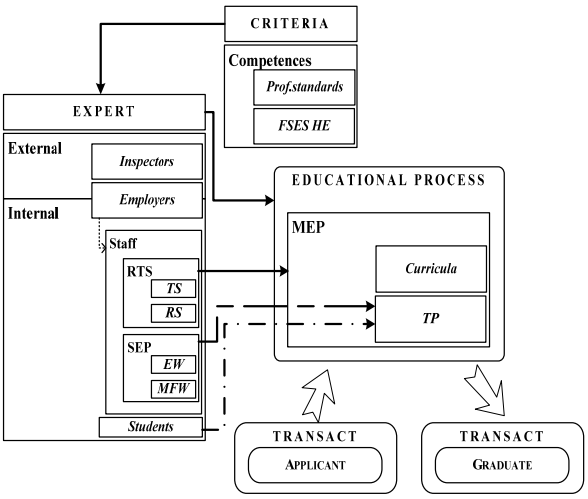

Fig. 1. Conceptual scheme of the formation of MEP based on professional competencies.

The FSES requirements on different directions require up to 30% of teachers training the students must be potential employers; in this connection the entity of Employers is an entity of the border between the inner and outer class of Experts, passing becoming the entity Staff.

The entity Staff includes Research and Teaching Staff (RTS), which includes Teaching Staff of (TS) and Researchers (RS); Staff supporting Educational Process (SEP) includes Engineering Workers (EW) and Methodical Field Workers (MFW) involved in the educational process.

The entity Students covers only those higher education students who have completed the study of a discipline or a discipline module [10].

The class Criteria includes the entity Competences of professional standards [11 – 13] and FSES HE.

The connection between classes Criteria and Experts is due to the subordination of the class Experts to the class Criteria.

The class Educational process includes the entity of MEP with elements of Curricula and Teaching Providing (TP).

RTS forms MEP and SEP is involved to the development of TP. The Students indirectly affect the change of TP after having been questioned on the discipline or the discipline module [10].

In the conceptual scheme (Fig. 1) elements of Transact are included. In the theory of mass servicing this concept is used to describe dynamic objects which while passing through the system changes its properties [14 - 17].

For the Educational Process such an object is an applicant being at the input of the process, who then acquires properties of a student according to the attributes and at the output process properties of a graduate with relevant direction competencies.

For this system the role of a specialist is left outside and is not considered.

Fig. 2 shows the conceptual scheme of data streams in the Information System (IS) with selected basic IT processes and types of operations [4].

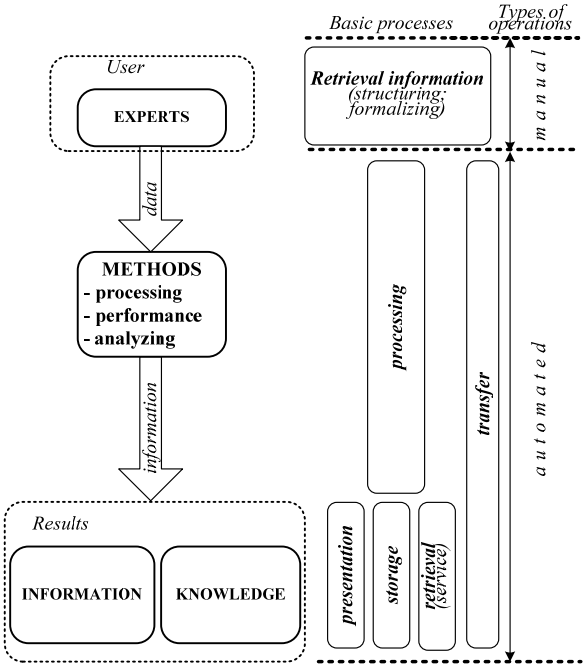

Fig. 2. Conceptual scheme of the data streams in IS.

Users are defined in the IS. The User is someone or something that sends commands and messages to the information system and receives messages from the information system. It is a part of the medium. Some users may have an authority to establish rules of a behavior or restriction for the information system [8]. The role of user Experts is to carry out a manual operation of the process of extracting information verbally stated, namely: structuring and formalizing i.e. representing data in the format of the machine representation.

All other processes are automated, i.e. they are performed by computers.

The methods, by means of which data are processed, refer to the collection of data for the distribution to entities of the class Teaching process, comparing criteria values [10], and analyzing the results. This processing lets obtain the resulting information and knowledge in accordance with the purpose, namely: to create the information technology of the process of the formation of main educational programs based on professional competencies.

The result of the work of the system is the information and the knowledge applied to improve the quality of the educational process. The Process of Presentation involves creating a user interface that displays the result obtained. Storage and Retrieval (service) processes provide the storage of created databases, knowledge bases, filtering operations, search, deleting outdated information.

The Transfer process consists of using local networks and data networks [18 – 20]. This system uses an application layer, which provides an access to shared network resources, network administrating, all kinds of electronic communication: e-mail, conferences, dialogue, etc.

The sequence of operations of the information technology of MEP forming process begins with determining the state of MEP (Fig. 3).

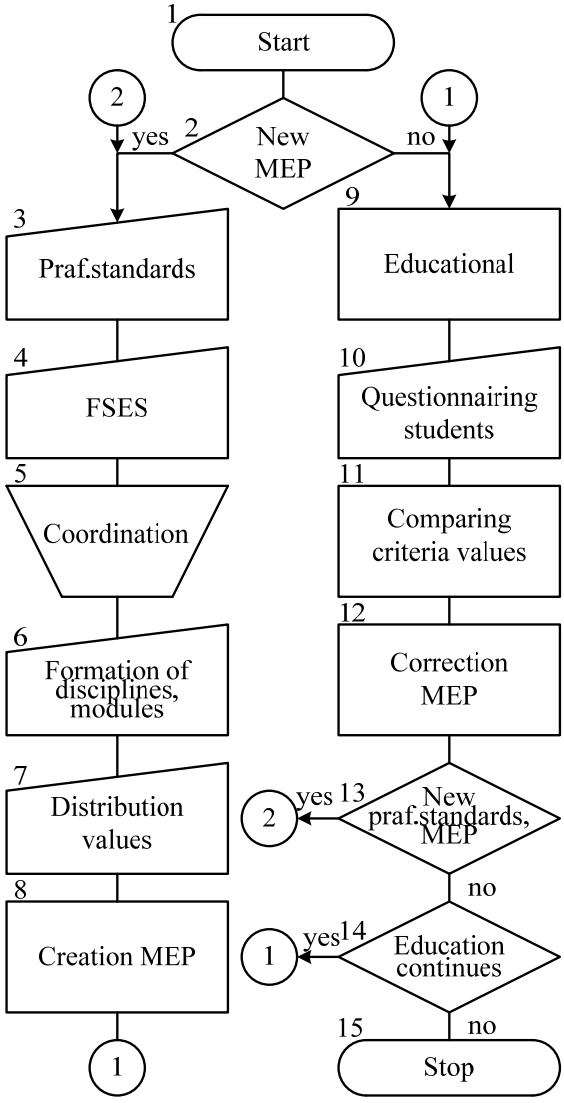

Fig. 3. Block diagram of the work algorithm of IS of the formation of MEP at the conceptual level.

The start of the algorithm of the information system block diagram is conditional. In case of opening a new direction in a college or in case of changing professional standards, FSES criteria or requirements to the graduates of a higher educational institution the introduction of new criteria and their coordination (manual operation, the implementation of which is provided by a user interface) is performed. The

formation of disciplines and disciplines modules in accordance with the competencies and the distribution of their performance [10] in the curriculum (Fig. 1) is performed by Staff and then MEP is created.

In accordance with established MEP the training of students is provided. After completing the study of the discipline or the discipline module questionnairing the students on relevant indicators of acquired competences corresponding to distributed parameters is held (Competence area MEP and standard). If there is a difference Staff corrects teaching and methodological support of the discipline module as well as there may be done a redistribution of the competences and academic hours among disciplines, modules, i.e. the change of the curriculum.

At reforming the education system there appears a possibility of changing of MEP when new professional standards and FSES are approved or when new requirements of employers to the professional quality of graduates are introduced.

When the status of FSES is current (block 2) the actions are repeated according to the block diagram (Fig. 3) from connector 1. If the training program is no longer used (block 14) in the algorithm of work of information system the conditional stop (block 15) is indicated in block diagram.

IV CONCLUSION

The basic concepts and definitions of the theory of information technologies and systems given in the article, don't contradict current ISO and GOST.

The attention is focused on the need to harmonize the terminology database.

The concept of creating an information technology process of the formation of MEP with professional competencies as a conceptual diagram of the process of formation of MEP with the professional competencies and data flows in the information system is given.

The graphical presentation of the principle of the information system of the formation of MEP at the conceptual level is also shown.

V REFERENCES

[1] Олейникова О.Н. Модульные технологии и разработка образовательных программ: учебное пособие / О.Н. Олейникова, А. А. Муравьева, Ю. В. Коновалова, Е.В.Сартакова. Изд. 2-е, перераб. и доп., М.: Альфа-М; ИНФРА-М, 2010, 256 с.

[2] Севергин В. М. О предметах учений технологий // Технологический журнал. 1822. Т. 7. Ч. 1. С. 1.

[3] Философский словарь / под ред. И. Т. Фролова. — 7 изд., перераб. и доп. — М.: Республика, 2001. — 719 с.

[4] Термины и определения стандарта ISO/IEC 2382-1. Информационно-справочный портал MorePC.ru http://www.morepc.ru/informatisation/iso2381-1.html?print

[5] Федеральный закон от 27 июля 2006 г. N 149-ФЗ "Об информации, информационных технологиях и о защите информации" (с изменениями от 27 июля 2010 г., 6 апреля, 21 июля 2011 г., 28 июля 2012 г.). Принят Государственной Думой 8 июля 2006 года. Одобрен Советом Федерации 14 июля 2006 года

[6] Теория информационных процессов и систем: учебное пособие для студ. высш. учеб. заведений/ [Б. Я. Советов, В. А. Дубенецкий, В. В. Цехановский и др.]; под ред. Б.Я. Советова, М.: Издательский центр «Академия», 2010, 432 с.

[7] О'Коннор Дж. Искусство системного мышления: необходимые знания о системах и творческом подходе к решению проблем/ Дж. О'Коннор, И. Макдермот., М.: Альпина Бизнес Букс, 2006, 256 с.

[8] ГОСТ 34.320-96 Информационные технологии. Система стандартов по базам данных. Концепции и терминология для концептуальной схемы и информационной базы. Дата введения в действие 30.06.2001 .Код ОКС 35.100.01;35.240, 46с.

[9] Мотайленко Л.В., Антонов И.В. Методика и средство автоматизации работы выпускающей кафедры// Качество подготовки специалистов и научные исследования в Вузах: Доклады. Международная научно-практическая конференция – XIV Академические чтения МАН ВШ, Псков: Изд-во ППИ, 2009, С. 120 – 126

[10] L. Motaylenko, O. Poletayeva, S. Lyokhin Development of information model of forming basic educational programs in the light of professional competencies // Environment. Technology. Resources: Proceedings of the 9th International Scientific and Practical Conference, June 20—22, 2013. Vol. II. Rezekne, 2013.

[11] Олейникова О.Н., Муравьева А.А. Профессиональные стандарты: принципы формирования, назначение и структура. Методическое пособие, М.: АНО Центр ИРПО, 2011, 100 с.

[12] Профессиональные стандарты как основа подготовки конкурентоспособных работников: Методическое пособие / М. В. Привезенцев, О. Н. Олейникова, А. А. Муравьева и др., М.: Альфа-М, 2007, 160 с.

[13] Мотайленко Л.В., Полетаев Д.И. Методика интеграции образования и профессионального сообщества на базе информационных технологий// Информатика и образование: Научно-методический журнал, №9 (248), 2013, С. 51-54

[14] Авдеев О.Н., Мотайленко Л.В. Моделирование систем: учебное пособие, СПб.: СПбГТУ, 2001, 179с.

[15] Власов Л.В., Колесников Д.Н., Сорокин И.А. Имитационное моделирование систем массового обслуживания с использованием GPSS: Учеб. пособие, Л.: ЛПИ, 1989.

[16] Марков А.А. Моделирование информационно-вычислительных процессов: Учебное пособие для вузов, М.: Изд-во МГТУ им. Н.Э. Баумана, 1999, 360с.

[17] Нейлор Т. Машинные имитационные эксперименты с моделями экономических систем: Пер. с англ., М.: Мир, 1975.

[18] Барри Нанс. Компьютерные сети. /Пер. с англ., М.: БИНОМ, 1995.

[19] Олифер В.Г., Олифер Н.А. Компьютерные сети. Принципы, технологии, протоколы, СПб.: Питер, 1999

Советов Б.Я. Информационные технологии: учеб. для вузов/ Б.Я. Советов, В.В. Цехановский. – 3-е изд., стер., М.: Высш. шк., 2006, 263 с.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 155-158

Importance of data acquisition in problem based learning

Valdis Priedols, Armands Grickus
Liepaja University, Faculty of Science and Engineering.
Address: K. Valdemara iela 4, Liepaja, LV-3401, Latvia

Abstract. **This research paper demonstrates untraditional learning and teaching method that is developed from combination of experimentation, usage of computer simulations and problem based learning. Taking all previously mentioned methods together there can be created very successful learning environment which provides students to master electromagnetism more effectively. Research focuses on proper use of data acquisition modules and computer simulations in problem based learning (PBL) teaching method. Based on the results of the research experimental PBL in various cases provides better learning outcomes, but there are also a few occasions where the results aren't so pleasing. Overall PBL provides results that are at least as promising as results of other teaching and learning methods. Therefore this method will be utilized in Liepaja University to teach physics, especially electromagnetism.**

Keywords: **computer simulation, data acquisition, educational technology, electronic learning.**

I INTRODUCTION

Learning electricity and magnetism has always been difficult [16]. While studying, students tend to confuse basic concepts, misinterpret schematic representations, do not relate macro and micro relationships in electric circuits, and are unable to link electrostatics and electrodynamics.

The problem with magnetism seems to be even more pronounced due to the fact that humans are not equipped with sensors to gauge magnetism. While electricity can be observed of felt by electric shocks, there is almost no sensual indication of magnetic fields [21]. Unlike mechanical phenomena, which can be sensed visually, electromagnetism is in a realm of physics that is not covered by any one of the five human senses. This is the major reason for difficulties encountered by students of all ages when trying to make electromagnetic concepts concrete [20]. Mostly people that can use their visual imagery can properly master electromagnetism.

Based on research and observations technology instructions in physics should make scientific views more accessible and meaningful [11]. The main goal is that technology should be used in teaching and learning of science as one of learning possibilities and not to have less effective results than other learning styles.

Main reasons for educators to use technology in science are to improve students' physics ability and to improve students' negative reactions toward physics [14].

Most commonly known appropriate reasons when studies are provided by technology are: a) to use computer interfacing equipment to collect and process data of a measurement or experiment; b) to perform experimental or theoretical modelling and to use computer simulations; c) to use research and presentation programs for gathering and displaying information [19].

As an ultimate teaching and learning method problem based learning (PBL) is by far considered as one of the best methods in education [8]. There can be a combination of instructions to solve problems. These instructions consist of specific computer based tasks to be solved through experimentation, paper and pencil exercises and class discussions.

The aim of this research is to consider virtual teaching and learning through the use of experimental PBL and PBL based on computer simulations [23]. This indicates that technology has an important role in virtual teaching and learning.

Hopefully there exist enough positive effects so that combined PBL can be developed in Liepaja University.

II MATERIALS AND METHODS

Common interfacing devices, which are used in the study process, are probes that are plugged into graphing calculators or measuring interfaces that are usually connected to computers. With aid of these devices students are capable of collecting and analyzing data of several measurements.

Data acquisition (DAQ) modules are simply connectable to the universal serial bus (USB) port of personal computer (PC) using standard cable. PC automatically identifies the module and installs the

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.176

driver. With aid of USB port data can be transferred more quickly and there can be performed multiple input/output (I/O) operations simultaneously.

Most of the data acquisition modules are compact and portable allowing users to realize measurements out of the lab [3]. There are no overall minuses because the choice of the data acquisition module depends on the needs of every user.

USB-DAQ

Usually all measurements in physics in Liepaja University are realized with aid of measuring interface Cobra3 – basic unit, provided by PHYWE. A lot of measurements can be taken with this interface, but there exist certain problems, that prevent students from using it. To use the interface in specific experiments a special measuring module has to be used which has to be purchased and does not come in the original packaging. Another considerable fact is that using the original packaging experimentation cannot be done in places where there is no electricity.

Considering all these facts a decision has been made to use a low cost data logger provided by Firmware Factory. It has several necessary options. In normal operation it operates autonomously, collecting data and providing control as required [2]. When plugged into a USB port, DAQ appears on the PC like a removable disk and it can be used as storage. It has a wide range of peripheral I/O. In addition to functioning as a flash drive, DAQ has a data acquisition – process control engine and a real time clock that allow it to function autonomously when not connected to PC.

As evidence shows it is significant that all measurements are taken with appropriate device which is easy to use.

Computer simulations

Computer simulations are applications of special interest in physics teaching because they can support powerful modeling environments involving physics concepts and processes [4].

A number of applications are available and was considered based on experience and results of other researches, aiming to stimulate students' active engagement and offering the opportunity to work under conditions that are extremely difficult, costly or time consuming to be created in the classroom or even the physics lab. The use of such applications has developed a new research field in physics education, since it radically changed the framework under which physics teaching is being understood and implemented. Computer simulations are considered as a part of teaching physics through PBL [7].

If educators use tools that are virtual then there must be a platform that combines together all the options that are necessary. Local e-learning platform has to be used in teaching and learning process.

III RESULTS AND DISCUSSION

Combined PBL, advantages and disadvantages

Today's information community expects graduates not only to have a specific knowledge base but also be able to apply this knowledge to solve complex problems in an efficient way. Educational research has shown that successful problem solvers possess an organized and flexible base and master the skills to apply this knowledge for problem solving.

PBL can be recognized from other teaching – learning styles with several characteristics such as learning is student centered, learning occurs in small groups under the guidance of a tutor. The tutor is a facilitator or guide. One of the most recognizable characteristic of PBL is that problems are encountered before any study has occurred. Students deal with real life problems, so they acquire skills that will be necessary in real life situations.

PBL can be used to study almost every subject and there is no significant proof that this learning method is more effective than conventional learning. But there are several positive signs: students found PBL to be more nurturing and enjoyable than conventional learning, PBL graduates performed as well as students who had received conventional learning.

Active learning can be also called as PBL because it's learning environment encourage students to engage in solving problems, sharing ideas, giving feedback and teaching each other [24]. Negative side of active learning is that students' attention to what the instructor was saying decreased as the lecture proceeded. But based on evidence integrating active learning strategies as part of the formal learning sessions can advance students learning.

Advantage of using experimental PBL is that students feel responsible for providing answers to the problems provided in the laboratory instructions, and those they encounter themselves. They develop skills to solve these problems, and therefore become more competent. Also laboratory time is used efficiently, with all participants fully engaged in the learning process. Students leave the lab more mature and confident about their skills. They know they can solve other problems when they encounter them in other classes or later in industry. Students develop understanding of how data acquisition system works and develop their own skills to solve hardware and software problems [17]. They also learn data acquisition techniques and software applications to solve real world problems.

Disadvantage is that students may initially struggle with PBL when this approach is new to them. They may not be prepared to answer questions posed by the instructor, they may want the instructor to solve all problems for them. Students may struggle with software and hardware for data acquisition. Learning how to use new software in one lecture can be a

challenge also laboratory assignment may not be completed during the given time.

Computer simulations in physics teaching

Learning physics is often considered to be a difficult pursuit. In most cases students can no relate theory with real life, they cannot t name examples that relate to real life problems. A variety of computer applications have been developed and used in teaching physics to improve learning process [12].

Among the various applications, computer simulations have achieved special importance in physics teaching and learning. Simulations offer new educational environment which aim to enhance teachers' instructional potentialities. Computer simulations offer a great variety of opportunities for modelling concepts and processes.

They are open learning environments that provide students with the opportunity to develop their understanding about phenomena and physical laws through a process of making hypothesis and testing ideas, isolate and manipulate parameters and therefore helping them to develop an understanding of the relationship between physical concepts and variables, investigate phenomena which are difficult to experience in a classroom or lab setting because it is extremely complex, technically difficult or dangerous, money consuming or time consuming, or happen too fast [13].

There are commonly known two types of computer models in physics. The first one is exploratory models, which are constructed by experts to represent domain knowledge. Usually they are micro – worlds that simulate physical processes and laws. Such micro – worlds encourage students explore and interact with them, handle parameters and observe their results.

The second is expressive models, which allow students express their own ideas on a domain. They provide learners with tools to define relationships between concepts, explore the consequences of relationships that students have defined and learn through an active process of representing their own models.

Computer simulations are used to enhance traditional instruction. Computer simulations are used as supplement or alternative to traditional teaching [5].

There are few results that students who used computer simulations in their study were able to get higher scores in tests, but those results don't prove that the knowledge they learned is consistent [6].

Simulations can also be used as a remedy to prepare for laboratory activities. This method shows results of increasing interest in specific courses and overall improvement of academic performance.

There is a danger that with computer simulation, students will see no need to take responsibility for their own understanding, to verify, or to challenge. This can result in students learning science passively.

Studies on how computer simulation improves learning performance have produced diverse results, with some of them finding that simulation based learning does not significantly improve the test results of learners.

High performance data acquisition on USB

The opinion for plugging data acquisition module into PCI bus to get higher performance is no longer so commonly seen. The reason for that is the usage of USB ports that perform as fast as PCI based cards.

Utilizing the USB as a data portal provides many advantages for data acquisition (DAQ)[1]. For example the software installation is much simpler, because users do not have to open their PC and configure the hardware. System software will automatically form a logical connection from DAQ channel to a destination within the application software and hardware setup [22]. Applications simply need to identify the logical pipe they wish to connect with and the system software ensures that data travels to the right destination. Application software does no need to know the details of peripheral addressing, interrupt placement, or other installation dependent parameters as it must for PCI plug-in cards.

The USB also has electronic advantages. The bus can power the peripheral so that the DAQ system does not need its own power source. This kind of specification is very useful if it is necessary to do measurements in different kinds of environment.

Moodle

Most common e-learning platform also known as virtual learning environment that is used in teaching process is Moodle [15]. Virtual learning environment provides different learning materials such as videos, music files, images, links to other web sites, animations, simulations. There are also tools provided by Moodle to test the level of knowledge. Moodle provides online space for discussion through the form of chats and forums.

Based on evidence Moodle can be used to support learning but not to replace learning process [15]. The structure and quality of materials that are put in Moodle must be strongly considered otherwise the learning process could be meaningful. Hardest part of good learning materials is to create learning materials that develop ability to solve problems that represent different situation or problems because students have difficulties with applying laws and equations to real life problems.

In several regions Moodle is most commonly used by teachers and professors [15].

IV CONCLUSION

Science and technology develop through the exchange of information, much of which is presented as moving images, diagrams, illustrations and models that summarize information and help understand

scientific data and phenomena. To fully understand the nature of scientific phenomena, students should be exposed to different types of models, such as verbal, mathematical and visual models.

Different teaching strategies can be used to provide suitable education support to enhance the effectiveness of learning. Technologies are often used by teachers to improve teaching and learning process.

Instructional computer simulations and other applications of the computer in the classroom will appropriately continue to be an important part of teaching physics.

The laboratory experiments that students carry out as part of the course are of exploratory nature and are often accompanied by real time data collection. Most of these experiments cannot be done with usual methods because most of them require a very long time of observation or in opposite the process that must be measured is very fast. Experiments are supplemented with 2D and 3D visualizations and simulations [18]. The visualizations allow students to make abstract ideas concrete. Most of physics concepts can be modeled with mathematical relationships [10]. Computer generated models of these relationships are provided for learners in every field of physics [9].

Based on research it is possible to develop virtual teaching and learning in Liepaja University. Computer simulations are already used in learning process in course of electromagnetism. There does not exist absolute arguments that simulations and PBL cannot be used in teaching and learning process. So these methods will be more deeply considered and properly used in the field of physics along with all the necessary experimentation and usage of computer simulations.

V REFERENCES

[1] S. Thanee, S. Somkuarnpanit and K. Saetang, Eds., *FPGA – Based Multi Protocol Data Acquisition System with High Speed USB Interface: International Multi Conference of Engineers and Computer Scientists 2010 Vol II,* March 17-19, 2010, Hong Kong, China.

[2] Biswajit Ray, Eds., *An Instrumentation and Data Acquisition Course for Electronics Engineering Technology Students: 2003 American Society for Engineering Education Annual Conference & Exposition,* 2003, Bloomsburg, USA.

[3] R. J. Robinson and J. Welli, Eds., *Introducing Data Acquisition and Experimental Techniques to Mechanical Engineering Students in the Freshmen Year: 2002 American Society for Engineering Education Annual Conference & Exposition,* 2002, Rochester, USA.

[4] A. Jimoyiannis and V. Komis, "Computer simulations in physics teaching and learning: a case study on students understanding of trajectory motion," Computers & Education, vol. 36, pp. 183-204, Nov. 2000.

[5] N. Rutten, W. R. van Joolingen and J. T. van der Veen, "The learning effects of computer simulations in science education," Computers & Education, vol. 58, pp. 136-153, Jul. 2011.

[6] L. P. Rieber, S. Tzeng and K. Tribble, "Discovery learning, representation, and explanation within a computer-based simulation: finding the right mix," Learning and Instruction, vol. 14, pp. 307-323, 2004.

[7] K. Chang, Y. Chen, H. Lin and Y. Sung, "Effects of learning support in simulation-based physics learning," Computers & Education, vol. 51, pp. 1486-1498, Jan. 2008.

[8] F. Dochy, M. Segers, P. van den Bossche and D. Gijbels, "Effects of problem-based learning: a metaanalysis," Learning and Instruction, vol. 13, pp. 533-568, 2003.

[9] W. Christian and F. Esquembre, "Modeling Physics with Easy Java Simulations," The Physics Teacher, vol. 45, pp. 475-480, Nov. 2007.

[10] K. Perkins, W. Adams, M. Dubson, N. Finkelstein, S. Reid and C. Wieman, "PhET: Interactive Simulations for Teaching and Learning Physics," The Physics Teacher, vol. 44, pp. 18-23, Jan. 2006.

[11] N. Finkelstein, W. Adams, C. Keller, K. Perkins and C. Wieman, "High-Tech Tools for Teaching Physics: the Physics Education Technology Project," Journal of Online Learning and Teaching, Sep. 2006.

[12] C. E. Wiemann, W. K. Adams and K. K. Perkins, "PhET: Simulations That Enhance Learning," Science, vol. 322, pp. 682-683, Oct. 2008.

[13] T. D. Jong and W. V. Joolingen, "Scientific discovery learning with computer simulations of conceptual domains," SAGE Publications, vol. 68, pp. 179-202, Nov. 2007.

[14] J. Bryan, "Technology for Physics Instruction," Contemporary Issues in Technology and Teacher Education, vol. 6(2), pp. 230-245, 2006.

[15] T. Martin-Blas and A. Serrano-Fernandez, "The role of new technologies in the learning process: Moodle as a teaching tool in physics," Computers & Education, vol. 52, pp. 35-44, Jun. 2008.

[16] Y. J. Dori and J. Belcher, "How does technology-Enabled Active Learning Affect Undergraduate Students Understanding of Electromagnetism Concepts," The Journal of the Learning Sciences, vol. 14(2), pp. 243-279, 2005.

[17] C. Wiemann and K. Perkins, "Transforming Physics Education," Physics Today, vol. 58(11), pp. 36-50, Nov. 2005.

[18] N. D. Finkelstein, W. K. Adams, C. J. Keller, P. B. Kohl, K. K. Perkins, N. S. Podolefsky and S. Reid, "When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment," The American Physical Society, vol. 1, pp. 1-8, Oct. 2005.

[19] R. N. Steinberg, "Computers in teaching science: To simulate or not to Simulate," Phys. Ed. Res, vol. 68, pp. S37-S41, 2000.

[20] M. E. Gredler, "Games and simulations and their relationships to Learning," Handbook of Research on Educational Communications and Technology, pp. 571-581, 2004.

[21] "Improving students understanding of electricity and magnetism," March 20, 2012. [Online]. Available: http://d-scholarship.pitt.edu/11767/1/Jing_Li's_Doctoral_Thesis_4_a%26s_format_revised7.pdf. [Accessed: Jan. 30, 2015].

[22] "Benefits & Potential Dangers of Using USB for Test & Measurement Applications," [Online]. Available: http://www.controldesign.com/assets/12WPpdf/121024-data-translation-benefits-dangers-using-usb.pdf. [Accessed: Jan. 30, 2015].

[23] "Problem based learning of Data Acquisition and Computer-Based control Applications in MET Courses," [Online]. Available: https://www.ni.com/pdf/academic/us/journals/lv02_52.pdf. [Accessed: Jan. 30, 2015].

[24] Y. J. Dori and J. Belcher, "Learning Electromagnetism with Visualizations and Active learning," in *Visualizations in Science Education,* vol. 1, Models and Modeling in Science Education, J. K. Gilbert, Ed. Springer Netherlands, 2005, pp. 187-216.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 159-163

Structuration of courses at studying disciplines of programming

Natalya Prokofyeva, Marina Uhanova, Oksana Zavyalova, Sabina Katalnikova
Riga Technical University, Faculty of Computer Science and Information Technology.
Address: ¼ Meza str., Riga, LV-1048, Latvia

Abstract. **The article describes the methodology of learning programming for students of various engineering disciplines (chemists, electricians, programmers, and mechatronicians). The courses "Computer Training" and "Programming Languages" taught at the Faculty of Computer-Science and Information Technology of Riga Technical University are used as examples. The purpose of these courses is to prepare specialists able to use computers effectively and without error in their future careers. The course structure depends on the particular specialty area. The course structure consists of three parts: the main (theoretical) part, the laboratory part and the practical test. In this paper we show that for better development of course material it is necessary to consider the solution of specific practical problems that may be encountered by engineers of particular profile, as well as to provide current and final control of students.**

Keywords: **methodology of course structuration, computer training, programming language.**

I INTRODUCTION

Many international scientific conferences (such as IEEE ICALT, IADIS e-Learning, IASTED CATE, etc.) and e-journals [1][2][3] are dedicated to research in the field of computer-based learning and knowledge assessment.

The current stage of educational development in Latvia is connected with the transition of higher education to the educational standards of the third generation based on the methodology of competence approach [4][5].

With the transition to the new standards, university teachers face a predictable range of questions - "Why do I have to teach?", "What to teach?", "How should I do the teaching?" and "How to evaluate student performance?" [6][7][8][9]. The answers have universal bearing, because they give the guidance for constructing and mastering each study discipline.

The Authors of this article would like to dwell on the methodology of structuring educational course content. The importance of the discussion is determined by the fact that functional and conscious acquisition of learning material mainly depends on its structure, in which the structure and character of the activities aimed at mastering the material is already implied.

Firstly, we want to focus your attention on the important principle of contemporary methodology – the principle of complementarity. It is connected with the introduction of innovations as an addition to the existing traditional components of study content and strengthening the links between them in the context of modelling a new system of educational course content. In this way it defines the priority of systemic and integrative approaches to selection and construction of content, which ensure that created courses focus on integration, systematization and structuring of a plurality of content elements into a rational system of educational content.

The rationality of educational content can be reached by founding "the core" of the content, which also defines the subsequent filling of courses. It is important to proceed from submission of content in any educational discipline, which is admitted in pedagogy – that is scientific knowledge, practical skills and also experience and professional skills in a specific area of knowledge. Moreover, we should consider general pedagogical principles of optimizing the amount and difficulty of educational material (theoretical and practical significance; matching the material to the age and individual characteristics of students, etc.)

Considering the above, "the core" can be presented through the unity of the following invariant parts of educational course content: informational and scientific part; laboratory and practical part; examination and verification part. Obviously, the filling of identified invariant parts with concrete elements will depend on requirements of educational standards, goals and objectives of learning in concrete disciplines, its peculiarities and other factors.

Courses in programming for chemists, electricians, programmers and mechatronicians are taught at the Faculty of Computer-Science and Information

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.179

Technology of Riga Technical University. The following article describes the methodology of structuring of the courses "Computer Training" and "Programming Languages" for students in these fields.

II THE CONTENT OF "COMPUTER TRAINING" AND "PROGRAMMING LANGUAGES" COURSES

The "Computer training" and "Programming languages" courses form the basis for the implantation of information technology into the educational system. The courses have an important place in the multilevel system of preparation of experts in different fields. They not only presuppose the organization of practical work using a computer, but also introduce students to the principles of algorithmic computational processes and to the technology of creating application programs.

The course "Computer Training" in RTU is taught to the first-year students of two faculties – the Faculty of Material Science and Applied Chemistry and the Faculty of Power and Electrical Engineering.

In the first theoretical part students learn techniques for algorithm development and programming, get acquainted with the syntax of high-level programming languages by the example of VBA (Visual Basic for Applications). It includes the following topics: elaborating branching algorithms and programs; writing cyclic programs and the use of nested loops; working with strings of characters and text files; using Microsoft Excel object model.

In addition, examples of tasks are selected for students, some of which are cited as the examples during lectures, some of which the students have to carry out on their own during the laboratory works and some of which are offered as optional tasks of high complexity for enthusiastic students.

In the laboratory and practical part a number of practical works are offered. For example, a part of laboratory work devoted to branching programs and algorithms is the development of macro for calculating stationary sedimentation rate of the particles by using the Reynolds number. Practical application of the cyclic program is illustrated by using numerical methods for calculating definite integrals. The estimation of the value of the definite integral is necessary in solving many engineering tasks, e. g. for calculating the parameters of the chemical process such as heat transfer.

The use of nested loop is shown not only on the example of matrix problems and the examples of processing of array elements, but also on the example of calculating the composition of the mixture obtained by mixing various substances.

The tasks associated with the calculation of the mixture usually are reduced to the calculation of the product vectors and matrices and are encountered quite often in chemistry. Other more common tasks include the opposite to given task nd require from chemical technologists the calculation of proportions of initial materials in order to obtain the mixture of a given composition. Given tasks are reduced to solving a system of linear equations. The development of macros in VBA for solving systems of linear equations is offered to students for independent work as an optional task of high complexity. One example of the development of programs for working with character strings is the task of finding a relative molecular mass of a substance. The students are offered to write in VBA the program that introduces the chemical formula of the substance (for example H_2SO_4) and calculates the molecular weight. It is proposed to try to implement the error handling in case when the user enters the name of a nonexistent element or enters the name of the element using the wrong case of letters. The programs written during laboratory work are performed in Microsoft Excel environment that facilitates the creation of user-friendly interface and allows using the advantage of Excel in the automation of calculations.

It is known that the pledge of quality training of students is to ensure the feedback, that is, the organization of knowledge and skills control acquired in the learning process. That is why examinations are held regularly on this study course after completing each topic.

The course "Programming Languages" is taught to the students of the Faculty of Computer Science and Information Technology and of the Faculty of Transport and Mechanical Engineering. The aim of this course is to introduce students to new software in the area of information technology. Thus, the main goal of the discipline is to provide theoretical knowledge about programming languages and their use in solving applied tasks. As a result of the program, students are expected to master the principles of organization of computing processes as well as programming techniques for high level programming languages.

The course content includes the following topics: classification of programming languages and their use; development of algorithms of and programs in C (work in the editor DEV C ++); work with dynamic variables; work with records and text files. In addition to traditional topics teachers give students the task – independently prepare abstracts on contemporary topics in the field of computer and information technologies and read a report in front of their classmates.

Laboratory practical part of the course consists of classical traditional tasks in studying any programming language: implementation of branched and cyclic algorithms; basis of modular programming, learning functions and procedures; working with text and text files.

For several years, in conducting classes on "Programming Languages", teachers have been using the ORTUS educational environment [10] for

knowledge evaluation. With the help of this environment different types of examinations, tests and independent work checks can be incorporated into the learning process. The teacher can create a test by setting the number of times of passing the test, time and place of delivery (such as in class or remotely). It is possible to envisage the formation of random questions, as well as the rotation of answer options in question. All types provide comments, answers, as well as the ability to use pictures, tables, etc. It is possible to create tests from different categories of questions, to take into account the difficulty of questions accord more points to complicated questions [6].

The teachers successfully implement the ORTUS environment in the traditional model of the educational process as it allows to make learning more interactive and increases the importance of independent student work. This course develops the basic skills of programming and it is important not only for professional training but also for the formation of information culture of future engineers.

Thus, in the courses "Computer Training" and "Programming Languages" a whole range of skills are acquired by future specialists (chemists, electricians, programmers mechatronicians), which are necessary for the widespread use of information and computer technologies as well as the appropriate software in the future professional activity.

III METHOD OF STRUCTURING AND ITS RESULTS

As is known, every learning course comprises three components: the main (theoretical) part (ideas and knowledge); the laboratory-practical part (skills, experience); the evaluation part (evaluation of skills). All these parts are obligatory and traditionally are taught by all teachers in higher education.

It has been proved that the use of computer systems in laboratory and practical part increases students' academic performance and motivation [11].

The authors suggest improving the structure of the course content, supplementing it with new elements that are directly related to the solution of applied tasks of relevant specialties – chemists, electricians, etc.

For example, during the learning process of nested loops theme on the subject "Computer Training", in one of the tasks students are encouraged to develop a program for calculating the proportions of the oxides in the cement charge. The cement charge is prepared by mixing clay marl (composition: CaO 12-25%, MgO 0-2%, SiO_2 25-40%, Al_2O_3 5-15%, Fe_2O_3 about 8%), lime (composition: CaO 48-55%, MgO 4-8%), nepheline (composition: Na_2O + K_2O 19-20%, Al_2O_3 29-30%, SiO_2 43-44%). In the composition of raw materials volatile substances are not typically included. Students are encouraged to create an Excel spreadsheet, similar to that shown in Fig. 1.

	A	B	C	D	E	F	G
1	the composition of raw materials						
2		CaO	MgO	SiO_2	Al_2O_3	Fe_2O_3	Na_2O+K_2O
3	clay	0.2	0.01	0.3	0.1	0.08	0
4	lime	0.49	0.05	0	0	0	0
5	nepheline	0	0	0.43	0.29	0	0.19
6							
7		the proportion of raw materials					
8		clay	lime	nepheline			
9		0.4	0.5	0.1			
10							
11		the proportion of oxide in the mixture					
12		CaO	MgO	SiO_2	Al_2O_3	Fe_2O_3	Na_2O+K_2O
13							

Fig. 1. Example of Excel spreadsheet

To find the proportion of oxides in the mixture it is necessary to multiply the vector containing the proportions of raw material by the matrix of raw material composition. As a result, the vector will be obtained 1x6 containing a proportion of oxides in the mixture. Macro that performs multiplication of this vector by a matrix and displays the results in Table Excel (in cells B13: G13) can look like this:

```
Public Sub Sostav()
For j = 2 To 7
  For i = 1 To 3
    Cells(13, j) = Cells(13, j) + _
      Cells(9, i + 1)*Cells(i + 2, j)
  Next
Next
End Sub
```

To solve this task, students need to know, firstly, how to work in the Microsoft Excel environment, secondly, to be able to describe the cyclic algorithm in the language of VBA (Visual Basic for Applications) and to use macros for solving a specific task. In such a way the laboratory and practical part is complemented by a new element (module), which allows students of chemical faculty to acquire knowledge in the area of information technology, as well as in their professional area.

So the subjects "Computer Training" and "programming languages" include not only the main components ("core") of training courses, but are supplemented with new content elements Mi (modules). The general scheme of structuring training courses is presented in Figure 2.

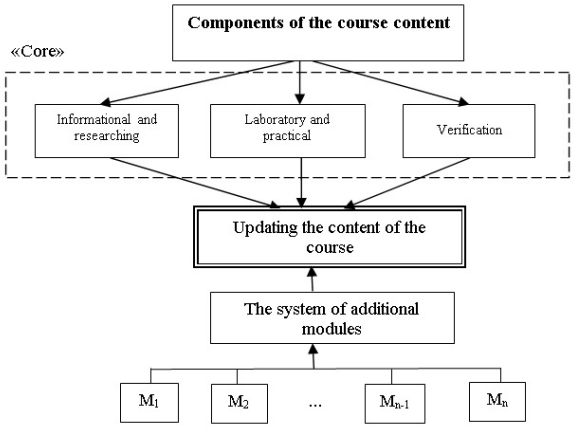

Fig. 2. The components of the course content

The modules can consist of applied tasks for chemists, electricians and mechatronicians. Those may also be the tasks for independent work, e. g., to write and defend an essay on one of the topics in the area of information technology, etc.

Thus, the authors hypothesize that a balanced complement of courses with new content must significantly improve student performance, as well as the quality of education in these disciplines.

In order to prove or disprove the hypothesis, an experiment was conducted and statistical data collected for five academic years at the Faculty (Table. 1). Students are divided into two main groups: the students who were trained under the traditional method (year 1, 2) and the students trained and evaluated under the proposed method (year 3, 4, 5).

TABLE 1.

THE NUMBER OF STUDENTS PARTICIPATING IN THE EXPERIMENT

Subject	Academic years					Students in total
	1	2	3	4	5	
Computer training	95	80	93	79	115	462
Programming languages	282	266	242	287	280	1357

The general formula for determining the coefficient of assessing the quality of student groups in the discipline of the training period (semester) has the form [12]:

$$K = \frac{1}{3}\left[\frac{1}{2}(A_{val} + A_{exam}) + \frac{1}{2}(O_{val} + O_{exam}) + \frac{1}{40}(Q_{val} + Q_{exam})\right]$$

Where

K - the evaluation of the quality coefficient;

A_{vol} - the average score of a group of students according to the results of current knowledge control during the semester;

A_{exam} - the average score of student groups according to exam results;

O_{val} - the overall assessment of student groups per semester, determined as follows: "5" - 90% of students have a positive assessment, with 50% of students having an assessment of "10" and "9"; "4" - 90% of students are evaluated positively, with 50% of students having a rating of "10", "9", "8" and "7"; "3" - 80% of students have a positive evaluation; "2" - not satisfying conditions for assessing the "3";

O_{exam} - the overall assessment of student groups for the exam, defined as O_{val};

Q_{val} - the percentage of good and excellent ratings in the group of students according to the results of the current knowledge control during the semester;

Q_{exam} - the percentage of good and excellent ratings in a group of students according to examination results.

The results of the experiment are shown in Tables 2 and 3.

As can be seen from the tables, in teaching students under the traditional scheme the coefficient of assessing the quality of training of students K is significantly lower than when applying advanced methods in teaching these subjects.

Thus, the balanced addition of courses with new content elements not only improves student performance, but also greatly improves the quality of education and knowledge control.

TABLE 2.

COEFFICIENT OF ASSESSING THE QUALITY OF STUDENT TRAINING ON THE SUBJECT "COMPUTER TRAINING"

YEARS	A_{VOL}	A_{EXAM}	O_{VAL}	O_{EXAM}	Q_{VAL}	Q_{EXAM}	K
1	5,57	6,56	3	3	42,22	49,00	3,78
2	5,71	6,52	3	3	40,51	48,15	3,78
3	5,97	7,20	3	4	45,83	64,04	4,28
4	6,28	7,20	4	4	58,05	67,45	4,63
5	6,79	7,37	4	4	52,43	73,45	4,74

TABLE 3.

COEFFICIENT OF ASSESSING THE QUALITY OF STUDENT TRAINING ON THE SUBJECT «PROGRAMMING LANGUAGES»

YEARS	A_{VOL}	A_{EXAM}	O_{VAL}	O_{EXAM}	Q_{VAL}	Q_{EXAM}	K
1	5,34	5,79	3	3	38,15	42,10	3,53
2	5,71	6,12	3	3	41,08	43,13	3,68
3	5,97	6,80	3	4	46,78	53,26	4,13
4	6,12	7,15	3	4	56,78	62,18	4,37
5	6,07	7,09	3	4	49,38	60,23	4,27

That is the actual task in the formation of the skills engineers will use in their future professional activities.

IV CONCLUSION

This technique of structuring training courses (Fig. 1) can be used for all courses taught in higher education institutions, as the quality control of training students encourages systematic, independent and creative learning activity, self-control of achieved results and their regular and objective evaluation, as well as the cultivation of responsibility for the results of one's academic work.

V REFERENCES

[1] The Electronic Journal of e-Learning (EJEL) [Online]. Available: http://www.ejel.org/main.html [Accessed: March, 2015].

[2] Journal of Educational Technology & Society [Online]. Available: English version: http://www.ifets.info/, Russian version: http://ifets.ieee.org/russian/periodical/journal.html [Accessed: March, 2015].

[3] International Academy, Research, and Industry Association (IARIA) [Online]. Available: http://www.iariajournals.org/ [Accessed: March, 2015].

[4] United Nations Educational, Scientific and Cultural Organization [Online] / – http://en.unesco.org/ [Accessed: March, 2015].

[5] Joint Council of Europe/UNESCO (Lisbon) Convention [Online]. http://conventions.coe.int/Treaty/en/Treaties/Html/165.htm [Accessed: March, 2015].

[6] L. Rastrigin, M. Erenshtein, *Adaptive teaching with a model*. Riga, Zinatne: 1986.

[7] L. Nicetsky, *Requirements specification for development of the RPI computer-based teaching system first version*. Riga, Riga Polytechnic Institute (RTU): 1977.

[8] L. Zaitseva, L. Novitsky, V. Gribkova, *Computer-aided teaching systems development and using*. Riga, Zinatne: 1989.

[9] N. Prokofjeva, *The Methodical and Technical Aspects of Students Knowledge Control*, Proceedings of the 13th Education and Virtuality International Conference, VIRT - 2011, Ukraine – Yalta, 2011, pp. 263 – 272.

[10] N. Prokofjeva, A. Anohina-Naumeca, O. Lebedeva, *Administration of Knowledge Assessment at Riga Technical University*. Proceedings of the 8th International Multi-Conference on Computing in the Global Information Technology (ICCGI 2013), July 21-26, 2013, Nice, France, IARIA, pp.34-39, ISBN: 978-1-61208-283-7. (indexed in ThinkMind Digital Library).

[11] N. Prokofjeva, D. Zagulova, S. Katalnikova, "Possibilities of Use of Computer Systems in Educational Process by Preparation of IT of Experts," Educational Technology & Society, 2013, Vol.16, No.4, pp.511-520. ISSN 1436-4522. [Electronic resource] / – http://ifets.ieee.org/russian/depository/v16_i4/pdf/18.pdf [Accessed: March, 2015].

[12] O. Grigorash, A. Trubilin, "Methods for estimating the quality of training of students and effectiveness of the training activity of teachers and department," Polythematic network electronic journal of the Kuban State of Agrarian University (The science journal KubGAU). Krasnodar: KubGAU, 2013. - №92 (08).

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 164-169

Fuzzy multiple criteria decision making approach in environmental risk assessment

Andrejs Radionovs, Oleg Uzhga-Rebrov
Rezekne Higher Education Institution, Faculty of Engineering.
Address: Atbrīvošanas aleja 115/K-3, Rezekne, LV-4601, Latvia

Abstract. **Being able to evaluate risks is an important task in many areas of human activity: economics, ecology, etc. In case of a sufficient amount of source information the risk is evaluated using statistical methods. However, in reality the sufficiency of statistical data in risk assessment is more exceptional than normal. In such cases experts' assessment make the only source of data. Experts are able to provide the necessary for analysis data due to their professional knowledge and experience. Certain amount of factors, which is to be evaluated by an expert (experts), significantly affects the process of experts' assessment. If a big number of relevant factors occur, an expert may face a problem of defining links between "factors" and "outcome". Fuzzy multiple criteria decision making approach can be used to solve the problem. Ecological risk assessment towards human health in case of gaseous substances escape at a chemical factory using hierarchical method and fuzzy multiple criteria decision making approach has been analyzed in the article.**

Keywords: **Fuzzy logic, risk assessment, fuzzy representation of knowledge.**

I INTRODUCTION

Due to the increasing both of human industrial activity and factors promoting the environment pollution, a necessity of risk assessment becomes more essential. There exist quite a big amount of investigations devoted to the analysis of ecological risks, for example [1] and [2]. In order to make the ecological risks analysis complete, a lot of statistical data are needed. There are many papers describing the ecological risks analysis through the assessment of statistical data, for example [3]. Appraisal of a building and its insurance could be an example of statistical method. For instance, there is enough information available all over the world on the possibility of fires in various types of buildings and their market values. Estimated losses caused by fire can be easily calculated as average losses related to previous cases. Due to the rates obtained the risk can be easily evaluated and the cost of the insurance can be determined. Unfortunately, when speaking about the analysis of ecological risks the enough availability of statistical data is more exceptional than normal. Uniqueness of many situations and their rare occurrence does not allow using an effective statistical instrument for getting certain assessments. In order to cope with the lack of objective information, experts' judgments are often used. On the basis of their professional knowledge and experience, experts are able to provide the necessary for analysis data. Quite a big number of methods are meant for acquiring and using ambiguous probabilistic assessments, including interval probability, probability of second degree, etc. Difficulty of these methods and a bad interpretation of uncertain results make the methods imperfect. In 1965, L.Zadeh [4] proposed a principally new conceptual basis for dealing with imprecise information – fuzzy set theory. This theory was widely developed during the past years. Nowadays, fuzziness is used practically in all fields of scientific and practical activity, including risk assessment. The method contains many advantages, for example: includes qualitative variables in the risk assessment; works with fuzzy information; operates with linguistic variables; allows fast modeling of complex dynamic systems; however, there exist disadvantages, too: it is reasonable to use the method purely with a small amount of the risk influencing factors. Increasing of number of the mentioned factors leads to a complicated process of setting the rules, and therefore makes the possibility to make a mistake more common, and an expert may face a problem of determining the links ″factors - final result″. So, what should we do if it is necessary to evaluate risk with a big number of factors? In such case hierarchical fuzzy multi-criteria decision making approach may be used.

There exists quite a big amount of papers describing the usage of this method in various fields, e.g. For example, paper [5] deals with using the method to diagnose the disease aphasia.

Hierarchical fuzzy multi-criteria decision making approach in the risk assessment towards human health in case of the gas substances escape at a chemical

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.169

factory has been used in the present article for the ecological risks assessment.

II FUZZY LOGIC AND FUZZY SETS

Nowadays there are a lot of textbooks and reference books available on the theory of fuzzy sets and fuzzy logic. Work [6] can be used as an example, but a more detailed and complete description of the theory can be found in [7]. Fuzzy sets theory represents a number of mathematical principles for knowledge representation based on degrees of a membership. The definition "fuzzy logic" has appeared in the sixties with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh, when he published his work "Fuzzy Sets" in order to provide a model for inexact and not precise concepts and subjective judgments.

Fuzzy Sets deals with degrees of membership to certain class and degrees of truth. Fuzzy logic uses fuzzy rules that should receive desire results from input linguistic data or variables.

The linguistic variable is a variable, which values sentences in a natural language. For example, linguistic variable "Age" contains values: young, old, etc.

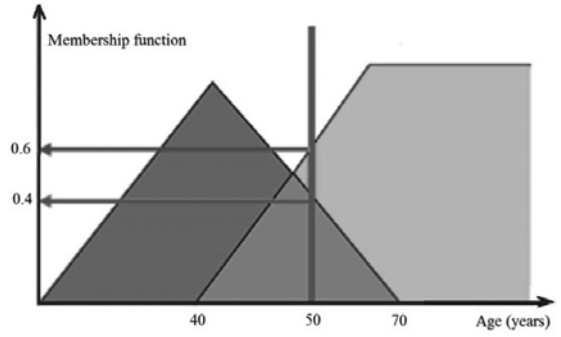

Fig. 1. Fuzzy age-classification scheme

The advantage of fuzzy methods is their reflection in human mind and its ability to store and process information that is uncertain, imprecise, and resistant to classification [8]. Fuzzy logic is an alternative to the Boolean logic where every proposition must be either true or false. However, fuzzy logic implies that things can be at the same time "true" and "false" with a certain membership degree [9]. For example, in Boolean logic a 50 year-old person can be young or old depending on the scale. But in fuzzy logic (as it is shown in Fig. 1) a man can be simultaneously old with degree of membership of 0.6 and young with 0.4.

Fuzzy set theory provides a way to use imprecise and uncertain information generated by the system and human judgments in a precise way. When the available environmental data availability do not provide an appropriate statistical information, fuzzy approaches can solve this problem, since it works well for addressing poorly defined parameters and linguistic variables [10].

One of the advantages of Fuzzy logic is its ability to work with different kinds of parameters (e.g. environmental, health), quantitative, qualitative and crisp values.

III MULTIPLE CRITERIA DECISION MAKING APPROACH

Many factors influence the risk assessment of human health in case of the gas substances escape at a chemical factory. Factors are represented by Fig. 2. Experts have to determine 3^{12}=531441 rules by using a one-level fuzzy risk assessment methodology with 12 input variables and 3 definite for each variable fuzzy sets. This phenomenon was called by Bellman [11] the "curse of dimensionality". The rule base rapidly overloads the memory and makes the fuzzy system unusable. The problem can be solved if the 12-variable fuzzy system is reduced into a hierarchical fuzzy system. For example, 8 hierarchically connected low-dimensional fuzzy systems as shown in Fig. 3, the number of fuzzy rules reduces to $3*(3^3)+5*(3^2)$=126 rules in total. This result is gained by dividing a big system into components, and by using the mechanism of fuzzy process towards each system component and level.

Further an example of fuzzy methodology towards one of the system's components is shown by a dotted line at Fig.3.

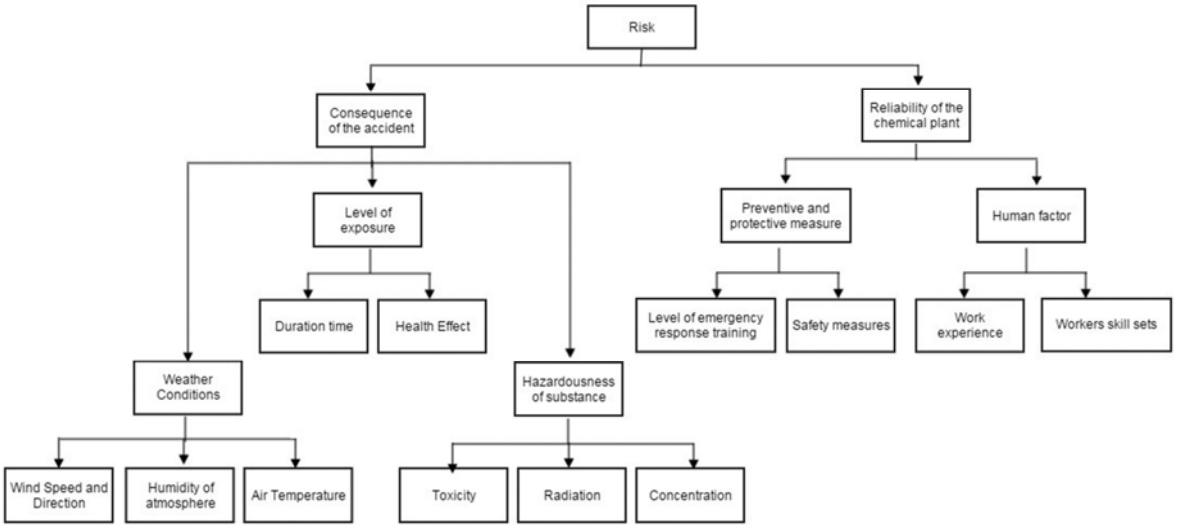

Fig. 2. Risk assessment scheme for escape of gas substances in chemical plants

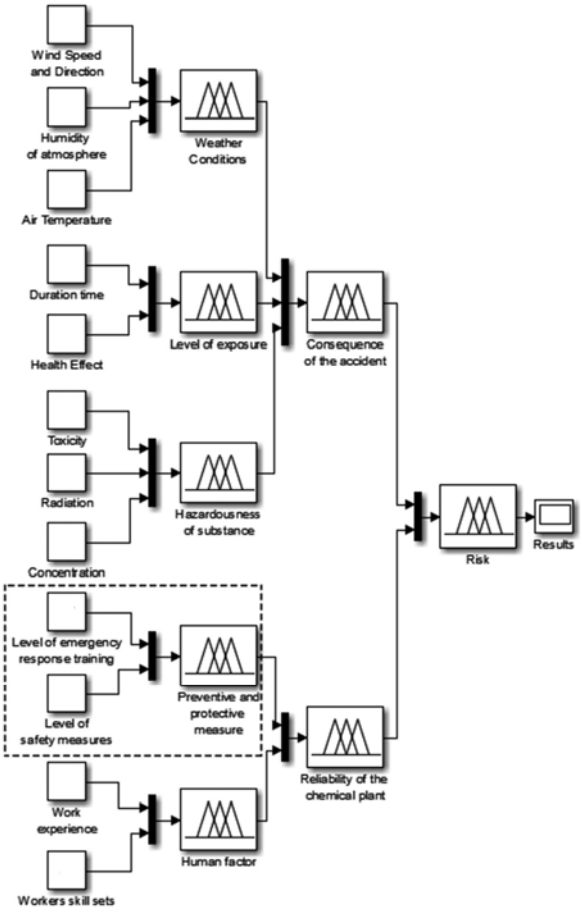

Fig. 3. Hierarchical Fuzzy Rule-Based Model Architecture

IV FUZZY PROCESS FOR ONE SYSTEM'S COMPONENT

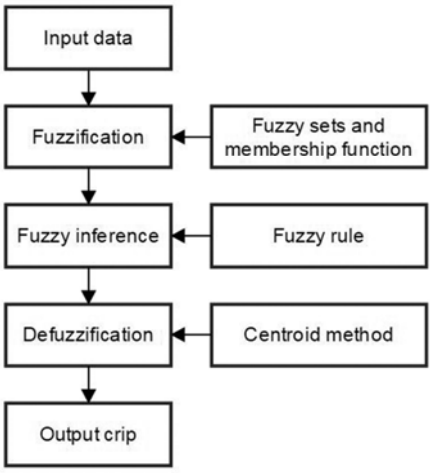

Fig. 4. Representation of fuzzy methodology

Projecting of uncertain system mainly includes two operations: knowledge base derivation and the selection of fuzzy inference process to perform fuzzy reasoning [12].

Successful projecting of uncertain system for certain Application Domain is a complicated process of several levels where the designer faces a big amount of alternative strategies of realization [13].

Fig. 4 shows that the proposed uncertain model includes various levels. Parameters of input and output should be determined at the first step, and then transformed from original values into linguistic categories by creating fuzzy sets for each of them. This process is called fuzzification. Secondly, it is necessary to determine the rules. These rules will

allow system to "reason" or going from the input variables to the output variable. For this purpose Mamdani's fuzzy inference method will be used. For the final step, the obtained value must be defuzzificated or converted back to crisp value in order to be used in the next level of fuzzy process as shown in Fig. 3.

All the steps have been implemented by using fuzzy toolbar available in Matlab and described further.

A. Definition of Input

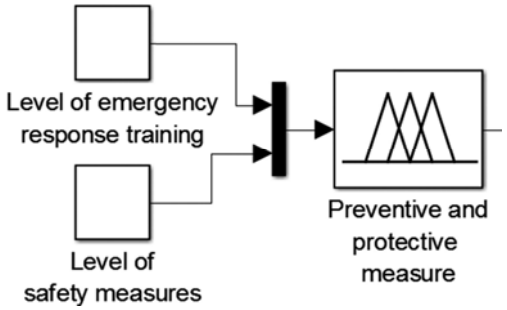

Fig. 5. Part of risk assessment scheme

At the first step the input and output variables (shown in Fig. 5) must be defined and then converted from the original values in to the linguistic categories by creating fuzzy sets for each of them. As it is shown in Tables I and II, fuzzy sets for *Level of emergency response training* and *Level of safety measures* inputs variables were made.

TABLE I

CRITERIA OF EMERGENCY RESPONSE TRAINING LEVEL

Emergency response training levels (E)	Criterion
I First Responder Awareness	Personnel likely to discover a release and call for assistance. They would leave the area, call for help, and keep others to get out of the area. Min.2-4 hours training
II Responder Operations	Personnel are to respond in a defensive manner without actually trying to stop the release. Their primary function is to contain the release from a safe distance, keep it from spreading, and prevent exposures. May be involved in decontamination. Min. 8-hours training
III Hazardous Material Technician	Their main purpose is rescue or to stop the release. They will approach to stop the release by plugging, patching or shutting down the process. Min. 24-hours training

TABLE II

CRITERIA OF SAFETY MEASURES LEVELS

Safety measures levels of the chemical plant (S)	Criterion
High	The plant fully corresponds to the safety requirements;
Medium	Generally, the plant corresponds to necessary safety conditions;
Low	The plant does not correspond to necessary safety requirements;

B. Output Definition

Parameter *Preventive and protective measure* represents the result of hierarchical analysis at the present level (shown by Figure 5). As it is shown in Table III, for output fuzzy sets were created.

TABLE III

CRITERIA PREVENTIVE AND PROTECTIVE MEASURE LEVEL

Preventive and protective measure level (P)	Criterion of evaluation
H-High	Gas escape: highly trained personnel, modern safety techniques are used.
M-Medium	Medium level of personnel training in force-majeure circumstances, insufficient safety measures
L-Low	Weak personnel training in case of chemical escape, no safety measures implemented.

C. Fuzzification

In the fuzzification process, the membership functions defined on the input, and output variables are applied to actual values to determine the degree of truth for each rule. Gaussian type of membership function has been used in the present paper. Fig. 6 presents the fuzzy sets and its membership function for one of input variables (*chemical plant safety measures*) used in this part of fuzzy risk assessment.

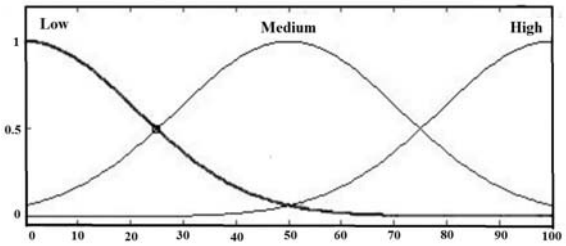

Fig. 6. Graphs of membership functions of chemical plant safety measures levels

D. Setting Up the Rules

Fuzzy logic is a decisional system based on linguistic rules [14]. Therefore, next step is to connect input and output fuzzy sets by rules. One type of fuzzy inference method is called the Mamdani's inference method or max-min inference method. This method is the most commonly seen fuzzy methodology. It was proposed in 1975 by E. Mamdani [15] as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators.

The mostly used rule format is: "IF x and y THEN z", where x and y is the premise and z is the

consequence. For example: "IF *emergency response training levels* is *III* AND *Safety measures level* is *High* THEN *Preventive and protective measure level* is *High*.

All the rules can be seen in Table IV.

TABLE IV

A SYSTEM OF RULES OF FUZZY INFERENCE

Emergency response training level (E)	Safety measures levels (S)		
	High	Medium	Low
I	M	L	L
II	H	M	L
III	H	H	M

E. Defuzzification

The *defuzzification* process is the conversion of the fuzzy output set to a crisp number. *Defuzzification* process and can be done by using different methods. Some of them are: max, mean-max membership principle, centroid method, medium, weighted average method, center of the largest area method and other methods [16].

In this paper a centroid method has been selected as the most common and widely used method of *defuzzification*. According to the centroid method the crisp value of the output variable is computed by finding the variable value of the center of gravity of the membership function for the fuzzy value. More complete description of this algorithm is available in [17].

The *defuzzification* process results are represented at the top of Fig. 7. For example, if parameter *Emergency response training level* makes 75 units and parameter *Safety measures level* – 63, *then Preventive and protective measure* level will be 56.5 units.

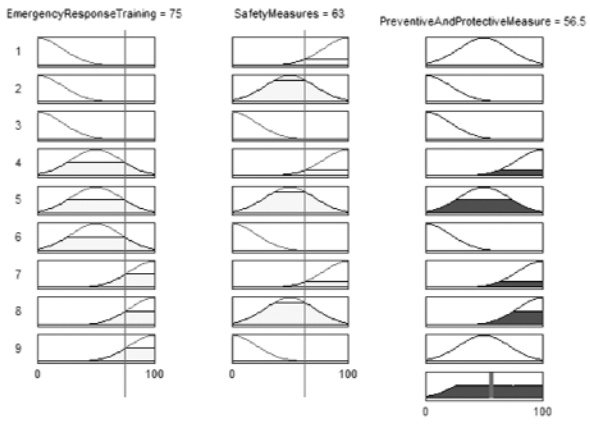

Fig. 7. Application of centroid method to the input variables

V RISK LEVEL

Considering the described before Fuzzy Process at each unit of the system illustrated by Fig. 3, we get a final result, which in the present paper represents the risk level.

The risk level is represented as a surface as it is shown in Fig. 8. as a three-dimensional representation of the risk, consequence of the accident and reliability of the chemical plant.

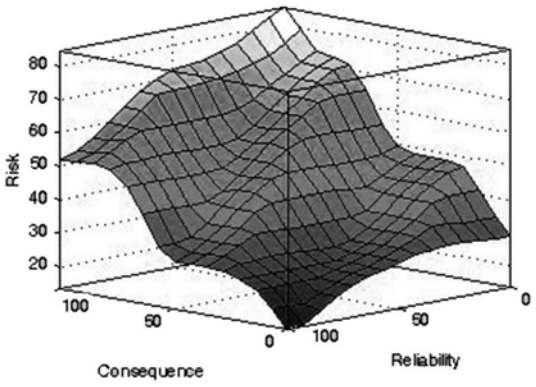

Fig. 8. Three-dimensional representation of the risk, consequence of the accident and reliability of the chemical plant

VI CONCLUSION

The application of hierarchical fuzzy multi-criteria decision making approach to the assessment of risk of environment and people's health it in case of potential accident and escape of gaseous substance allows to significantly facilitate the process of risk assessment when a big amount of source data are available, to represent the results in a simple and understandable way, which is appropriate for determining the needed activities for certain results of analysis.

One of the main advantages of using linguistic variables in hierarchical form is that such expressions are more intuitive that makes it easier for experts to give their evaluations in unclear and complex situations, in which numerical estimations or crisp estimations are hard to obtain. Also using the hierarchical method a number or rules that experts must define is considerably reduced. In this work rules are reduced from 531441 to 126 rules.

This methodology can be used not only by public authorities but also by plant managers, since it is a method that allows evaluating the risk level of the site and enables to understand whether the safety measurements are suitable. This method can be used as a preliminary risk assessment tool to expose situations where more complete analysis is needed. Also this methodology can be used in the case when a thoughtful decision of a risk level reduction is needed.

VII REFERENCES

[1] Carlon, C., Critto, A., Marcomini, A., Nathanail, P. "Risk based characterization of contaminated industrial site using multivariate and geostatistical tools", Environmental Pollution, vol. 111 (3), 2001, pp. 417-427.

[2] Burningham, K., Thrush, D. "Pollution concerns in context: A comparison of local perceptions of the risks associated with living close to a road and a chemical factory", Journal of Risk Research, vol. 7 (2), 2004, pp. 213-232.

[3] Martin B., Pearson A., Bauer B. "An Ecological Risk Assessment of Wind Energy", The Nature Conservancy Helena, Montana, pp.146, 2009
[4] L. Zadeh, "Fuzzy Sets" Information Control, 1965, 8, pp. 338 – 353.
[5] M.R. Akbarzadeh-T, M. Moshtagh-Khorasani "A hierarchical fuzzy rule-based approach to aphasia diagnosis.", Journal of Biomedical Informatics, Vol. 40, No. 5, 2007, pp. 465–475
[6] M. Ganesh, Introduction to Fuzzy Sets and Fuzzy Logic. PHI Learning Pvt. Ltd, 2006, 256 p.
[7] G.J. Klir, Y. Bo, Fuzzy Sets and Fuzzy Logic. Prentice Hall, 1995, 592 p.
[8] T.E. McKone, A.W. Deshpande,, 2005. Environ. Sci. Technol. 39 (2) 42A.
[9] L.A. Zadeh, 1983. The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems, Fuzzy Sets and Systems 11, 199.
[10] L. Zadeh, "Fuzzy logic and approximate reasoning" Synthese, Vol 30, 1975, pp. 407 – 428.
[11] Bellman, R., *Adaptive Control Processes*, Princeton University Press, Princeton, 1966.
[12] Cordón O., Herrera, F., Peregrín A., "Looking for the Best Defuzzification Method Features for each Implication Operator to Design Accurate Fuzzy Models" ,Technical Report, DECSAI, 1999.
[13] Garibaldi, J. M. and Ifeachor, E.C., "Application of simulated Annealing Fuzzy Model Tuning to UmbilicalCord Acid-base Interpretation", IEEE Transactions on Fuzzy Systems , Vol.7, No.1. 1999.
[14] R. E. Bellman, L. A. Zadeh, "Decision making in a fuzzy environment, „Management Science, vol. 17, no. 4, 1970, pp. B141-B164
[15] Mamdani, E.H. and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," International Journal of Man-Machine Studies, Vol. 7, No. 1, 1975, pp. 1-13.
[16] Klir, G. J., and Yuan, B., *Fuzzy Sets, Fuzzy Logic and Fuzzy System*, World Scientific Singapore, 1996.
[17] Ross, T. J., "Properties of membership functions, fuzzification, and defuzzification" in Fuzzy Logic with Engineering Applications, Third Edition, John Wiley & Sons, Ltd, Chichester, United Kingdom, 2010, pp. 89–116.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 170-174

Application Of The Interlaced Sweep Method For The Solution Of Problems In Field Theory

Aloizs Ratnieks, Marina Uhanova
Riga Technical University, Faculty of Computer science and Information technology.
Address: ¼ Meza str., Riga, LV-1048, Latvia

Abstract. **For solution of problems in field theory the method of sweep is very popular. The authors suggest a very effective method of interlaced sweep. The essence of the interlaced sweep method lies in the fact that matrix of the linear algebraic equations system is broken into parts and the solution of separate blocks is sought by direct methods. Usually for each line of the grid a separate block is created. The system of block equations has a tridiagonal matrix where only elements of the main diagonal and two neighboring diagonals are different from zero. The system of equations with such a matrix is easily solved by the method of elimination of unknowns.**

By solving the problems by the finite element method, the nodes of touching of neighboring elements can be placed on curved lines, and the sweep on these lines can be performed observing the principle of interlaced sweep. By following this method, the neighboring lines should not belong to the same half-step.

Keywords: **sweep method, differential equation, system of equations.**

I INTRODUCTION

The solution of wide range of practically important problems can be reduced to calculation of various fields which are described by differential equations in partial derivatives [1]. By application of the finite difference method or of the finite element method, the solution of differential equation in partial derivatives is reduced to solution of system of equations. The matrix of this system usually has very high order where only small proportion of coefficients differs from zero [2].

The solution of this system by the method of elimination of unknown according to Gauss scheme is inefficient because it leads to increase of number of unequal to zero coefficients in the matrix.

For solution of difference equations problems in field theory the method of sweep is very popular. There are many different modifications of this method. The speed of convergence of different sweep method modifications is influenced by relative value of the matrix's diagonal elements relating to the previous step, to alternation of sweep directions, to type of displacement used, to periodic change of displacement for reverse and other factors [3]. The authors suggest a very effective method of interlaced sweep which is neither a method of simultaneous displacements, nor a method of successive displacements along the lines.

II DESCRIPTION OF THE INTERLACED SWEEP METHOD

The essence of the sweep method lies in the fact that matrix of the linear algebraic equations system received as a result of the finite difference approximation of differential equation in partial derivatives, is broken into parts and the solution of separate blocks is sought by direct methods. Moreover, the equations included in the block are met exactly. Usually for each line of the grid a separate block is created. The equations system of block has a tridiagonal matrix where only elements of the main diagonal and two neighboring diagonals are different from zero. All other elements of this matrix are equal to zero. The system of equations with such a matrix is usually solved by the method of elimination of unknowns.

In the case of interlace sweep application in two-dimensional area, each iteration consists of two half steps which alternate:

- sweep on odd gridlines;
- sweep on even gridlines.

Let us denote:

k – number of grid line;

n – number of grid node on the line;

U_{kn} – component of the required vector of the grid node with indices k and n;

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.190

PX_{kn} – coefficient of the matrix corresponding to the grid node with indices k and n of the line directed along X - axis;

PY_{kn} – coefficient of the matrix corresponding to the grid node with indices k and n of the line directed along Y - axis;

F_{kn} – right-hand member of the equation which corresponds to the grid node with indices k and n;

j – serial number of iteration;

ni – current iteration number;

kr – relaxation coefficient;

h – step of change of grid nodes coordinates.

When solving problems for one-dimensional field by the sweep method, all diagonal elements of the matrix correspond entirely to the current step, i.e. relaxation coefficient is equal to 1.

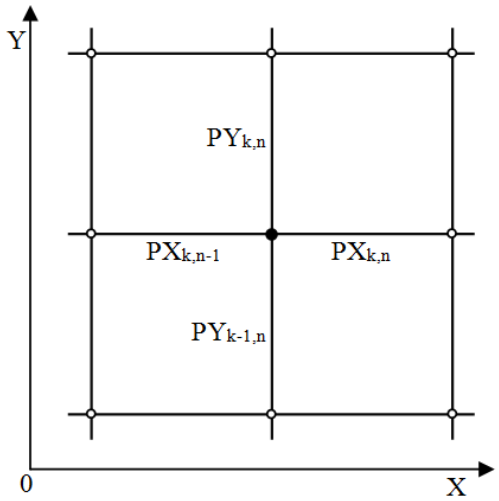

Fig. 1 Designations of grid elements.

To compensate partially the errors resulting from the transfer of some matrix coefficients of the initial equation system to the previous half-step and the increase of the convergence speed in solving the two-dimensional and three-dimensional problems, the recurrence formulas of iterative process can be changed so that the components of the required vector are determined by a factor of relaxation directly inside the block. In other words, each component of the required vector would be calculated based on the assumption that all other components of the vector have been calculated with the same coefficient of relaxation.

To achieve this goal, the recurrence formula of the half-step of the iteration process for n - the grid node depicted in Fig. 1 and sweep along the grid line directed along the x-axis must be written in the form (1).

$$PX_{k,n-1}\cdot U_{k,n-1}^{j+1/2} - DX_{kn}\cdot U_{kn}^{j+1/2} + PX_{kn}\cdot U_{k,n+1}^{j+1/2} =$$
$$= F_{kn} - PY_{k-1,n}\cdot U_{k-1,n}^{j} - PY_{kn}\cdot U_{k+1,n}^{j} +$$
$$+\left(1-\frac{1}{kr}\right)\cdot\left(PX_{k,n-1}+PX_{kn}\right)\cdot U_{kn}^{j-1/2}$$

where:

$$DX_{kn} = \left(PX_{k,n-1}+PX_{kn}\right)/kr + PY_{k-1,n} + PY_{kn} \qquad (1)$$

From the above recurrence formula of the iterative process follows that the coefficients of the main diagonal of the matrix in one-dimensional area are substituted by corresponding coefficients of the main diagonal of the matrix for initial problem for two or three dimensional area. Therefore, only that part of diagonal elements should be divided be the coefficient of relaxation which corresponds to the diagonal elements of the matrix of one-dimensional area. In the case of a sweep along the grid line which is directed along the y-axis, the coefficients of the main diagonal of the matrix of one dimensional area are equal to the sum of coefficients $PY_{k-1,n}$ and $PY_{k,n}$ with the opposite sign.

One part of the diagonal element of the matrix of one-dimensional area, obtained as a result of division by the coefficient of relaxation refers to the current half-step, but the remaining part refers to previous iteration. If the coefficient of relaxation is equal to one then all diagonal elements of the matrix of the initial problem relate entirely to the current half-step.

III DESCRIPTION OF CONTROL EXAMPLES

To evaluate the effectiveness of interlaced sweep method in solving two-dimensional problems the following equation (2) was being solved in the square area with dimensions 32×32, containing 900 internal grid nodes.

$$\frac{\partial}{\partial x}\left(\sigma\frac{\partial U}{\partial x}\right)+\frac{\partial}{\partial y}\left(\sigma\frac{\partial U}{\partial y}\right)=f \qquad (2)$$

This equation has exact analytical solution (3):

$$U(x,y)=x^3y^3 \qquad (3)$$

by

$$\sigma=\frac{1}{xy}, \qquad f=3\left(x^2+y^2\right) \qquad (4)$$

The potentials in the boundary nodes of the grid must be calculated by the function $U(x,y)=x^3y^3$. To solve this equation the following program was written:

```
#include "stdafx.h"
#include<stdio.h>
#include <conio.h>
#include<math.h>

double u[32][32], px[32][32], py[32][32];
double d[32], b[32], f[32][32];
double x, y, h, maxr, c, kr, km;
int i, j, m, k, n, ni;

void iterx(void);
int _tmain(int argc, _TCHAR* argv[])
{
  kr = 1.74;
  km =  1.0 - 1.0 / kr;

  // calculation of the parameters
  // of the grid
  h = 0.1;
  y = 1.0;
  for(i = 0; i < 32 ; i++ ){
    x = 1.0;
    for(j = 0; j < 32; j++ ){
      u[i][j] = x * x * x * y * y * y;
      px[i][j] = 1.0 / (y * (x + h / 2.0));
      py[i][j] = 1.0 / (x * (y + h / 2.0));
      f[i][j] = 3 * h * h * (x * x + y * y);
      x = x + h;
    }
    y = y + h;
  };

  // assign 0.0 to the potentials of
  // all internal nodes of the grid
  for(i = 1; i < 31; i++)
    for(j = 1; j < 31; j++)
       u[i][j] = 0.0;

  // solution of the problem
  ni = 0;
  do {
    maxr = 0;
    // sweep on odd gridlines
    for(k = 1; k < 31; k += 2) iterx();
    // sweep on even gridlines
    for(k = 2; k < 31; k += 2) iterx();
    ni++;
    // output of the results
    printf("\n%+15.5e", maxr);
    // _getch();
  }while(maxr >  0.005);
  printf("\n%+6.3f%+8d", kr, ni);
  printf("\n%+16.5e%+16.5e%+16.5e",
    u[10][10], u[15][15], u[20][20]);
  _getch();
  return 0;
}

// sweep function along X axis
void iterx(void){
// calculation of diagonal coefficients
// and free members
  for(n = 1; n < 31; n++){
    d[n] = - (px[k][n-1] + px[k][n]) / kr -
              py[k-1][n] - py[k][n];
    b[n] = f[k][n] - py[k-1][n]*u[k-1][n] -
      py[k][n] * u[k+1][n] +
      km * (px[k][n-1] + px[k][n])*u[k][n];
    }

  // correction of free members at the
  // beginning and at the end of the line
  b[1] = b[1] - px[k][0] * u[k][0];
  b[30] = b[30] - px[k][30] * u[k][31];
  // direct sweep
  for(n = 2; n < 31; n ++){
    d[n] = d[n]-px[k][n-1]*px[k][n-1]/d[n-1];
    b[n] = b[n]-b[n-1]*px[k][n-1]/d[n-1];
  }

  // reverse sweep
  c = b[30] / d[30];
  if(fabs(u[k][30] - c) > fabs(maxr))
      maxr = fabs(u[k][30] - c);
  u[k][30] = c;
  for(n = 29; n > 0; n --){
    b[n] = b[n] - px[k][n] * u[k][n + 1];
    c = b[n] / d[n];
    if(fabs(u[k][n] - c) > fabs(maxr))
      maxr = fabs(u[k][n] - c);
    u[k][n] = c;
  }
}
```

By solving the problem using the Microsoft Visual Studio Express Edition 2012, the absolute value of maximum correction to the solution has decreased in 10^6 times in 47 iterations by kr=1.74. The total error of the finite approximation and other factors in the middle point of the area with coordinates (2.5, 2.5) is equal to 0.016%.

To evaluate the effectiveness of the interlaced sweep method in solving three-dimensional problems the following equation (5) was being solved in the area having the form of the cube with dimensions 32×32×32, containing 27000 internal grid nodes.

$$\frac{\partial}{\partial x}\left(\sigma\frac{\partial U}{\partial x}\right)+\frac{\partial}{\partial y}\left(\sigma\frac{\partial U}{\partial y}\right)+\frac{\partial}{\partial z}\left(\sigma\frac{\partial U}{\partial z}\right)=f \qquad (5)$$

This equation has an exact analytical solution (6):

$$U(x,y,z)=x^3y^3z^3 \qquad (6)$$

at

$$\sigma = \frac{1}{xyz}, \qquad f = 3\left(x^2y^2 + x^2z^2 + y^2z^2\right) \qquad (7)$$

The potentials in the boundary nodes of the grid must be calculated by the function $U(x,y,z) = x^3y^3z^3$.

To solve the problems for three-dimensional area the principle of interlaced sweep must be observed in all planes which have sweep lines.

When solving the problems the absolute value of maximum correction for solution is reduced in 10^6 times in 47 iterations by kr=7.9. The total error of the finite approximation and other factors in the midpoint of the field with coordinates (2.5, 2.5, 2.5) were equal to 0.017%.

Relatively large value of the coefficient of relaxation is explained by the fact that in three-dimensional case just four coefficients of the initial system matrix relate to the previous half-step and only small part of the coefficient which belongs to the main matrix diagonal in the initial system is divided by the coefficient of relaxation.

To solve the problems in two and three dimensional fields, the authors decided to modify the sweep method for sequential displacement along the grid lines by means of substitution of recurrence formulas of the iteration process for the sweep method by replacement formulas for two and three dimensional area where only one part of the diagonal element corresponding to the matrix of the one-dimensional field is divided by the coefficient of relaxation. This modification of the sweep method has increased the speed of convergence of iteration process in two and three dimensional fields in several times.

The execution of iterations by solving the control example by all above mentioned methods was interrupted when the maximum module of correction to the solution decreased in 10^6 times.

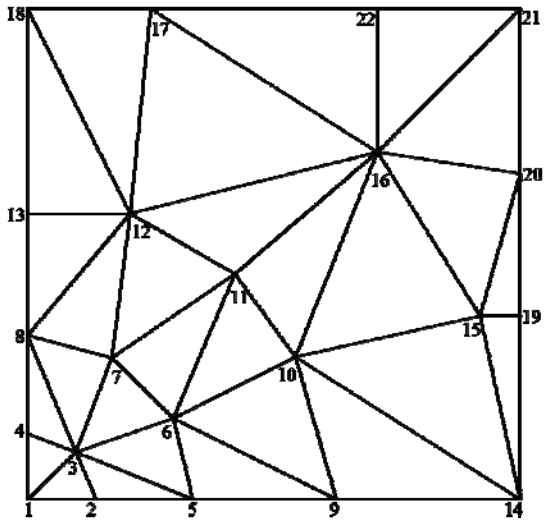

Fig. 2 Triangular grid.

The results of the comparison of the interlaced sweep method with such methods as: the sweep method for successive displacement along the grid lines, the modified sweep method, and the method of successive over-relaxation for two-dimensional area are shown in the table 1.

Similar results for the three-dimensional area are shown in the table 2.

TABLE 1

THE RESULTS OF SOLVING THE CONTROL EXAMPLE IN TWO-DIMENSIONAL AREA

Name of the method	Coefficient of relaxation	The number of iteration	error
Successive over-relaxation	1.77	98	0.28%
Sweep for successive displacement along the lines	1.0	368	0.06%
Modified sweep method	1.76	54	0.01%
Interlaced sweep	1.74	47	0.016%

TABLE 2

THE RESULTS OF SOLVING THE CONTROL EXAMPLE IN THREE-DIMENSIONAL AREA

Name of the method	Coefficient of relaxation	The number of iteration	error
Successive over-relaxation	1.81	95	0.006%
Sweep for successive displacement along the lines	1.0	416	0.05%
Modified sweep method	7.9	65	0.012%
Interlaced sweep	7.9	47	0.017%

IV CONCLUSION

The number of operations to be performed at each point of the grid when solving the problems by the methods listed in the table differs slightly, and the speeds of convergence differ in several times. The modification of the sweep method for sequential displacement along the grid lines has increased the speed of its convergence in 6 times, but the method of interlaced sweep turned out to be more effective. This proves the high efficiency of the interlaced sweep method.

By solving the problems by the finite element method, the nodes of touching of neighboring elements (Fig. 2) can be placed on curved lines, and the sweep on these lines can be performed observing the principle of interlaced sweep. By following this method, the neighboring lines should not belong to the same half-step.

V REFERENCES

[1] O. Zienkiewicz, R. L Taylor and D. D. Fox, *The Finite Element Method for Solid and Structural Mechanics.* ELSEVIER, 2011.

[2] William F. Ames, *Numerical Methods For Partial Differential Equations.* Academic Press Inc, 1992.

[3] В. Вазов, Дж. Форсайт. *Разностные методы решения дифференциальных уравнений в частных производных.* ИЛ, М., 1963.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 175-179

Job Shop Production Planning under Uncertainty: A Monte Carlo Rollout Approach

Axel Simroth[1], Denise Holfeld[1], Renè Brunsch[2]
[1]Fraunhofer Institute for Transportation and Infrastructure Systems IVI
[2]apromace data systems GmbH

Abstract. **The Monte Carlo Rollout method (MCR) is a novel approach to solve combinatorial optimization problems with uncertainties approximatively. It combines ideas from Rollout algorithms for combinatorial optimization and the Monte Carlo Tree Search in game theory. In this paper the results of an investigation of applying the MCR to a Scheduling Problem are shown. The quality of the MCR method depends on the model parameters, search depth and search width, which are strong linked to process parameters. These dependencies are analyzed by different simulations. The paper also deals with the question whether the Lookahead Pathology occurs.**

Keywords: **optimization under uncertainty, robust optimization, Monte Carlo Rollout method, combinatorial problems, Scheduling Problem.**

I INTRODUCTION

Many challenges have to be solved, when applying optimization methods to industrial problems. Possible inaccuracies in information are one of the major challenges. In general, information is not completely known a priori, like it is assumed in mathematical optimization theory. Often information changes over time or can only be estimated. Even solutions must be provided in real time. Both aspects are relevant for many industrial applications like the Job Scheduling Problems.

The Monte Carlo Rollout method (MCR) has been used successfully in various industrial applications to solve combinatorial problems approximately. Here, the problem is considered, to assign different jobs to different machines as well as the order of the production. Machine breakdowns can occur, which lead to an interruption of the production ([1]). There are several possibilities for the distribution of jobs to the machines. Which distribution should be chosen to minimize the delay? The knowledge on uncertainties should be included into the planning. Therefore a stochastic model is necessary to model uncertainties, with the aim to assign probabilities to different possible future developments.

II PROPOSED APPROACH

In order to make decisions in real time, a heuristic H is used to determine a solution based on current information in many cases. The idea is to improve the solution by a Monte Carlo Rollout (MCR) approach as meta-heuristic. In this way, the heuristic is combined with a stochastic model for simulating future developments. This leads to better and more robust solutions for optimization problems under uncertainties. The MCR approach combines ideas from Rollout algorithms for combinatorial optimization and the Monte Carlo Tree Search in game theory. Therefore, we want to introduce all three methods.

Rollout algorithms ([2], [3]) can be used for optimization problems that have a sequential structure, i.e. which can be solved by making a sequence of consecutive decision steps with a limited number of alternative decisions in each step. With an infinite number of alternatives a finite selection must be made or an abort criterion must be selected. These optimization problems can be represented as a search tree. Each node is regarded as a decision point and each leave of the tree represents a solution. By means of the Rollout method, each alternative decision is evaluated in order to choose the best alternative. The Rollout algorithm iteratively explores all different alternatives in the current decision step. It uses a so-called base heuristic for making decisions in the steps following the current decision. The base heuristic usually is a fast, rather simple but solid heuristic for the problem at hand, that solves the problem in a sequential manner. With the help of this base heuristic the Rollout algorithm gets an evaluation of the alternative at a leaf of the decision tree, namely at that leaf that would be reached if the base heuristic would be applied after choosing the alternative considered.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.617

After evaluating all alternatives in the current step, the one that leads to the best results is chosen. In this way only a small part of the tree is examined and the heuristic guides into the most promising part of the tree. If the problem contains *N* decisions, after *(N-1)* applications of the Rollout algorithm a path from root to leaf will be obtained. This path represents the approximated solution of the problem. In Fig.1 the process of a Rollout evaluation for *n* alternatives is shown schematically. For sequentially consistent heuristics, it can be shown, that the Rollout approach leads to no worse results than the pure application of sequentially consistent heuristic.

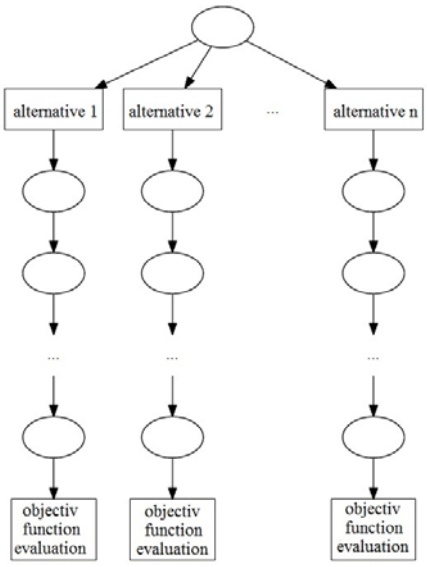

Fig. 1. Schematic representation of a Rollout algorithm.

Monte Carlo Tree Search (MCTS) ([4], [5], [6]) is used for problems where no good heuristic was found to evaluate a decision. Instead of using a noisy and possible misleading heuristic evaluation, the alternative decisions are evaluated by means of random games. Often there is game-specific information about the quality of moves which can be used to weight the possible moves. The MCTS algorithm includes four steps which are shown in Fig.2. The steps are Selection, Expansion, Simulation and Back-propagation and they are repeated several times to generate the MCTS tree.

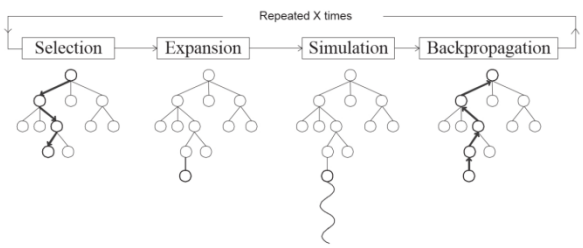

Fig. 2. Schematic representation of a MCTS algorithm.

In the first step a node which is not been included into the MCTS tree yet is selected. This selected node is added to the MCTS tree in step 2, which is the starting point of step 3. To avoid an inexact heuristic evaluation the evaluation of a node is carried out by simulation, also called play-outs. Therefore, all further decisions are taken at random. Such random game ends in a leaf which can be rated clearly. In the last step the MCTS tree is updated with the information from the simulation, e.g. by using the Mimimax algorithm, like [7]. These 4 steps are repeated up to a maximum. The maximum can be the number of nodes in the MCTS tree or the runtime of the algorithm.

The Monte Carlo Rollout method combines both approaches to handle sequential optimization problems that are afflicted with uncertainties. The evaluation of an alternative decision by solving the problem further with a simple and fast base heuristic is adapted from the Rollout approach. The uncertainties are covered through the random selection of future situations, by means of a random player as in the Monte Carlo Tree Search. So, the optimization problem with uncertainties is modeled as a two-player game. The first player is the decision maker, which decides on the base of a simple heuristic. The second player is the random player, which creates new future situations by random. The game where both players move consecutively is called MCR. With a set of different MCRs, an alternative is tested by evaluating a set of random future scenarios. So, the long-term behavior and robustness against uncertainties of this alternative could be analyzed. The MCR method is shown schematically in Fig.3.

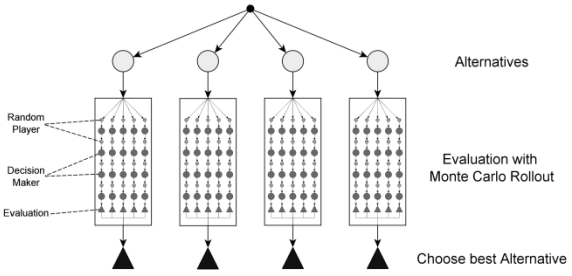

Fig. 3. Schematic representation of a MCR algorithm.

III APPLICATION

Considering the application of the Monte Carlo Rollout method to a Scheduling Problem, more precisely to a Stochastic Dynamic $(m \times l)$ Job Shop Problem with due date. Here, several plans are available, each with a static heuristic evaluation, which results from an objective function. Over time, however, machine breakdowns occur, which lead to delays. The MCR method is used for decision support to choose a robust plan against random future failures.

Considering jobs $J_1, \dots, J_n$ with given processing duration $d(J_i) \geq 0$. Each job J_i consists of sub jobs $J_{ij}, j = 1, \dots, m$ with $0 \leq d(J_{ij}) \leq d(J_i)$ and $\sum_{j=1}^{m} d(J_{ij}) = d(J_i)$. In the $(m \times l)$ Job Shop Problem

n jobs on m different types of machines are to be processed. Each job consists of m sub jobs with true positive processing durations. Each sub job is processed on one machine type. Each of the m types of machines has capacity l. Each job J_i has a fixed predetermined due date $F(J_i)$. This is transferred to its sub jobs $F(J_{ij})$, according to the follow-up relations. The aim of the scheduling is to determine the execution time for each sub-job of all jobs, taking into account capacity constraints. A scheduling plan is feasible, if all jobs are completed by the due date and all capacity limits are observed.

To minimize the total delay is an appropriate objective function. So, every feasible and complete plan is evaluated by this objective function and the plans are comparable. This problem can be seen as a sequential decision process and is represented by a decision tree. The root of the tree is an empty plan. Each edge adds a sub job. Hence, each knot is a sub plan and each leave is a plan. If a job is added to the plan, the execution time and the assignment to a machine for each sub-job have to be determined. Decisions regarding the execution time of a job are selected in order to minimize the objective function, taking into account follow-up conditions and capacity constraints.

Heuristics follow only a single path of the decision tree and result in sub optimal planning. Possible heuristics are:

H_1: Starting each sub job as early as possible and choose the earliest predetermined due date of all unassigned jobs $I \subseteq \{1, \dots, n\}$:

$$J_k \text{ with } k = \arg\min_{i \in I} F(J_i)$$

H_2: Starting each sub job as early as possible and choose the latest possible execution time of all unassigned jobs $I \subseteq \{1, \dots, n\}$:

$$J_k \text{ with } k = \arg\min_{i \in I} F(J_i) - d(J_i)$$

So far, a deterministic problem has been described. Machine breakdowns and the related uncertainties of occurrence of breakdowns lead to a stochastic and dynamic problem. To model the appearance of machine failures a stochastic model is necessary. This should take into account the operating time and the various demands of different jobs to the machine. In practice, a bathtub effect could be observed for the failure rate. Either the failure occurs immediately, due to defects, or later, due to wear. To reproduce this effect the Weibull distribution is used to calculate the probability that a machine breakdown occurs during the execution of a sub job. The Weibull distribution is a continuous probability distribution over the positive real numbers. It is frequently used for the study of lifetimes and failure rates. Two parameters are necessary. Scaling parameters $\lambda > 0$ is chosen according to the characteristic lifetime of a machine. The shape parameters $k > 0$ generates falling failure rates for $0 < k < 1$ and rising failure rates for $k > 1$. In order to model the bathtub effect two Weibull distributions are combined. First with shape parameter $k_1 = 0.5$ to generate falling failure rates, second with $k_2 = 5$ to generate rising failure rates. Figure 4 illustrates the failure rate generated in this way and Figure 5 the resulting cumulative distribution function of the combined Weibull distribution.

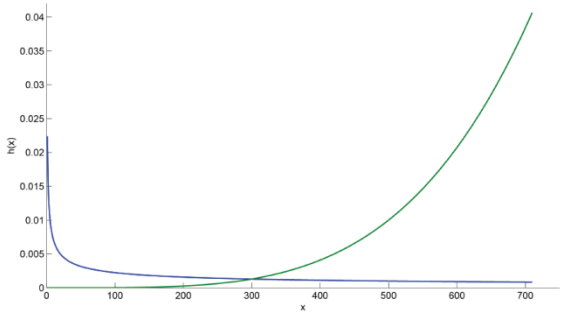

Fig. 4. Failure rates of distributions with $k_1 = 0.5$ and $k_2 = 5$ and same $\lambda = 500$.

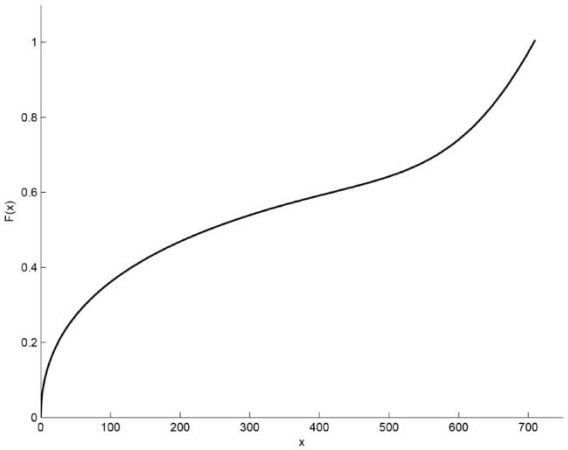

Fig. 5. Cumulative distribution function of the combined Weibull distribution.

Machine breakdowns cause interruptions in the execution of sub jobs for the duration of repairing d_r. An affected sub job can not be terminated on a different machine. Reschedule is necessary to repair not feasible plans. A repair heuristic will serve as base heuristic for the Monte Carlo Rollout algorithm. Sub jobs which have not yet started can be moved between machines to ensure all capacity constraints. Furthermore, the objective function is to be minimized.

Considering the problem as a game leads to a two persons game - the repair heuristic as first player and a random disturbance as second player. Player two starts. Thus, the stochastic dynamic $(m \times l)$ Job Shop Problem has a sequential character. This enables the

use of the MCR method to determine an expected evaluation of a plan in terms of an objective function.

Starting from a plan π_0 successive, random disturbances are generated and after each disturbance a reaction using the repair heuristic follows. In detail, in each step of the random player the next machine breakdown is fixed with help of the stochastic model. The next decision step is based on the repair heuristic, which has a finite number of alternatives. The number of alternatives arises from the possible exchange options. For each alternative MCR simulations (playoffs) with search depth d are performed. So, at least after d steps, or if the end of the planning interval is reached, the playoff terminates and the objective function value is stored. For each alternative, b playoffs are played and the mean of all corresponding objective function values yield the MCR rating of this alternative. In contrast to the evaluation of the initial plan with the objective function, possible machine break downs are considered here, too. The decider chooses the alternative with the best MCR rating.

IV RESULTS AND DISCUSSION

It was considered a (2×2) Job Shop Problem with due date. Both types of machines are available twice. Each job J_i consists of two sub jobs J_{i1} and J_{i2} with $m(J_{i1}) \neq m(J_{i2})$ for all jobs $i = 1, \dots, n$. All generated instances cover n jobs and the conditions/operating hours of all machines. To build feasible plans the two heuristics H_1 and H_2 are used as well as the heuristics using the expected values $\mathbb{E}H_1$ and $\mathbb{E}H_2$. The plan with the best objective function value is used as reference.

Instance parameters have a large impact. More precisely, impacts on the success of the MCR method have: repairing time, maximal operating time of machines, initial conditions of the machines and the due date of the jobs. The first three parameters correlate with the due date. The closer the due dates are to the planned execution end times, the less room for schedule changes. At such time-critical plans almost any disturbance means a delay. The smaller the repair time, the better it is absorbed by the buffer between execution end time and due date. The greater the repair time, the greater is the resulting delay. Also the shape parameter of the Weibull distribution in the underlying stochastic model has a significant impact. As larger the shape parameters as fewer distributions occur. Also, the larger the operating hours of a machine, the more likely are the occurrence of failures.

The results demonstrate the potential of the MCR method. Significant improvements are possible, but related to search depth d and search width b. Figure 6 represent exemplary the mean distance to the best plan (created by H_1, H_2, $\mathbb{E}H_1$ or $\mathbb{E}H_2$) averaged over 1000 instances for search depths $d \leq 3000$ and different search widths, from $b = 1$ to $b = 50$. The dashed line is the mean value of the objective function by using only the heuristic.

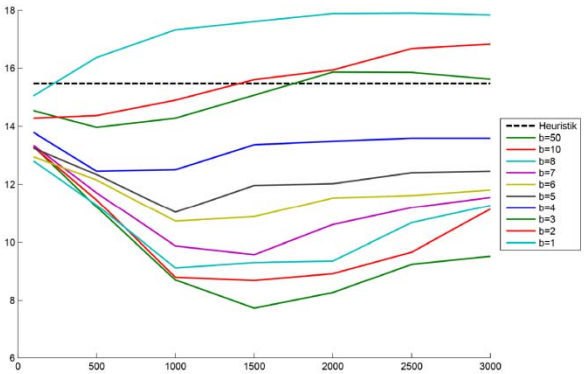

Fig. 6. Mean distance to the best plan averaged over 1000 instances with different d and b ($\lambda = 3000, d_r = 30$).

As expected, an increase of search width b results in better results. The results demonstrate that widths $b < 4$ are not practical. Worse results are obtained with increasing depth and some of them are even poorer than the heuristic results. The impact of search width is not as large as expected. So, an increase of the width is useful, but no more above a certain value b^*. In the selected model, $b^* = 20$ is independent from shape parameter λ, being seen in Figure 7. With increasing width more possible realizations of each plan are considered. Are the initial operating hours of a machine large, it is very likely that the machine failures. Due to the time-sensitive plans, the time of the failure is insignificantly, since any disruption leads to delays. Therefore a few simulations per plan are sufficed.

The search depth d is specified in time units. Only increasing the depth, without adapting the width, has a negative influence on the results, see each of the first lines in Figure 7. The greater the depth, the more gain in information about future events. But at the same time there are a much larger number of possible scenarios. Therefore, the approximation becomes worse with constant width.

The Lookahead Pathology occurs for many different parameter combinations. Here, starting with a search depth $d = 1$ an increasing depth leads first to improved results and later to poorer results, see Fig. 6. This effect is strongly dependent on the shape parameter λ of the Weibull distribution. As larger λ is as fewer disturbances occur. In this model, with a larger λ one disturbance per machine in average occurs. The search depth should not exceed the period of expected first failures. Whether there is a second failure of a machine is random according to the stochastic model. The influence of a breakdown immediately after the repairing is very large. Misinterpretations of the MCR method are likely.

Therefore, the optimal depths are $d^* = 1500$ for $\lambda = 3000$ and $d^* = 2000$ for $\lambda = 4000$. In contrast, with $\lambda = 2000$ several breakdowns per machine on average occur. In this case, a potential second breakdown is detected with increasing depth d. Here, the Lookahead Pathology does not occur.

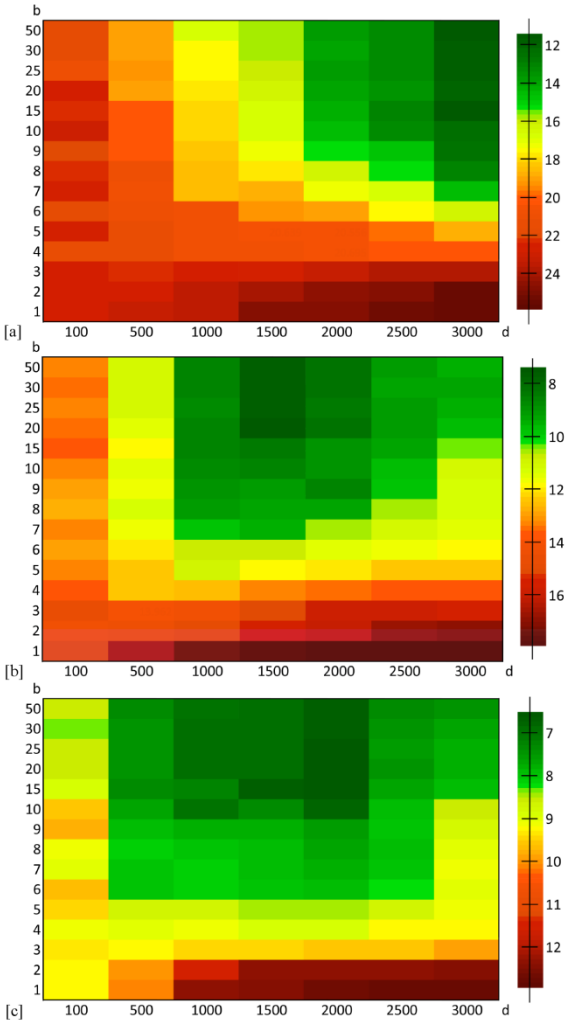

Fig. 7 Mean distance to the best plan averaged over 1000 instances with different d and b ($d_r = 30$). Shape parameter of the Weibull disturbance [a] $\lambda = 2000$, [b] $\lambda = 3000$, [c] $\lambda = 4000$.

V CONCLUSION

We have shown the potential for improvement by use of the Monte Carlo Rollout method by means of a Job Shop Problem. The results of a heuristic based on current information could be improved significantly. By using the MCR method as meta-heuristic the heuristic is combined with a stochastic model for simulating future developments. Better and more robust solutions with respect to future developments can be achieved. It has been shown, that an appropriate choice of search depth and search width is needed. The investigation has shown that increasing the width has a positive influence on the result. However, it may exist a width beyond which, the results does not improve anymore. With an increasing search depth, more information on future developments is included. But to interpret these correctly, the width must be increased, too. The Lookahead Pathology could be detected and should be considered, too.

VI ACKNOWLEDGMENTS

We acknowledge the financial support from the German Federal Ministry of Education and Research (BMBF) under the project PLUSS.

VII REFERENCES

[1] M. Rosner: Untersuchung des Monte-Carlo-Rollout-Verfahrens fuer Optimierungsprobleme unter Ungewissheit, Diplomarbeit TU Dresden, 2013.

[2] D.P. Bertsekas, J.N. Tsitsiklis, C. Wui: Rollout Algorithms for Combinatorial Optimization, Journal of Heuristics, 3, 245–262, 1997.

[3] D.P. Bertsekas, D. A. Castanon: Rollout Algorithms for Stochastic Scheduling Problems, Journal of Heuristics, 5, 89-108, 1999.

[4] G. Chaslot, S. Bakkes, I. Szita, P. Spronck: Monte-Carlo Tree Search: A New Framework for Game AI, Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference, 5, 216-217., 2008.

[5] L. Kocsis, C. Szepesvari: Bandit based Monte-Carlo Planning, ECML-06. Number 4212 in LNCS, Springer, 5, 282-293, 2006.

[6] B. Brügman: Monte Carlo Go,1993

[7] C. Browne, E.J. Powley, D. Whitehouse, S.M. Lucas, P.I. Cowling, P.Rohlfshagen, S. Tavener, D. Perez, S.Samothrakis, S. Colton: A survey of Monte Carlo tree search methods, IEEE Trans. Comput. Intell. AI Games, 4, 1, 2012.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 180-185

NDVI Short-Term Forecasting Using Recurrent Neural Networks

Arthur Stepchenko[1], Jurij Chizhov[2]
[1]Ventspils University College
[2]Riga Technical University

Abstract. **In this paper predictions of the Normalized Difference Vegetation Index (NDVI) data recorded by satellites over Ventspils Municipality in Courland, Latvia are discussed. NDVI is an important variable for vegetation forecasting and management of various problems, such as climate change monitoring, energy usage monitoring, managing the consumption of natural resources, agricultural productivity monitoring, drought monitoring and forest fire detection. Artificial Neural Networks (ANN) are computational models and universal approximators, which are widely used for nonlinear, non-stationary and dynamical process modeling and forecasting. In this paper Elman Recurrent Neural Networks (ERNN) are used to make one-step-ahead prediction of univariate NDVI time series.**

Keywords: **Artificial Neural Networks, Elman Recurrent Neural Networks, Normalized Difference Vegetation Index.**

I INTRODUCTION

Human activities affect ecosystems, including the natural vegetation cover. Vegetation cover change is important factors that affect ecosystem condition and function. A change of vegetation cover may have long term impact on sustainable food production, freshwater and forest resources, the climate and human welfare. Documenting changes occurring in vegetation cover at periodic intervals is very important to providing information about the stability of vegetation.

The use of satellite-based remote sensor data has been widely applied to provide a cost-effective means to develop land cover coverages over large geographic regions. Vegetation cover is an evident part of land cover. Change detection has become a widespread application of remotely sensed data because of repetitive wide coverage, short revisit intervals and good image quality. Change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times. The main prerequisite in using remote sensing data for vegetation change detection is that changes in land cover result in changes in radiance values and changes in radiance due to land cover change are large with respect to radiance change caused by others factors such as differences in atmospheric conditions, differences in soil moisture and differences in sun angles [1].

Vegetation indices calculated from satellite images can be used for monitoring temporal changes associated with vegetation. Vegetation Indices (VIs) are combinations of Digital Numbers (DNs) or surface reflectance at two or more wavelengths designed to take out a particular property of vegetation. Each of the VIs is designed to emphasize a particular vegetation property. Analyzing vegetation using remotely sensed data requires knowledge of the structure and function of vegetation and its reflectance properties. This knowledge enables the linking of vegetative structures and their condition to their reflectance behavior in an ecological system of interest [2]. The normalized difference vegetation index (NDVI) is developed for estimating vegetation cover from the reflective bands of satellite data. The NDVI is an indicator which quantifies the amount of green vegetation. Past studies have demonstrated the potential of using NDVI to study vegetation dynamics. The NDVI data layer is defined as:

$$NDVI = (NIR - R)/(NIR + R), \qquad (1)$$

where *NIR* represents the spectral reflectance in near infrared band and *R* represents red band. Greener and dense vegetation has low red light reflectance and high near infrared reflectance, and thus high NDVI values. The NDVI real values, by definition, would be between -1 and +1, where increasing positive values indicate increasing green vegetation, but low positive values and negative values indicate non-vegetated surface features such as water, barren land, rock, ice, snow, clouds or artificial materials [3]. The NDVI also has the ability to reduce external noise factors such as topographical effects and sun-angle variations.

Time series analysis of remotely sensed data, as shown earlier, has gained special attention supported

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.167

by availability of wide-coverage, high temporal satellite data. NDVI time series data has been employed to predict a NDVI variable beyond the time span. Univariate autoregressive integrated moving average (ARIMA) models are widely used for a univariate time series forecasting, also for the NDVI time series [10]. However, these models are parametric and are based on the assumption that the time series been forecasted are linear and stationary and have a limited ability to capture nonlinearity and nonstationary in time series. The difficulty of forecasting arises from the inherent non-linearity and non-stationarity in the NDVI index. Many previous studies propose that non-linear machine learning approaches such as neural network models perform better than traditional time series linear models with minimum initial assumptions and high forecasting accuracy. Therefore, neural networks are used as an alternative to traditional statistical forecasting methods.

II MATERIALS

A. Study Area

Ventspils Municipality is located in the western part of Courland, Latvia with total area of 2472 km2 (Fig. 1).

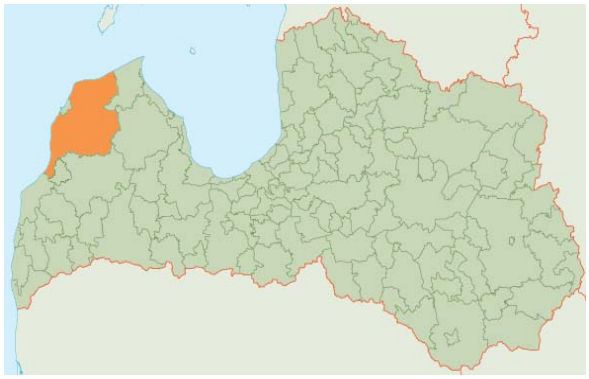

Fig. 1. Location of the Ventspils Municipality.

The climate in the study site is determined by temperate climate zone with significant maritime features. Approximately half of the area of Ventspils Municipality is covered by forests. Latvia lies on the border between two different forest types: the northern coniferous zone and the broad-leaved trees of the temperate zone, so the tree species, characteristic for the both forest types, can be found in the landscape.

B. NDVI Data Set

This study explored the use of multi-temporal, smoothed MODIS Terra and Aqua NDVI composite data with spatial resolution 250 m and produced on 16 day intervals (Fig. 2).

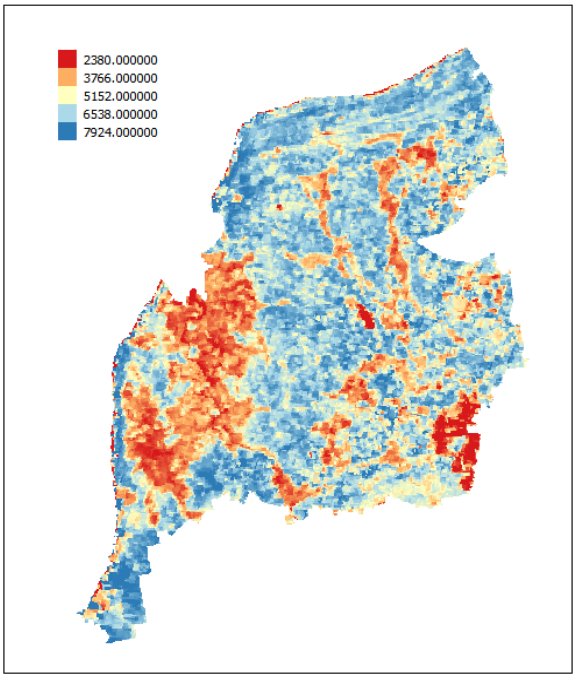

Fig. 2. MODIS Terra NDVI satellite image.

The NDVI data set consists of 624 smoothed NDVI images that obtained every 16 days over 14 years. Mean values of these images are used as NDVI time series observations (Fig. 3).

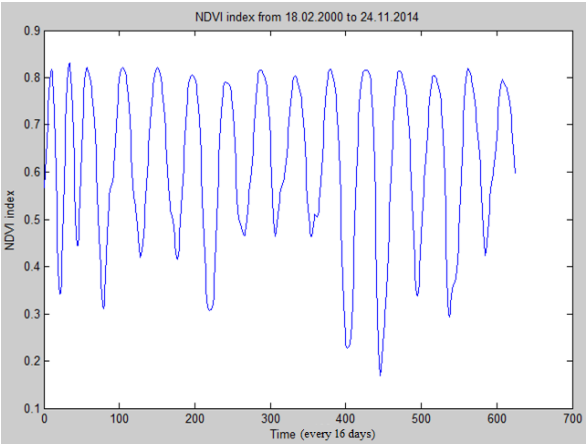

Fig. 3. Smoothed NDVI time series from 18.02.2000 till 24.11.2014.

The NDVI time series data provide a seasonal trajectory – time series show pronounced seasonal oscillations, which correspond to the vegetation phenological cycles where maximum NDVI values are observed between May and August. Variations in the NDVI values are seen to be 0.2 to 0.9 units. NDVI trends are not always monotonic but can change. A positive trend can change for example into a negative one and reversely.

III THEORETICAL BACKGROUND

A. Artificial Neural Networks

Artificial neural networks (ANNs) are a form of artificial intelligence, which attempts to mimic the function of real neurons found in the human brain [4]. ANNs are one of the most accurate and widely used forecasting models that are used in forecasting social, economic, business, engineering, foreign exchange, stock problems and others. Structure of artificial neural networks makes them valuable for a forecasting task with good accuracy.

As opposed to the traditional model-based empirical and statistical methods such as regression and Box-Jenkins approaches, which need prior knowledge about the nature of the relationships between the data, artificial neural networks are self-adaptive methods that learn from data, and only few a priori assumptions about data are needed [5].

Neural networks learn from examples and can find functional relationships among the data even if relationships are unknown or the physical meaning is the baffling [4]. Therefore, ANNs are well suited for problems, whose solutions require knowledge that is difficult to specify but for which there are enough data or observations.

Artificial neural networks can generalize. After learning from the input data (a sample or pattern), ANNs can often correctly process the early unseen sample even if the sample data are noisy. Neural networks are less sensitive to error term assumptions and they can tolerate noise, chaotic components better than most other methods. Artificial neural networks are also universal function approximators. It was proved that a neural network can approximate any continuous function with any accuracy. The ANN performs the following unknown function mapping:

$$y_{t+1} = f\left(y_t, y_{t-1}, ..., y_{t-p}\right), \qquad (2)$$

where y_t is the observation at time t [5].

An individual neuron receives weighted inputs from previous layers, which are summed in each node using a combination function. The result of this combined summation is passed through a transfer function to produce the nodal output of the processing element. The combination function and transfer function together constitute the activation function. The most widely used activation function for the output layer is the linear function as a non-linear activation function may introduce distortion to the predicated output. The sigmoid (logistic), exponential (hyperbolic) tangent, quadratic or linear functions are often used as the hidden layer transfer function. The neural network (2) can approximate any continuous function when the number of hidden nodes q is sufficiently large [6]. In practice, if a network structure has a small number of hidden nodes, then it works well in "out-of-sample" forecasting on data that were not used in training. To improve the accuracy of the neural network, each data point in the input neurons needs to be normalized – rescaled within the range of [−1, 1] or [0, 1] and standardized to scale data and transformed to make the time series stationary. Transformation can be implemented as taking logarithmic returns of the time series, differencing the time series, etc. The multilayer perceptron is trained with error-correction learning, which means that the desired response for the system must be known. The error correction learning works in the following way: from the system response $d_i(n)$ at node i at iteration n, and the desired response $y_i(n)$ for a given input pattern, an instantaneous error $e_i(n)$ is defined by:

$$e_i(n) = d_i(n) - y_i(n). \qquad (3)$$

Using the theory of gradient-descent learning and the standard Back Propagation (BP) algorithm, each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e. the weight from node j to node i (w_{ij}) can be calculated by:

$$w_{ij}(n+1) = w_{ij}(n) + \eta\delta_i(n)x_j(n), \qquad (4)$$

where, x_j is a transform function at node j, i and j indicate different layers. The local error $\delta_i(n)$ can be directly computed from $e_i(n)$ at the output node or can be computed as a weighted sum of errors at the internal nodes. The constant η is called the step size. There are some major disadvantages of this gradient descent approach, one of them is stuck into local minima, which can be mostly avoided by using a learning rate but that sometime may cause serious problem of overshooting, there also another problem of very slow convergence of the learning algorithm which severely depends upon choosing right value for learning rate [8]. Regarding to this issues, there are some more methods available to use in aid of standard back propagation learning, such as Broyden-Fletcher-Goldfarb-Shanno method or BFGS. BFGS method is a classical quasi-Newton second-derivative line search method, and also is one of the most effective and fastest algorithms of the unconstrained optimization problems at present. In quasi–Newton methods, the idea is to use matrices which approximate the Hessian matrix and/or its inverse, instead of exact computing of the Hessian matrix (as in Newton-type methods). The direction of search is based on an $n \times n$ direction matrix S which serves the same purpose as the inverse Hessian H^{-1} in the Newton method. This matrix is generated from available data and is contrived to be an approximation of H^{-1}. Furthermore, as the number of iterations is increased, S becomes progressively a more accurate representation of H^{-1}, and for convex

quadratic objective functions it becomes identical to in *n + 1* iterations. The BFGS method exposes superlinear convergence [9].

B. Elman Recurrent Neural Network

Although feed-forward neural networks are used in many forecasting applications, another type of neural networks – Elman recurrent neural network (ERNN) – is also used in forecasting applications with good accuracy. According to the general principle of the recurrent networks, there is a feedback connection from the outputs of some neurons in the hidden layer to neurons in the context layer that stores the delayed hidden layer outputs (Fig. 4).

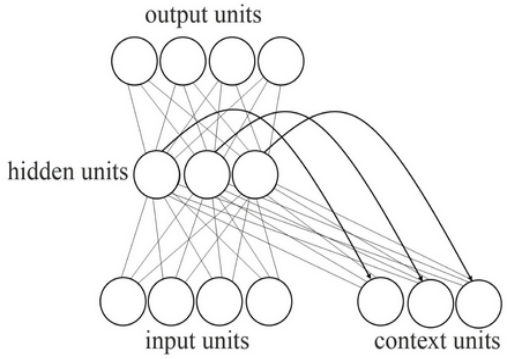

Fig. 4. Elman recurrent neural network architecture.

The most important advantage of ERNN is a robust feature extraction ability, when the context layer stores useful information about data points in past. Since ERNN contains the context layer, it is possible to improve forecasting accuracy by using ERNN instead of feedforward neural network (FNN) [7].

IV EXPERIMENTAL PROCEDURE

The aim of this experiment is to investigate the capability and accuracy of recurrent neural networks in the NDVI time series forecasting. This study implements an Elman recurrent neural network model to predict the NDVI index. In the first stage mean values of every NDVI image were calculated and the NDVI time series was created. The second stage employed ERNN for forecasting.

The data set was divided into two sets, training and testing data set by 80/20 principle, namely, 80% of the NDVI data (from February 18, 2000 to March 12, 2012, a total of 500 observations) were used as a training data set and others of the NDVI data (from March 5, 2012 to November 24, 2014, a total of 124 observations) were used as a testing data set.

There are no fixed rules for the selection of input variables for developing ANN model. ANN model used in this study is the four-layer Elman recurrent network. To achieve optimal weights of ERNN, BFGS quasi-Newton backpropogation algorithm provided by the MATLAB neural network toolbox is used to train the network. In this study, the hyperbolic tangent function and a linear function are used as activation functions for the hidden and output layers, respectively. The number of epochs that are used to train is set to 1000. The training of ERNN was stoped when the error achieves 10^{-6} or when the number of epochs reached 1000. As the number of hidden neurons is concerned, there is currently no theory to determine how many nodes in the hidden layer are optimal. In the present study, the number of hidden nodes was progressively increased from 2 to 20. A program code was written in MATLAB language for the development of the ERNN model.The optimal complexity of ERNN model, that is, the number of hidden nodes, was determined by a trial-and-error approach.

With respect to measure prediction performance of the proposed ERNN model, we introduce loss estimators, namely; RMSE, MAPE and DS. The square root of the mean of the square of all of the errors (RMSE) is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}, \qquad (5)$$

where, $\hat{y}_i$ – forecasted value, y_i – observed value, N – number of observations. The MAPE (Mean Absolute Percent Error) measures the size of the error in percentage terms and is given by:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|100. \qquad (6)$$

Directional symmetry is a statistical measure of a model's performance in predicting the direction of change, positive or negative, of a time series from one time period to the next and is given by:

$$DS = \frac{100}{N-1}\sum_{i=2}^{N} d_i, \qquad (7)$$

where,

$$d_i = \begin{cases} 1, if\,(y_i - y_{i-1})(\hat{y}_i - \hat{y}_{i-1}) \geq 0 \\ 0, else \end{cases}. \qquad (8)$$

These loss estimators are used in chapter *V* to determine proposed ERNN model accuracy.

V RESULTS

In the several experiments were found, that optimal number of hidden nodes is six. Optimal ERNN topology is shown in Fig. 5.

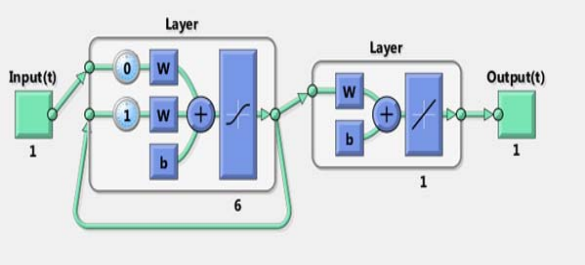

Fig. 5. Optimal ERNN topology.

ERNN network using two neurons was not converging and stopped after 1000 epochs (Fig. 6).

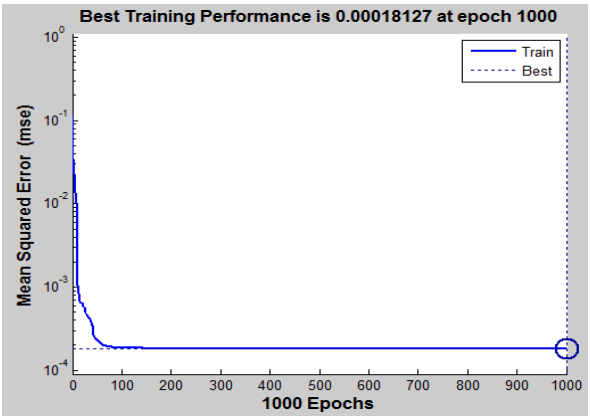

Fig. 6. ERNN convergence using 2 neurons in hidden layer.

ERNN network using three neurons was converging after 433 epochs (Fig. 7). However, this model doesn't give the best result for test data. There is an overfitting effect that can be found in the neural network modeling process. An overfitted model has a good accuracy on training data, but poor accuracy on "out of the sample" data.

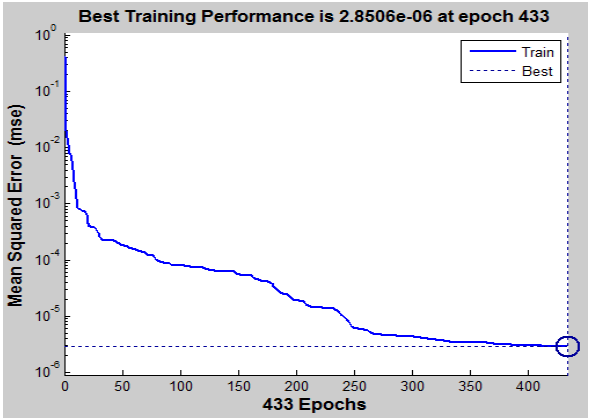

Fig. 7. ERNN convergence using three neurons in hidden layer.

ERNN convergence for best model on test data using six neurons in hidden layer was obtained after 613 epochs (Fig. 8).

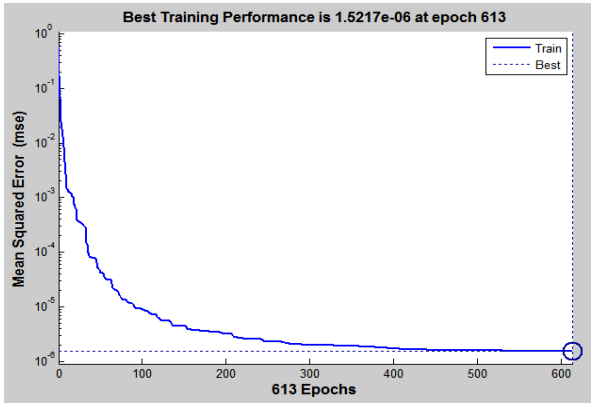

Fig. 8. ERNN convergence using six neurons in hidden layer.

For optimal ERNN architecture, the RMSE value is 0.0374, the MAPE value is 4.9959% and the DS statistic is 90.5882% for training data (table 1).

TABLE 1.

LOSS ESTIMATORS ON TRAINING DATA SET

RMSE	MAPE	DS
0.0374	4.9959%	90.5882%

Actual and predicted values of NDVI time series training data is shown in Fig. 9.

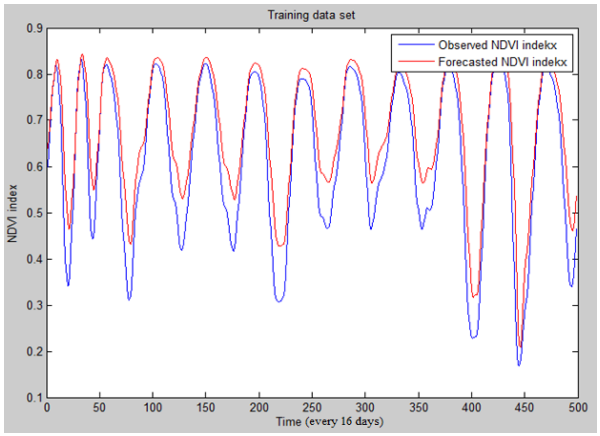

Fig. 9. Actual and predicted values of NDVI time series training data.

The RMSE value is 0.0352, the MAPE value is 1.2883% and the DS statistic is 86.2385% for testing data (table 2).

TABLE 2.

LOSS ESTIMATORS ON TESTING DATA SET.

RMSE	MAPE	DS
0.0352	1.2883%	86.2385%

Actual and predicted values of NDVI time series testing data is shown in Fig. 10.

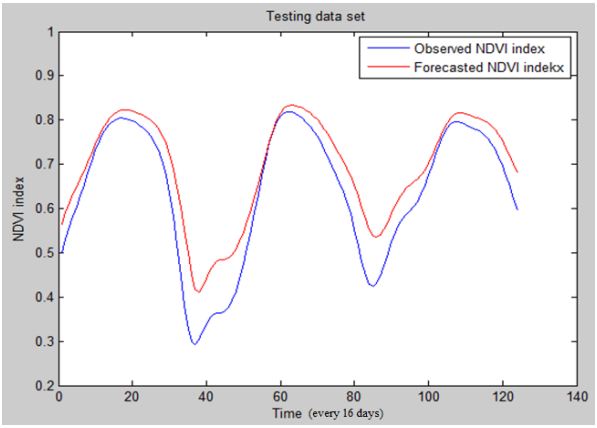

Fig. 10. Actual and predicted values of NDVI time series testing data.

VI SUMMARY AND CONCLUSIONS

In this paper one-step-ahead predictions of the Normalized Difference Vegetation Index (NDVI) data recorded by satellites over Ventspils Municipality in Courland, Latvia is obtained, using Elman recurrent neural network (ERNN). NDVI is an important variable for vegetation forecasting. Artificial Neural Networks (ANN) are computational models and universal approximators, with good generalization ability, that are widely used for nonlinear, non-stationary and dynamical process, such as NDVI time series, modeling and forecasting. Therefore ANN give better results than traditional statistical forecasting methods, such as autoregressive integrated moving average (ARIMA) parametric model, that are based on the assumption that the time series been forecasted are linear and stationary. Since ERNN contains the context layer, it is possible to improve forecasting accuracy by using ERNN instead of feed-forward neural networks. Using optimal ERNN architecture, the RMSE error is 0.0352, the MAPE error is 1.2883% and directional symmetry is 86.2385% on the test data. Therefore the study concludes that the forecasting abilities of ERNN provides a potentially very useful method for NDVI time series forecasting.

VII SUGGESTIONS FOR FUTURE RESEARCH

Some fruitful avenues for future studies are possible. First, in this study was used a univariate NDVI time series, future studies may consider multivariate NDVI time series. Second, in this study was used a simple recurrent neural network, ERNN. In contrast, future research may consider other recurrent neural network models, such as long short term memory neural network. Third, instead of one-step-ahead prediction multistep-ahead prediction can be implemented.

VIII REFERENCES

[1] M. M. Badamasi, S. A. Yelwa, M. A. AbdulRahim and S. S. Noma, "NDVI threshold classification and change detection of vegetation cover at the Falgore Game Reserve in Kano State, Nigeria," Sokoto Journal of the Social Sciences, vol. 2, no. 2, pp. 174-194.

[2] N. B. Duy and T. T. H. Giang, "Study on vegetation indices selection and changing detection thresholds selection in Land cover change detection assessment using change vector analysis," presented at International Environmental Modelling and Software Society (iEMSs), Sixth Biennial Meeting, Leipzig, Germany, 2012.

[3] E. Sahebjalal and K. Dashtekian, "Analysis of land use-land covers changes using normalized difference vegetation index (NDVI) differencing and classification methods," African Journal of Agricultural Research, vol. 8, no. 37, pp. 4614-4622, September 26, 2013.

[4] A. Shabri and R. Samsudin, "Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model," Mathematical Problems in Engineering, vol. 2014, article ID 201402, July 2014.

[5] G. Zhang, B. E. Patuwo and M. Y. Hu, "Forecasting with artificial neural networks: the state of the art," International Journal of Forecasting, vol. 14, no. 1, pp. 35-62, March 1998

[6] M. Khashei and M. Bijari, "An artificial neural network (p, d,q) model for timeseries forecasting," Expert Systems with Applications, vol. 37, no. 1, pp. 479-489, January 2010.

[7] C. H. Aladag, E. Egrioglu and C. Kadilar, "Forecasting nonlinear time series with a hybrid methodology," Applied Mathematics Letters, vol. 22, no. 9, pp. 1467-1470, September 2009.

[8] S. M. Jadhav, S. L. Nalbalwar and A. A. Ghatol, "Artificial Neural Network Models based Cardiac Arrhythmia Disease Diagnosis from ECG Signal Data, "International Journal of Computer, vol. 44, no. 15, April 2012.

[9] A. Ghosh and M. Chakraborty, "Hybrid Optimized Back propagation Learning Algorithm For Multi-layer Perceptron, " International Journal of Computer Applications, vol. 60, nr. 13, December 2012.

[10] M. Manobavan, N. S. Lucas, D. S. Boyd and N. Petford, "Forecasting the interannual trends in terrestrial vegetation dynamics using time series modelling techniques, "Presented at the ForestSAT Symposium Heriot Watt University, Edinburgh, United Kingdom, 5th - 9th August 2002.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 186-192

The analytical solution of the 3D model with Robin's boundary conditions for 2 peat layers

Ērika Teirumnieka[1], Ilmārs Kangro[1], Edmunds Teirumnieks[1], Harijs Kalis[2]
[1] Rezekne Higher Education Institution,
Atbrīvošanas aleja 115, LV – 4601, Rēzekne, Latvia,
Phone: +371 4625150, Fax +371 4625901;
e-mail: Erika.Teirumnieka@ru.lv, ilmars.kangro@ru.lv, Edmunds.Teirumnieks@ru.lv
[2] Institute of Mathematics, Latvian Academy of Sciences and University of Latvia,
Zeļļu 8, Riga LV – 1002, Latvia,
Phone +371 67033721, Fax +371 67820113; e-mail: kalis@lu.lv

Abstract. **In this paper we consider averaging methods for solving the 3-D boundary value problem in domain containing 2 layers of the peat block. We consider the metal concentration in the peat blocks. Using experimental data the mathematical model for calculation of concentration of metal in different points in every peat layer is developed. A specific feature of these problems is that it is necessary to solve the 3-D boundary-value problems for elliptic type partial differential equations of second order with piece-wise diffusion coefficients in every direction and peat layers.**

The special parabolic and exponential spline, which interpolation middle integral values of piece-wise smooth function, are considered. With the help of this splines is reduce the problems of mathematical physics in 3-D with piece-wise coefficients to respect one coordinate to problems for system of equations in 2-D. This procedure allows reduce the 3-D problem to a problem of 2-D and 1-D problems and the solution of the approximated problem is obtained analytically.

The solution of corresponding averaged 2-D initial-boundary value problem is obtained also numerically, using for approach differential equations the discretization in space applying the central differences. The approximation of the 2-D non-stationary problem is based on the implicit finite-difference and alternating direction (ADI) methods. The numerical solution is compared with the analytical solution.

Keywords: **diffusion problem, special splines, analytical and numerical solution.**

I INTRODUCTION

The task of sufficient accuracy numerical simulation of quickly solution 3-D problems for mathematical physics in multilayered media is important in known areas of the applied sciences. To achieve this goal we consider two methods: special finite difference scheme and averaging method by using integral parabolic and exponential splines. For engineering calculation the concentration of metals in peat layered blocks is chosen the averaging method.

The finite-difference method is used only for solving the obtained 1-D and 2-D problems. The layered peats block are modelled in [5], [4].

A. Buikis ([1], [3]) considered different assumptions for averaging methods along the vertical coordinate. These methods were applied for the mathematical simulation of the mass transfer process in multilayered underground systems. It is necessity to solve the 3-D initial-boundary-value problems for parabolic type partial differential equations of second order with piece-wise parameters in multilayer domain. The special spline, which interpolation middle integral values of piece-wise smooth function, is defined. With the help of this splines is reduce the problems of mathematical physics in 3-D with piece-wise coefficients to respect one coordinate analytically to problems for system of equations in 2-D.

The solutions of corresponding 2-D initial-boundary value problem are obtained numerically, using the implicit finite difference approximation and

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.618

alternating method of Douglas and Rockford. The 3-D problem is reduced to 2-D and 1-D problems used integral parabolic and exponential splines.

II MATHERIALS AND METHODS

1. A mathematical model

The process of diffusion the metal in the peat block is consider in 3-D parallelepiped

$$\Omega = \{(x,y,z): 0 \le x \le L_x, 0 \le y \le L_y, 0 \le z \le L_z\}$$

The domain Ω consists of multilayer medium. We will consider the non-stationary 3-D problem of the linear diffusion theory for multilayered piece-wise homogenous materials of 2 layers in the form

$$\Omega_i = \{(x,y,z): x \in (0,L_x),\ y \in (0,L_y),\ z \in (z_{i-1},z_i)\},\ i = 1;2$$

where $H_i = z_i - z_{i-1}$ is the height of layer Ω_i, $z_0 = 0, z_2 = L_z$.

We will find the distribution of concentrations $c_i = c_i(x,y,z,t)$ in every layer Ω_i at the point $(x,y,z) \in \Omega_i$ and at the time t by solving the following 3-D initial-boundary value problem for partial differential equation (PDE):

$$\frac{\partial c_i(x,y,z,t)}{\partial t} = \frac{\partial}{\partial x}\left(D_{ix}\frac{\partial c(x,y,z,t)}{\partial x}\right) + \frac{\partial}{\partial y}\left(D_{iy}\frac{\partial c(x,y,z,t)}{\partial y}\right) + \frac{\partial}{\partial z}\left(D_{iz}\frac{\partial c(x,y,z,t)}{\partial z}\right) + f_i(x,y,z,t)$$

$$x \in (0,L_x),\ y \in (0,L_y),\ z \in (0,L_z),$$

$$t \in (0,t_f),\ i = 1;2$$

$$\partial c_i(0,y,z,t)/\partial x = \partial c_i(x,0,z,t)/\partial y = 0$$

$$D_{1z}\partial c_1(x,y,0,t)/\partial z - \beta_z(c_1(x,y,0,t) - c_{0z}(x,y)) = 0$$

$$D_{ix}\partial c_i(L_x,y,z,t)/\partial x + \alpha_{ix}(c_i(L_x,y,z,t) - c_{iax}(y,z)) = 0,\ i = 1;2$$

$$D_{iy}\partial c_i(x,L_y,z,t)/\partial y + \alpha_{iy}(c(x,L_y,z,t) - c_{iay}(x,z)) = 0,\ i = 1;2$$

$$D_{2z}\partial c_2(x,y,L_z,t)/\partial z + \alpha_z(c_2(x,y,L_z,t) - c_{az}(x,y)) = 0 \qquad (1.1),$$

$$c_1(x,y,z_1,t) = c_2(x,y,z_1,t),$$

$$D_{1z}\partial c_1(x,y,z_1,t)/\partial z = D_{2z}\partial c_2(x,y,z_1,t)/\partial z$$

$$c_i(x,y,z,0) = c_{i0}(x,y,z), i = 1;2,$$

$$x \in (0,L_x),\ y \in (0,L_y),\ z \in (0,L_z),$$

$$t \in (0,t_f),\ i = 1;2$$

where $c_i = c_i(x,y,z,t)$ are the concentrations functions in every layer, $f_i(x,y,z,t)$ - the fixed source function, D_{ix}, D_{iy}, D_{iz} are constant diffusion coefficients, $\alpha_{ix}, \alpha_{iy}, \alpha_z, \beta_z, i = 1;2$ are the constant mass transfer coefficients for the 3rd type boundary conditions, $c_{az}, c_{iay}, c_{iax}, c_{0z}, i = 1;2$ are given concentration values on the boundaries, t_f is the final time, $c_{i0}(x,y,z), i = 1;2$ are given initial concentrations:

where $c_i = c_i(x,y,z,t)$ are the concentrations functions in every layer, $f_i(x,y,z,t)$ - the fixed source function, D_{ix}, D_{iy}, D_{iz} are constant diffusion coefficients, $\alpha_{ix}, \alpha_{iy}, \alpha_z, \beta_z, i = 1;2$ are the constant mass transfer coefficients for the 3rd type boundary conditions, c_{az}, c_{iay}, c_{iax}, $c_{0z}, i = 1;2$ are given concentration values on the boundaries, t_f is the final time, $c_{i0}(x,y,z), i = 1;2$ are given initial concentrations:

1) The homogenous 3rd type conditions by $x = L_x, y = L_y, z = L_z, z = 0$,

2) The symmetrical conditions by $x = 0, y = 0$. The values c_i and the flux functions $D_{iz}\partial c_i / \partial z$ must be continuous on the contact line between the layers $z = z_1$.

2. The averaged method in z-direction with the integral exponential spline

Using averaged method respect to z with exponential or hyperbolic trigonometric functions we have

$$c_i(x,y,z,t) = c_{iz}(x,y,t) + m_{iz}\frac{0.5H_i\sinh(a_i(z-\bar{z}_i))}{\sinh(0.5a_iH_i)} + e_{iz}\left(\frac{H_i^2}{4}\left(\frac{\sinh^2(a_i(z-\bar{z}_i))}{\sinh^2(0.5a_iH_i)} - A_{i0z}\right)\right),$$

where $c_{iz}(x,y,t) = H_i^{-1}\int_{z_{i-1}}^{z_i} ci(x,y,z,t)dz$,

$$A_{i0z} = \frac{\sinh(a_iH_i)/(a_iH_i) - 1}{\cosh(a_iH_i) - 1}, \qquad \bar{z}_i = (z_{i-1} + z_i)/2,$$

$z \in [z_{i-1}, z_i], i = 1;2$.

If parameters $a_i > 0$ are tending to zero, then we obtain the limit as a parabolic spline because of $A_{i0z} \to 1/3$ [1]:

$$c_i(x,y,z,t) = c_{iz}(x,y,t) + m_{iz}(x,y,t)(z-\bar{z}_i) + e_{iz}(x,y,t)((z-\bar{z}_i)^2 - H_i/12), i = 1;2.$$

The unknown functions $m_{iz}(x,y,t), e_{iz}(x,y,t)$ it is possible to find out from the boundary conditions (2.1) in the following form in z-direction:

1) For $z = 0$,

$$d_1(m_{1z} - e_{1z}H_1) - \beta_z(c_{1z} - 0.5m_{1z}H_1 + e_{1z}H_1(0.5A_{11z}H_1 - c_{0z})) = 0,$$

2) For $z = L_z$,

$d_2(m_{2z}+e_{2z}H_2)+$
$\alpha_z(c_{2z}+0.5m_{2z}H_2+e_{2z}H_2(0.5A_{21z}H_2-c_{az}))=0,$
3) For $z=z_1$, $c_{1z}+0.5m_{1z}H_1+e_{1z}H_1 0.5A_{11z}H_1=$
$c_{2z}-0.5m_{2z}H_2+e_{2z}H_1 0.5A_{21z}H_2,$
$d_{1z}(m_{1z}-e_{1z}H_1)=d_{2z}(m_{2z}-e_{2z}H_2),$
where $d_{iz}=0.5D_{iz}a_i\coth(0.5a_iH_i)$,
$A_{i1z}=0.5(1-A_{i0z})\to 1/3$, if $a_i\to 0$, $i=1;2$.
From boundary conditions by $z=z_1$ follows:
$m_{iz}=m_{0i}(c_{2z}-c_{1z})+m_{i1}e_{1z}H_1+m_{i2}e_{2z}H_2,$
$i=1;2$, where $m_{01}=2d_{2z}/(H_2s_1)$,
$m_{11}=-(d_{1z}+A_{11z}H_1d_{2z}/H_2)/s_1,$
$m_{12}=d_{2z}(A_{21z}-1)/s_1$, $s_1=d_{1z}+d_{2z}H_1/H_2,$
$m_{02}=m_{01}d_{1z}/d_2$, $m_{21}=(m_{11}+1)d_{1z}/d_{2z},$
$m_{22}=m_{12}d_{1z}/(d_{2z}+1).$
From boundary conditions by $z=0$, $z=L_z$ follows:
$e_{1z}H_1a_{11}+e_{2z}H_2a_{12}=\beta_z(c_{1z}-c_{0z})-b_1(c_{2z}-c_{1z}),$
$e_{1z}H_1a_{21}+e_{2z}H_2a_{22}=-\alpha_z(c_{2z}-c_{az})-b_2(c_{2z}-c_{1z}),$
where $b_1=m_{01}(0.5H_1\beta_z+d_{1z}),$
$b_2=m_{02}(0.5H_2\alpha_z+d_{2z})$
$a_{11}=d_{1z}(m_{11}-1)+0.5\beta_zH_1(m_{11}-A_{11z}),$
$a_{22}=d_{2z}(m_{22}+1)+0.5\alpha_zH_2(m_{22}-A_{21z}),$
$a_{12}=d_{1z}m_{12}+0.5\beta_zH_1m_{12},$
$a_{21}=d_{2z}m_{21}+0.5\alpha_zH_2m_{21}.$

$$f_{iz}(x,y,t)=H_i^{-1}\int_{z_i-1}^{z_i}f_i(x,y,z,t)dz,$$

$$c_{iax}^v(y)=H_i^{-1}\int_{z_i-1}^{z_i}c_{iax}(y,z)dz,$$

$$c_{iay}^v(x)=H_i^{-1}\int_{z_i-1}^{z_i}c_{iay}(x,z)dz,$$

$$c_{iz,0}^v(x,y)=H_i^{-1}\int_{z_i-1}^{z_i}c_{i0}(x,y,z)dz,\ i=1;2.$$

3. The averaged method in y-direction
Using averaged method with respect to y we have

$$c_{iy}(x,t)=L_y^{-1}\int_0^{L_y}c_{iz}(x,y,t)dy,$$
$$c_{iz}(x,y,t)=c_{iy}(x,t)+$$
$$m_{iy}(x,t)\frac{0.5L_y\sinh(a_i(y-0.5L_y))}{\sinh(0.5a_iL_y)}+$$
$$e_{iy}(x,t)\left(\frac{L_y^2}{4}\left(\frac{\sinh^2(a_i(y-0.5L_y)}{\sinh^2(0.5a_iL_y)}-A_{i0y}\right)\right),$$

With the unknown functions $m_{iy}(x,t),e_{iy}(x,t)$ we can determine these functions from the boundary conditions (3.1) in following form:

Then $e_{iz}H_i=c_{1z}b_{i1}+c_{2z}b_{i2}+c_{0z}b_{i3}+c_{0z}b_{i4},$
$i=1;2$, where $b_{11}=(a_{22}(\beta_z+b_1)-b_2a_{12})/\det,$
$b_{12}=(a_{12}(\alpha_z+b_2)-b_1a_{22})/\det,$
$b_{21`}=(-a_{21}(\beta_z+b_1)+b_2a_{11})/\det,$
$b_{22`}=(-a_{11}(\alpha_z+b_2)+b_1a_{21})/\det,$
$b_{23}=a_{21}\beta_z/\det$, $b_{24}=a_{11}\alpha_z/\det,$
$\det=a_{11}a_{22}+a_{12}a_{21}.$
Now the initial-boundary value $2D$ problem is in following form

$$\begin{cases}\dfrac{\partial c_{iz}(x,y,t)}{\partial t}=\dfrac{\partial}{\partial x}\left(D_{ix}\dfrac{\partial c_{iz}(x,y,t)}{\partial x}\right)+\\ \dfrac{\partial}{\partial y}\left(D_{iy}\dfrac{\partial c_{iz}(x,y,t)}{\partial y}\right)+f_{iz}(x,y,t)+\\ B_{iz}\begin{pmatrix}c_{1z}(x,y,t)b_{i1}+c_{2z}(x,y,t)b_{i2}+\\ c_{0z}(x,y)b_{i3}+c_{az}(x,y)b_{i4}\end{pmatrix},\\ x\in(0,L_x),\ y\in(0,L_y),\ t\in(0,t_f),\\ \partial c_{iz}(0,y,t)/\partial x=\partial c_{iz}(x,0,t)/\partial y=0,\\ D_{ix}\dfrac{\partial c_{iz}(L_x,y,t)}{\partial x}+\alpha_{ix}\left(c_{iz}(L_x,y,t)-c_{iax}^v(y)\right)=0,\\ D_{iy}\dfrac{\partial c_{iz}(x,L_y,t)}{\partial y}+\alpha_{iy}\left(c_{iz}(x,L_y,t)-c_{iax}^v(x)\right)=0,\\ c_{iz}(x,y,0)=c_{iz,0}(x,y),\end{cases}\quad(2.1)$$

where $B_{iz}=2d_i/H_i$,
$m_{iy}(x,t)=e_{iy}(x,t)L_y=-b_{i5}(c_{iy}(x,t)-c_{iay}^v(x)),$
where $b_{i5}=\alpha_{iy}/(0.5L_y\alpha_{iy}(1+A_{i1y})+2d_{iy}),$
$A_{i1y}=0.5(1-A_{i0y}),$

$$A_{ioy}=\frac{\sinh(a_iL_y)/((a_iL_y)-1)}{\cosh(a_iL_y)-1},$$

$d_{iy}=0.5L_yD_{iy}a_i\coth(0.5a_iL_y).$
Then the initial-boundary value problem (2.1) is in the following form

$$\begin{cases}\dfrac{\partial c_{iy}(x,t)}{\partial t}=\dfrac{\partial}{\partial x}\left(D_{ix}\dfrac{\partial c_{iy}(x,t)}{\partial x}\right)+f_{iy}(x,t)+\\ B_{iy}b_{i5}c_{iy}(x,t)+\\ B_{iz}\left(c_{1y}(x,t)b_{i1}+c_{2y}(x,t)b_{i2}\right)+\\ B_{iz}\left(c_{0z}^v(x)b_{i3}+c_{az}^v(x)b_{i4}\right)+B_{iy}b_{i5}c_{iay}^v(x),\\ x\in(0,L_x),t\in(0,t_f),\ \partial c_{iy}(0,t)/\partial x=0,\\ D_{ix}\partial c_{iy}(L_x,t)/\partial x+\alpha_{ix}\left(c_{iy}(L_x,t)-c_{iax}^{vv}(y)\right)=0,\\ c_{iy}(x,0)=c_{iy,0}(x),\end{cases}\quad(3.1)$$

where $B_{iy}=2D_{iy}/H_i$,

$$f_{iy}(x,t) = L_y^{-1}\int_0^{L_y} f_{iz}(x,y,t)dy,$$

$$c_{az}^{v}(x) = L_y^{-1}\int_0^{L_y} c_{az}(x,y)dy,$$

$$c_{0z}^{v}(x) = L_y^{-1}\int_0^{L_y} c_{0z}(x,y)dy$$

$$c_{iax}^{vv} = L_y^{-1}\int_0^{L_y} c_{iax}^{v}(y)dy,$$

$$c_{iy,0}(x) = L_y^{-1}\int_0^{L_y} c_{iz,0}(x,y)dy,\ i=1;2.$$

4. The averaged method in x-direction

It is possible to make the averaging also with respect to x

$$c_{ix}(t) = L_x^{-1}\int_0^{L_x} c_{iy}(x,t)dx,$$

$$c_{iy}(x,t) = c_{ix}(t) +$$

$$m_{ix}(t)\frac{0.5L_x\sinh(a_i(x-0.5L_x))}{\sinh(0.5a_iL_x)} +$$

$$e_{ix}(t)\left(\frac{L_x^2}{4}\left(\frac{\sinh^2(a_i(x-0.5L_x)}{\sinh^2(0.5a_iL_x)} - A_{i0x}\right)\right)$$

with the unknown functions $m_{ix}(t), e_{ix}(t)$. We can determine these functions from the boundary conditions (4.1) in the following form:

$$m_{ix} = e_{ix}L_x = -b_{i6}\left(c_{ix}(t) - c_{iax}^{vv}\right),$$

where $b_{i6} = \alpha_{ix}/\left(0.5L_x\alpha_{ix}\left(1+A_{i1x}\right)+2d_{ix}\right)$.

$$A_{i1x} = 0.5\left(1-A_{i0x}\right),$$

$$A_{iox} = \frac{\sinh(a_iL_x)/\left((a_iL_x)-1\right)}{\cosh(a_iL_x)-1}.$$

Then the initial-boundary value problem (3.1) is in following form

$$\begin{cases} \dfrac{\partial c_{ix}(t)}{\partial t} = -\left(B_{ix}b_{i6} + B_{iy}b_{i5}\right)c_{ix}(t) + \\ f_{ix}(t) + B_{iz}\left(b_{i1}c_{1x}(t) + b_{i2}c_{2x}(t)\right) + \\ B_{iz}\left(c_{0z}^{vv}b_{i3} + b_{i4}c_{az}^{vv}\right) + \\ B_{iy}b_{i5}c_{iay}^{vv} + B_{ix}b_{i6}c_{iax}^{vv},\ t\in(0,t_f), \\ c_{ix}(0) = c_{ix,0}, \end{cases} \quad (4.1)$$

where $f_{ix}(t) = L_x^{-1}\int_0^{L_x} f_{iy}(x,t)dx$,

$$c_{az}^{vv} = L_x^{-1}\int_0^{L_x} c_{az}^{v}(x)dx,\ c_{ix,0} = L_x^{-1}\int_0^{L_x} c_{iy,0}(x)dx,$$

$$c_{0z}^{vv} = L_x^{-1}\int_0^{L_x} c_{0z}^{v}(x)dx.$$

Therefore we have from (4.1) the initial problem for ODEs of the first order:

$$\begin{cases} \dot{u}_1(t) = c_{11}u_1(t) + c_{12}u_2(t) + e_1(t), \\ \dot{u}_2(t) = c_{21}u_1(t) + c_{22}u_2(t) + e_2(t), \\ u_1(0) = u_{10}, u_2(0) = u_{20}, \end{cases} \quad (4.2)$$

where $u_i(t) = c_{ix}(t)$, $\dot{u}_i = \dfrac{\partial u_i(t)}{\partial t}$,

$u_i(0) = c_{ix,0}, i=1;2$,

$$e_i(t) = B_{iz}\left(c_{0z}^{vv}b_{i3} + b_{i4}c_{az}^{vv}\right) + B_{iy}b_{i5}c_{iay}^{vv} + B_{ix}b_{i6}c_{iax}^{vv} + f_{ix}(t),$$

$$c_{11} = B_{1z}b_{11} - \left(B_{1x}b_{16} + B_{1y}b_{15}\right),$$

$$c_{22} = B_{2z}b_{22} - \left(B_{2x}b_{26} + B_{2y}b_{25}\right),$$

$$c_{21} = B_{2z}b_{21},\ c_{12} = B_{1z}b_{12}.$$

The solutions of (4.2) can be obtained with the help of classical methods [6]. For the averaged stationary solution follows the analytical solution in the form

$c_{1x} = \left(e_2c_{12} - e_1c_{22}\right)/d$, $c_{2x} = \left(e_1c_{21} - e_2c_{11}\right)/d$,

where $d = c_{11}c_{22} - c_{12}c_{21}$.

5. Analytical model for estimating the parameters a_1, a_2

We consider the special 1-D diffusion problem in the z - direction for $f = \alpha_x = \alpha_y = 0$,

$c_{az}(x,y) = C_a\cos(\pi x/L_x)\cos(\pi y/L_y)$,

$c_{0z}(x,y) = C_0\cos(\pi x/L_x)\cos(\pi y/L_y)$. Then the stationary solution of (2.1) is in the form $c_i(x,y,z) = c_i(z)\cos(\pi x/L_{ix})\cos(\pi y/L_{iy})$, where the function $c_i(z)$ is solution for following boundary-value problem:

$$\begin{cases} \dfrac{\partial}{\partial z}\left(\dfrac{\partial c_i(z)}{\partial z}\right) - b_i^2c_i(z) = 0,\ z\in(z_{i-1},z_i), \\ D_{1z}\partial c_1(0)/\partial z - \beta_z\left(c_1(0) - C_0\right) = 0, \\ D_{2z}\partial c_2(L_z)/\partial z + \alpha_z\left(c_2(L_z) - C_a\right) = 0 \\ c_1(H_1) = c_2(H_2), D_{1z}\dfrac{\partial c_1(H_1)}{\partial z} = D_{2z}\dfrac{\partial c_2(H_1)}{\partial z} \end{cases}, \quad (5.1)$$

where $b_i = \pi\sqrt{\left(D_{ix}/L_x^2 + D_{iy}/L_y^2\right)/D_{iz}}, i=1;2$.

Therefore the exact solution is

$c_1(z) = P_1\sinh(b_1z) + P_2\cosh(b_1z)$,

$c_2(z) = P_3\sinh(b_2z) + P_4\cosh(b_2z)$, where the constants P_1, P_2, P_3, P_4 are the functions of $D_{ix}, D_{iy}, D_z, L_x, L_y, L_z, \alpha_z, \beta_z, C_0, C_a$. The averaged values are

$$c_1^v = H_1^{-1}\int_0^{H_1} c_1(z)dz = (H_1)^{-1}\frac{P_1}{b_1}\left(\cosh(b_1H_1) - 1\right)$$

$$+ (H_1)^{-1}\frac{P_2}{b_1}\sinh(b_1H_1),$$

$c_2^y = H_2^{-1}\int_{H_1}^{L_z} c_2(z)dz =$

$(H_2)^{-1}\left(\frac{P_3}{b_2}(\cosh(b_2L_z)-\cosh(b_2H_1))\right)+$

$(H_2)^{-1}\left(\frac{P_4}{b_2}(\sinh(b_2L_z)-\sinh(b_2H_1-1))\right).$

This form of solution remained also for discrete approximation $c(z_j)=C_j$, $z_j=(j-1)h$, $h=\frac{L_z}{N}$, $j=\overline{1,NP}$, $H_1=hN1$ by using exact finite diference scheme (FDS) from N.Bahvalov [2]:

$c_j = P_1\sinh(b_1z_j)+P_2\cosh(b_1z_j)$, $j=\overline{1,N1}$,

$c_j = P_3\sinh(b_2z_j)+P_4\cosh(b_2z_j)$,

$j=\overline{N1,N1+1}$.

1) $C_2 - C_1\left(1+h\frac{\beta_z}{D_{1z}}\right)+hC_0\frac{\beta_z}{D_{1z}}=0$, $z=z_1=0$,

2) $C_N - C_{N+1}\left(1+h\frac{\alpha_z}{D_{2z}}\right)+hC_a\frac{\alpha_z}{D_{2z}}=0$,

$z=z_{N+1}=L_z$,

3) $P_1\sinh(b_1H_1)+P_2\cosh(b_1H_1) = P_3\sinh(b_2H_1)+P_4\cosh(b_1H_1)$, for

$z=H_1$, $j=N1$,

4) $\frac{D_{1z}(C_{N1}-C_{N1-1})}{h}=\frac{D_{2z}(C_{N1+1}-C_{N1})}{h}$,

for $z=H_1$, $j=N1$.

Coefficients P_1,P_2,P_3,P_4 are obtained from:

$P_1=b_6P_2-b_5C_0$, $P_3=b_4P_4-b_3C_a$,

b_3,b_4,b_5,b_6 are the functions g_i dependent on a variety of variables:

$b_3=g_1(h,\alpha_z,D_{2z})$, $b_4=g_2(h,\alpha_z,D_{2z},b_2,L_z)$,

$b_5=g_3(h,\beta_z,D_{1z},b_1)$, $b_6=g_4(h,\beta_z,D_{2z},b_1)$.

Coefficients P_2,P_4 are the functions g_5,g_6 dependent of variables

C_0,C_a,H_1,H_1+h,H_1-h,

$\alpha_z,\beta_z,D_{1z},D_{2z},b_1,b_2,L_z$.

For comparing we use the averaged method with respect to z with exponentials spline. Then

$\alpha_{ix}=\alpha_{iy}=0$, $c_{iz}=c_{ix}$,

$e_i=B_{iz}(C_0b_{i3}+C_ab_{i4})$, $i=1;2$,

$c_{11}=B_{1z}b_{11}-D_{1z}b_1^2$, $c_{12}=B_{1z}b_{12}$, $c_{21}=B_{2z}b_{21}$,

$c_{22}=B_{2z}b_{22}-D_{2z}b_2^2$, and using the averaged method in z -direction the functions c_{az},c_{0z} are replaced with C_a,C_0.

We have following numerical results ($L_z=1, H_1=0.6, H_2-0.4, L_x=L_y=1$, $C_0=0.3$, $C_a=2.0$, $D_{1z}=10^{-3}$, $\alpha_z=20$, ($a_1=a_2=0.0001\approx 0$ - for parabolic spline). $D_{1x}=D_{1y}=10^{-4}$, $b_1=1.4050$, $\beta_z=10$, $b_2=4.4429$, $D_{2z}=5\cdot10^{-4}$, $D_{2x}=D_{2y}=5\cdot10^{-4}$) for maximal error δ and averaged values c_{1z},c_{2z} depending on a_1,a_2 $(c_1^y=0.3558, c_2^y=0.9809)$. The numerical results are given in the TABLE 1

The solution c_{1z}, $z\in[0,0.6]$, c_{2z}, $z\in[0.6,1.0]$ for 3 methods $(N=20)$ is represented in Figs. 1, 2. For FDS we have $\delta=0.0336(N=20)$, $0.0160(N=40)$, $0.0078(N=80)$.

TABLE 1

THE MAXIMAL ERROR ∂ AND AVERAGED VALUES c_{1z}, c_{2z} DEPENDING ON a_1,a_2

a_1	a_2	∂	c_{1z}	c_{2z}
0	0	0.1252	0.3022	0.9263
0.5	0.5	0.1235	0.3030	0.9273
1	1	0.1185	0.3051	0.9304
2	2	0.0996	0.3132	0.9421
3	3	0.0710	0.3250	0.9597
4	4	0.0450	0.3390	0.9812
5	5	0.0528	0.3528	1.0046
3	4	0.0378	0.3408	0.9842
2	4	0.0316	0.3422	0.9867
1	4	0.0283	0.3431	0.9884
1.2	4	0.0281	0.3430	0.9882

Using averaged method with respect to z with 2 unknown parameters in every layer we have

$c_i(x,y,z,t)=c_{iz}(x,y,t)+m_{iz}\frac{0.5H_i\sinh(a_i(z-\bar{z}_i))}{\sinh(0.5a_iH_i)}$

$+e_{iz}\left(\frac{H_i^2}{4}\left(\frac{\sinh^2(a_{0i}(z-\bar{z}_i)}{\sinh^2(0.5a_{0i}H_i)}-A_{i0z}\right)\right)$,

$A_{i0z}=\frac{\sinh(a_{0i}H_i)/(a_iH_i)-1}{\cosh(a_{0i}H_i)-1}$, $i=1;2$.

We have the exact solution for $a_1=b_1$, $a_2=b_2$, $a_{01}=\frac{b_1}{2}$, $a_{02}=\frac{b_2}{2}$ for every other parameters. $\alpha,\beta,C_0,C_a,H_1,H_2,L_z,D_{1z},D_{2z}$, b_1,b_2.

III RESULTS AND DISCUSSION

6. Some numerical results

A uniform grid in the space $((M+1)\times(N+1))$: $\{(y_i, x_j), y_i = (i-1)hy\ x_j = (j-1)hx\}$, $i = \overline{1, M+1}$, $j = \overline{1, N+1}$, $M\ hy = L_y, N\ hx = L_x$ with a time t moments $t_n = n\tau, n = 0,1,\ldots$. Subscripts (i, j, n) refer to y, x, t indices with the mesh spacing and for approximation the functions $c_{kz}(x, y, t), k = 1;2$ we have the grid function with following values: $U1^n_{i,j} \approx c_{1z}(x_j, y_i, t_n)$, $U2^n_{i,j} \approx c_{2z}(x_j, y_i, t_n)$ or $Uk^n_{i,j} \approx c_{kz}(x_j, y_i, t_n)$, $k = 1;2$.

For grid function's $Uk^n_{i,j}$ calculation Tomas algorithm in x and y directions was used for realization the alternating direction method (ADI) of Douglas and Rachford (1955) and 3-point difference equations for every direction were constructed.

The numerical results are obtained for [5]:

$z_m = mh_z, m = \overline{0,30}$, $h_z = \dfrac{L_z}{30}$,

$D_{1x} = D_{1y} = 10^{-4}$, $D_{2x} = D_{2y} = 5\cdot 10^{-4}$,

$L_z = 3$, $a_1 = 4.5$, $a_2 = 3.0$,

$H_2 = 1.2$, $L_x = L_y = 1$ $\alpha_z = 20$, $\beta_z = 10$,

$\alpha_{1x} = 2$, $\alpha_{2x} = 2$, $\alpha_{1y} = 4$, $\alpha_{2y} = 4$,

$c^v_{1ax} = 2.5$, $c^v_{2ax} = 2.5$, $c^v_{1ay} = 2.5$, $c^v_{2ay} = 2.5$,

$N = 22$, $M = 20$, $a_1 = 4.5$, $a_2 = 3.0$.

On the top of earth $(z = L_z)$ the concentration $c[mg/kg]$ of metals is measured in following nine points in the (x, y) plane:

$c(0.1, 0.2) = 3.69$; $c(0.5, 0.2) = 4.43$;
$c(0.9, 0.2) = 3.72$; $c(0.1, 0.5) = 4.00$;
$c(0.5, 0.5) = 4.63$; $c(0.9, 0.5) = 4.11$;
$c(0.1, 0.8) = 3.71$; $c(0.5, 0.8) = 4.50$;
$c(0.9, 0.8) = 3.73$.

This date are smoothing in matrix c_{az} by 2D interpolation with MATLAB operator, using the spline function. In Fig. 2 we can see the distribution of concentration c for Ca in the (x, y) plane by $z = L_z$. On the below of peat block $z = 0$ the elements of matrix c_{0z} have a constant value $1.30\,mg/kg$.

For the initial condition the averaged solutions $c_{kz}(x, y)$, $k = 1;2$ are chosen. We have the stationary solution with $\tau = 10$, $t_f = 200$ the maximal error 10^{-6}, the maximal value of $c_{1z}(x, y)$ 2.500 for averaged method and for ADI method (following results we can see in Figs 3, 4.).

Depending on the number of the grid points (N, M) we have following maximal values for averaged method and ADI methods:

4.6311 ($N = 22$, $M = 20$), 4.6298 ($N = 12$, $M = 10$), 4.6324 ($N = 32$, $M = 30$).

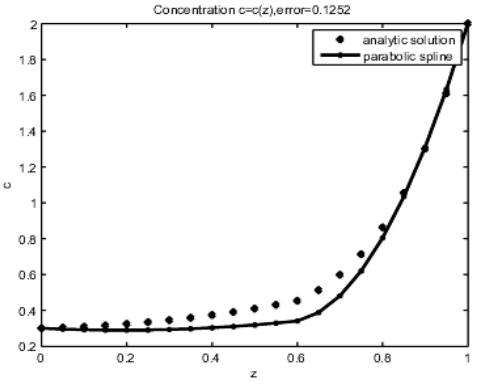

Fig. 1. Solution $c_1(z), c_2(z)$ for $\alpha_z = 20$, $\beta_z = 10$, $a_1 = 1.2$, $a_2 = 4.0$, using parabolic spline.

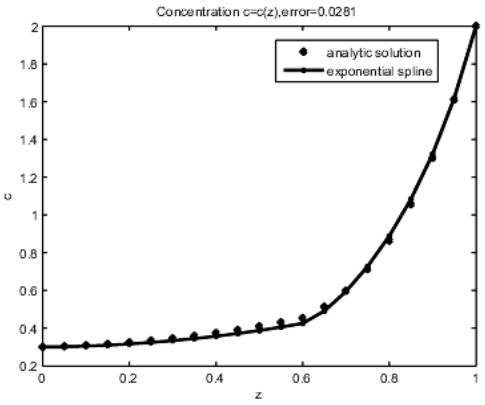

Fig. 2. Solution $c_1(z), c_2(z)$ for $\alpha_z = 20$, $\beta_z = 10$, $a_1 = 1.2$, $a_2 = 4.0$, using exponential spline.

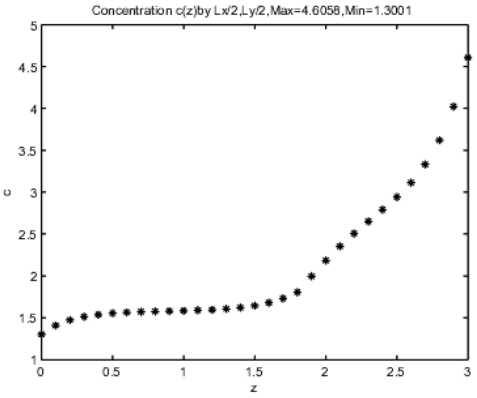

Fig. 3. Graphic of averaged concentration $c(z)$ for $x = L_x/2, y = L_y/2$.

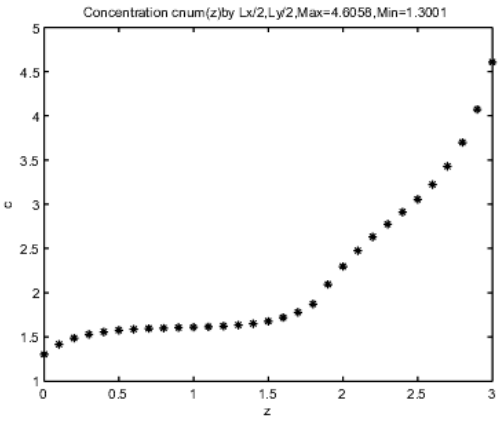

Fig. 4. Graphic of numerical concentration $c(z)$ for

$$x = L_x / 2, y = L_y / 2 .$$

IV CONCLUSIONS

The calculations were carried out with the mathematical model, was confirmed by the results of the measurements carried out on the surface of the Earth and the deeper layers – the highest concentrations of metals are at the top surface of the Earth, then its very rapidly decreases with increasing depth of peat layer.

Elements concentration in peat profiles confirms with respect to the possibility of using trace elements concentration as an indicator of the region and global environmental pollution.

The 3-D diffusion problem in peats block is reduced to 2-D and 1-D problems used the integral parabolic and exponential splines. The 1-D differential and discrete problems are solved analytically.

Increasing the number of parameters from one to two for the exponential spline formula, in the process of solving one dimension task the exact solution for calculating the concentration in z -direction was obtained.

The solutions of corresponding averaged non stationary 2-D initial-boundary value problem are obtained numerically using alternating-direction implicit (ADI) method of Douglas and Rachford. The numerical solution is compared with the designed analytical solution.

The relative error in the calculation of the concentration of heavy metals (FE, CA) with the averaged method which has been designed by authors, and the grid method does not exceed 2% -3%. It is perfectly acceptable (valid for use) accuracy to tackle practical engineering technical problems.

V REFERENCES

[1] A. Buikis. *The analysis of schemes for the modelling same processes of filtration in the underground.* Riga, Acta Universitatis Latviensis, 592 1994, pp. 25-32 (in Latvian).

[2] N. S. Bahvalov, N. P. Zhitkov, G. M. Kobelhkov. *Numerical mthods.* M:, Nauka, 1987 (in Russian).

[3] A. Buikis. *The approximation with splines for problems in layered systems.* Riga, Acta Universitatis Latviensis, 592 1994, pp. 135-138 (in Latvian).

[4] I. Kangro, H. Kalis, A. Gedroics, E. Teirumnieka, E. Teirumnieks. *On mathematical modllingof metals distributions in peat layers. Mathematical Modelling and Analysis,* 19:4 2014, pp. 568-588. Taylor & Francis, Mortimer House, 37-41, London W1T 3JH, UK.

[5] E. Teirumnieka, E. Teirumnieks, I. Kangro, H. Kalis, A. Gedroics. *The mathematical modeling of Ca and Fe distribution in peat layers. Proceed. of the 8-th int. scientific practical conference "Environment. Technology. Resources.",* pp. 40-47, Rezekne higher education institution, June 20-22, 2 2011.

[6] C. Henry Edwards, David E. Penny. *Differential equations and boundary value problems, computing and modelling.* Pearson, Prentice Hal, third edition, 2008 (in Russian).

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 193-197

Using the Concept of Fuzzy Random Events in the Assessment and Analysis of Ecological Risks

Oleg Uzhga-Rebrov[1], Galina Kuleshova[2]
[1]*Rezekne Higher Education Institution*,
[2]*Riga Technical University*

Abstract. **In many cases, the assessment and analysis of ecological risks is a complicated task, which is first of all related to obtaining reliable initial information. As a rule, ecological risks are due to unrepeated unique situations; from this it follows that sufficient statistical data on whose basis reliable evaluation of specific risks is made, are not available. On the other hand, unfavourable impacts on the external environment can affect the components of an ecosystem differently. The complexity of correlations among the components of an ecosystem significantly complicates an analysis of possible impacts on the components of a specific system.**

When statistical data are missing or insufficient, experts who perform the required assessment on the basis of their knowledge and experience but often also using their intuition, are the only source of initial data. Here, however, the problem of reliability of expert evaluations arises. If other sources of information are missing, we have to accept subjective evaluations of experts as a basis, without an opportunity to evaluate the degree of their confidence.

In this kind of situation, it seems to be validated to introduce the extent of uncertainty into the evaluations of parameters of ecological risks. This can be accomplished by using fuzzy initial evaluations. This paper focuses on the concept of fuzzy random events and shows favourable chances of using that concept in the assessment and analysis of ecological risks.

Keywords: **Function of membership to a fuzzy random event, fuzzy random event, operations on fuzzy events, probability distribution function, probability of a fuzzy random event.**

I INTRODUCTION

Any risk can be characterized by two components (dimensions): [1] potential losses that might occur due to the presence of unfavourable circumstances; (2) chances of occurrence of unfavourable circumstances (factors) which are commonly specified by means of probabilistic evaluations.

Depending on the character of unfavourable circumstances (factors) and related to them potential losses, risks can be classified as economic, political, social, military etc. Quite a wide group is composed of ecological risks. Those risks can be roughly divided into risks for economic activity of humans due to inefficient use of natural resources (flooding of large areas caused by building power stations, territory degradation due to the felling of the forests and predatory usage of lands). Another field of ecological risks is related to harmful impacts on an environment as a result of human economic activity. This kind of risk occurs mainly due to harmful extras that negatively affect the components of ecosystems.

To assess any risk, its two components have to be assessed: losses that might occur as a result of one of another human activity or effect of unfavourable external and internal factors, and probabilities of those losses. When sufficient statistical data are available, there is no difficulty to assess the risks. A visual example of this is the assessment of risks of fire and car accident performed by insurance companies. Evaluations of such risks are obtained on the basis of statistical material; due to that, their reliability is very high. Quite a different situation is with assessing unique ecological risks. Here statistical data are not available, as a rule. The only source of information is specialists - experts who accomplish requested assessment on the basis of their knowledge and experience but sometime –intuition. It is impossible to evaluate the reliability of expert evaluations a priori since the results of expert evaluations can be heavily affected by different heuristics and nonobjectivities. A detailed analysis of such heuristics and nonobjectivities is provided in [4]. There are also proposed some techniques that enable one to a priori evaluate potential biases that may have place in the evaluations of some experts. By means of such techniques it is only possible to slightly correct evaluations of experts under the possibility of

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.172

potential common mistakes, however total reliability of those evaluations remains at the sign of question.

Recently, techniques have been developing intensively that enable obtaining and using uncertain expert evaluations. A smaller or greater part of uncertainty is artificially incorporated into the initial evaluations. Otherwise, instead of one deterministic evaluation, say, the probability of random event occurrence, a generalized uncertain evaluation is integrated which includes a set of deterministic evaluations with different degrees of uncertainty. The obtained uncertain evaluations are processed and aggregated; as a result, uncertain resulting evaluations are derived. This kind of approach seems to be more preferable as compared to the use of deterministic initial and resulting evaluations regarding which there is no any confidence that they adequately represent the existing state of things.

To model the uncertainty of expert evaluations, fuzzy set theory is widely used. This common approach can be divided into two major directions: (1) using the concept of fuzzy random events and (2) using fuzzy evaluations of the probabilities and losses. This paper discusses the concept fuzzy random events, defines probability calculations for such events and analyses the possibility of using this kind of event in the assessment and analysis of ecological risks.

II WHAT IS A FUZZY RANDOM EVENT?

Let us define the notion of a fuzzy random event using a simple and visual example. Let us assume that an accident may occur at the chemical plant, as a result of which a harmful substance can get into an external environment. That substance might cause a substantial damage to some components of the surrounding ecosystem.

To assess the risks related to the potential accident at the plant, it is necessary to assess probabilities of different level of environment pollution. For that purpose, let us introduce a nondimensional factor of pollution extent:

$$c = \frac{c_r}{c_{mpc}} \tag{1}$$

where c_r - real concentration of harmful substance in the environment; c_{mpc} - maximum allowable concentration of that substance.

Since c is a continuous random variable, probability distribution function $F(c)$ can be constructed for it. When statistical data are not available, the construction of the requested distribution function can be made on the basis of expert evaluations. Methods of construction of subjective distribution functions are considered in detail in [6], [7]. Not going into technical details, let us assume that based on the data provided by an experienced expert, distribution function $F(c)$ is constructed for the continuous random variable c defined in expression (1). A graph of that distribution function is shown in Fig.1.

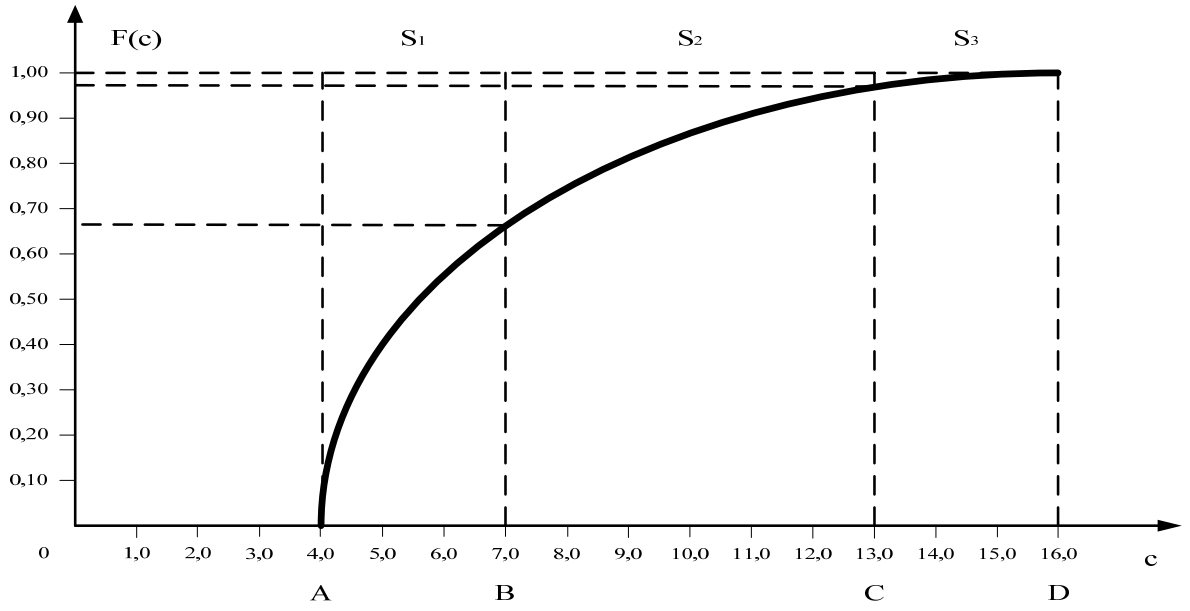

Fig.1. A graph of distribution function $F(c)$, constructed using expert evaluations of an expert

For further analysis it is worth distinguishing some generalised categories of pollution of the environment. In Fig.1 these categories are depicted as S_1, S_2, S_3 and can be interpreted as follows:

S_1 - low pollution: $c \in [4,0;7,0]$;

S_2 - moderate pollution: $c \in [7,0;13,0]$;

S_3 - heavy pollution: $c \in [13,0;16,0]$.

From the graph of distribution function $F(c)$ shown in Fig.1 it is easy to determine probabilities of each of pollution categories:

$$p(S_1) = F(B) - F(A) = 0,65 - 0,00 = 0,65;$$

$$p(S_2) = F(C) - F(B) = 0,97 - 0,67 = 0,30;$$

$$p(S_3) = F(D) - F(C) = 1,00 - 0,95 = 0,05.$$

The problem with the above-mentioned separation of pollution level into categories is related to the determination and interpretation of borders between the categories. Let us assume that to lessen the risk, specific actions are required for each category. Let us also assume that for a certain scenario of possible development of the situation a conclusion is made that potential pollution is evaluated as $c = 0,66$, but for alternative scenario an evaluation of the pollution level $c = 0,74$ is obtained. Hence, the planned actions aimed at lessening the risk, formally have to be different for each scenario, which seems to be improbable taking into account that the difference between evaluations $c(.)$ is insignificant.

In this kind of situation it is worth to somehow „wash away" the borders between the categories. A correct technique for that purpose is to use fuzzy categories instead of unambigously specified ones. One possible way to represent the aforementioned categories in a fuzzy form is depicted in Fig. 2.

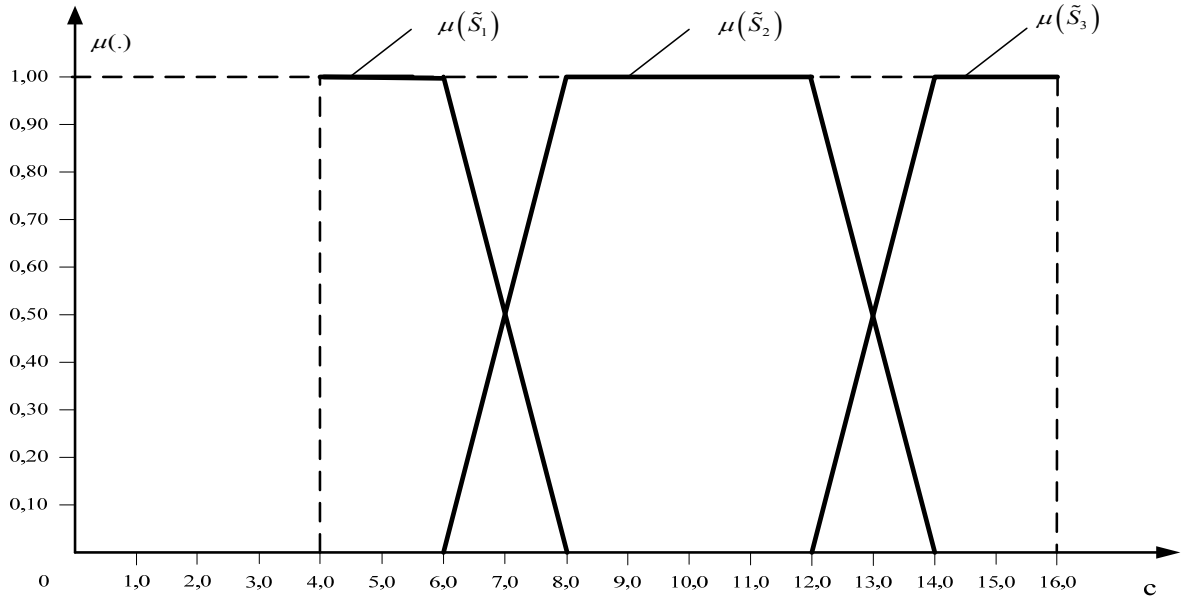

Fig.2. Representation of deterministic categories of environment pollution level in a fuzzy form

The falling of real pollution level into one of categories S_1, S_2, S_3 in Fig.1 can be interpreted as a deterministic random event. Then the falling of real level of pollution into one of fuzzy categories $\tilde{S}_1$, $\tilde{S}_2$, $\tilde{S}_3$ in Fig.2 can be viewed of as fuzzy random event: real level of pollution falls into a fuzzy category $\tilde{S}_i$ with the membership degree $\mu(\tilde{S}_i)$, $i = 1,2,3$.

Let us then assume that the forecasted level of environment pollution is characterised by value $c = 0,65$. Using graphs of membership functions $\mu(\tilde{S}_1)$ and $\mu(\tilde{S}_2)$ it is easy to determine that this pollution level with the membership degree $\mu(\tilde{S}_1) = 0,73$ belongs to the fuzzy category $\tilde{S}_1$, but with the membership degree $\mu(\tilde{S}_2) = 0,27$ belongs to the fuzzy category $\tilde{S}_2$. This is the principal difference between deterministic categories and fuzzy categories. In the first case, real value of the relevant

random variable may only belong to a single category, buti n the last case – to several fuzzy categories.

How the fuzzy borders between the categories are formed? In the interval $c \in [4,0;6,0]$ pollution is unambigously related to category S_1, while in the interval $c \in [6,0;8,0]$ the pollution may be either ascribed to category S_1, or – category S_2 with different degree of membership. In the interval $c \in [8,0;12,0]$ pollution is unambigously related to category S_2. Pollutions in the intervals $c \in [12,0;14,0]$ and $c \in [14,0;16,0]$ can be interpreted in similar way.

Let us briefly consider the theory of fuzzy events. The notion of a fuzzy event and its probability was first stated by L. Zadeh in his underlying work [10].

Definition. Let (R^n, Θ, P) be a probabilistic space, in which Θ is σ-field of Borel sets in R^n, but P is the probabilistic evaluation in R^n. Then *a fuzzy event* in R^n is a fuzzy set A in R^n, whose membership function, μ_A $(\mu_A : R^n \to [0,1])$, is Borel-measurable

If a fuzzy event $\tilde{A}$ is related to a continuous probabilistic function in the domain of definition of $\tilde{A}$, the probability of this event can be calculated as follows [10]:

$$P(\tilde{A}) = \int_{R^n} \mu_{\tilde{A}}(x) dP = E(\mu_{\tilde{A}}). \qquad (2)$$

Strict mathematical details of fuzzy event theory can be found in [2].

Fuzzy events are in essence fuzzy subsets specified in the domain of definition of relevant random variable. Say, in the example considered above fuzzy events $\tilde{S}_1$, $\tilde{S}_2$ and $\tilde{S}_3$ are typical fuzzy subsets with trapezoidal membership functions $\mu(\tilde{S}_i)$, $i = 1,2,3$; from this it follows that all operations on fuzzy sets can be applied to fuzzy events. More detailed information about the operations on fuzzy sets is provided in [3], [5] and [8].

Let us calculate the probabilities of the considered above fuzzy categories using expression (2).

$$P(\tilde{S}_1) = \int_{4,0}^{6,0} \mu_{\tilde{S}_1}(c) dP + \int_{6,0}^{8,0} \mu_{\tilde{S}_1}(c) dP = (F(6) - F(4)) + \frac{1}{2}(F(8) - F(6)) =$$

$$= (0,55 - 0) + \frac{1}{2} * (0,75 - 0,55) = 0,55 + \frac{1}{2} 0,20 = 0,65.$$

Here $F(.)$ is the value of distribution function at the corresponding value of c.

$$P(\tilde{S}_2) = \int_{6,0}^{8,0} \mu_{\tilde{S}_2}(c) dP + \int_{8,0}^{12,0} \mu_{\tilde{S}_2}(c) dP + \int_{12,0}^{14,0} \mu_{\tilde{S}_2}(c) dP =$$

$$= \frac{1}{2}(F(8) - F(6)) + (F(12) - F(8)) + \frac{1}{2}(F(14) - F(12)) =$$

$$= \frac{1}{2}(0,75 - 0,55) + (0,95 - 0,75) + \frac{1}{2}(0,98 - 0,95) = \frac{1}{2} * 0,20 + 0,30 + \frac{1}{2} * 0,03 = 0,315;$$

$$P(\tilde{S}_3) = \frac{1}{2} \int_{12,0}^{4,0} \mu_{\tilde{S}_3}(c) dP + \int_{14,0}^{16,0} \mu_{\tilde{S}_3}(c) dP = \frac{1}{2}(F(14) - F(12)) + (F(16) - F(14)) =$$

$$= \frac{1}{2}(0,98 - 0,95) + (1,00 - 0,98) = \frac{1}{2} * 0,015 + 0,02 = 0,035.$$

It is easy to see that the sum of probabilities of all fuzzy events is equal to 1. This is due to the fact that for all distribution functions whose graphs are shown in Fig.2 the following requirement holds

$$\sum_i \mu_{\tilde{S}_i}(c) = 1 \qquad (3)$$

for all values $c \in [4,0;16,0]$.

It should be noted that in the fundamental work [10] condition (3) has not been formulated as an obligatory requirement; the requirement was formulated later so as to ensure that the values of probabilities of fuzzy events are similar to the values of common probabilities for the complete group of events. (for

more details see [4]). If membership functions of relevant fuzzy events are constructed so that condition (3) is not satisfied, the resulting values of probabilities can be renormated properly to ensure that condition (3) holds.

III RESULTS AND DISCUSSION

In the example discussed above, both at deterministic and fuzzy borders between the categories, category *„low pollution"* has the highest probability of occurrence but category „heavy pollution" has the lowest probability of occurrence. In both cases it is only related to the shape of graph of distribution function $F(c)$ in Fig.1.

It is not difficult to calculate the values of probabilities of fuzzy events $P\left(\tilde{S}_i\right), i=1,2,3$ by expression (2) because graphs of membership functions $\mu\left(\tilde{S}_i\right), i=1,2,3$ in Fig.2 have the trapezoidal shape and operations of integration are extremely simple. In real tasks, graphs of functions of membership in fuzzy events (categories) are, as a rule, of trapezoidal or triangular shape, which sufficiently simplifies calculation of probabilities of respective fuzzy events. In case if the graphs of membership functions are of different shape, they have to be represented in analytical form; for that purpose approximation may be used.

Where the concept of fuzzy random events can be used? A good example of such application is discussed in [9], where the author solves the task of modelling and practical evaluation of processes in real fuzzy stochastic system. Another widely used application of that concept is tasks of fuzzy classification of the objects. The theoretical grounds of solving such tasks are considered in [1].

The concept of fuzzy random events can also be successfully applied in the case of discrete random variables. If some statistical data about realizations of random variable X in the past are available, deterministic values of probabilities of falling the values of this variable into the specified fuzzy categories can be evaluated. For the calculation of probabilities of fuzzy random events (of falling the next realisation of X in the corresponding fuzzy categories), expression (2) is transformed as follows:

$$P\left(\tilde{A}_i\right)=\sum_{x_j\in\tilde{A}} p\left(x_j\right)\mu_{\tilde{A}_i}\left(x_j\right), i=1,...,n\ . \quad (4)$$

What are potential advantages of using the concept of fuzzy random events in the assessment and analysis of ecological risks? First, multiple studies have shown that humans better think in terms of fuzzy categories, i.e., they better express their judgements and evaluations in a fuzzy environment. Second, the specifics of ecological risk assessment in most of cases is that the initial data are either absent or insufficient to employ standard statistical techniques. Effective modelling of the uncertainty of results of expert evaluations and further use of such uncertain information allow one to more successfully assess and analyze various ecological risks.

IV CONCLUSIONS

This work analyses possibilities of using the concept of fuzzy random events in the assessment and analysis of ecological risks. Taking into account intensive development of industry, problems of risk assessment and prevention from negative impact on the environment are becoming ever more important both in the developed and developing countries. The lack of proper objective data makes it necessary to widely use expert evaluations; however the requirement of deterministic evaluations from experts is in evident contradiction with the requirement of reliability of such evaluations. Due to that, it seems necessary to use uncertain expert evaluations. The modelling of expert evaluation uncertainty can be performed in different ways. The technique discussed in this paper, i.e., using fuzzy random events, exemplifies one of possible directions. It can be successfully employed for practical evaluation and analysis of different types of ecological risks.

V REFERENCES

[1] J. van den Berg and U. Kaumak. „On the Notion of Statistical Fuzzy Entropy", in *Soft Methodology and Random Information Systems*, Advances in Soft Computing, Heidelberg: Physica Verlag, 2004, pp.535 – 542.

[2] S. Gudder, "What Is Fuzzy Probability Theory?" Foundations of Physics, Vol. 30, No. 10, 2000, pp. 1663 – 1678.

[3] G.J. Klir and B.Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ: Prentice Hall, 1995.

[4] M.G. Morgan and M. Henrion, *Uncertainty. A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, 1998.

[5] K. Tanaka, *An Introduction to Fuzzy Logic for Practical Applications*. New York etc.: Springer, 1997, 138 p.

[6] O. Užga-Rebrovs, *Lēmumu analīze*. Rēzekne: RA izdevniecība, 1997.

[7] O. Užga-Rebrovs *Komerciālo lēmumu analīze. I, II daļa.* Rēzekne: RA izdevniecība, 2001.

[8] O.I. Uzhga-Rebrov, *Upravlenije neopredelennostjami. Chastj 4. Kombinirovanije neopredelennostej*, Rēzekne: RA Izdevniecība, 2014. (In Russian).

[9] A. Walaszek-Babiszevska, "Probability Measures of Fuzzy Events and Linguistic Fuzzy Modeling – Forms Expressing Randomness and Imprecision", in *Proceedings of 7th WSEAS Int. Conf. on Artifical Intelligence, Knowledge Engineering and Data Bases (AIKED'08)*, University of Cambridge, UK, Feb. 20 – 22, 2008, pp. 207 – 213.

[10] L.A. Zadeh, „Probability Measures of Fuzzy Events", Journal of Mathematical Analysis and Applications,Vol. 23, No 2, 1968, pp. 421 – 427.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 198-204

Path Planning Methods in Chemical Engineering

Edvards Valbahs[1], Peter Grabusts[1], Ilo Dreyer[2]
[1] Rezeknes Augstskola, Faculty of Engineering.
Address: Atbrivosanas aleja 115, Rezekne, LV-4601, Latvia
[2]Riga Technical University, Faculty of Material Science and Applied Chemistry.
Address: 14/24 Azenes Street, Riga, LV-1048, Latvia

Abstract. **Usually, when the practical motion planning and the shortest path are discoursed, mainly the limited number of tasks is observed. Almost all the tasks associated with the path from one point in 2D or 3D space to another point can be attributed to the usual issue in the practical application. Motion planning and the shortest path have vivid and indisputable importance as human activity in such areas as logistics and robotics. In our work we would like to draw particular attention to the field of application seems to be unnoticeable for the task such as motion planning and the shortest path problem. Due to quite simple examples used, we would like to show that the task of motion planning can be used for simulation and optimization of multi-staged and restricted processes which are presented in chemical engineering accordingly. In the article the simulation and optimization of three important chemical-technological processes for the chemical industry are discussed. The work done gave us the possibility to work out software for simulation and optimization of processes that in some cases facilitates and simplifies the work of professionals engaged in the field of chemical engineering.**

Keywords: **chemical industry, path planning, Travelling salesman problem, Simulated Annealing.**

I INTRODUCTION

The article is connected to the path planning and "Travelling salesman problem" (TSP) [2] used for decision and process automation in unusual or unexpected field of application, like chemical engineering. Practical importance of path planning is significant in transport, path planning tasks among different geographical objects. However, we would like to mention some unnoticeable cases.

Formation of genetic map – the map includes chromosomes that consist of definite marker sequence. In the given case the markers are unique DNA segments within one genome. In laboratory conditions it is possible to determine and distinguish different markers that allow forming genome maps. In order to prepare a map it is necessary to know the sequence of markers in a genome. The task of searching genome marker sequence can be modelled as TSP task. Marker sequence can be envisaged as a path along all markers. A group of researchers from National Institute of Health developed the methods and software for marker sequence searching problem [13].

Telescope orientation and searching new planets – orientation of large telescopes is rather complicated and takes much energy. To decrease the period of orientation, TSP problem is applied (for example, mentioned 200 galaxies can be investigated with a telescope during one night) [15], [22].

In crystallography – positioning crystal sample in x-ray diffractometer is a procedure happening for a long time. Due to one experiment you can include up to 30 000 different positions. When applying path planning methods, it is possible to decrease the length of experiment by 46% [14].

In electronics for example, boring holes in silicon plate, practically all the electronic silicon devices have silicon plates containing different microcircuits. Path planning application allows improving the productivity of assembly lines up to 10% [19]. One can also name the other applications in electronics: brazing silicon plates, developing computer microcircuits, searching silicon plates from defects.

Data organization in groups of elements with similar properties is the principal element in data interpretation. TSP problem was successfully adapted to solving the problem [16], [23]. TSP was successfully applied to distribute the collection of musical compositions.

German company BOWE CARDTEC develops and delivers identification and credit cards. Customers of the company book various cards. In order to make these cards, it is necessary to reconfigure equipment from one working task to the other working task and this is rather prolonged operation. To improve this

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.188

process, path planning methods are applied and the entire process was modelled as path planning problem. Working tasks were presented like cities, but the time required to make the reconfigurations was presented in a kind of distance among the cities. Path planning problem gives the possibility to improve the productivity up to 65% [10].

Our previous article was dedicated to mobile agent that moved across 2D closed premises and path planning was successfully applied [12]. Now we would like to show that path planning can be successfully applied for some chemical engineering processes. More detailed motivations and goals are described in the next section of our article.

II GOALS

The aim of the article is to describe some chemical engineering approaches to the production of substances as multistaged process. The processes have to be fulfilled in succession, controlling the considerable parameters of processes. The quality and quantity of selected product depends on sequence of the process stages and control of characteristic parameters a great deal. Actually in chemical engineering one or several alternative stages can be selected for any stage of process which deals with substances production. Alternative stages differ from each other with completely definite parameters set which characterizes selected stage. Thus, according to previously mentioned technological processes can be characterized as follows:

- certain branching – each alternative branch can be viewed as alternative path;
- selection among alternative stages – it is possible to select one solution from several paths;
- stage characterized with parameters – there is a possibility to characterize alternative stages with definite parameters set.

Taking into account the above mentioned marks, it can be confirmed that chemical engineering processes are characterized with a tree or graph (i.e. the data structure in a kind of a tree or graph). Therefore, chemical engineering processes can be viewed within path planning task. Our aim is to discuss chemical engineering processes as graph-type data structure for the following substances or groups of substances (that allows programming decision-making and process analysis):

- Borane – it can be used in the area of high-energy fuels for jet planes and rockets. The thermal decomposition of diborane (B_2H_6) has been used to produce coatings of pure elementary boron for neutron-detecting devices and applications requiring hard, corrosion-resistant surfaces. Boranes can be used as vulcanizing agents for natural and synthetic rubbers, and are particularly effective in the production of silicone rubbers [6], [11], [17];
- Malonic acid derivatives – widely used in biochemistry and biology [4], [24].

The above mentioned compounds have been selected, for there are a lot of scientific sources that describe the obtaining of given compounds and in the same time describe all the possible alternatives as well.

III ASSUMPTIONS

To fulfil the previously mentioned goals, it is necessary to reflect the processes of obtaining borane and aluminium oxide in a kind of graphs. The structure of graphs allows performing further analysis of processes. Correlation among the substances involved in the process and process stage both can be presented as a table. The process of obtaining borane is described in Table 1.

TABLE 1.

BORANE OBTAINING PROCESS

Substances before conversion (or current stage of the process)		Characteristic parameters for stage: temperature of reaction (C^o)	Product of stage	
Substance marking on graph	Name of substance and formula		Name of substance and formula	Substance marking on graph
1	B4H10 - tetraborane	200	B5H9 - pentaborane	2
		60	B5H11 - pentaborane	5
2	B5H9 - pentaborane	-	-	-
3	B2H6 - diborane	180	B4H10 - tetraborane	1
		130	B5H9 - pentaborane	2
		280	B10H14 - decaborane	4
		115	B5H11 – pentaborane	5
4	B10H14 - decaborane	95	B4H10 - tetraborane	1
5	B5H11 - pentaborane	25	B10H14 - decaborane	4
		100	B4H10 - tetraborane	1

For example, the two substances (two different pentaboranes) can be obtained from tetraborane (Table 1), but at different temperatures (in this case the temperature is an important parameter of current stage). Moreover, tetraborane can be obtained either from decaborane, pentaborane or diborane (Table 1).

The graph (Fig. 1) related to borane obtaining process (Table 1) is built up, where graph nodes are substances and graph edges are the temperature in each stage. Each stage has its own weight and is relevant to the temperature.

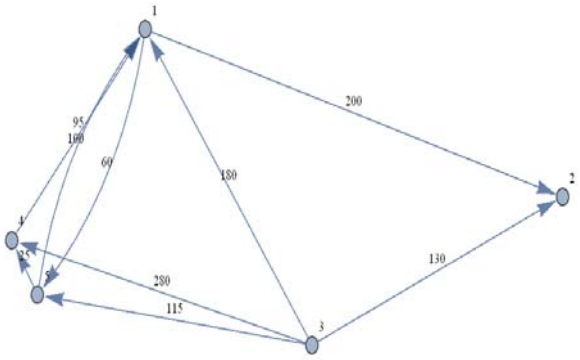

Fig. 1. Graph for borane obtaining process

It is needed to mention that temperature plays a very important role in chemical technology. Actually, all chemical processes are taking place at definite temperature, which stimulates or suppresses the processes. It means that the temperature within the process should be sustained, i.e. supplied with additional amount of warmth or it can be removed. All manipulations with amount of warmth are directly related to power consumption. Actually, in chemical technology with amount of warmth energy efficiency of process can be characterized. Chemical engineer, designing technological process, is always interested in decreasing heat consumption, i.e. selecting a process where the final product is obtained with lower consumption of heat as far as possible, increasing energy efficiency of obtaining substance accordingly.

As to malonic acid derivatives, the design principle of previously mentioned graph remains with its own specific features, related to stage parameters, i.e., outcome of reaction Table 2.

TABLE 2.
MALONIC ACID DERIVATIVES OBTAINING PROCESS

SUBSTANCES BEFORE CONVERSION (OR CURRENT STAGE OF THE PROCESS)	CHARACTERISTIC PARAMETERS FOR STAGE: (100% - OUTCOME OF REACTION) (%)	PRODUCT OF STAGE
SUBSTANCE MARKING ON GRAPH		*SUBSTANCE MARKING ON GRAPH*
A1	40	A2
A2	35	A3
A3	31	B3
B1	55	B2
B2	30	B3
B3	10	C2
	30	D5
C1	2	D4
C2	50	C1
D1	24	D2
D2	45	D3
D3	4	D4
D4	33	D5
D5	5	D6
D6	-	-

Table 3 shows the names of malonic acid derivatives in accordance with the names of vertexes in the graph (Fig. 2).

TABLE 3.
NAMES OF MALONIC ACID DERIVATIVES IN THE GRAPH

SUBSTANCE MARKING ON GRAPH	NAME OF SUBSTANCE AND FORMULA (WHERE R IS ALKANE DERIVATIVE)
A1	di(R) 2-((R)oxymethylidene)propanedioate
A2	R 4-hydroxy-2-methylpyrimidine-5-carboxylate
A3	R 4-chloro-2-methylpyrimidine-5-carboxylate
B1	2-((R)oxymethylidene)propanedinitrile
B2	R (2E)-3-[(Z)-(1-aminoethylidene)amino]-2-cyanoprop-2-enoate
B3	R 4-amino-2-methylpyrimidine-5-carboxylate
C1	4-amino-2-methylpyrimidine-5-carboxamide
C2	4-amino-2-methylpyrimidine-5-carbonitrile
D1	4,6-dihydroxy-2-methylpyrimidine-5-carboxamide
D2	4,6-dichloro-2-methylpyrimidine-5-carbonitrile
D3	4-amino-6-chloro-2-methylpyrimidine-5-carbonitrile
D4	5-(aminomethyl)-2-methylpyrimidin-4-amine
D5	(4-amino-2-methylpyrimidin-5-yl)methanol
D6	5-(chloromethyl)-2-methylpyrimidin-4-amine

Weights of edges at the graph are proportional to the products of the reaction (Fig. 2).

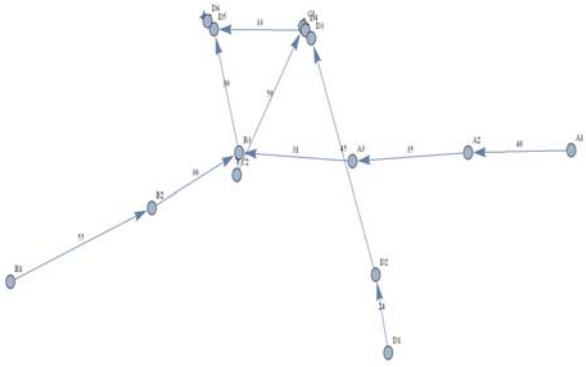

Fig. 2. Graph for malonic acid derivatives obtaining process

Outcome of reaction is amount of product obtained as a result of chemical process/reaction (it is a very important parameter in chemical engineering) and it is always less than 100 per cent (it characterizes the effectiveness of reaction). It can be seen that the higher outcome of the stage, the better result is achieved according to the parameter. But we should take into consideration that for path planning it is better to have a total distance which must be the smallest as far as possible. Thus, admittedly, value of edges in the given case is difference between 100 per cent and outcome of current stage.

IV RESULTS

Discussing processes of forming borane and aluminium oxide, it can be assumed that all stages take one and the same prolongation of time. If time of reaction is unchangeable in all stages, then difference of stages including amount of warmth will depend only on temperature. It is possible to find the most energy-efficient path studying the graph of process and looking for the shortest path from start vertex to final one. Fig. 3 shows the most energy-efficient path from diborane (vertex 3) to decaborane (vertex 4) in the process of obtaining boranes (Table 1).

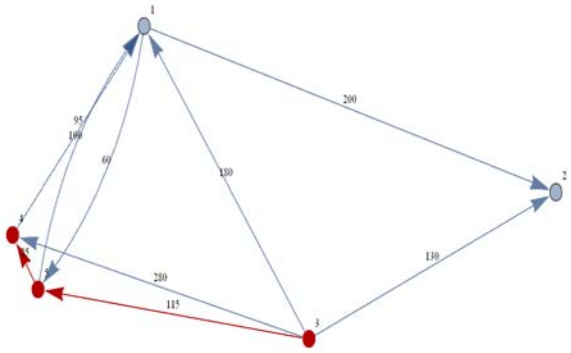

Fig. 3. The most energy-efficient path from diborane to decaborane

The production of decaborane (vertex 1) from tetraborane is more beneficial in comparison with the production from diborane (it can be vividly seen from Fig. 4 that the total length of the path is shorter compared with the previous example).

If it is necessary to obtain the group of substances tetraborane, pentaborane (vertex 5) and decaborane as energy-efficient path, it is needed to consider the part of the graph which does not contain vertex 2 (Fig. 5).

The shortest path among all the vertexes in the graph can be obtained as a part of TSP, i.e. Simulated annealing (SA) can be used to find the energy-efficient path in the graph [1], [3], [5], [7], [18], [20], [21].

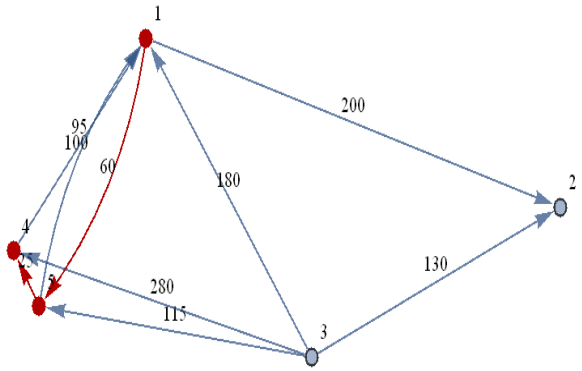

Fig. 4. The most energy-efficient path from diborane to decaborane

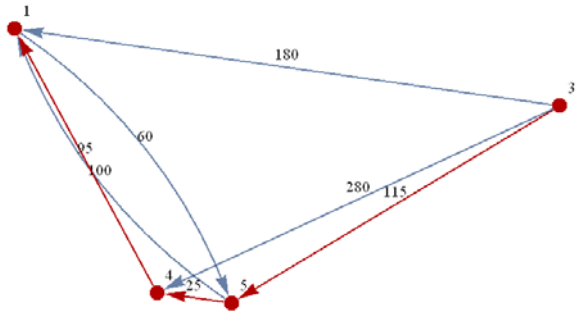

Fig. 5. The graph presents the part of boranes obtaining process with the shortest path among all the vertexes

Outcome of reaction and temperature are of great practical importance. Outcome of reaction shows what amount of substance of the initial mixture of substances is obtained at the end of conversion to the product of chemical reaction. Outcome of reaction reveals the efficiency of substances consumption in chemical processes. Therefore, it can be confirmed that the given parameter determines exactly economic indicators of the processes (as the substance can be also bought without production, and it is necessary to decide either it should be bought or produced and it will be reflected in terms of money). For stages of the process on obtaining derivatives of malonic acid different alternatives can be viewed which considerably differ by outcome of reaction, but not by temperature. Let us look at the graph for malonic acid derivatives obtaining process. A graph structure can be extended by assigning a weight to each edge as difference between 100 per cent and outcome of reaction, and by this way the shortest path is explained to be looked for in the graph. As the shortest path is searched the less the edge is the better result is. And on the contrary, it is better when the outcome of reaction reaches to 100 per cent. In order to apply path planning methods, in this case the weight of edge should be calculated as difference between 100 per cent and outcome of reaction, 100 per cent being theoretical maximal value of outcome of reaction. The

shortest path from di(R) 2-((R)oxymethylidene) propanedioate (vertex A1 and where R is alkane derivative) to 5-(chloromethyl)-2-methylpyrimidin-4-amine (vertex D6) for malonic acid derivatives obtaining process (Table 2) is presented in Fig. 6.

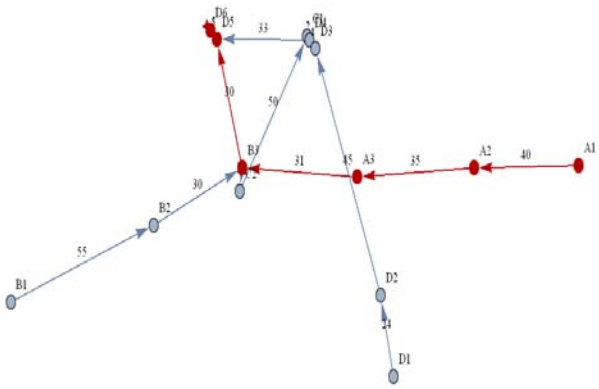

Fig. 6. The shortest path from di(R) 2-((R)oxymethylidene) propanedioate (where R is alkane derivative) to 5-(chloromethyl)-2-methylpyrimidin-4-amine (total path length is 141 per cent)

The shortest path from 2-((R)oxymethylidene) propanedinitrile (vertex B1 and where R is alkane derivative) to 5-(chloromethyl)-2-methylpyrimidin-4-amine (vertex D6) is depicted Fig. 7.

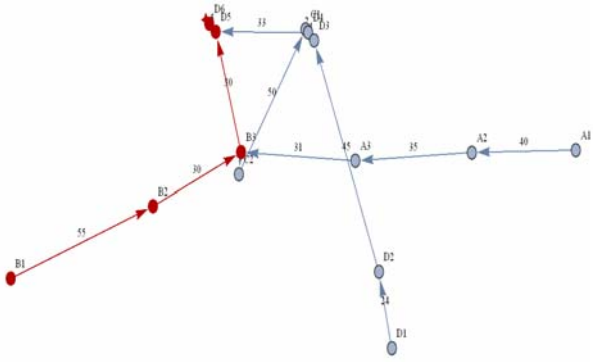

Fig. 7. The shortest path from 2-((R)oxymethylidene) propanedinitrile (where R is alkane derivative) to 5-(chloromethyl)-2-methylpyrimidin-4-amine (total path length is 120 per cent)

Fig. 8 shows the shortest and most efficient path for the outcome of the reaction (compared with the previously mentioned cases) from 4,6-dihydroxy-2-methylpyrimidine-5-carboxamide (vertex D1) to 5-(chloromethyl)-2-methylpyrimidin-4-amine (vertex D6) .

The shortest path from 4,6-dihydroxy-2-methylpyrimidine-5-carboxamide (vertex D1) to 5-(aminomethyl)-2-methylpyrimidin-4-amine (vertex D4) is shown in Fig. 9.

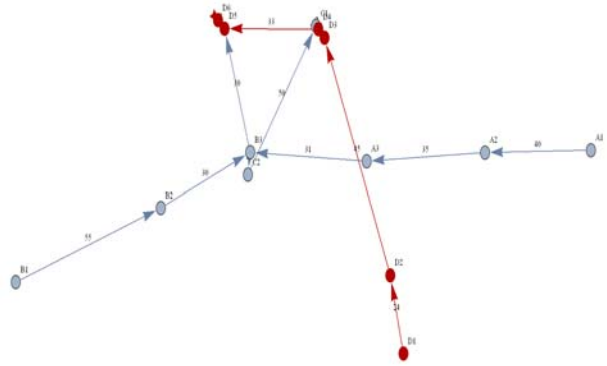

Fig. 8. The shortest path from 4,6-dihydroxy-2-methylpyrimidine-5-carboxamide to 5-(chloromethyl)-2-methylpyrimidin-4-amine (total path length is 111 per cent)

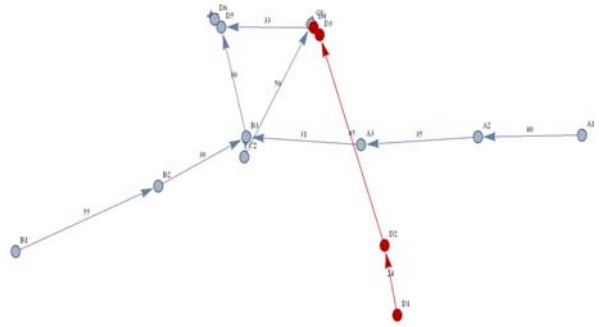

Fig. 9. The shortest path from 4,6-dihydroxy-2-methylpyrimidine-5-carboxamide to 5-(aminomethyl)-2-methylpyrimidin-4-amine (total path length is 73 per cent)

The shortest and most efficient path for the outcome of the reaction (compared with the previously mentioned cases) from R 4-amino-2-methylpyrimidine-5-carboxylate (vertex B3 and where R is alkane derivative) to 5-(aminomethyl)-2-methylpyrimidin-4-amine (vertex D4) is depicted in Fig. 10.

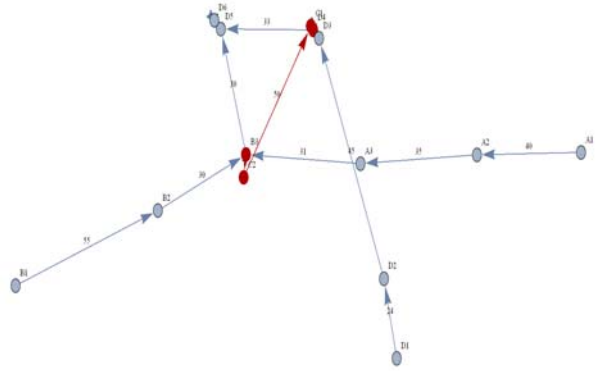

Fig. 10. The shortest path from R 4-amino-2-methylpyrimidine-5-carboxylate (where R is alkane derivative) to 5-(aminomethyl)-2-methylpyrimidin-4-amine (total path length is 62 per cent)

V CONCLUSION

Submitted examples show that TSP and path planning can be successfully used in automatizing the analysis and decision-making for borane, malonic acid derivatives and other processes for production of chemical substances. Automated decision-making and analysis give the possibility to facilitate the design process. Optimal conditions of the process in selected technological processes are rather easy to be found without using any software, which are done by engineers in industry and searching these optimal conditions extends the design process. If a process contains more alternative stages than those in the selected processes, the design process is expected to be protracted, which will also increase the price of technological process implementation. Design practices can be improved even with a partial atomization of the designing process.

We would like to mention that many linked substances (vertexes) are involved in chemical engineering processes. The methods of obtaining substances often change (e.g. production of bioactive substances). Chemical engineering processes are dynamic systems that change from time to time. A lot a scientific researcher is conducted and as a result known products can be obtained in more efficient processes in comparison with technologies that are well know nowadays.

Working out new technologies it is necessary to take into a count the competitive period of time for chemical technology or process developed. It is needed to know the parts of technology that will have to be changed in the nearest future in order to keep the competitiveness of technology. Observing technological process from the point of view of such approach gives the possibility to make necessary changes more efficiently [8]. Taking in to account the previous mentioned it is impossible to find the shortest path (more effective solution) by heart in systems or technology said [9]. In addition, we would like to mention the example of robot scientist Eve which was developed by the researches and engineers from Cambridge and the University of Manchester. Robot Eve is successfully involved in the developing process of new medicines. This work is an excellent combination of robotics, artificial intelligence and pharmacology [25]. We would like to demonstrate that currently, artificial intelligence methods can contribute a considerable for chemical engineering.

The programme product developed in the context of this work simplifies the process of designing for borane and malonic acid derivatives obtaining processes. It is easy to track optimization process using the TSP and path planning because all the selected examples of chemical engineering processes in this article are simple. In our views, TSP, path planning algorithms and methods can be successfully applied to other chemical engineering processes optimization tasks with much more complexity (with the large number of alternative stages and other wider group of parameters, i.e. not only for temperature).

As it was mentioned in the Introduction, TSP and the shortest path problem have specific field of application which seems to be unnoticeable:

- formation of genetic map;
- telescope orientation and searching new planets;
- positioning crystal sample in x-ray diffractometer;
- boring holes in silicon plate;
- data organization in groups of elements;
- decreasing wallpaper strips;
- cutting templates in glass production;
- investigating the problem in evolution process.

We suppose that the mentioned list can be added with the new application of TSP and the shortest path problem, i.e. application in chemical engineering. As regards the chemical engineering, the implementation of TSP and the shortest path problem give the opportunity of developing new methodology for chemical engineering.

VI REFERENCES

[1] R. Azencott, *Simulated annealing: parallelization techniques.* John Wiley and Sons, 1992.

[2] W.J. Cook, *In Pursuit of the Traveling Salesman*. Princeton: Princeton University Press, USA, 2011.

[3] D. D. Davendra, Traveling Salesman Problem, Theory and Applications. Rijeka: InTech, 2010.

[4] P. Patnaik, *Handbook of Inorganic Chemicals*. McGraw-Hill, 2003.

[5] E. Aarts and J. Korst, *Simulated annealing and Boltzmann machines: A stochastic approach to combinatorial optimization and neural computing*. John Wiley and Sons, 1989.

[6] N.N. Greenwood and A. Earnshaw, *Chemistry of the Elements*. 2nd ed., Elsevier - Butterworth–Heinemann, 1997.

[7] R. H. J. M. Otten and L. P. P. P. Ginneken, *The Annealing Algorithm*. Kluwer Academic Publishers, 1989.

[8] G. Towler and R. Sinnott, *Chemical Engineering Design, Second Edition: Principles, Practice and Economics of Plant and Process Design*. 2nd ed., Butterworth–Heinemann, 2012.

[9] D. Voet, J.G. Voet and C.W. Pratt, *Principles of Biochemistry.* 4th ed., John Wiley & Sons, Inc., 2012.

[10] D.L. Applegate, R. Bixby, V. Chvátal and W. Cook, *The Traveling Salesman Problem: A Computational Study*. Princeton University Press, 2006.

[11] F. A. Cotton, G. Wilkinson, C.A. Murillo and M. Bochmann, M, *Advanced Inorganic Chemistry*. 6th ed., John Wiley & Sons: New York, 1999.

[12] E. Valbahs and P. Grabusts, *Path Planning Using Non - Euclidean Metric, 8th EUROSIM Congress on Modelling and Simulation EUROSIM2013*, Cardiff, Wales, United Kingdom, 10-12 September, 173-178, 2013.

[13] R. Agarwala, D.L. Applegate, D. Maglott, G.D. Schuler and A.A. Schäffer, "A Fast and Scalable Radiation Hybrid Map Construction and Integration Strategy," Genome Research, 10, 350–364, 2000.

[14] R.G. Bland and D.F. Shallcross, "Large traveling salesman problems arising from experiments in x-ray crystallography," Operations Research Letters, 8, 123-133, 1989.

[15] S. Carlson, "The Amateur Scientist: Telescopes and a problem for traveling salesmen in space," Scientific American, 276, 121 – 123, 1997.

[16] S. Climer and W. Zhang, "Rearrangement clustering: Pitfalls, remedies, and applications," Journal of Machine Learning Research, 7, 919–943,2006.
[17] M.A. Fox and K. Wade, "Evolving patterns in boron cluster chemistry," Pure Appl. Chem., 75 (9), 1315–1323, 2003.
[18] V. Granville, M. Krivanek and J-P. Rasson, "Simulated annealing: A proof of convergence," IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(6), 652–656, 1994.
[19] M. Grötschel, M. Jünger and G. Reinelt, "Optimal Control of Plotting and Drilling Machines: A Case Study," Zeitschrift für Operations Research, 35(1), 61-84, 1991.
[20] L. Ingber, "Simulated annealing: Practice versus theory," Math. Comput. Modelling, 18, 29–57, 1993.
[21] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, "Optimization by Simulated Annealing," Science, vol. 220, 671-680, 1983.
[22] E. Kolemen and N.J. Kasdin, "Optimal Trajectory Control of an Occulter-Based Planet-Finding Telescope," Advances in Astronautical Sciences, AAS 07-037, Vol. 128, 215- 233, 2007.
[23] J.K. Lenstra, "Clustering a data array and the traveling salesman problem," Operations Research, 22, 413–414, 1974.
[24] N. Weiner, "Malonic acid. Organic Syntheses," Coll. Vol. 2, 376, 1943.
[25] "Robot scientist Eve will allow scientists to save time and money in the development of new drugs," [Online]. Available:http://www.supersait.org/robot-scientist-eve-will-allow-scientists-to-save-time-and-money-in-the-development-of-new-drugs.html#more-410 [Accessed: February 23, 2015].

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 205-211

Application of the Ontology Concept for the Needs of Theoretical Mechanics

Aivars Vilkaste, Peter Grabusts
Rezekne Higher Education Institution, Faculty of Engineering.
Address: Atbrivosanas aleja 115, Rezekne, LV-4601, Latvia

Abstract. **The concepts of theoretical mechanics have been sufficiently well studied, they are used both for educational purposes and practical application. Nowadays, the use of ontologies is developing rapidly, thus allowing acquiring knowledge about the specific field of application. The author of this research analyses appropriateness of ontologies for the needs of theoretical mechanics. The study shows the use of graphs in the work with taxonomy concepts and describes the key notions of the ontologies. The author has tried to develop the concept of domain ontology for better understanding of the key notions of theoretical mechanics with a help of Protégé, which would be useful for students in the process of studying theoretical mechanics.**

Keywords: **ontology, Protégé, taxonomy, theoretical mechanics.**

I INTRODUCTION

One of the many engineering activities is the research of material objects, i.e. any physical matter that has some mass. Given that all material objects interact with each other and due to this effect they can move in relation to each other, it can be concluded that there are different systems of material objects whose existence and development can be modulated depending on the environment in which the system is functioning or depending on the impact of other systems, individual objects or forces on this system. Theoretical mechanics studies material objects, it explores mechanical movements and mutual positions of material objects and their positions of equilibrium. Theoretical mechanics sees material objects as material points, absolutely solid bodies or their combinations in mechanical system with a certain weight, despite the fact that they differ in shape, size, material and mass. Thus, the object of the study is an abstract model instead of a real physical formation. Theoretical mechanics forms a common theoretical basis for all the technical objects, it is closely related and is developing together with them. Resistance of materials, theory of elasticity, fluctuation theory, machinery and mechanical theory, structural mechanics, theory of plasticity, hydrodynamics, technological machinery theory of various fields and many other theories are based on the general rules of theoretical mechanics. Theoretical mechanics is a basis for general technical preparedness which allows analyzing and evaluating operation of different systems and use of the obtained results, in order to solve various technical problems and tasks.

The studied material objects and systems of these objects are interconnected and interdependent. Interaction of material objects and object-systems is usually very diverse and complex, so there is a need to structure and display a wide range of various information, in order to ensure a common understanding of the studied objects and their systems, as well as to find rational and effective solutions for specific problems or tasks.

II SOME NEEDS OF THEORETICAL MECHANICS

Theoretical mechanics studies mechanical motions of material objects and their positions of equilibrium, as well as reciprocal mechanic impact of material objects.

The following aspects should be taken into account:

1. Object of study is a material point (absolutely solid body with a constant distance between any two points thereof).
2. Mechanical system is a set of any material points that mechanically interact with each other.
3. Mutual interaction of material objects and mechanical systems is characterized by the physical concept of force $\vec{F}$, which is the cause of the speed, direction of a motion or deformation.
4. Material objects and mechanical systems are affected by external forces $\vec{F}_e$ and internal forces $\vec{F}_i$.

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.505

5. Depending on the material object or mechanical system, the same force may be both external and internal.
6. System of forces is formed by several forces that affect the same material object.

It should be also taken into account that any material body in mechanics is seen as a mechanical system consisting of a set of material points. Besides, an absolutely solid body is considered to be a constant mechanical system because the distance between its points always remains the same.

There are various forces in natural environment: Ampere force, Archimedean force, repulsive force, external forces, friction force, lift force, elasticity force, gravitational force, internal forces, Coulomb force, motion stimulating force, motion refraining force, total force, Lorentz force, interaction force, molecular forces, attractive force, repulsive force, reaction force, spring force, gravity, weight, impact force, hauling force, etc. (see Fig. 1) [1, 3].

Fig.1. Forces

Besides, the study covers fundamental and non-fundamental forces, forces of direct and indirect impact, as well as conservative and non-conservative forces, etc. The existing forces have both common and different characteristics and can be grouped by selecting different features, for example grouping according to their performance (see Fig. 2). Some of the forces are studied in theoretical mechanics.

Fig. 2. Distribution of forces by their performance

It can be seen, that when addressing the tasks of theoretical mechanics:

- it is important to accurately determine the type of the force and its compliance with the mechanical action;
- to anticipate the effects of the impact of the force or the system of forces on the material point or mechanical system.

Henceforth, the authors shall explore the needs of the theoretical mechanics by using ontology.

III POTENTIAL OF ONTOLOGIES

Over the past twenty years, the term “ontology” has been introduced and is commonly used in the engineering science, it is used to describe the models with different levels of detail (structuring) in an understandable way, as well as to demonstrate wide and complex information through conceptual schemes. Ontologies allow divided usage of unified understanding of information structure, reuse of the knowledge of problem domain, making clear assumptions in a specific problem domain, separation of knowledge of problem domain from operational knowledge, as well as analysis of the knowledge of problem domain.

The term "ontology" comes from philosophy. In philosophy, ontology is a study of the nature of being (“ontos” – being, “logos” – study), defining characteristics of the existence, where general basics, principles, structure and regularities are described. Each discipline has its own ontology, assumptions about what things exist or may exist in reality, what could be the conditions of their existence, and what they might be dependent on.

Originally, the term “ontology” was introduced in some areas of the research of artificial intelligence: knowledge engineering, natural language processing and knowledge representation. In late 1990's, the term “ontology” became widespread in other areas: intelligent information integration, information retrieval from the Internet, knowledge management, and e-commerce [11]. Ontology is also widely used in semantic webs.

In 1991, R. Neches and his colleges were one of the first to define ontology that is being used in the IT field and may be used in engineering science as well: “An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary” [12].

This definition shows the composition of ontologies and provides some guidelines for building them: identification of basic terms and relations, identification rules for combining terms, provision of definition of terms and relations. Basing on this definition of ontology, it can be concluded that ontology contains not only ontology terms that are precisely defined, but also the knowledge that can be derived from them. T. R. Gruber defines the ontology in the following way: “An ontology is a formal explicit specification of a shared conceptualization” [7]. Ontologies are often equated with taxonomic hierarchies of classes. Thus, the aim of ontologies is to accumulate knowledge in a general and formal way.

Literature sources provide more than ten different definitions of ontology. Although each author offers his definition of ontology, they are all based on the T. R. Gruber’s definition of ontology, which is modified or updated.

Ontologies are often interrelated and are used to create knowledge bases, because "an ontology is a hierarchically structured set of terms for describing a problem domain that can be used as a skeletal foundation for a knowledge base" [11]. V. Devidzic states that "an ontology is a set of terms and relations where all concepts are defined and interpreted in a declarative way" [9]. M. Uschold and R. Jasper define that "an ontology may take a variety of forms, but necessarily it will include a vocabulary of terms, and some specification of their meaning. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the possible interpretations of terms" [15].

When analyzing definitions of ontologies, N. Guarino and other scientists [13] note than an ontology should be "human-understandable and machine-readable", as well as "a shared view on conceptualization of the problem domain".

Ontologies can be classified in different forms. One of the most popular forms of classification is the one proposed by Guarino, who classified types of ontologies according to their level of dependence on a particular task or point of view [6]:

- *Top-level ontologies*: describe general concepts like space, time, event, which are independent of a particular problem or domain.
- *Domain-ontologies*: describe the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology.
- *Task ontologies*: describe the vocabulary related to a generic task or activity by specializing the top-level ontologies.
- *Application ontologies*: they are the most specific ones. Concepts often correspond to roles played by domain entities. They have a limited reusability as they depend on the particular scope and requirements of a specific application.

V. Solovyov, B. Dobrov, V. Ivanov and V. Lukasevic have formulated basic characteristics of ontologies [4]:

- at logical level, each ontology complies with a specific theory, but sometimes there is a specific fixed model;
- ontologies are created basing on the module principle: when identifying a new ontology, the previous ontologies can be used;
- ontologies have to be easily understandable and usable.

B. Dobrov and V. Lukashevic have found a formal designation for ontologies [4]:

$$O = \{X, R, F\} \quad (1)$$

where O – stands for ontology of domain, X – a set of domain concepts, R – a set of relations of these concepts, F – a set of function interpretations of concepts and relations between them.

V. Graudina in her article on ontologies [11] has analyzed 12 different ontology classification schemes and has elaborated a scheme classification. Ontologies may be classified in accordance with their expressiveness, weight, degree of formalization, or in line with the generality of ontologies, generality of the studied problem, creation process and the reflected worldviews. Taking into account the classification by the generality of ontology, we can talk about ontology classification by dependence on a specific application or task, by asking ontological knowledge of languages, possibilities of knowledge visualization and other parameters. One of the types of ontologies are application ontologies that depend on the specific application, objectives and tasks, as well as the methods for their solution, they contain the knowledge needed for the specific applications. Application ontologies describe the concepts that depend on the task ontology and on the domain ontology.

Methods and methodology for the creation of ontologies are studied by the ontology engineering. It describes the set of designing principles, development processes and activities, as well as supports technology and systematic methodologies, which in turn facilitate the development and use of the complete cycle of ontologies, i.e. designing, implementation, evaluation, validation, maintenance, distribution, mapping, integration, sharing and reuse. Ontology engineering aims to make explicit the knowledge and procedures contained in applications. It offers guidelines for solving interoperability problems caused by semantic barriers. Ontology engineering provides [10]:

1) Determination of the concept class in the ontology.
2) Implementation of taxonomy in classes (subclass - class - superclass).
3) Elaboration of a structure of concepts and situations.
4) Determination of characteristics and the value of these characteristics.
5) Procedures for conclusions and situation transformation.

N. F. Noy and D. L. McGuinness [8] offer seven steps for creation of ontologies, however, they do not consider that this is the only correct method for creation of ontologies. Ontology building process is iterative, modification and supplementation take place throughout the whole building process. However, before the creation of an ontology, the author suggests to answer some basic questions:

1) What is the domain the ontology will cover?
2) What are the aims of using the ontology?
3) What are the types of questions that should be answered with a help of the ontology?
4) Who will use and maintain the ontology?

The answers to these questions may change during the ontology designing process, but they always help limiting the scope of the model.

Step 1. Determine the domain and scope of the ontology.

Step 2. Consider reusing existing ontologies.

Step 3. Enumerate important terms in the ontology.

Step 4. Define the classes and the class hierarchy.

Step 5. Define the properties of classes – slots.

Step 6. Define the facets of the slots.

Step 7. Create instances.

Taking into account the definition, classification of ontologies and their building methodology, it can be stated that the concept of ontology may be applied in order to study the mechanical movements of material objects and their positions of equilibrium, as well as reciprocal position of material objects, which is the task of theoretical mechanics.

Thus, an ontology is an explicit representation of knowledge. It is a formal, explicit specification of shared conceptualizations, representing the concepts and their relations that are relevant for a given domain of discourse [7].

Ontologies are widely applied in the Semantic Web, but there is no information about the use of ontologies in theoretical mechanics.

Considering the fact that the ontology should be "human-understandable and machine-readable" [6], as well as "easy understandable and usable" [4], the author shall show the ontology building process in two ways. Firstly, the use of graphs and graph theory, because information that is correctly visually encoded is more easy to understand and process than any other type of information. Secondly, use of the open-source platform Protégé [14], which allows building models of knowledge in a given field.

IV USE OF GRAPHS IN TAXONOMY DEFINING

Graph is a system that consists of a finite number of objects of two types (vertices and edges), and each edge is connected with two vertices. Graph may be drawn by depicting its verticals with nodes, while each edge is depicted by an arbitrary line connecting the relevant vertices (see Fig. 3).

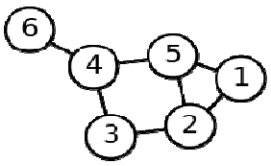

Fig. 3. Graph with 6 vertices and 7 edges

Counting and implementation of graph models make the analysis and solution of tasks demonstrative and effective [2], which is important when building application ontology. Visualization of relations by graph helps analyzing them both theoretically and experimentally, by using information technologies [2].

It should be noted that the graph concept can be generalized in the following way: subsets of the set of graph edges are taken as graph edges, besides these subsets (edges) can reoccur. In this case, we obtain a hypergraph that is an incidence system of objects (hypergraph vertices) and a finite system consisting of non-empty sets of these objects (faces or edges of the hypergraph) (see Fig. 4).

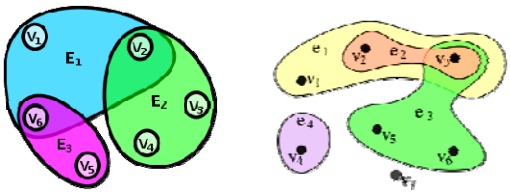

Fig. 4. Hypergraph models

Mathematical description of a graph is the following: graph is a pair of sets $G = (V; E)$, where elements of the set V are vertices $V = \{v_i\}$, $i \in I = \{1, 2, \dots, n\}$, while elements of the set E are edges that connect the vertices $E = \{e_j\}$, $j \in J = \{1,2, \dots, m\}$; each edge is a subset of the set of vertices V. $E \subset V \times V$, where $V \times V$ is Descartes multiplication of sets and each edge is a pair (v_1, v_2) or (v_2, v_1). Definition of hypergraph is similar: hypergraph is a pair of sets $H = (V; E)$, where $V \neq \{\emptyset\}$ and $E \subseteq P(V)$. Edges of the hypergraph (elements of the set E) correspond to arbitrary subsets of the set V of vertices.

The graph theory defines orientated and non-orientated hypergraphs and, traditionally, both vertices and edges are colored in a non-orientated graph [2]; it assigns positive integers to vertices or edges or colors vertices and edges of the graph. This principle may be applied to hypergraphs too, as it is a generalization of the graph concept. A colored, partly-orientated hypergraph is called semantic hypergraph that shall be used for ontology building. The above mentioned facts allow defining ontology as a semantic hypergraph:

$$H = O(X, R, I) \qquad (2)$$

where X is a set of concepts (vertices), R – a set of relations (edges), and I – a set of predicates that define the incidence (connectivity) in the hypergraph.

Thus, the model of graph theory may be used for demonstrative implementation of taxonomy in classes (subclass – class – superclass).

V POSSIBILITIES OF ONTOLOGY IN THE PROCESS OF DEFINING THE CONCEPTS OF THEORETICAL MECHANICS

When addressing the problems of theoretical mechanics, it is important to determine the type of the force and its compliance with the mechanical action.

In view of this, the author shall demonstrate a fragment of the ontology by using a model of graph theory.

When creating an ontology of forces, all existing forces may be seen as its elements, however it should be taken into account that the theoretical mechanics studies interaction of material objects and mechanical systems. Then, when creating the ontology, the author shall demonstrate the direct and indirect forces. There shall be two subclasses ("Contact forces" and "Action - at - a - Distance Forces") in the class named "Forces". The author shall demonstrate the relations between the concept classes, subclasses and their elements in a form of organized taxonomy. For the needs of taxonomy, the relations *A_kind_of* shall be used. The fragment of taxonomy for the ontology "Force" can be seen in the Figure 5.

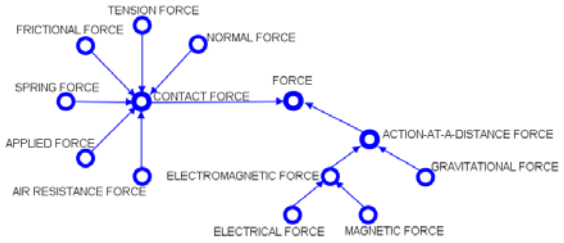

Fig. 5. Taxonomy fragment

It should be noted that one object may contain a part of another object. For example, material object is a part of mechanical system, which can also be seen as a material object, the weight of the material object is part of the friction force. This type of relations is named *Part_of*, and it is sufficient for such concepts to list their components. In many cases, it is important to show the relations between the components. In such cases, the structure concepts are often used and the relations *Have_structure* are applied to them. One of the elements of the subclass *Contact forces* is *Frictional Force*, whose structure may be demonstrated in a form of semantic web by using a hypergraph (see Fig. 6).

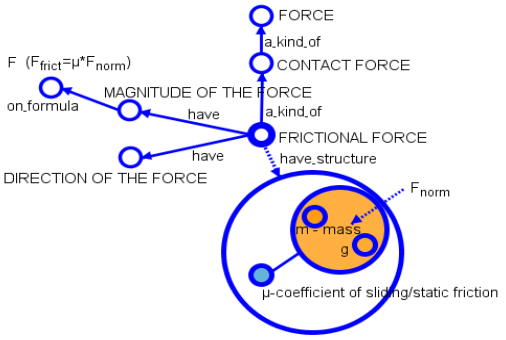

Fig. 6. The concept of *Frictional Force* structure fragment

It can be seen that the model of graph theory used for the study makes the information needed for the theoretical mechanics of ontologies easily perceivable and understandable.

Basing on the assumptions about the needs of tasks of theoretical mechanics, it would be useful to develop a general approach to the framework that could be helpful for this type of task class (see Fig. 7) [5].

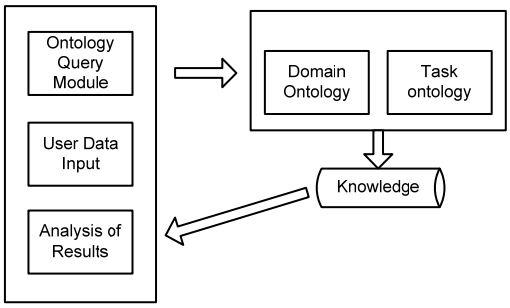

Fig. 7. Prototype of ontology-based framework for the needs of tasks of theoretical mechanics

Developing framework Protégé OWL tool is used for construction of this prototype [14]. Protégé is a special tool, which is thought to create and edit ontology, but OWL (Web Ontology Language) is a language through which it is possible to define the ontology. OWL ontology may include descriptions of classes, their characteristics and their instances. OWL formal semantics describes how, by using these data, to get information which was not openly described in ontology, but which is derived from the data semantics.

Protégé is a free, open source ontology editor and a knowledge acquisition system. Protégé provides a graphic user interface to define ontologies. In Protégé environment a number of knowledge-modeling structures and actions that support ontology creation, visualization and editing of different display formats are implemented. It also includes deductive classifiers to validate that models are consistent and to infer new information based on the analysis of ontology.

Protégé is an extensible knowledge model. Protégé's primitive - the component of its knowledge model - provide classes, instances of these classes, frame representing attributes of classes and instances.

Ontology development with the help of Protégé starts with the definition and description of classes hierarchy, and then instances of these classes and different types of relationships (properties in Protégé) in order to put more meaningful information within the ontology are assigned [5].

Looking at the possibilities of applying ontologies for the needs of theoretical mechanics, at the beginning of the research the author studied only the domain ontology. In order to demonstrate it, similarly to using the model of graph theory, the class "Forces" was chosen (see Fig. 8 and Fig. 9).

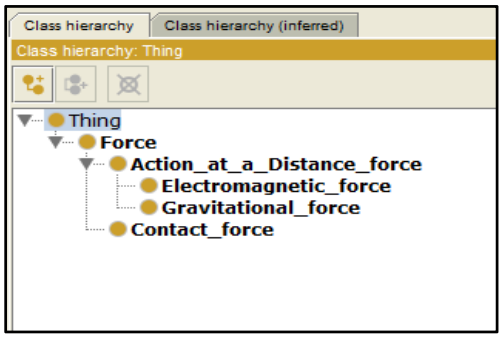

Fig. 8. Subclasses of the class „Force" in the „Class hierarchy" tab of Protege

It is assumed that a class "Force" has been introduced and that this class contains two subclasses *Contact force* and *Action – at – a- Distance force*.

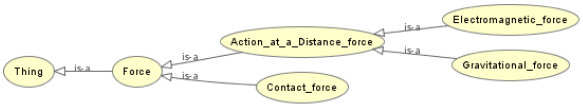

Fig. 9. Visualization of subclasses of the class *Force* in OWLViz tab of Protégé

The subclass *Contact force* contains the following elements: *Applied force*, *Air resistance force*, *Normal force*, *Spring force*, *Frictional force*, *Tension force*. The subclass *Action – at – a – Distance force* contains *Gravitational force* and *Electromagnetic force*. The subclass *Electromagnetic force* contains *Electrical force* and *Magnetic force* (see Fig. 10).

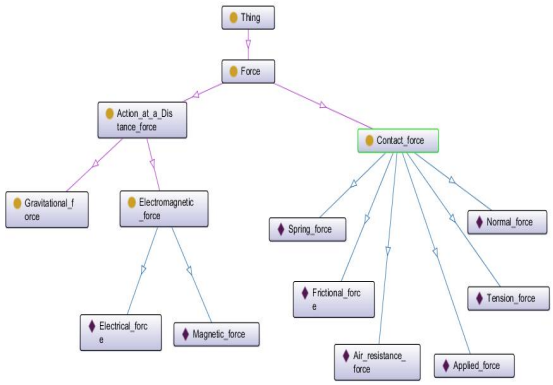

Fig. 10. Visualization of subclasses of the class *Force* in OntoGrapf tab of Protégé

This example demonstrates that, by using the Protégé, it is possible to create an efficient description of ontologies that may be used for the needs of theoretical mechanics. It can be seen that it is possible to quickly identify the forces that mechanically and directly affect material objects or mechanical systems. Ontology building is a laborious process. For the needs of theoretical mechanics, it is possible to create many coherent ontologies. Due to the fact that the creation of ontologies of theoretical mechanics is in the early stage, the work on building and developing of ontology of theoretical mechanics concepts is to be continued.

VI CONCLUSION

Process of development of any ontology is iterative, modification and supplementation is an inevitable part of it. In the future, one of the possible new directions needed for the domain ontology of theoretical mechanics might be related to introduction of new concepts, definition and description of new classes, formation of relations between the concepts and their characteristics.

The results of the study show that, for the forces and their systems studied in theoretical mechanics, it is possible to build various ontologies that would exist and operate simultaneously. The abovementioned examples reveal that, by using ontologies, is possible to demonstrate the links both between material objects and forces, as well as between objects and forces that operate in mechanical systems. It is important to make the newly-created ontologies matching, thus creating a common approach to the needs of theoretical mechanics.

It can be seen that the use of ontology helps us to understand the concepts (forces in this case), their characteristics and existing relations, that create awareness of the forces affecting material objects and their systems, causing their movement or immobility. It provides an efficient solution for the tasks of theoretical mechanics.

The results of this study shall be useful for students while studying theoretical mechanics, they shall also provide understanding of the basic concepts of theoretical mechanics and ontologies.

VII REFERENCES

[1] J.S. Walker, *Physics*. Upper Saddle River, NJ, Pearson / Prentice Hall, 2004.

[2] J.L.Gross and J.Yellen, *Graph theory and its applications*. Boca Raton, Landon, N.Y. Chapman & Hall / CRC, 2006, pp. 2 – 52, 534 – 540.

[3] D. Kleppner and R. Kolenkow, *An Introduction to Mechanics*. 2nd ed.,Cambridge University Press, 2014, pp. 81 – 110.

[4] V.Solovjov, B.Dobrov, V.Ivanov, and N.Lukashevich, *Ontologies and thesauri*. Moskwa, 2006. (In Russian)

[5] P. Grabusts, *The Concept of Ontology for Numerical Data Clustering*, *9th International Conference "Environments. Technology. Resources"*, Rezekne, Latvia, 2013.

[6] N. Guarino, *Formal Ontology in Information Systems, 1st International Conference on Formal Ontology in Information Systems, FOIS*, Trento, Italy, IOS Press, pp.3 – 15, 1998.

[7] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Systems Laboratory Computer Science Department Stanford University Stanford, California, Tech. Rep. KSL 92 – 71, 1993.

[8] N.F.Noy and D.L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford knowledge systems laboratory, Stanford University, Stanford, California, Tech. Rep. KSL-01-05, 2001.

[9] V. Devedzic, "Understanding Ontological Engineering," Communications of ACM, vol.45, no.4, pp.136. – 144, April 2002.

[10] A.Gladun and J.Rogushina, "Application of the ontology for enterprises system," Journal "Corporate Systems" , No1, 2006. (In Russian)

[11] V. Graudiņa, "Variety of Definitions and Classifications of Ontologies," Scientific Journal of Riga Technical University, vol.22, pp. 181- 192, 2005.

[12] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. Swartout, "Enabling technology for knowledge sharing," AI Magazine, vol.12, no. 3, pp. 36 – 56, 1991.

[13] N.Guarino and P. Giaretta, "Ontologies and Knowledge Bases Towards a Terminological Clarification", [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.320.8006&rep=rep1&type=pdf [Accessed: Feb.15, 2015].

[14] Protégé project homepage, [Online]. Available: http://protege.stanford.edu/products.php [Accessed: March 30, 2015].

[15] M. Uschold and R. Jasper, "A Framework for Understanding and Classifying Ontology Applications," [Online]. Available: http://www.cs.man.ac.uk/~horrocks/Teaching/cs646/Papers/uschold99.pdf [Accessed: March 3, 2015].

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 212-216

Analytic-Numerical Modeling and Investigation of Nanostructures' Dynamics on Material Surfaces After Laser Irradiation

Uldis Žaimis[1, a], Sharif E. Guseynov[1, 2, b]
[1]*Institute of Mathematical Sciences and Information Technologies, Liepaja University, 14 Liela Street, Liepaja LV-3401, Latvia;*
[2]*Institute of Aeronautics, Faculty of Transport and Mechanical Engineering, Riga Technical University, 1A Lomonosov Street, Building-1, Riga LV-1019, Latvi*a
[a]*uldiss.z@inbox.lv*, [b]*Sh.E.Guseinov@inbox.lv*

Abstract. **The work is devoted to modeling the formation and behavior of solid nano-sized particles on the surface of materials. In the simulation, it is assumed that the main processing technology of surface nanostructures is laser irradiation, which causes the Brownian motion of nanoparticles, due mainly to thermal fluctuations: if the temperature around the nanoparticles is uniformly distributed, the time average of the Brownian fluctuations is zero; however, if there is a temperature gradient around the nanoparticles, the thermal fluctuations affect the nanoparticle in different ways from different sides, and there is a force like the thermophoretic force, biasing the average position of the nanoparticle. When building a 1D model of the formation and flow behavior of nanoparticles, three important assumptions are introduced: the impact of nanoparticles on the process of irradiation is negligible; the impact of nanoparticles on each other as compared to the effect of laser irradiation on them is also negligible; and nanoparticles after laser irradiation can move both forward and backward and at every fixed period of time, moving the nanoparticles does not impose any steric constraints. Under the above assumptions, a 1D continuous model is built, implicit and explicit finite difference schemes to solve it are developed; their convergence and order of convergence are studied; an output condition ensuring the stability of the explicit difference scheme is obtained, the unconditional stability of implicit difference scheme is proven, and software for computer implementation of some of the obtained analytical and numerical results developed.**

Keywords: **Surface nanostructures, laser irradiation, mathematical modeling, single-phase model, explicit and implicit finite difference schemes**

I INTRODUCTION

Two of the most important trends in the field of modern nanotechnology are the development of new high-precision, low-cost methods for the synthesis of nanostructures in terms of dissipative systems and the development of decision analytical models that reflect the fundamental properties of such systems and processes. Self-organized formation of nanostructures (e.g., surface nanostructures in semiconductors) is possible in the given circumstances and supported from outside the dissipative thermodynamic process (for example, molecular beam epitaxy). Spontaneous surface nanostructures (regularly arranged pores, cracks, "shoots," protruding islands, pyramidal or conical shapes, and the like) were also observed after optical laser irradiation [1]. This method of thermal effect on the surface of some semiconductor materials has the fundamental properties of dissipative methods of the synthesis of nanomaterials. The following points are important:

1. In all cases of the synthesis, the self-organization (self-assembly) of atoms (molecules) in a regular structure is observed and due to various factors including the instant phase transition under conditions of extreme temperature and the grouping of impurity atoms in the liquid phase at the surface of the plate.
2. Self-organized structures exhibit all the signs of the nano-objects in mind, registering a quantum size effect in all cases of synthesis.
3. The system of irradiation synthesis, in our opinion, has the highest order of self-

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.193

organization, other than the epitaxial growth system, since it "rights" chemical composition, thickness and the pattern of each layer defined in-system, and it cannot be programmed from the outside at each step.

However, the fundamental limitation of laser irradiation to date is the lack of developed techniques of controlled fusion. In other words, as a result of nanorelief, laser radiation is generated in an unpredictable manner. Based on the data input process (type of material, shape and intensity of the laser pulse and the initial temperature of the material), the nature of the future "picture" cannot be rightly determine. It should also be noted that the effective search control mechanisms of irradiation synthesis are very difficult in the absence of analytical and computational models that describe the key structural properties of the dissipative system.

II CONSTRUCTIONS OF MATHEMATICAL MODELS

A. Analytical model with the Dirichlet boundary conditions, and analysis of model

Consider the following 1D quasi-linear model:

$$\begin{cases} u_t(x,t)+\varepsilon\cdot u_{tt}(x,t)=\nabla\left(k\left(u(x,t)\right)\cdot\nabla u(x,t)\right)+f(x,t), \\ 0<x<l,\ 0<t\le T<\infty, \\ \left\{u(x,t)\big|_{t=0+0}=u_0(x),\ 0\le x\le l,\right. \\ \left\{u_t(x,t)\big|_{t=0+0}=u_1(x),\ 0\le x\le l,\right. \\ \left\{u(x,t)\big|_{x=0+0}=u_2(t),\ 0\le t\le T,\right. \\ \left\{u(x,t)\big|_{x=l-0}=u_3(t),\ 0\le t\le T,\right. \end{cases} \quad (1)$$

where $0<k\left(u(x,t)\right)\in C^1\left\{(0,l)\times(0,T]\right\}$; $0<f(x,t)\in C\left\{(0,l)\times(0,T]\right\}$; $u_0(x)\in C[0,l]$; $u_1(x)\in C[0,l]$; $u_2(t)\in C[0,T]$; $u_3(t)\in C[0,T]$ are given functions of their arguments; ε, l and T are constants. For $0<T\equiv const.<\infty$, there is a time during which we investigate the dynamics of solid particles on the surface of a nano-sized material after laser irradiation. The first equation in (1) describes the temperature field of a physical system consisting of a one-dimensional (in general, non-homogeneous) material rod of length l. The physical meanings of the other model variables are:

- $k\left(u(x,t)\right)$ is the thermal conductivity of the material rod;
- $0<\varepsilon\equiv const.\ll 1$ is a control parameter of the model.

The coefficient of thermal conductivity $k\left(u(x,t)\right)$ has positive temperature dependence, thus creating a positive feedback as a thermodynamic system under consideration (the so-called blow-up regime). To simplify the analysis, we can fix it as a constant, $k\equiv const.>0$, or a simpler function such as, $k\left(u(x,t)\right)=k(x)>0\ \ \forall x\in[0,l]$. Mathematical models often use similar systems of the indicated ratios if the value is changed in a relatively small range. In this case we have:

$$\begin{cases} u_t(x,t)+\varepsilon\cdot u_{tt}(x,t)=\nabla\left(k(x)\cdot\nabla u(x,t)\right)+f(x,t), \\ 0<x<l,\ 0<t\le T<\infty, \\ \left\{u(x,t)\big|_{t=0+0}=u_0(x),\ 0\le x\le l,\right. \\ \left\{u_t(x,t)\big|_{t=0+0}=u_1(x),\ 0\le x\le l,\right. \\ \left\{u(x,t)\big|_{x=0+0}=u_2(t),\ 0\le t\le T,\right. \\ \left\{u(x,t)\big|_{x=l-0}=u_3(t),\ 0\le t\le T,\right. \end{cases} \quad (2)$$

where $u_0(x)\in C[0,l]$; $u_1(x)\in C[0,l]$; $u_2(t)\in C[0,T]$; $u_3(t)\in C[0,T]$; $0<k\left(u(x,t)\right)\in C^1\left\{(0,l)\times(0,T]\right\}$ are given functions of their arguments; ε, l and T are constants. In the proposed analytical model, the control parameter ε is the most interesting option from the perspective of current research in the field of nanostructure formation during irradiation. There is a hypothesis that the value ε depends on the irradiation mode (the frequency, power and pulse shape of the laser) and the physiochemical parameters of the irradiated material. In general, it is an unknown function of some physical and chemical quantities. Varying ε during the numerical experiment, we can indirectly simulate the formation of nanostructures in various conditions whose presence in the analytical model is not yet possible to specify in any explicit analytical way.

Thus, the 1D equation in the model is more general than the classical equation of mass and heat transfer and diffusion considered, for example, in [2]. The opportunity to study the processes of formation of nanostructures in a numerical experiment which more realistically reproduces the conditions of laboratory experience is the motivation for the inclusion of the term $\varepsilon\cdot u_{tt}(x,t)$ in the model.

Let us analyze the dissipative properties of the model (1). Differential equation models include $f(x,t)>0$ and therefore characterize the open thermodynamic system. It can be argued that the term u_t simulates the infinite growth of the temperature field in this system, "creating" extreme temperature conditions – a necessary condition for the instantaneous phase transitions with an increase in surface nanostructures. The term $\varepsilon\cdot u_{tt}(x,t)$ describes the process of decay, the scattering of thermal energy in the system, preventing it from overheating and thus its destruction. The expression $u_{tt}(x,t)$ can be interpreted as a heat wave. The control parameter ε must cause asymmetry in the decision.

B. Model with the Neumann boundary conditions

The boundary conditions in the proposed model (1) and (2) are the Dirichlet boundary conditions, i.e.,

boundary conditions of the first type. The potential problem of the boundary conditions is due to the fact that the collection of the required source data may be problematic with the lack of electronic measuring instruments for recording the speed of flow of nanoparticles; $grad\left(u(x,t)\right)$. Dirichlet boundary conditions for these models require raw data in the form of the quantity and not the flow of the nanoparticles. In the case where the problem of recording the velocity of the flow is technically easy to implement, it is guided by the methods described, for example, in [3], as boundary conditions are preferable to be set by the Neumann condition (i.e., the boundary conditions of the second type), rather than the Dirichlet conditions. In this case, we could consider the models (1) and (2) with Neumann boundary conditions:

$$\begin{cases} u_t(x,t)+\varepsilon\cdot u_{tt}(x,t)=\nabla\left(k\left(u(x,t)\right)\cdot\nabla u(x,t)\right)+f(x,t), \\ 0<x<l,\ 0<t\le T<\infty, \\ \begin{cases} u(x,t)\big|_{t=0+0}=u_0(x),\ 0\le x\le l, \\ u_t(x,t)\big|_{t=0+0}=u_1(x),\ 0\le x\le l, \end{cases} \\ \begin{cases} u_x(x,t)\big|_{x=0+0}=u_2(t),\ 0\le t\le T, \\ u_x(x,t)\big|_{x=l-0}=u_3(t),\ 0\le t\le T, \end{cases} \end{cases} \tag{3}$$

where $0<k\left(u(x,t)\right)\in C^1\left\{(0,l)\times(0,T]\right\}$; $u_0(x)\in C[0,l]$; $u_1(x)\in C[0,l]$; $u_2(t)\in C[0,T]$; $u_3(t)\in C[0,T]$; $0<f(x,t)\in C\left\{(0,l)\times(0,T]\right\}$ are given functions of their arguments and ε, l and T are constants. In the case of $k\left(u(x,t)\right)=k(x)>0\ \ \forall x\in[0,l]$, we do the same as in (2).

III CONSTRUCTION OF THE DISCRETE MODELS AND ANALYSIS

A. Development and investigate of explicit difference scheme

We construct stepped grid for spatial variable $x\in[0,l]$:

$$\Omega_x^{(1)}\overset{def}{\equiv}\left\{x_j:\ x_j=j\cdot h_x,\ 0\le j\le N,\ h_x=\frac{l}{N},\ N\in\mathbb{N}\right\},$$

$$\Omega_x^{(2)}\overset{def}{\equiv}\left\{x_{j+\frac12}:\ x_{j+\frac12}=\left(j+\frac12\right)h_x,\ 0\le j<N,\ h_x=\frac{l}{N},\ N\in\mathbb{N}\right\}.$$

Next, we construct regular grid for time variable $t\in[0,T]$:

$$\Omega_t\overset{def}{\equiv}\left\{t^n:\ t^n=n\cdot h_t,\ 0\le n\le M,\ h_t=\frac{T}{M},\ M\in\mathbb{N}\right\}.$$

Now entering the discrete function $U_j^n=u\left(x_j,\ t^n\right)$, we can approximate the model (2) in a discrete area in the following specific manner:

$$\frac{U_j^{n+1}-U_j^n}{h_t}=-\frac{V_{j+\frac12}^{n+1}-V_{j-\frac12}^{n+1}}{h_x}+F_j^n, \tag{4}$$
$$j=\overline{1,(N-1)},\ n=\overline{0,M};$$

$$\varepsilon\cdot\frac{V_{j+\frac12}^{n+1}-V_{j-\frac12}^{n}}{h_t}+V_{j+\frac12}^{n+1}=-k_{j+\frac12}^{n}\cdot\frac{U_{j+1}^n-U_j^n}{h_x}, \tag{5}$$
$$j=\overline{1,(N-1)},\ n=\overline{0,M};$$

$$\begin{cases} U_j^0=u_{0,j},\ \ \forall j=\overline{0,N}, \\ U_j^1=h_t\cdot u_{1,j}+u_{0,j},\ \ \forall j=\overline{0,N}; \end{cases} \tag{6}$$

$$\begin{cases} U_0^n=u_2^n,\ \ \forall n=\overline{0,M}, \\ U_N^n=u_3^n,\ \ \forall n=\overline{0,M}, \end{cases} \tag{7}$$

where $V_i^m=k\left(x_i\right)\cdot\left.\frac{\partial u\left(x,t^m\right)}{\partial x}\right|_{x=x_i}$, i.e. V_i^m is an approximation of the function $k(x)\cdot\frac{\partial u(x,t)}{\partial x}$; $F_j^n=f\left(x_j,t^n\right)$.

If it is possible to use the second type of boundary conditions we could consider the model (2) with the boundary conditions of the second kind (i.e., the Neumann conditions):

$$\begin{cases} \left.\frac{\partial \mathrm{u}(\mathrm{x},\mathrm{t})}{\partial \mathrm{x}}\right|_{\mathrm{x}=0}=\mathrm{u}_2(\mathrm{t}),\ 0\le \mathrm{t}\le \mathrm{T}, \\ \left.\frac{\partial \mathrm{u}(\mathrm{x},\mathrm{t})}{\partial \mathrm{t}}\right|_{\mathrm{x}=1}=\mathrm{u}_3(\mathrm{t}),\ \ 0\le \mathrm{t}\le \mathrm{T}. \end{cases}$$

For this case an approximation is of the form:

$$\begin{cases} \frac{\mathrm{h_x}}{2}\times\overset{\circ}{\mathrm{U}}_{\mathrm{t0}}+\overset{\circ}{\mathrm{U}}_0^{\,\mathrm{n+1}}=\mathrm{u}_2^{\mathrm{n}},\quad \mathrm{n}=0,1,...,\mathrm{M}, \\ \frac{\mathrm{h_x}}{2}\times\overset{\circ}{\mathrm{U}}_{\mathrm{tN}}+\overset{\circ}{\mathrm{U}}_{\mathrm{N}}^{\,\mathrm{n+1}}=\mathrm{u}_3^{\mathrm{n}},\quad \mathrm{n}=0,1,...,\mathrm{M}, \end{cases}$$

where we have the following notation:

$$\begin{cases} \overset{\circ}{U}_{ta}\overset{def}{\equiv}\frac{U_{j(=a)}^{n+1}-U_{j(=a)}^n}{h_t} \qquad \left(a=\{0;N\}\right), \\ \overset{\circ}{U}_a^{\,n+1}\overset{def}{\equiv}\frac{U_{j(=a)+1}^{n+1}-U_{j(=a)}^{n+1}}{h_x} \qquad \left(a=\{0;N\}\right). \end{cases}$$

We return to the difference scheme (4)-(7). Substituting (5) into (4) and using the new designations

$$U_t\overset{def}{\equiv}\frac{U_j^{n+1}-U_j^n}{h_t};$$

$$U_{\bar t}\overset{def}{\equiv}\frac{U_j^n-U_j^{n-1}}{h_t};$$

$$U_{\bar x}\overset{def}{\equiv}\frac{U_j-U_{j-1}}{h_x};$$

$$U_x \overset{def}{\equiv} \frac{U_{j+1}-U_j}{h_x},$$

we obtain the following explicit finite difference scheme for the mathematical model (2):

$$\begin{cases} \varepsilon \cdot U_{\overline{tt}}^n + U_t^n = \partial_x\left(k \cdot \partial_{\overline{x}} U^n\right) + F_j^n, \quad j = 1,2,...,N-1; \\ \\ \begin{cases} U_j^0 = u_{0,j}, \quad j = 1,2,...,N; \\ U_j^1 = h_t \cdot u_{1,j} + u_{0,j}, \quad j = 1,2,...,N; \end{cases} \\ \begin{cases} U_0^n = u_2^n, \quad n = 0,1,...,M; \\ U_N^n = u_3^n, \quad n = 0,1,...,M \end{cases} \end{cases} \tag{8}$$

The cconvergence and stability of the scheme are not presented here due to limited space.

B. Development and investigate of implicit difference scheme

In this section, an implicit difference scheme is developed on a discrete area $\Omega_t \times \left\{\Omega_x^{(1)} \times \Omega_x^{(2)}\right\}$ for the approximation of the mathematical model (2), where nets Ω_t, $\Omega_x^{(1)}$ and $\Omega_x^{(2)}$ are constructed as in the previous section:

$$\Omega_x^{(1)} \overset{def}{\equiv} \left\{ x_j : \; x_j = j \cdot h_x,\, 0 \le j \le N,\, h_x = \frac{l}{N},\, N \in \mathbb{N} \right\};$$

$$\Omega_x^{(2)} \overset{def}{\equiv} \left\{ x_{j+\frac{1}{2}} : \; x_{j+\frac{1}{2}} = \left(j+\frac{1}{2}\right)h_x,\, 0 \le j < N,\, h_x = \frac{l}{N},\, N \in \mathbb{N} \right\};$$

$$\Omega_t \overset{def}{\equiv} \left\{ t^n : \; t^n = n \cdot h_t,\, 0 \le n \le M,\, h_t = \frac{T}{M},\, M \in \mathbb{N} \right\}.$$

As in Section 3.1, where we approximate mathematical model (2), we obtain:

$$\frac{U_j^{n+1}-U_j^n}{h_t} = -\frac{V_{j+\frac{1}{2}}^{n+1} - V_{j-\frac{1}{2}}^{n+1}}{h_x}, \tag{9}$$

$$\forall j = \overline{1,(N-1)}; \quad \forall n = \overline{0,M};$$

$$\varepsilon \frac{V_{j+\frac{1}{2}}^{n+1} - V_{j+\frac{1}{2}}^{n}}{h_t} + V_{j+\frac{1}{2}}^{n+1} = -k_{j+\frac{1}{2}}^{n+1} \frac{U_{j+\frac{1}{2}}^{n+1} - U_j^{n+1}}{h_x} + F_j^n, \tag{10}$$

$$j = \overline{0,(N-1)}; \quad n = \overline{0,M};$$

$$\begin{cases} U_j^0 = u_{0,j}, \quad \forall j = \overline{0,N}, \\ U_j^1 = h_t \cdot u_{1,j} + u_{0,j}, \quad \forall j = \overline{0,N}; \end{cases} \tag{11}$$

$$\begin{cases} U_0^n = u_2^n, \quad \forall n = \overline{0,M}, \\ U_N^n = u_3^n, \quad \forall n = \overline{0,M}, \end{cases} \tag{12}$$

Substituting (9) with (11) into (10), we obtain the following implicit, finite difference scheme approximating the continuous mathematical model (2):

$$\begin{cases} \varepsilon \cdot U_{\overline{tt}}^n + U_t^n = \partial_x(k \cdot \partial_{\overline{x}} U^{n+1}), \\ \forall j = 1,2,...,N-1; \\ \begin{cases} U_j^0 = u_{0,j}, \quad j = 0,1,...,N; \\ U_j^1 = h_t \cdot u_{1,j} + u_{0,j}, \quad j = 0,1,...,N; \end{cases} \\ \begin{cases} U_0^n = u_2^n, \quad n = 0,1,...,M; \\ U_N^n = u_3^n, \quad n = 0,1,...,M. \end{cases} \end{cases} \tag{13}$$

C. The main properties of the constructed difference schemes

- The developed explicit difference scheme (4) - (7), which approximates the mathematical model (2), is easier to implement with a computer than the developed implicit difference scheme (9) - (12) approximating the same mathematical model;
- The order of convergence of the two difference schemes is the same. The two difference schemes' estimate in the energy norm is $\left\|u\left(t^n\right) - U^n\right\|_\infty \le Const \cdot \left(h_t + h_x^2\right)$, for each of which is developed its own *Const* and only these constants are different, not the orders of convergence;
- The explicit difference scheme (4) - (7) is a conditionally stable difference scheme, and for its stability it is sufficient to choose the steps h_x and h_t, respectively of a spatial grid $\Omega_x^{(1)}$ and regular grid Ω_t such that:
  - when $k(x) \equiv const, \; \forall x \in (0,l)$, the inequality $\left\|u\left(t^n\right) - U^n\right\|_\infty \le \overline{C} \cdot \left(h_t + h_x^2\right)$, called a sufficient condition for stability of (4) - (7) holds, where $\overline{C}$ is some constant, and $\|\bullet\|_\infty$ is the energy norm, determinated as $\|A\|_\infty \overset{def}{\equiv} \max\limits_{0 \le j \le N} |A_j|$;
  - when $k(x) \not\equiv const, \; \forall x \in (0,l)$, the inequality $B > \frac{1}{4} \cdot C$ holds, where $B \overset{def}{\equiv} \frac{\varepsilon + \frac{h_t}{2}}{h_t^2} \cdot E$ and $CU^n \overset{def}{\equiv} -\partial_x(k \cdot \partial_{\overline{x}} U^n)$;
- The implicit difference scheme (9) - (12) is an unconditionally stable difference scheme for the first case when $k(x) \equiv const, \; \forall x \in (0,l)$, and for the second case, when $k(x) \not\equiv const, \; \forall x \in (0,l)$.

Otherwise, for the computer implementation of the implicit difference scheme (9) - (12), steps h_x and h_t

respectively of a spatial grid $\Omega_x^{(1)}$ and regular grid Ω_t can be chosen arbitrarily, guided only by the consideration of the size of $h_t + h_x^2$ and without fear of problems in the theory of difference schemes (in general and in the theory of numerical methods) or the volatility of the developed methods with respect to the original initial and boundary data.

IV NUMERICAL EXPERIMENTS

The theory outlined above, based on an explicit difference scheme, is a computer program written in the software MatLAB for numerical experiments. Using coefficients $\varepsilon = 1$ and $k = 0.3$ and initiating systems with short-term external influence at time $t = 0.04,$ the outcome of Fig. 1 is obtained, which shows the projection of the material over time. In this example we used the Dirichlet boundary conditions.

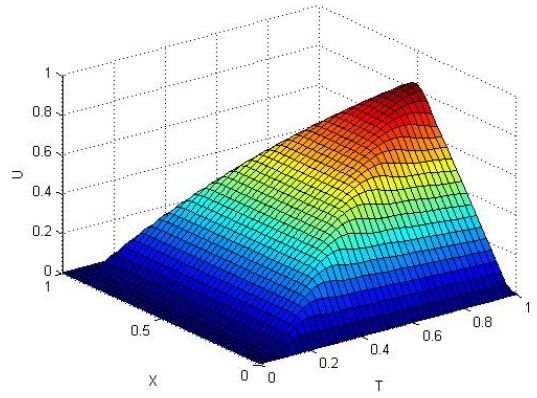

Fig. 1. Illustration of the process of forming a nanoscale spire: realization of the discrete model (4)-(7) under $\varepsilon = 1$ ("bad case").

V CONCLUSIONS

This work concentrated on developing a model of the effects of solid nanoparticles on the surface of materials. The 1D continuous analytical model was constructed with both Neumann and Dirihlet boundary conditions. Both implicit and explicit finite difference schemes were constructed and their convergence and stability investigated. Finally, an example of a numerical experiment was done with computer software developed form the theory.

VI ACKNOWLEDGMENTS

For the second co-authors the present article was executed within the framework of the following two Research Projects: 1) The State Research Programme "NextIT": *Next generation Information and Communication Technologies*; 2) The European Regional Development Fund Project No. 2014 / 0029 / 2DP / 2.1.1.1 / 14 / APIA / VIAA / 088: *"Development of an experimental long flight distance unmanned aerial vehicle prototype for multi-purpose environmental monitoring"*.

VII REFERENCES

[1] A. Medvid, *Nano-cones Formed on a Surface of Semiconductors by Laser Radiation: Technology, Model and Properties*, Proceedings of the International Conferences on *Nanowires Science and Technology*, February 2010. Croatia: INTECH. pp. 402-424.

[2] J. Kaupuzs, Sh. E. Guseynov, J. S. Rimshans, and A. Medvid, "Modeling of surface structure formation after laser irradiation." International Journal of Recent Advances in Fluid Mechanics, Heat & Mass Transfer and Biology. pp. 49-56. 2011.

[3] A. N. Tikhonov, and A. A. Samarsky, Equations of Mathematical Physics. Moscow: Lomonosov Moscow State University Publishing, 1999.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 217-222

Automatic Transformation of Relational Database Schema into OWL Ontologies

Imants Zarembo
Rezekne Higher Education Institution

Abstract. **Ontology alignment, or ontology matching, is a technique of mapping different concepts between ontologies. For this purpose at least two ontologies are required. In certain scenarios, such as data integration, heterogeneous database integration and data model compatibility evaluation, a need to transform a relational database schema to an ontology can arise.**

To conduct a successful transformation it is necessary to identify the differences between relational database schema and ontology information representation methods, and then to define the transformation rules. The most straight forward but time consuming way to carry out transformation is to do it manually. Often this is not an option due to the size of data to be transformed. For this reason there is a need for an automated solution.

The automatic transformation of OWL ontology from relational database schema is presented in this paper; the data representation differences between relational database schema and OWL ontologies are described; the transformation rules are defined and the transformation tool's prototype is developed to perform the described transformation.

Keywords: **ERD, OWL, transformation.**

I INTRODUCTION

Relational databases are one of the most popular storage solutions for all kinds of data. To fully realize vision of the semantic web, the gap between legacy relational databases must be narrowed. The problem of determining the rules of how to map an entity from a legacy relational database to an equivalent entity in an ontology is commonly referred to as the database-to-ontology mapping problem. It is important to note that the term mapping has been interchangeably used for two different problems: mapping and transformation. A mapping maps the entity of one data source to at most one entity of another. A mapping can be viewed as a collection of mapping rules all oriented in the same direction. A transformation is the process of expressing the entities of data source with respect to the entities of another data source. In this paper, the mapping as a transformation is considered.

The transformation process is the first step towards relational database integration into the semantic web. Other tasks such as ontology matching requires that the relational database data model is transformed into an ontology before ontology matching can take place.

The use of ontologies is important for many rapidly expanding technologies as they provide the critical semantic foundation. In the context of computer and information science, the ontology defines a set of representative primitives using which a domain of knowledge is modeled. In comparison with relational databases, Web Ontology Language (OWL) is fully built upon formal logic, which allows logical reasoning. The reasoning can help to validate the consistency of the conceptual model and to find the properties and the relationships, which were not previously apparent. Another important benefit of using ontologies are the semantic search capability. Semantic search improves search accuracy by understanding the contextual meaning of terms as they appear in the searchable data.

This paper presents an automatic transformation of OWL ontology from relational database schema. The transformation process applies the transformation rules to the entity relationship diagrams and creates appropriate ontology entities.

II WEB ONTOLOGY LANGUAGE

The OWL 2 Web Ontology Language, informally OWL 2, is an ontology language for the Semantic Web with formally defined meaning [1]. OWL ontologies provide classes, properties, individuals, and data values and are stored as Semantic Web documents. OWL ontologies can be used along with information written in Resource Description Framework (RDF). OWL ontologies themselves are primarily exchanged as RDF documents.

OWL allows for greater machine interpretability of Web content than the content supported by XML, RDF and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. OWL has three increasingly-expressive sublanguages:

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.170

OWL Lite, OWL DL and OWL Full. OWL DL is used for the purposes of this paper to support the maximal expressiveness without losing computational completeness.

III RELATIONAL DATABASE SCHEMA

Many legacy systems have been documented using Entity Relationship Diagrams (ERD) or Extended ER diagrams.

An Entity Relationship Diagram is a representation of data within a domain, a visual form of relational database. It consists of entities and relationships between these entities.

For legacy systems to be used as a component of emerging semantic web, these systems need to be upgraded. Both ERD and OWL represent entities and their relationships which provides opportunity of transformation of ERD to OWL ontology.

IV RELATED WORK

There is a considerable amount of research on the topic of transforming and mapping of relation databases to ontologies. Approaches vary dramatically, ranging from manual transformation to relational database structure duplication using an ontology [2]. Although relational database transformation and mapping techniques have different goals in mind, similar results can be achieved by mapping relational databases to RDF datasets. The mapping produces correspondences between a relational database and an ontology, while transformation uses relational database to create a new object – an ontology.

In their paper [3] Zhang and Li analyse mappings and construct rules based on relational database to generate ontology concepts, properties, axioms and instances. The authors of [4] propose discovering mappings between relational database schema and ontology by exploiting mappings based on virtual documents and mapping consistency validation. Alalwan et al. in [5] propose automatic transformation system based on Structure Query Language (SQL) and metadata to extract semantic aspects which otherwise could not be inferred from the SQL. In [6] the authors propose an approach of learning OWL ontology from data in relational database. The approach proposes a technique that can be used to acquire ontology from relational database automatically by using a group of learning rules instead of using a middle model. The approach presented in [7] proposes to extract database metadata information from relational database using reverse engineering technique, and then to analyse the corresponding relationship between relational database and OWL ontology. The author of [8] maps relational data into ontology and fills ontology with data from relational database using mapping rules. Astrova et al. [9] propose to map relation database constructs (tables, columns, data types, constraints and rows) to an ontology.

World Wide Web Consortium (W3C) offers several recommendations in relation to mapping of relational databases to RDF. Recommendation “A Direct Mapping of Relational Data to RDF” [10] provides simple direct mapping from relational data to RDF. In [11] RDB to RDF mapping language is described, which allows creating customized mappings from relational databases to RDF datasets in the form of graphs.

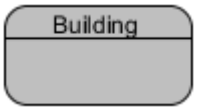

Fig. 1. ER class example

```
<owl:Class rdf:about="#Building">
</owl:Class>
```

Fig. 2. OWL class example

The D2RQ Mapping Language [12] provides a declarative mapping language for mapping relational database schemas to RDF vocabularies and OWL ontologies. The mapping defines a virtual RDF graph that contains information from the database.

V MATERIALS AND METHODS

Relational databases and ontologies are designed with different goals in mind, however they are similar. To successfully transform data model represented by relational database schema into OWL ontology (Fig. 3), a set of transformation rules must be defined. During the transformation the differences in data models can lead to data or semantic meaning loss.

The transformation rules must take into account the differences between both information representation methods. The rules defined in this paper describe how to transform each of the major relational database constructs into an OWL ontology.

The created ontology is described in OWL-DL language, which is recommended by W3C for publishing and sharing ontologies on the web. OWL-DL is based on Description Logics [13] and has computational properties for reasoning systems.

A. Classes

Ontology class (Fig. 2) creation depends on particular database table cases (Fig. 1) [14]. Three different cases are taken into account in the proposed approach:

Case 1

For the first case to occur, a table must be used only to relate other tables in a many-to-many relationship. (Fig. 4).

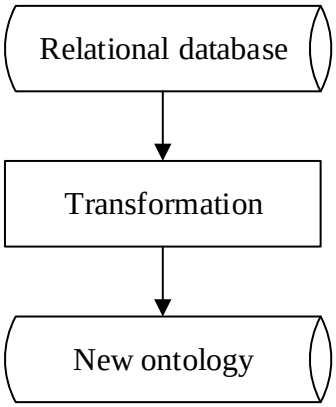

Fig. 3. Transformation process

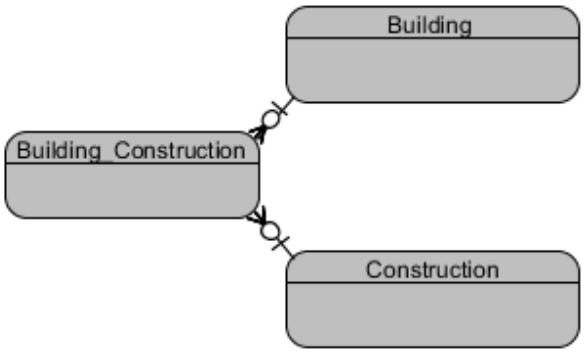

Fig. 4. ER class relationship example

```
<owl:ObjectProperty
  rdf:about="Building_Construction">
  <rdfs:range
   rdf:resource="#Building"/>
  <rdfs:domain:resource="#Construction"/>
</owl:ObjectProperty>
```

Fig. 5. OWL class relationship example

The relationship can be divided into two disjoint subsets of columns, each take part in a referential constraint with the related tables (Fig. 5).

For that reason all columns in the table are considered foreign and primary keys, because their combination uniquely defines the rows of the table.

Case 2

For the second case to occur, a table must be related to another table by a referential integrity constraint. In this case all primary keys of the table are considered foreign keys.

Case 3

If none of the previous cases occur, the third case is considered.

B. Attributes

An entity can have one of several types of attributes [16]:

- Simple attributes;
- Composite attributes;
- Multi-valued attributes.

Simple attributes

Simple attribute of entity (Fig. 6) can be mapped into datatype property of corresponding OWL class (Fig. 7). The domain of the datatype property is set to the entity it belongs and range is the actual datatype of that attribute. The ER diagram data type values can be mapped to XSD datatypes in OWL using the mappings shown in Table 1.

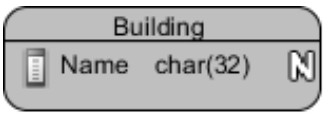

Fig. 6. ER simple attribute example

```
<owl:DatatypeProperty
 rdf:about="hasBuildingName">
  <rdfs:range
   rdf:resource="#string"/>
</owl:DatatypeProperty>
```

Fig. 7. OWL simple attribute example

TABLE I

MAPPINGS BETWEEN SQL AND XSD DATATYPES

SQL	XSD
integer/int	Xsd:integer
float	Xsd:float
char/varchar/vchar	Xsd:string
time	Xsd:time
date	Xsd:date
datetime	Xsd:datetime
boolean	Xsd:boolean

Composite attribute

There are two ways to map a composite attribute to OWL datatype property:

- To transform only simple component attributes of composite attribute to datatype properties of corresponding OWL class, but ignore the composite attribute itself.
- To transform composite attribute to datatype property and then map its component attributes to subproperty of corresponding datatype property.

The first approach is preferable for relational database transformation because relational schema has only instances of simple, component attributes. Composite attributes are ignored.

Multi-valued attribute

Databases cannot handle multi-valued attributes efficiently, while ontologies have an adequate way to deal with them. In the proposed approach multi-valued attributes are put in separate relations to avoid tuple duplication.

Multi-valued attributes are mapped to datatype property, but without a “functional” tag.

C. Generalization and specialization

Subtype relations in the ERD (Fig. 8) can be transformed to subClassOf in the OWL ontology (Fig. 9). OWL subClassOf represents the generalization or specialization hierarchy.

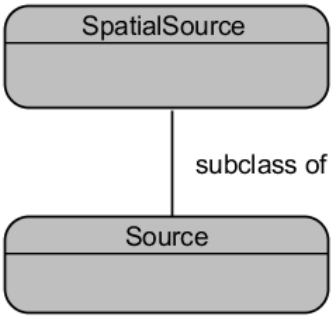

Fig. 8. ERD subclass example

```
<owl:Class rdf:about="SpatialSource">
  <rdfs:subClassOf
   rdf:resource="#Source"/>
</owl:Class>
```

Fig. 9. OWL subclass example

D. Enumerator

In ERD enumerators are simple classes, often structurally indistinguishable from other classes. For that reason it is not always possible to automatically detect if the class contains an enumerator.

To transform ERD enumerator class to OWL ontology, the class must first be specifically marked and the transformation algorithm must be able to identify the type of marking to perform a successful transformation. Types or markings can vary. One of the more common approaches is to mark table name with special suffix or postfix, which can be uniquely identified by the transformation tool.

Enumerations can be transformed into named data types in OWL (Fig. 10). The value of this built-in OWL property must be a list of individuals that are the instances of the class. First existence of the data type must be declared, then the definition of the data type can be given using the "DatatypeDefinition" axiom. The allowed values of the Enumeration are listed in the "DataOneOf" statement.

```
<owl:Class rdf:ID="CoveregeType">
  <rdfs:subClassOf>
    <owl:Class>
      <owl:oneOf
       rdf:parseType="Collection">
        <owl:Thing rdf:about="#forest"/>
        <owl:Thing rdf:about="#water"/>
        <owl:Thing rdf:about="#grass"/>
      </owl:oneOf>
    </owl:Class>
  </rdfs:subClassOf>
</owl:Class>
```

Fig. 10. OWL enumerator example

E. Cardinality

In OWL owl:cardinality is a built-in property that links a restriction class to a data value belonging to the range of XML Schema datatype nonNegativeInteger. OWL 2 contains six different cardinality axioms, three for object properties and three for data properties:

- “ObjectMinCardinality”;
- “ObjectMaxCardinality”;
- “ObjectExactCardinality”;
- “DataMinCardinality”;
- “DataMaxCardinality”;
- “DataExactCardinality”.

```
<owl:ObjectProperty rdf:about="Diameter">
  <rdfs:domain
   rdf:resource="#NetworkSegment"/>
  <rdfs:range>
    <owl:Restriction>
      <owl:onProperty
       rdf:resource="#hasDiameter"/>
        <owl:qualifiedCardinality
         rdf:datatype="integer">1
        </owl:qualifiedCardinality>
      <owl:onDataRange
       rdf:resource="#float"/>
    </owl:Restriction>
  </rdfs:range>
</owl:ObjectProperty>
```

Fig. 11. OWL exact cardinality example

In OWL 2 a cardinality constraint puts a restriction on the number of values that the property can take, in the context of this particular class description.

```
<owl:ObjectProperty
 rdf:about="BuildingName">
  <rdfs:domain rdf:resource="#Address"/>
  <rdfs:range>
    <owl:Restriction>
      <owl:onProperty
       rdf:resource="#hasBuildingName"/>
        <owl:minQualifiedCardinality
         rdf:datatype="integer">0
        </owl:minQualifiedCardinality>
      <owl:onDataRange
     rdf:resource="#string"/>
```

```
    </owl:Restriction>
  </rdfs:range>
  <rdfs:range>
    <owl:Restriction>
      <owl:onProperty
       rdf:resource="#hasBuildingName"/>
        <owl:maxQualifiedCardinality
         rdf:datatype="integer">1
       </owl:maxQualifiedCardinality>
      <owl:onDataRange
     rdf:resource="#string"/>
    </owl:Restriction>
  </rdfs:range>
</owl:ObjectProperty>
```

Fig. 12. OWL exact cardinality example

A restriction containing a “MinCardinality” constraint describes a class of all individuals that have at least N semantically distinct values for the property concerned, a “MaxCardinality” constraint describes a class of all individuals that have at most N semantically distinct values for the property concerned.

An “ExactCardinality” constraint (Fig. 11) describes a class of all individuals that have exactly N semantically distinct values for the property concerned, where N is the value of the cardinality constraint [1].

In the case when "MinCardinality" and "MaxCardinality" constraints have the same values they can be replaced with the single "ExactCardinality" constraint to make ontology more concise (Fig. 12). In the case when the cardinality upper limit is 1, the property can be marked functional by adding a "FunctionalObjectProperty" or "FunctionalDataProperty" to object or data properties of the ontology accordingly.

If an attribute is mandatory its cardinality is always 1. If an attribute is not mandatory its minimal cardinality is 0 and maximal cardinality 1.

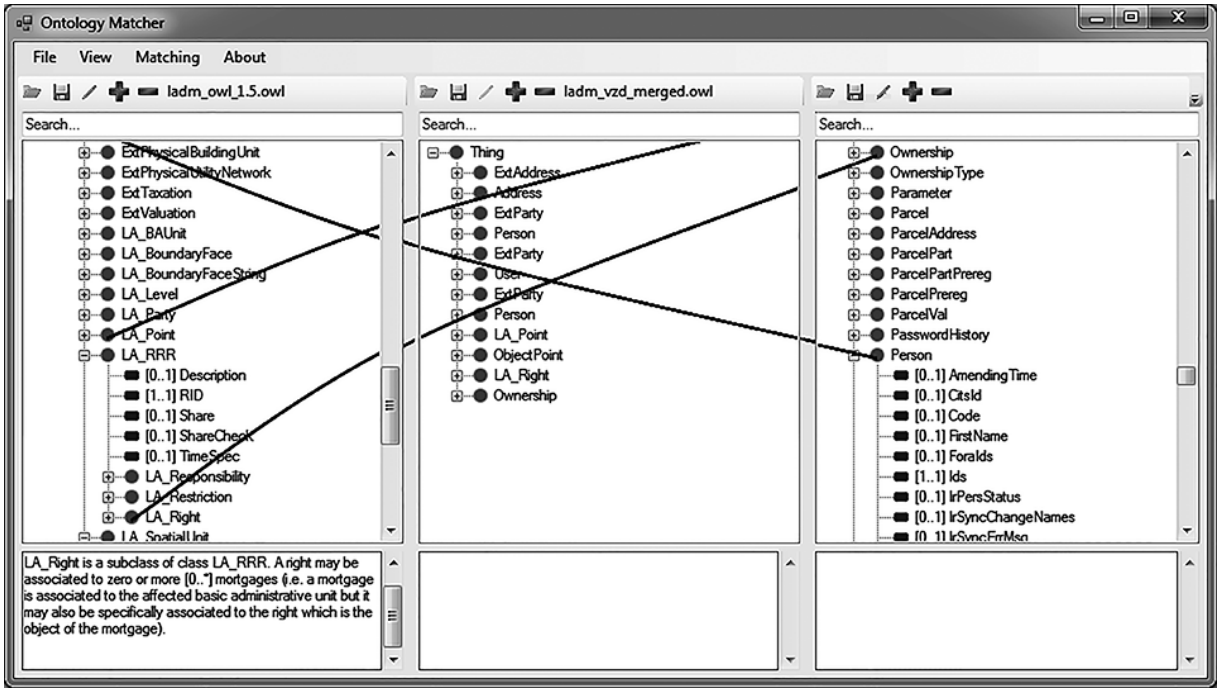

Fig. 13. Screenshot of ontology matching tool prototype

VI MATCHING TOOL

The transformation process was carried out using a tool prototype specifically developed for this task. In Fig. 13, the screenshot of the tool can be seen with two ontologies opened and several concepts matched. The tool consists of three parts:

1. The first part is designed for automatic transformation of UML geospatial profile class diagrams into OWL ontologies [17], using similar approach to the one proposed in this paper.
2. The second part is designed for automatic transformation of relational database schemas into OWL ontologies, using the approach proposed in this paper.
3. The third part is designed for manual and automatic ontology matching. Ontology matcher consists of several name based ontology matching algorithms [15] that can be used automatically [18].

VII RESULTS AND DISCUSSION

The transformation process employed to transform relational database data model, represented with ER diagrams, into an OWL ontology can be described with the following steps:

1. Relational database model is prepared for further processing as an XML Metadata Interchange (XMI) format file. XMI is an Object Management Group (OMG) standard for exchanging metadata information via Extensible Markup Language (XML).
2. Next, an XMI file is parsed by the transformation tool.
3. The transformation rules are applied and the ontology is generated.
4. The resulting ontology is then saved to OWL format file.
5. The OWL ontology is ready for visualization and further use by specially designed tools (e.g. Protégé).

The transformed ontology was put to the test using the ontology matching tool prototype. The tests consisted of automatically matching two ontologies using name based ontology matching algorithms. The matching results showed average similarity of 0.5. The full results were published in [18].

The solution presented in this paper can be useful to software engineering practitioners who are concerned with semantic web technology integration with legacy databases and to those who need to transform relational database data models to ontologies for ontology matching purposes.

VIII CONCLUSIONS

The approach of relational database schema transformation into OWL ontology was presented in this paper. Some of the similarities and differences of ERD and OWL ontologies were analysed. The transformation rules defined. The prototype tool for automatic ERD transformation to OWL ontologies developed and the process of the transformation described.

The proposed approach allowed to successfully transform relational database components like tables, fields and relationships between tables to an ontology.

Automatic ontology creation from ER diagrams is the first step towards legacy database integration with semantic web technology.

Further work on this topic includes studying relational database denormalization techniques and applying them to the transformation process. This would help to improve transformed ontology hierarchical structure, which is lacking in the proposed approach when transforming normalized relational databases.

IX REFERENCES

[1] W3C. “OWL 2 Web Ontology Language Document Overview (Second Edition)”, 2012. [Online]. Available: http://www.w3.org/TR/owl2-overview/. [Accessed March 10, 2015].

[2] D. E. Spanos, P. Starvrou and N. Mitrou, Bringing Relational Databases into the Semantic Web: A Survey, Semantic Web, vol 3, pp. 169-209, Apr. 2012.

[3] L. Zhang and J. Li, Automatic Generation of Ontology Based on Database, Journal of Computational Information Systems, pp. 1148-1154, Apr. 2011.

[4] W. Hu and Y. Qu, *Discovering Simple Mappings Between Relational Database Schemas and Ontologies, ISWC'07/ASWC'07 Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, pp. 225-238, 2007.

[5] N. Alalwan, H. Zedan and F. Siewe, “Generating OWL Ontology for Database Integration”, Advances in Semantic Processing, 2009. SEMAPRO '09, pp. 22-31, Oct. 2009.

[6] M. Li, X. Y. Du and S. Wang, *Learning Ontology from Relational Database, Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference*, pp. 3410-3415, Aug. 2005.

[7] S. Zhou, H. Ling, M. Han and H. Zhang, “Ontology Generator from Relational Database Based on Jena”, Computer and Information Science, pp. 263-267, 2010.

[8] Z. Telnarova, Relational database as a source of ontology creation, Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference, pp. 135-139, Oct. 2010.

[9] I. Astrova, N. Korda and A. Kalja, *Rule-Based Transformation of SQL Relational Databases to OWL Ontologies, In Proceedings of the 2nd International Conference on Metadata & Semantics Research*, 2007.

[10] W3C. “A Direct Mapping of Relational Data to RDF”, 2012. [Online]. Available: http://www.w3.org/TR/rdb-direct-mapping/. [Accessed May 11, 2015].

[11] W3C. “R2RML: RDB to RDF Mapping Language”, 2012. [Online]. Available: http://www.w3.org/TR/r2rml/. [Accessed May 11, 2015].

[12] “The D2RQ Mapping Language”, 2012. [Online]. Available: http://d2rq.org/d2rq-language [Accessed May 11, 2015].

[13] F. Baadar, I. Horrocks and U. Sattler, “Description Logics as Ontology Languages for the Semantic Web”, Festschrift in honor of Jörg Siekmann, Lecture Notes in Artificial Intelligence, 2003.

[14] R. Ghawi and N. Cullot, “Database-to-Ontology Mapping Generation for Semantic Interoperability”, VLDB ‘07, September 23-28, 2007.

[15] J. Euzenat and P. Shvaiko, *Ontology Matching, 2nd edition*. Springer-Verlag, Berlin Heidelberg (DE), 2013.

[16] M. Fahad, “ER2OWL Generating OWL Ontology from ER Diagram”, Intelligent Information Processing IV, pp. 28-37, 2008.

[17] I. Zarembo, S. Kodors, *Automatic Transformation of UML Geospatial Profile to OWL Ontologies*, *Virtual Multidisciplinary Conference (QUAESTI 2013)*, ISBN: 978-80-554-0826-2, Slovak Republic, pp. 225-229, Dec. 2013.

[18] I. Zarembo, A. Teilans, A. Rausis, J. Buls, “Assessment of Name Based Algorithms for Land Administration Ontology Matching,” ICTE in Regional Development, vol. 43, pp. 53-61, Dec. 2014.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 223-227

CALCULATION TEMPERATURE AND PRESSURE OF THE ROTARY ENGINE

Yu.N. Zhuravlyov, S.N. Semyonov, J.N. Lukyanov, A.L. Perminov, S.I. Tikhonov, M.A. Donchenko

Abstract. **The principles of calculating the temperature and pressure in the working chambers of the rotary vane engine with an external supply of heat are considered. The mathematical model for calculating the pressure and the temperature in the chamber with heat transfer between the working fluid and the chamber walls is built. The plots of the dependence of the pressure and the temperature in the chamber on the angle of rotation of the output shaft at the minimum and maximum temperature of the walls are obtained.**

Keywords: **the rotary vane, the heat transfer with the walls, the chamber temperature, the chamber pressure.**

I INTRODUCTION

The main areas of economic development in the XXI century is the search for promising energy conversion technologies based on high-performance thermodynamic cycles for the use of renewable energy sources. In the opinion of many foreign experts, promising direction is the development and widespread adoption of power generation systems based on external combustion engines [6]. One of such engines is a rotary-vane engine with an external supply of heat (RVE), developed the staff of the Pskov State University Y.N. Lukyanov, M.A. Donchenko and others [4,5].

Vanes group of the RVE (Fig. 1) consists of a cylindrical housing 1 in which installed two coaxial rotor (external and internal), two sealings housing 2 and two end covers 4. The outer rotor and the inner rotor are consists of two shafts 6, 7 respectively, the four pistons 5, four pressure plates 8 and four seals 3. Two rotors form a four working chambers of variable volume. A four power strokes are proceeds in each of the chambers consistently: receipt of the working fluid to the heater through the windows in the end cap 4, the expansion of the working fluid, post working fluid to the cooler through the windows in the end cap 4 and the compression of the working fluid.

Thus, working chamber the engines consists by the following components: two shafts 6, 7, two end caps 4, two pistons 5 and the housing 1.

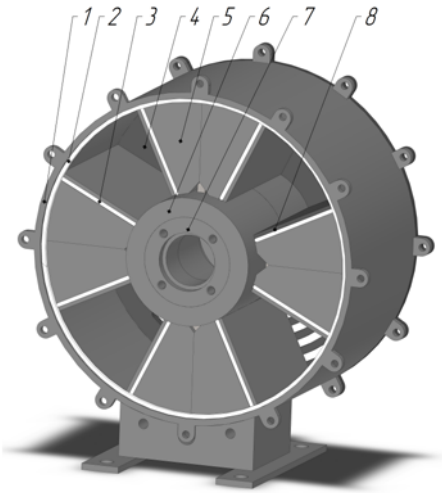

Fig. 1. Vanes group RVE:
1 – housing; 2 – sealings housing; 3 – sealings pistons; 4 – end covers; 5 – pistons; 6, 7 – shafts; 8 – pressure plates.

When designing a new heat engine, one of the first stages of calculation is necessary to determine the temperature and pressure in working chambers.

Existing methods of calculating the parameters of the operating cycle heat engine with external heat supply based usually on the isothermal mathematical model [1]. In some works at calculations of a running cycle the processes compression and expansion in the isolated volume consider adiabatic [7]. However, it should note that when using a heat engine occurs at constant heat transfer between the working fluid and the walls of the chamber, which has noticeable effect on the temperature and pressure of the working fluid in the chamber. In this paper, we propose a

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.195

mathematical model to calculate the main parameters of the operating cycle engines with an external supply of heat for cycles of expansion and contraction of the working fluid in the isolated volume, which takes into account the heat transfer between the working fluid and the parts that make up the working chamber.

II DIFFERENTIAL EQUATIONS PRESSURE AND TEMPERATURE CHANGES IN THE CHAMBER

The processes of expansion and contraction of the working fluid in the isolated volume described by the first law of thermodynamics, written for the case of constant mass [2]:

$$Mc_V dT + pdV + dQ_w = 0, \qquad (1)$$

where M – the mass of the working fluid in the chamber; c_V– heat capacity at constant volume of the working fluid; T– the temperature of the working fluid in the chamber; p– the pressure of the working fluid in the chamber; V– chamber volume; dQ_w– heat given (received) the working fluid by heat exchange with the walls of the chamber.

Considering in order that the working fluid in the chamber is ideal gas we write the equation of state ideal gas:

$$pV = MRT, \qquad (2)$$

where R – gas constant.

After a series of transformations of the equations (1) and (2) in accordance with [2], we obtain two differential equations:

$$\frac{dT(\alpha)}{d\alpha} = -\frac{(k-1)T(\alpha)}{V(\alpha)}\frac{dV(\alpha)}{d\alpha} - \frac{1}{Mc_V}\frac{dQ_w(\alpha)}{d\alpha} \qquad (3)$$

$$\frac{dp(\alpha)}{d\alpha} = -\frac{kp(\alpha)}{V(\alpha)}\frac{dV(\alpha)}{d\alpha} - \frac{p(\alpha)}{Mc_V T(\alpha($$

$$\frac{dQ_w(\alpha)}{d\alpha}, \qquad (4)$$

where k – adiabatic exponent; α– angle of rotation of the output shaft of the engine.

Equation (3), is a differential equation for the variation of temperature in the chamber. Equation (4), is a differential equation of the pressure change in the chamber.

The volume of the working chambers RVE is given by:

$$V(\alpha) = (\psi(\alpha) - \psi_p)c, \qquad (5)$$

where $\psi(\alpha)$ – angle between the axes of the pistons; ψ_p– the angular size of the piston; c– design parameter.

According to [2], the angle between the axes of the pistons is determined by the expression:

$$\psi(\alpha) = 2(a + b\cos 2\alpha), \qquad (6)$$

where $a = \pi/4$; $b = \pi/4 - \psi_{\min}/2$ and where ψ_{min} – the minimum value of the angle ψ.

With regard to (5), (6) can be written $\frac{1}{V(\alpha)}\frac{dV(\alpha)}{d\alpha} = -\frac{4b\sin 2\alpha}{2a + 2b\cos 2\alpha - \psi_n}$.

Consider the summand $\frac{dQ_w(\alpha)}{d\alpha}$. According to [2]:

$$\frac{dQ_w(\alpha)}{d\alpha} = \frac{\beta F(T(\alpha) - T_w)}{\omega}, \qquad (7)$$

where β – the average value of heat transfer coefficient; F– heat exchange surface area; T_w– surface temperature of parts forming a working chamber; $\omega = \frac{d\alpha}{dt}$ – angular velocity.

If the temperature of the surfaces of parts constituting working chamber different, the heat dQ_w is determined by the following relation:

$$\frac{dQ_w(\alpha)}{d\alpha} = \sum_{i=1}^{n}\frac{\beta_i F_i(T(\alpha) - T_{wi})}{\omega}, \qquad (8)$$

where β_i – the average value of heat transfer coefficient surface separate part; F_{ic}– heat exchange surface area separate part; T_{wi} – surface temperature separate part; n – number of surfaces, forming a working chamber; i – serial number surface forming a working chamber.

For RVE relation (8) takes the form:

$$\frac{dQ_w(\alpha)}{d\alpha} = \frac{\beta_1 F_1(T(\alpha) - T_{w1})}{\omega} + \frac{\beta_2 F_2(T(\alpha) - T_{w2})}{\omega} + 2\frac{\beta_3 F_3(T(\alpha) - T_{w3})}{\omega} + 2\frac{\beta_4 F_4(T(\alpha) - T_{w4})}{\omega}, \qquad (9)$$

where β_1, F_1, T_{w1} – heat transfer coefficient, surface area and surface temperature housing 1 (see. Fig. 1); β_2, F_2, T_{w2}– heat transfer coefficient, surface area and surface temperature shafts 6, 7 (see. Fig. 1); β_3, F_3, T_{w3} – heat transfer coefficient, surface area and surface temperature end covers 4 (see. Fig. 1); β_4, F_4, T_{w4}– heat transfer coefficient, surface area and surface temperature pistons 5 (see. Fig. 1).

Adding equations (3) and (4) in accordance with equation (8) yields:

$$\frac{dT(\alpha)}{d\alpha} = -\frac{(k-1)T(\alpha)}{V(\alpha)}\frac{dV(\alpha)}{d\alpha} - \frac{1}{M \cdot c_V}\sum_{i=1}^{n}\frac{\beta_i F_i\left(T(\alpha) - T_{wi}\right)}{\omega} \quad (9)$$

$$\frac{dp(\alpha)}{d\alpha} = -\frac{k \cdot p(\alpha)}{V(\alpha)}\frac{dV(\alpha)}{d\alpha} - \frac{p(\alpha)}{Mc_V T(\alpha)}\sum_{i=1}^{n}\frac{\beta_i F_i\left(T(\alpha) - T_{wi}\right)}{\omega} \quad (10)$$

Equations (9) and (10) must be supplemented by the initial conditions.

The main difficulty is the determination of the unknown β_i and T_{wi}, appearing in the equations (9, 10).

The heat transfer coefficient is a complex function of various quantities characterizing the process of heat transfer. In general this relationship can be expressed as [3]:

$$\beta = f\left(u, T_w, T, \lambda, c_p, \rho, \mu, a, \Phi, l_1, l_2 \ldots.\right), \quad (11)$$

where ω – the rate of the working fluid, T_w – wall temperature, T– the temperature of the working fluid, λ– thermal conductivity, c_p– heat capacity at constant pressure of the working fluid, ρ – the density of the working fluid, μ – the dynamic viscosity of the working fluid, a – thermal diffusivity, Φ – function shape of the body, l_1, l_2 – the dimensions of the body.

In most cases, analytically determine the coefficient of heat transfer is not possible. That is why, it is determined experimentally or by numerical simulation. Method of determining the heat transfer coefficient for parts vanes group RVE numerical simulation methods is given in [8].

Each working chamber of the engine is formed by the following components (see. Fig. 1): the two shafts 6, 7, two end caps 4, the housing 1 and the piston 5.

Wall temperature parts, forming a working chamber T_{wi}, varies during engine operation from T_{wimin} to T_{wimax}.

III DETERMINATION OF THE TEMPERATURE OF PARTS OF THE WALLS FORMING THE CHAMBER

The minimum temperature T_{wi} of the wall partsforming a working chamber, have at the time the engine is started and it is equal to the ambient temperature T_{atm}. After which they exchange a heat with the working fluid warmed and gradually heated to a temperature T_{wimax}, at which further operation of the engine is not changed.

For an approximate determination T_{wimax} can be assumed that the heat exchange of the working fluid with the walls working parts forming the chamber stationary. Then, as the temperature of the working fluid should be adopted mean-temperature working fluid in one cycle:

$$T_{ave} = \int_{\alpha=0}^{2\pi} T(\alpha)_{lim} d\alpha, \quad (12)$$

where $T(\alpha)_{lim}$ – the temperature of the working fluid calculated without heat exchange with the walls of the working fluid.

Further, it is necessary to determine the approximate density of the heat flow through the parts forming a working chamber. Methods for determining the density of heat flux through parts with by stationary heat transfer depends on the shape of the items (a flat, cylindrical or spherical) and given in [3]. Knowing the density of the heat flow in details and the heat transfer conditions on its borders, we can determine the temperature of the wall part which is in contact with the working fluid.

Next, determine the temperature field in the body parts 1 (see. Fig. 1) for the design scheme shown in Fig. 2.

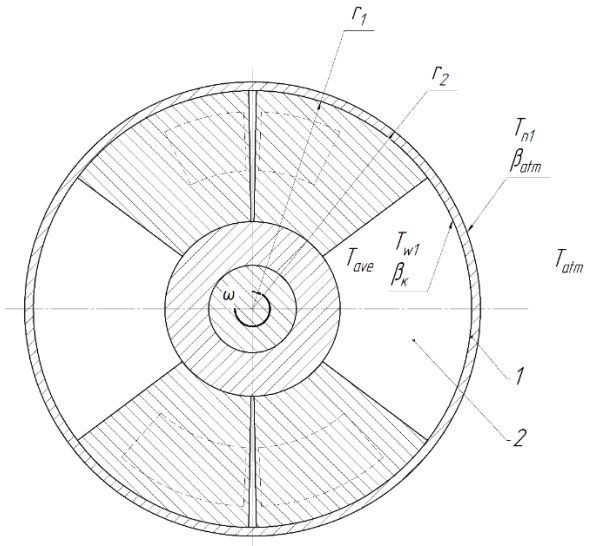

Fig. 2. Design scheme:
1 – housing; 2 – working chamber.

Heat flow density on the inner and outer surfaces of the housing 1 is given by [3]:

$$q_{in} = \frac{\lambda\left(T_{w1} - T_{n1}\right)}{r_1 \ln\left(\frac{r_2}{r_1}\right)} \quad (13)$$

$$q_{out} = \frac{\lambda(T_{w1} - T_{n1})}{r_2 \ln\left(\frac{r_2}{r_1}\right)}, \qquad (14)$$

where q_{in} – heat flow density on the inner surface of the housing; q_{out}– heat flow density on the outer surface of the housing; T_{w1}– the temperature of the inner surface of the housing; T_{n1}– temperature of the outer surface of the housing; r_1– the inner radius of the housing; r_2– the outer radius of housing; λ – thermal conductivity.

Adding to the equation (13), (14) the boundary conditions of the 3rd kind, written for the outer and inner surfaces of the case, we obtain a system of equations:

$$\begin{cases} q_{in} = \dfrac{\lambda(T_{w1} - T_{n1})}{r_1 \ln\left(\dfrac{r_2}{r_1}\right)} \\ q_{out} = \dfrac{\lambda(T_{w1} - T_{n1})}{r_2 \ln\left(\dfrac{r_2}{r_1}\right)} \\ q_{in} = \beta_h(T_{w1} - T_{ave}) \\ q_{out} = \beta_{atm}(T_{n1} - T_{atm}) \end{cases}, \qquad (15)$$

where β_h – heat transfer coefficient of from the working fluid to the inner surface of the housing; β_{atm} – heat transfer coefficient of from the outer surface of the housing to the environment.

From equations (15) is defined T_{w1}, which is the maximum wall temperature during operation of the engine. To determine the maximum wall temperature of the other parts forming the working chamber is necessary to solve a system of equations such as (15).

We shall show the calculation of the maximum temperature of the wall housing parts 1 (Figure 1) for the layout RVE 10 kW.

The input parameters for the calculation are: T_{ave}=398,2 K; T_{atm}=293 K; r_1=125 mm; r_2=130 mm; material - steel 40; λ=51 W/(m*K); working body - the air.

The numerical value of the heat transfer coefficient β_h defined in [8] and is equal to β_h=38,34 W/(m^2*K). The numerical value of the heat transfer coefficient β_{atm} determined by the method described in [3] and is equal to β_{atm}=10 W/(m^2*K).

For a given input parameters temperature on the inner surface of the housing is equal to T_{w1}=376,5 K.

Numerical values wall temperatures for all the items forming the working chamber shown in the table.

THE NUMERICAL VALUES OF THE WALL TEMPERATURE

Part Name	Part number, i	The wall temperature T_{wi}, K	
		minimal, T_{wimin}, K	maximum, T_{wimax}, K
Housing 1 (Figure 1)	1	293	376,5
The rotors 6, 7 (Figure 1)	2	293	388,7
The end cap 4 (Figure 1)	3	293	377,2
Piston 5 (Figure 1)	4	293	398,2

IV NUMERICAL SIMULATION

The numerical solution of differential equations (9) and (10) was carried out in the system with the help of Mathcad functions Odesolve designed to solve linear differential equations for the highest derivative Runge-Kutta methods.

Let us show the results of calculations for the compression cycle in RVE power of 10 kW. Compression cycle in RVE runs at an angle of rotation of the output shaft $0^\circ \le \alpha \le 45^\circ$. Pressure and temperature in the chamber at the beginning of the compression cycle are respectively, p_0=1 and T_0=293 K.

The input parameters for the calculation are:k=1,35; R=287 J/(kg·K); c_v=82 J/(kg·K); M=1,22 10^{-3} kg; ω=180 rpm; β_1=38,34 W/(m^2·K); F_1=0,019 m^2; β_2=38,34 W/(m^2·K); F_2=0,0077 m^2; β_3=35 W/(m^2·K); F_1=0,008 m^2; β_1=36,7 W/(m^2·K); F_1=0,009 m^2.

Solve the equation (9), (10) with T_{wi} of equal T_{wimin}, which corresponds to engine start, and with T_{wi} of equal T_{wimax}, which corresponds to the steady operation of the engine.

Graphs of temperature and pressure in the working chamber during compression of the working fluid at temperatures of walls T_{wimin} and T_{wimax} are shown in Figure 3.

It may be noted that a change in temperature of the walls of the working chamber from T_{wimin} to T_{wimax} the temperature in the chamber at the end of the compression stroke is changed from T(45°)=367,7 K to T(45°)=377,4 K, that is approximately 2,7%, the pressure is changed from p(45°)=2,44 atm to p(45°)=2,51 atm, that is approximately 2,8%. Thus we can conclude that transient temperature of the chamber walls during operation RVE layout power of 10 kW leads to a change in temperature and pressure in the working chamber is not more than 3%.

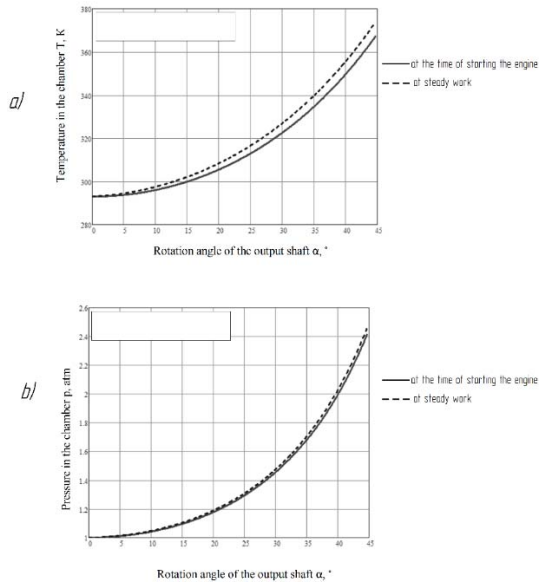

Fig. 3. Graphs of temperature and pressure:
a) the chamber temperature dependence of the rotation angle of the drive shaft; b) the chamber pressure dependence of the angle of rotation of the drive shaft.

V CONCLUSIONS

In this paper, the mathematical model for calculating the pressure and temperature in the chamber during the processes of expansion and contraction of the working fluid in an isolated volume considering heat transfer the working fluid from the chamber walls. Is given plot the schedule of dependence pressure and temperature in the chamber of the rotation angle of the output shaft at the minimum and maximum wall temperature. Found that transient temperature of the chamber walls during operation RVE layout power of 10 kW leads to a change in temperature and pressure in the working chamber is not more than 3%.

These calculations are needed to calculate the thermal stress of the engine parts, as well as to describe the gas exchange occurring in the rotary-vane engine.

VI REFERENCES

1. Reader T., Hooper C. *Stirling Engines*. M.: Mir, 1986, 464.
2. Petrychenko R.M. *Physical basis intracylinder processes in internal combustion engines*. L.: Publishing House of Leningrad. University, 1983, 244.
3. Tsvetkov F.F., Grigoriev B. A. *Heat and Mass Transfer: A manual for schools. - 2nd ed., Rev. and add*. M: Publishing House of the MEI, 2005, 550.
4. RF Patent 2374526 for an invention. Int. Cl. F16H25/04. *Mechanism for converting motion* / Lukyanov Y.N., Zhuravlev Y.N. et al. Publ. 27.11.2009. Bull. Number 33.
5. RF Patent 2387844 for an invention. Int. Cl. F01G1/077, F02G1/044. *Rotary-vane engine with an external supply of heat* / Lukyanov Y.N., Zhuravlev Y.N. et al. Publ. 27.04.2010. Bull. Number 12.
6. Kirillov N.G. *Production of Stirling engines - a new branch of mechanical engineering of the XXI century* // Turbines and diesel engines. Issue 2 (March-April). Yaroslavl: LLC "Turbomachines", 2010. P. 2-10.
7. Gorozhankin S.A. *Defining the parameters of the actual cycle Stirling engines based on their adiabatic model* // Modern industrial and civil construction. T. 2. № 4. Makeevka Univ. Donbas National Academy of Civil Engineering and Architecture, 2006, 187-194.
8. Semyonov S.N. *Calculation of the heat transfer coefficient in a rotor-and-blade engine with external heat supply* // Izvestija TulGU. Technical sciences. Vol. 6: in 2 hours. Ch. 1. Tula, 2013, 245-253.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 228-231

Artificial Neural Networks and Human Brain: Survey of Improvement Possibilities of Learning

Aleksejs Zorins, Peter Grabusts
Rezeknes Augstskola, Faculty of Engineering,
Atbrivosanas aleja 115, Rezekne, LV-4601, Latvia,
E-mail: Aleksejs.Zorins@ru.lv, Peteris.Grabusts@ru.lv

Abstract. **There are numerous applications of Artificial Neural Networks (ANN) at the present time and there are different learning algorithms, topologies, hybrid methods etc. It is strongly believed that ANN is built using human brain's functioning principles but still ANN is very primitive and tricky way for real problem solving. In the recent years modern neurophysiology advanced to a big extent in understanding human brain functions and structure, however, there is a lack of this knowledge application to real ANN learning algorithms. Each learning algorithm and each network topology should be carefully developed to solve more or less complex problem in real life. One may say that almost each serious application requires its own network topology, algorithm and data pre-processing. This article presents a survey of several ways to improve ANN learning possibilities according to human brain structure and functioning, especially one example of this concept – neuroplasticity – automatic adaptation of ANN topology to problem domain.**

Keywords: **Artificial Neural Networks, Brain Networks, Artificial Neural Network Learning Algorithms.**

I INTRODUCTION

There are numerous applications of Artificial Neural Networks (ANN) at the present time and there are different learning algorithms, topologies, hybrid methods etc. It is strongly believed that ANN is built using human brain's functioning principles but still ANN is very primitive and tricky way for real problem solving, because in any application should be found answers to the following questions [1]:

- Is the network complex enough to be capable to encode a solution?
- Is it possible to find solution in a reasonable amount of time?
- How can we guarantee that a trained network is matching closely enough our problem domain and hidden regularities in the data?

In the recent years modern neurophysiology advanced to a big extent in understanding human brain functions and structure [3,5,7,8,9,10], however, there is a lack of this knowledge application to real ANN learning algorithms. Each learning algorithm and each network topology should be carefully developed to solve more or less complex problem in real life. One may say that almost each serious application requires its own network topology, algorithm and data pre-processing.

This article presents a survey of several ways to improve ANN learning possibilities according to human brain structure and functioning, especially one example of this concept – neuroplasticity – automatic adaptation of ANN topology to the problem domain.

II NATURAL AND ARTIFICIAL NEURAL NETWORKS

Human brain consists of different types of cells called neurons which are connected with each other thus making a network of neurons. This is a very simplified model because in creation and training of ANN we ignore chemical processes, frequency of neurons, features of different brain parts and many other things that are not understood or even unknown (see Fig. 1).

That is why it would be worth mentioning that all "natural" neurons and neural networks used by scientists and also adopted here in this article are very primitive and far from a real world.

Keeping in mind this limitation Figure 2 gives a "basic neuron" scheme, adopted from [1].

 ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.165

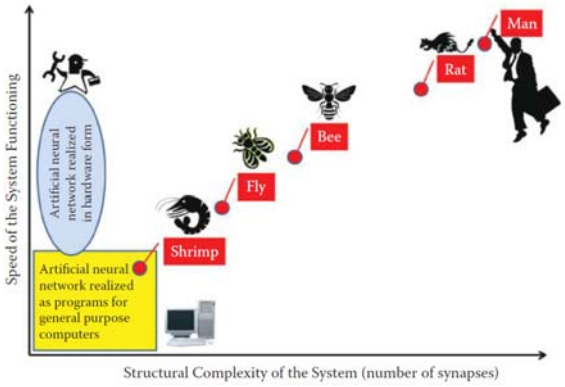

Fig. 1. ANN in comparison with real biological system considering speed of functioning and level of structural complexity [12]

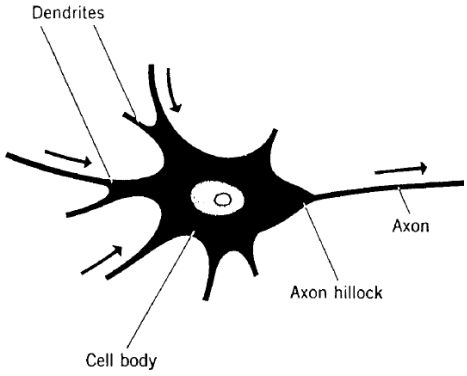

Fig. 2. Natural neuron scheme

The natural neural network (or brain) consists of billions interconnected neurons, where each neuron receives input signals from other neurons by its dendrites and produces an output to other neuron by its axon. The biological neural network complexity depends on the organism, e.g. a few hundred in some simple creatures and hundreds of billions in human brain [1].

Using this simplified biological neuron as a model on artificial neuron (software model) has been created and implemented in many types of ANN (see Fig. 3):

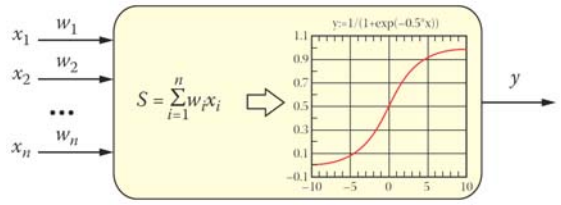

Fig. 3. The structure of widely used Multilayer Perceptron neuron (MLP) mathematical model [12]

In Fig. 3 artificial neuron receives its inputs $x_1 \ldots x_n$ weighted with $w_1 \ldots w_n$ (these weights represent the strength of connection). All inputs are combined and calculated with a sum function and then propagated to activation function (sigmoid in this case). After calculations neuron gives output value y, which is fed to other neurons or is treated like a final output signal of the network.

More than hundred years ago scientists discovered several regions in the brain, each of which is responsible for its own functions (see Fig. 4). Still there is only a beginning of understanding how we are thinking and how our brain is operating. Anyway Artificial Intelligence (AI) experts are trying to simulate some functional or structural features of the brain in ANN and other AI applications. For instance, the discovery of regions in the brain is implemented by committees of networks, where several networks with different configurations are devoted to solve one task as a team. Another application of this idea is hybrid networks when a complete artificial network consists of several different parts, e.g. backpropagation and Kohonen self-organizing maps.

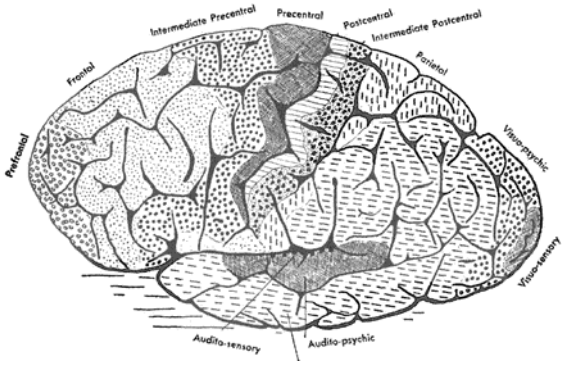

Fig.4. Anatomical parcellation of the human cerebral cortex, the left hemisphere as rendered by Alfred Campbell [9]

Figure 5 shows several examples of ANN with different structure and types. Accordingly, different learning algorithms are used to train such kind of networks: error back-propagation, self-organizing maps, recurrent learning, algorithms of hybrid types etc.

It should be noted that this article is not giving a complete review of all possible kinds of ANN, even a comprehensive book can hardly deal with this topic. The author is just showing the main principles and connection with basic brain and neurophysiology discoveries. This is an attempt to find new ways of improving existing ANN and its learning algorithms. Human brain is a wonderful instrument and we should be able to use it properly. ANN could help us in this task – modeling and solving different problems with this technology will lead us to a better understanding of our true nature.

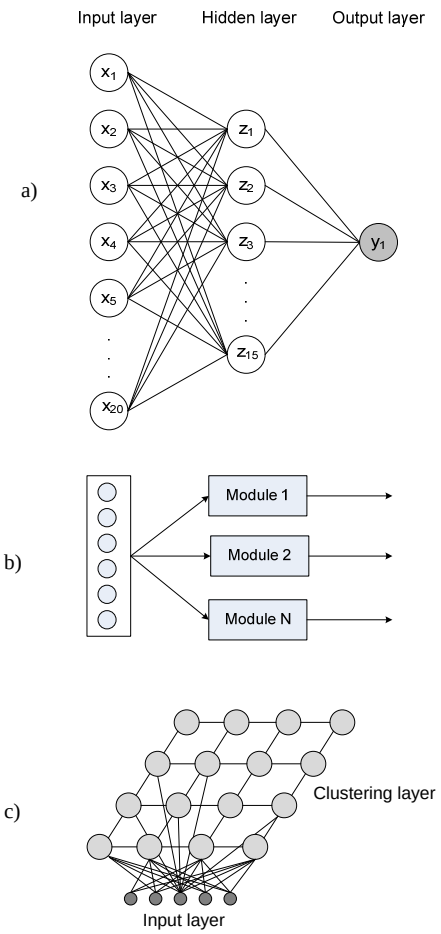

Fig. 5. ANN of different structure and learning algorithm: a) MLP with 3 layers and 36 neurons; b) committee of networks, where module 1-3 are ANN with different parameters; c) Kohonen self-organizing map.

III POSSIBILITIES OF IMPROVING ANN LEARNING ALGORITHMS

When we consider an artificial neural network to be constructed and trained for solving a real problem the question of complexity arise. Human brain consists of hundreds of billions of neurons but even most powerful computer can represent only a small part of it not even taking into consideration functional properties but just number of neurons.

ANN is trained individually for each particular task and its structure and parameters depends on this learning outcome. The next problem is learning time, which is required to train a network and find a good solution of a problem. The learning time, of course, depends on effectiveness of the algorithm, implementation technology with specific software and hardware resources available. One can build a big ANN with many neurons and train it on a specific learning dataset, however, there is a problem of so-called "over-fitting", when a network fully represents learning data remembering all these data in its weights and is failed to find hidden regularities of the task.

These are only some problems connected with ANN and some of them are more or less successfully solved. Here are some ideas of how to improve existing ANN learning algorithms according to brain theory.

The artificial neuron strength depends on its weight matrix (see Fig. 3). And the main idea of learning is to adjust these weights according to learning data. The learning time is highly dependent on initially generated weights and normalization of input data. Different normalization techniques are used to solve this issue [12].

Each real phenomenon can be fully described and understood using knowledge from many fields of science. In this case modern neurobiology could give many insights to AI scientists and especially those who are working with ANN

Human connectomics is a part of neurobiology that tries to map and analyze brain connectivity, across all scales, from the micro-level connections between neurons to macro-level between brain regions [3]. These data usually are obtained by incorporating magnetic resonance imaging (MRI) information into a form of a matrix (see Fig. 6). Then all nodes are fully described and a schematic representation of a brain region is obtained.

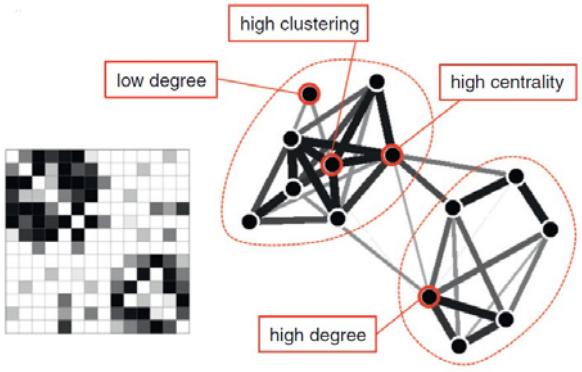

Fig.6. Schematic example of a weighted undirected connection matrix and the corresponding graph structure for a small network of 14 nodes. Nodes with high/low degree, high clustering and high centrality are indicated (adopted from [3])

When one obtains information about brain region the ANN could be designed on the base of this knowledge. In this case the ANN would be much closer to real biological neural network and probably will lead scientists to a better understanding of brain and will allow solving more complex problems than ever before.

Another very important feature of a human brain is neural plasticity or neuroplasticity [7]. This means the ability of brain to teach and to change itself according to situation. These changes include:

- Growing new neurons;
- Altering the distribution and location of neurons;

- Modifying new synaptic connections between existing neurons in the brain regions and between these regions.

This interesting feature of human brain is partially implemented in ANN. For example, the neural network pruning techniques have been developed in the recent years [5, 13]. The main idea is to remove unnecessary neurons from the network using certain parameter as an indicator. As a result fully connected ANN is optimized and the number of neurons is decreased thus getting smaller learning time and hardware resources (see Fig. 7).

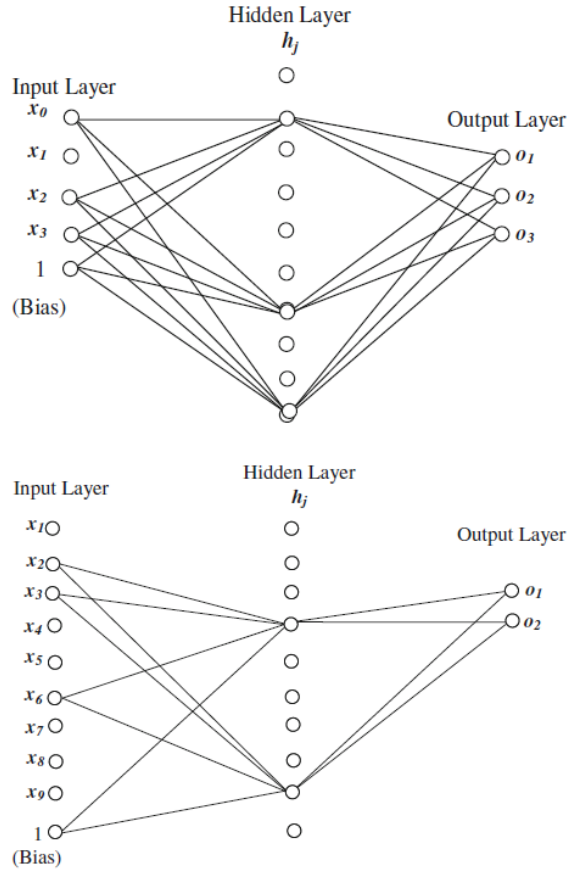

Fig. 7. Two examples of pruned networks (adopted from [5])

Figure show that in both cases the ANN is reduced by more than 50% (input and hidden layers) and is keeping its classification precision on 94% and 97% of correctly classified pattern respectively [5].

This approach is just removing unnecessary neurons and connections between them, still there may be a need for additional neurons and connections.

IV CONCLUSIONS AND FUTURE WORK

This article is a beginning of understanding how modern neurobiology could help artificial neural network researchers in their studies. We are just in the beginning of new complete and true science where every aspect of a whole is observed, described and fully understood.

There is a strong need for cooperation among scientists from different fields of science in order to get a better understanding who we are and where are we going. Artificial Neural Networks can help us to understand the functions of a brain from a different perspective, solving new tasks and real-world problems.

The directions of future study are:

- Development and creation of a new ANN structure based on MRI and neurobiological prototype of a human brain.
- Creation of a universal ANN learning algorithm which could optimize a neural network in a way of not only removing, but also adding necessary neurons and/or connections between neurons according to a problem solved.
- Development and implementation of a network committee model based on a brain functioning principles.

The ANN is a well-known technology and very widely developed and popular among AI researchers. However, there is always a place for new ideas and breakthrough. The author really hopes that the new development in neurobiology and ANN would come close together and help us in our everyday life.

V REFERENCES

[1] M. A. Arbib, *The Handbook of Brain Theory and Neural Networks*. Cambridge MA: The MIT Press, 2003

[2] G. Ascoli, *Computational Neuroanatomy: Principles and Methods*. Totowa, New Jersey: Humana Press, 2002.

[3] T. Behrens, O. Sporns, “Human Connectomic,” Current Opinion in Neurobiology, Vol. 22, pp 144-153, 2012.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, "Complex networks: Structure and dynamics," Physics Reports, vol. 424, pp. 175-308, Feb. 2006.

[5] M. Gethsiyal Augasta, T. Kathirvalavakumar, „A Novel Pruning Algorithm for Optimizing Feedforward Neural Network of Classification Problem,” Neural Processing Letters, Vol. 34, pp 241-258, Dec. 2011.

[6] M. Kaiser, “A tutorial in connectome analysis: Topological and spatial features of brain networks,” NeuroImage, Vol. 57, Issue 3, pp 892–907, Aug. 2011.

[7] Y. Perwej, F. Parwej, “A Neuroplasticity (Brain Plasticity) Approach to Use in Artificial Neural Network,” International Journal of Scientific & Engineering Research, Vol. 3, Issue 6, pp 1-9, Jun. 2012.

[8] M. Rubinov, O. Sporns, “Complex network measures of brain connectivity: Uses and interpretations,” NeuroImage, Vol. 52, Issue 3, pp 1059–1069, Sept. 2010.

[9] O. Sporns, “From simple graphs to the connectome: Networks in neuroimaging,” NeuroImage, Vol. 62, pp 881-886, 2012.

[10] O. Sporns, *Networks of the Brain*. Cambridge MA: The MIT Press, 2011.

[11] C. J. Stam, E. C. W. Van Straaten, “The organization of physiological brain networks,” Clinical Neurophysiology, Vol. 123, pp 1067–1087, 2012.

[12] R. Tadeusiewicz, R. Chaki, N. Chaki. *Exploring Neural Networks with C#*. New York: CRC Press, 2015.

[13] X. Zeng, D. S. Yeung, “Hidden neuron pruning of multilayer perceptrons using a quantified sensitivity measure,” Neurocomputing, Vol.69, pp 825–837, 2006.

Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 232-237

Niskanen Classical Model Implementation of the Office's Operations Evaluation

Ērika Žubule, Peter Grabusts
Rezekne University of Applied Science,
Atbrivoshanas alley 115, Rezekne, LV-4601, Latvia

Abstract. **The economic downturn has caused fiscal tension in the country in recent years. Budget revenues decreased and it indicated the need for the revision and control of expenditure. Therefore, the issues related to the evaluation opportunities of governmental expenditure are becoming particularly urgent. It is emphasized that a key factor in public sector's performance is the public good, which ensures services of government institutions, and encompasses both quantitative and qualitative aspects. However, economic researches on the public sector's performance efficiency focus on problems related to the determination of the public good as well as activate the possibility for its practical application. The US economist W.A. Niskanen mathematically grounded the bureau performance balance-shaping principles in the context of possibilities for evaluation of public administration, id est. the situation in which the institution's performance can be evaluated as effective. In order to evaluate efficiency aspects of the public sector's performance as well as their potential impact on the formation process of the state budget spending part, the authors of the research put forward the following objective – to perform mathematical calculations of bureau performance balance-shaping principles, grounded by the US economist W. A. Niskanen, and to assess the model's potential opportunities for practical application. As a research novelty, the developed practical application software of Niskanen's classical model of bureau performance is presented in MATLAB environment.**

Keywords: **governmental expenditure, Matlab, Niskanen model, public good, public sector's performance efficiency.**

I INTRODUCTION

The state budget reflects governmental expenditure and the income ensuring the expenditure. However, in the state finance theory and practice the biggest attention is paid to the expenditure, emphasizing the most important argument - the most efficient budget reforms are possible exactly in the sphere of governmental expenditure, because states' budget expenditure traditionally include so called unproductive expenses as well as uncontrollable expenses and the main task of governments is to reduce them.

It should be emphasized that it is especially important for a number of developed countries because they cannot ensure sufficient income from taxes if the economic situation gets worse. Consequently there is an increase in the budget deficit and loans become the main source of income. Thus, governmental expenditure and potential possibilities of reducing it have become topical in the aspect of state finances.

In the present situation the problems of forming budget expenditure have become highly topical in Latvian state finance sphere. They are related to following aspects:

- categorical terms of fiscal consolidation set a clear necessity for the reduction of budget expenditure;
- structural reforms of state administration aimed to limit governmental expenditure.

It should be stressed that in course of time different opinions were offered in order to reflect this issue in the context of planning budget expenditure. The most popular one is the model of budget maximization or the classical model of bureau performance offered by the American economist W.A. Niskanen [1]., in which the author studies the influence of clerks' performance on public finances; studies of French economists J.L. Migue and G. Belenger where the most essential attention is paid to unproductive budget expenditure that negatively influences the efficiency of budget expenditure [1],[2],[6]. At the beginning of the 21st century there appeared new opinions on the formation of budget expenditure where potential solutions were offered by Japanese economists H. Shibata and A. Shibata, marking them as possibilities of minimizing budget bureaucracy [4].

Thus, one of the most dominant issues in the state finance theory and practice is the issue of productivity of the state budget expenditure and its evaluation options. The aforementioned aspects are especially topical in the present situation of economic and

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.189

financial crisis when in the context of conditions of fiscal consolidation countries pay bigger attention to the management of state expenditure. Limited resources of state finances accentuate the problems of expenditure of state functions and their efficiency. Therefore, the most essential task is to carry out structural reforms in the state administration. The authors declare that they have to ensure not only the evaluation of the structure of state expenditure but also the revision of those management systems, which in course of time influenced the broadening of governmental functions and consequently the increase of states' budgets. Thus, the aspect of the results of public sector activities and the necessity of their evaluation becomes important.

The strengthening of the opinion that in the state administration there are limited possibilities to use evaluation elements of efficient and purposeful performance characteristic to the private sector caused the developed countries to gradually elaborate and introduce the system of productivity indices during the process of budget reforms with a purpose to ensure efficiency evaluation of the public sector administration. The conformity of the state administration performance to the needs of the society was set as its main criteria. According to the system of productivity indices the approach to the amounts of state budget finances and their efficiency should be justified, indicating the target of the use of financial funds, in the context of which the target results are planned and determined [3],[5].

It is emphasized that the most essential index in the evaluation of the public sector performance is the public benefit that ensures services of state institutions. It includes both quantitative and qualitative aspects. However, economic researches on the efficiency of the public sector performance indicate the problems of determining the public good as well as emphasize the possibility of its practical application [5].

In order to evaluate aspects of the efficiency of the public sector performance and their potential influence on the process of making state budget expenditure, the authors of the research set a target – to do mathematic calculations of the balance principles of bureau performance grounded by the American economist W.A. Niskanen and evaluate their possibilities of practical application.

The tasks of the research are subordinated to the aim of the research:

- To study the theoretical ground of W.A. Niskanen's classical model of bureau performance.
- To make the program realization of the practical application of the model.
- To summarize the results and draw conclusions.

In order to achieve the aim and implement the tasks the following research methods are used: content analysis, logically-constructive, mathematical modelling on the basis of the study of opinions existing in economic literature.

II THEORETICALLY MATHEMATIC SUMMARY OF W.A.NISKANEN'S CLASSICAL MODEL OF BUREAU PERFORMANCE

In the 90s of the 20th century the American economist W.A. Niskanen, after having studied the factors influencing the formation of the governmental budget expenditure, emphasized the importance of the civil service or bureaucracy, which is characterized by distinctive instincts of self-protection and self-preservation and all of it influence the possibilities of clerks in the formation of the state budget expenditure. In the context of evaluation possibilities of the state administration performance, the economist mathematically grounded the balance principles of bureau performance, that is, a situation when the performance of an institution can be evaluated as efficient [2].

The classical model of bureau performance offered by W.A. Niskanen accentuates main indices that influence efficiency of the institution's performance: expenditure, amounts of allotted budget funds, services of the institution and the public good in general ensured by results of the institution's performance. The following aspects are emphasized in the model:

- Bureau is a monopolist in providing certain services.
- Bureau work is financed from the budget funds and is allotted by authority organs: government, parliament.
- The bureau has an essentially better access to the information on service costs than authority organs and such asymmetry of information creates a situation when the bureau is allotted bigger funds than it really needs.

The following aggregates are introduced:

- B – budget funds allotted to the bureau (*Budget*) that the society sees as a benefit from the amount of provided services Q, therefore, B is a function from bureau activities or the amount of services provided by the bureau.
- C – costs of the bureau (*Costs)*, which also depend on Q.

Niskanen's model is based on the assumption that a complete information about $C=C(Q)$ is available only to the bureau, which tries to get the maximum amount of budget funds, but the dependence $B=B(Q)$ can be conceived only by those who grant finances.

The mathematical description of Niskanen's model is as follows. The formulas are introduced:

$$V = a - 2bQ \quad (1)$$

$$C = c + 2dQ \quad (2)$$

where V – the maximum good for consumers;

C – the maximum costs of the bureau;

Q – the amount of services provided by the bureau;

a – the government readiness to pay for services of the bureau;

b – the potential amplitude of changes of the parameter "*a*";

c – costs of the bureau;

d – the potential amplitude of changes of the parameter "*c*".

One can conclude that parameters *a* and *b* are related to the amounts of budget funds granted and used, but parameters *c* and *d* - to the costs ensuring the institution's work. Regarding given conditions:

$$B = aQ - bQ^2 \quad (3)$$

$$TC = cQ + dQ^2 \quad (4)$$

where *B* – the total budget of the bureau and *TC* – total minimal costs of the bureau.

Regarding these conditions the *Q* balance level is determined in the following way: maximization gives the highest limit of *Q*: *Q=a/b*. The limitation that *B>TC* gives the lowest level of *Q*: *Q = 2(a-c)/b+2d*. These *Q* levels are the same at *a=2bc/b – 2d*. Thereby, the balance level of services provided by the bureau is achieved at these conditions:

$$Q = \begin{cases} = \dfrac{a-c}{b+d} & a < \dfrac{2bc}{b-d} \\ = \dfrac{a}{2b} & a \geq \dfrac{2bc}{b-d} \end{cases} \quad (5)$$

Thus, W.A. Niskanen grounds mathematically the optimal amount of services provided by the bureau and accordingly granted budget funds and real costs of bureau actions. But it should be emphasized that the economist has accepted the following conditions as the most favorable ones: *a*= 100, *b*=1, *c*=75, *d*=0,3, and has set the parameter "*a*" as the most important.

W.A. Niskanen's classical principle of bureau performance balance at different values of the parameter "*a*" was verified in practice in the experimental part of the research. There is also included the calculation of the optimal value, which indicates the efficient work of the institution, that is, when the maximum good is provided to consumers of institution's services. It is related to the maximal difference of the total budget of the institution and ultimate costs (*B-C*).

III PROGRAM REALIZATION OF PRACTICAL APPLICATION OF NISKANEN'S MODEL

A program support in *Matlab* environment was made in order to illustrate the classical model of bureau performance grounded by W.A. Niskanen. The programming language *Matlab* (www.mathworks.com) was used to implement the program environment. At Niskanen's accepted conditions where *a*=100, *b*=1, *c*=75 and *d*=0,3, according to formulas (2), (3), (5) the following optimum values were acquired: *Q*=124,2857, *B*=4687,3469 and *C*=149,5714. It is graphically shown in Figure 1.

Figure 2 presents the interface of our program. After it is run, the main window will appear where fields for entering basic information are pictured (initial values are given by default).

The program requires entering parameters *a*, *b*, *c*, *d* indicated in Niskanen's model and their increase values. Then the calculation process of the values *Q*, *B*, *C* of the given model is started and the optimal values of bureau performance balance accentuated in the model and their appropriate graphical diagram appear on the screen. At the same time all parameters *a*, *b*, *c*, *d* of the model and the calculated values of *Q*, *B*, *C* are written in the text file, which can be used to analyze the model numerically.

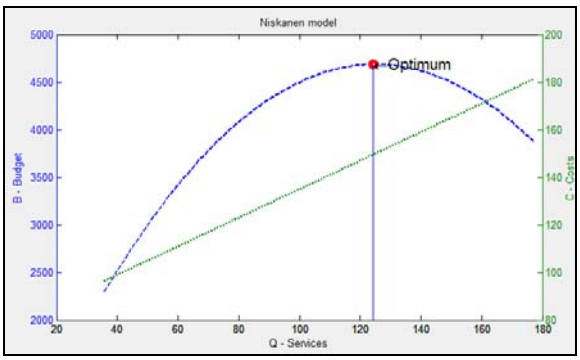

Fig. 1. Optimal values of the amount of services provided by the institution, budget funds and ultimate costs at Niskanen's determined

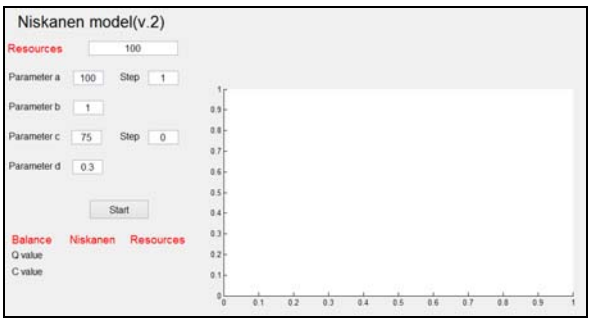

Fig. 2. Window of main basic information

If the data are entered correctly, one can begin the calculation procedure of values actualized in the model by clicking on *Start.* If the process is successful, the window will appear where the results of operation will be shown (Figure 3).

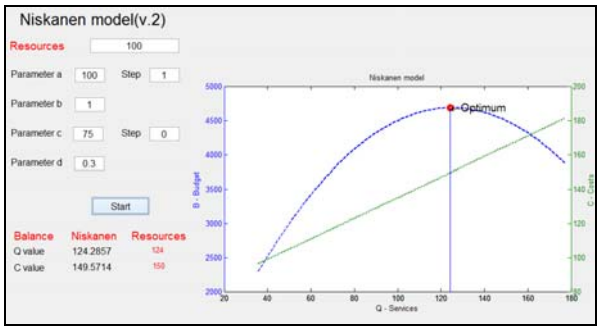

Fig. 3. Window of operation results

In the left bottom corner of the window the calculation results can be seen. They characterize Niskanen's principles of optimal balance of bureau performance at the initial values of *a*, *b*, *c*, *d*, that is, the values of *Q* and *C* (Figure 4).

Balance	Niskanen
Q value	124.2857
C value	149.5714

Fig. 4. Optimal values of the balance principle of the model

On the right side of the window the diagram is made, which shows the functions *B(Q)* and *C(Q)* (Figure 1). The dependency of *B* on *Q* is shown on the left side of the diagram; the calculated values of *B* appear on the left axis of coordinates. The function *B(Q)* is shown by the blue stripy line. Accordingly, the dependency of *C* on *Q* is shown on the right side of the diagram and the calculated values of *C* are given on the right axis of coordinates. The dotted green line displays the function *C(Q)*. Simultaneously the optimum point is being calculated. It shows the maximum *B-C* value (on the diagram it is shown as a red circle with a word *Optimum* next to it).

If the parameters of the model are changed, for instance, if the value 0,25 is entered in the box of *Parameter d*, the following results are acquired (Figure 5).

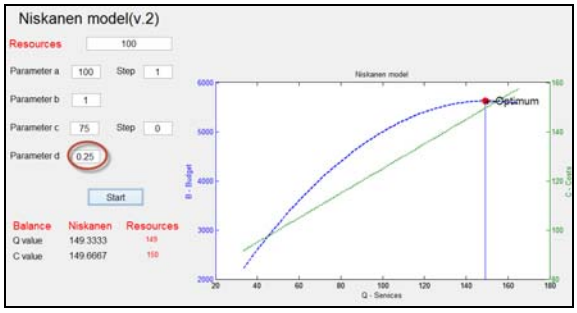

Fig. 5. Graphical interpretation of the model at the condition where *d*=0,25

After experimenting with different parameter values of *a*, *b*, *c*, *d*, it can be established that in the most cases Niskanen's model does not achieve a balance and the optimum is not calculated. The diagram is made but it does not correspond to the balance principles of the model. For instance, after changing the parameter value of *b* to 10, the following results are acquired (Figure 6).

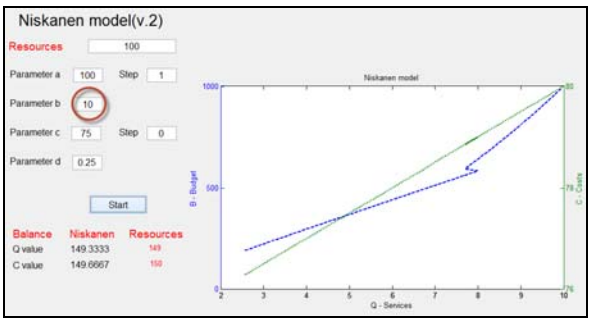

Fig. 6. Graphical interpretation of the model at the condition where *b*=10

Thus, it can be concluded that a trustable interpretation of Niskanen's model of bureau performance depends on the certain proportional changes of initial parameter values of *a*, *b*, *c* and *d*.

The results of program operation were written in the text file after each entering. The initial parameter values were determined: *a*=100,00 with a step = 1; *b*= 1; *c*= 75, with a step = 0; *d*= 0,30. At these conditions the optimal situation, when the difference between *B* and *C* is the biggest, is achieved when the value of "*a*" is 162,00, the value of *Q* - 124,29, *C* – 149,57, *B* – 4687,35, and the difference of *B-C* is 4537,78.

IV APPROBATION OF PROGRAM REALIZATION OF PRACTICAL APPLICATION OF NISKANEN'S MODEL

In order to approbate principles of Niskanen's classical model of bureau performance, the authors used financial indices of Latvian Ministry of Economics for the year 2011. They are structured in the following way (explanation of the law of the Republic of Latvia "On the state budget for the year 2011", part 5.3 [7]):

1. Resources for covering the costs (111,6 millions monetary units), including:
   - Income from charged services – 0,78.
   - Financial aid from abroad – 0,65.
   - Grants from general income – 110,2.
2. Expenditure – in total (114,8 millions monetary units).
3. Financial balance: - 3,2 millions monetary units (it is planned to "cover" it with the increase or reduction of funds balance from charged services and other income).

Interpreting the above mentioned quantitative indices one can say that the index "*a*" (*the readiness of the government to pay for bureau services*) is 110,2 millions monetary units, but the costs of the bureau or

the index "*c*" -114,8 millions monetary units are bigger than the government is ready to ensure.

The task is to use Niskanen's balance principles of bureau performance in order to find out which values of *Q* and *C* correspond to them. The result of the program realization were the following values of indices: *Q* – 136,9 millions monetary units, *C* – 164,8 millions monetary units (see Figure 7).

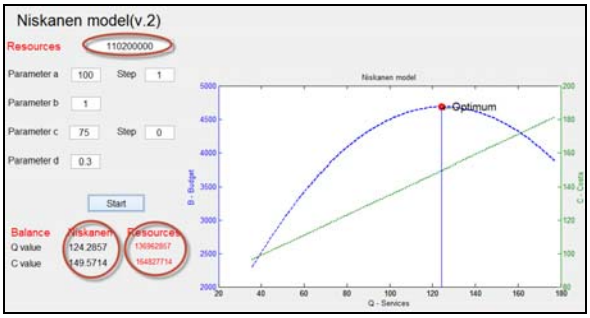

Fig. 7. Graphical interpretation of the financial indices of Latvian Ministry of Economics

It can be concluded that:

- According to the evaluation of the financial operations of the ministry, it needs 111,6 millions monetary units .
- 110,2 millions monetary units are granted from the state budget to ensure its work.
- On the basis of Niskanen's balance principles of bureau performance, the optimal amount of services provided by the ministry is 136,9 millions monetary units (*Q*), but the ministry could declare 166,9 millions monetary units as the maximum cost (*C*).

Thus, it can be said that the amount of services provided by the ministry in the situation of balanced performance exceeds the amount of funds determined by the ministry and envisaged in the budget for covering necessary costs. It means that bigger financial resources are needed to ensure the functions of the ministry than they are at the moment. However, it should be stressed that according to Niskanen's assumption, the ministry, when demanding budget funds, could operate with a bigger amount of necessary costs.

V THE RESULTS OF PROGRAM REALIZATION OF PRACTICAL APPLICATION OF NISKANEN'S MODEL OF BUREAU PERFORMANCE

1. After checking experimentally balance principles of the classical model of bureau performance with different values of parameters *a*, *c* and their increases *b*, *d*, it can be established that in the most cases Niskanen's model does not reach the balance and correspondingly it is not possible to calculate optimal values. Curves are graphically constructed but they do not show the optimal situation (Figure 6).

2. On the basis of Niskanen's balance principles of bureau performance at constant values of *b*, *c*, *d*, it can be concluded that after gradually increasing values of the parameter "*a*" and recording corresponding values of the amount of services (*Q*), maximal costs (*C*), total budget (*B*) and the difference between granted budget funds and costs (*B*-*C*), there can be seen a proportional increase in amounts of services provided by the bureau and its costs, as well as in the difference between common budget funds and maximal costs

3. The increase values of the parameter "*a*" are clarified, when, firstly, the difference between granted budget funds and costs is the biggest, which indicates the most favourable situation in bureau's work and, secondly, when the corresponding index begins to decrease, which indicates the negative character of the efficiency of bureau's work.

It can be concluded that the balance principle of bureau performance grounded by W.A.Niskanen works only at certain parameter values.

Possibility of practical application of Niskanens's model:

1. Institutions that receive funds from the state budget can determine the amount of services provided and consequently the efficiency of their performance, which will be showed by the balance point of provided services, when the difference between granted budget funds and service-ensuring costs increases and reaches the maximum.
2. Niskanen's model of balance principles of bureau performance can help to analyze the situation at certain parameters, creating a system-approach to the evaluation of state institutions' work and consequently to the improvement of making budget expenditure. Thus, it can serve as a base model for the improvement of planning state budget expenditure.

VI CONCLUSIONS AND SUGGESTIONS

The conclusions and suggestions of the research follow from the evaluation of the possibility to apply Niskanen's model of bureau performance in theoretical and practical ways.

1. Niskanen's classical model of bureau performance accentuates main indices that influence the efficiency of institution's work: costs, amounts of granted budget funds, services provided by the institution and the social good ensured by the performance of the institution in total.
2. Determining concrete initial parameter, that is, values *a*= 100,00 with a step=1,00; *b*=1,00; *c*=75,00 with a step= 0; *d*=0,30, the optimal situation, when the difference between *B* and *C* is the biggest, is achieved when the value of "*a*" is 162,00, the value of *Q* - 124,29, *C* – 149,57, *B* – 4687,35 and the difference *B*-*C* is 4537,78.

3. With a help of the program realization for checking Niskanen's balance principles of bureau performance, it is possible to determine the efficiency of provided services and bureau performance from an institution financed from the state budget. It will be showed by the optimal point of provided services when the difference between granted budget fund and service-ensuring costs increases and reaches the maximum.
4. With a help of Niskanen's model of balance principles of bureau performance it is possible to analyze the situation according to certain parameters, thus ensuring a system-approach to the evaluation of state institutions' work and consequently to the improvement of making budget expenditure. Thus, it can serve as a base model for the improvement of planning state budget expenditure.

VII REFERENCES

[1] W.A. Niskanen, *Bureaucracy and Representative Government*. Chicago, 1971.
[2] W.A. Niskanen, *Bureaucracy and Public Economics*. Cheltenham, Edward Elgar, 1994.
[3] J.E. Stiglitz, *Economics of the Public sector*. Princeton University, 1988.
[4] H. Shibata, A. Shibata, *The budget- minimizing bureaucrat, The Challenge of Fiscal Disparities for State and Local Governments.* The selected Essays of Ladd H., E.E.Cheltenham, UK, Horthampton, MA. USA, 1999.
[5] *Public Productivity Handbook*, Second Edition, ed.by M.Holzer, 2004.
[6] J.L. Migue and G. Belanger, *Toward a General Theory of Managerial Discretion. Public Choice*, vol.17, p. 37-46, 1994.
[7] The explanation of the law of the Republic of Latvia "On the state budget for the year 2011", part 5 (in Latvian).

PRINCIPLES OF PUBLICATION USE AND ETHICS

The Publisher states the following principles of publication use and ethics:

1. Only original works, that have not been previously published and in which the data do not infringe the copyright of another person or organization, are accepted for publishing.
2. Every submitted paper is reviewed anonymously by 2 reviewers.
3. Publisher has the right to use plagiarism detection software to screen submitted papers.
4. Authors retain copyright over their work, while allowing the conference to place this unpublished work under a Creative Commons Attribution License, which allows others to freely access, use.
5. Publisher undertakes to publish articles accepted after reviewing and, according to the scientific quality of an article, publish them electronically and include them in various international scientific paper data bases.
6. Publisher undertakes to use author's personal data only for the needs of identification of authors and co-authors, as well as to preserve data in accordance with the restrictions determined in the current regulations of the Republic of Latvia.

The author (-s) of original article confirms and guarantees that:

1. A paper has been written by the Author (s) who has (have) been stated in the Paper and in this statement and there is no any other person who would have any rights to this Paper.
2. The Paper is original; it has not been previously published and is notunder simultaneous consideration by another publications.
3. The Paper will notlater be submitted elsewhere.
4. There are references for all the pictures, data and other materials used in the Paper.
5. The results of the Paper are genuine; they do not infringe on the original work of another person or organisation. The Author (-s) have taken all the necessary measures to ensure the requirements of the sponsors' who finance the Paper.
6. If author finds a significant error after the publishing of the paper, it is the author's obligation to promptly notify the editor.
7. The Author (-s) has (have) been authorised to submit the Paper for publication; the intellectual property rights will not be infringed according to the 14th, 15th articles of the Copyright law and other normative acts of the Republic of Latvia and international normative acts binding for the Republic of Latvia.
8. The Author (-s) take (-s) the sole responsibility related to the third parties' objections and demands concerning the publishing or using the Paper.
9. The Author (s) take (-s) all the responsibility and promptly, irrevocably and unconditionally undertakes to independently at his (their) own expenses and risk if there will be objections or demands from the third parties' concerning the publishing or using the Paper.

Publication editor:

1. Ensures double-blind peer-review of the submitted paper.
2. Until the publishing, ensures the confidentiality of information concerning the submitted article.
3. Strives to prevent any possible conflict of interests between authors and reviewers or editors.

Reviewer:

1. Evaluates the content of the article according to the scientific article criteria, without regard to ethnic origin, gender, sexual orientation, citizenship, religious beliefs or political views of the authors.
2. Keeps confidential all the information related to submitted paper.
3. Notifies the editor if there is a suspicion of plagiarism or copyright infringement.
4. If the reviewer feels unqualified to review or knows that it is impossible to provide review in time, he should notify the Editor-in-Chief and ask to exempt him from reviewing.