

Vol. 1. Issue 2, 2015

ISSN: 2453-7314

# ***Central European Researchers Journal***





## A WORD OF WELCOME FROM THE EDITORS

Dear Colleagues, Readers and Authors,

We are very glad to propose to you next number of the “Central European Researchers Journal (CERES)”. As previous number the topics of this number cover different areas and problems of engineering. Most of authors of accepted papers are young researchers. They have been participated at the special Workshop on CERES: Modern Experience on Young Researchers Organization (<http://idt.fri.uniza.sk/idt2015/index.php?clanok=section> ). This workshop has been organized under the International Conference on Information and Digital Technologies (July 7th - 9th, 2015, Zilina, Slovakia) and supported by the Tempus CERES -Centers of Excellence for young REsearchers (544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES). All accepted papers have been reviewed and selected by members of Conference Programme Committee and journal Editorial Board. We hope that these papers will be interesting for readers and these publications will be useful for authors.

Some of papers in this number have been presented at the International Workshop “Information Technologies in Medical and Biomedical Application” that was held at University of Zilina 26-27 November 2015.

With best wishes

*Prof. Vyacheslav Kharchenko*  
*Prof. Elena Zaitseva*

**Central European Researchers Journal.** Volume 1. Issue 2

**Editor-in-chief:** Kharchenko Vyacheslav, Zaitseva Elena

**Editorial Board:** Belotserkovsky Alexei, Bezobrazov Sergei, Dmytrychenko Mykola, Drahansky Martin, Kachurka Pavel, Khakhomov Sergei, Levashenko Vitaly, Liauchuk Viktor, Lukashevich Marina, Melnychenko Oleksandr, Oliinyk Andrii, Slavinskaya Elena, Subbotin Sergey, Tatur Michail, Vojnar Tomas.

**Addres of the editorial office:** Central European Researchers Journal - editorial, Faculty of Management Science and Informatics, University of Zilina, Univerzita 8215/1, 01026, Zilina, Slovakia, [editorial@ceres-journal.eu](mailto:editorial@ceres-journal.eu)

Each paper was reviewed by reviewers.

**Publisher:** JMTM, s.r.o., Sad SNP 8, 010 01, Zilina, Slovakia, [publisher@jmtm.sk](mailto:publisher@jmtm.sk)

Published biannually

ISSN: 2453-7314

December 2015

## CONTENTS

<i>Andrey Khobnya, Viktor Liauchuk, Oleg Demidenko</i> Formalization of the quality of service mechanisms in packet-switched networks . . . . .	1
<i>Helena Zhivitskaya, Tanya Safronava</i> Fuzzy model for inventory control under uncertainty . . . . .	10
<i>Mary Tyagunova, Tanya Golub, Andrey Proskura</i> Integrated Approach to Improve Reliability of Neural Network . . . . .	14
<i>Sergey Subbotin</i> The Set of Data Transformations for Dimensionality Reduction in Technical and Biomedical Diagnosis . . . . .	23
<i>Valerii Lovkin</i> Risk-Oriented Decision Making During Integrated Investment Management under Uncertainty . . . . .	36
<i>Ruslan Vikhorev, Sergey Tyurin</i> Advanced logic cell FPGA . . . . .	42
<i>Olena Shytikova, Galyna Tabunshchuk</i> Automated decision support system for GTU tests process . . . . .	47
<i>Anastasiia Strielkina, Artem Tetskyi, Bogdan Selin, Oleksandr Solovyov, Dmitriy Uzun</i> Service for Vulnerabilities Analysis and Security Assessment of Open Source Systems . . . . .	53
<i>Sergei Khakhomov, Uladzislau Beizerau</i> Francisk Scorina Gomel State University in implementation of Tempus project “Eastern partnership in pedagogical innovations in inclusive education” (INOVEST) . . . . .	65
<i>Anna Trofymenko</i> Optimization of medical institutions in Ukraine . . . . .	74
<i>Liauchuk Veranika</i> Identification of explicit lexicon for opinion mining in the media . . . . .	81
<i>Tetiana Bragina, Galyna Tabunshchuk, Dmitro Moroka</i> Testing methods library for applications with web-based interfaces . . . . .	90
<i>Jana Jurecekova, Eva Babusikova, Jan Kliment, Dusan Dobrota</i> Genome-wide association study of prostate cancer . . . . .	95
<i>Eva Babusikova, Milos Jesenak, Dusan Dobrota, Jana Jurecekova</i> A new perspectives in a detection of markers of human diseases in Slovakia . . . . .	96

# Formalization of the quality of service mechanisms in packet-switched networks

Andrey Khobnya, Viktor Liauchuk, Oleg Demidenko

**Abstract**—This paper presents analysis and generalization of quality of service (QoS) mechanisms in packet-based networks. Decomposition of QoS mechanisms is presented. Each QoS mechanism can be split into four components each of which is responsible for the defined subtask. Method of formalization is presented in which each QoS component is presented by function of the defined arguments. This method was used for simulation of QoS mechanisms in next generation networks.

**Keywords**—network traffic, QoS, simulation.

## I. INTRODUCTION

The objectives of the simulation of the quality of service mechanisms in networks can be the following:

- optimization of the quality of service (QoS) control mechanisms settings in order to achieve better parameters of network performance;
- calculation of the quality of service characteristics using different settings in order to determine the reserves to increase the network load;
- determination of influence degree of various traffic classes to the network;
- calculation of the threshold load at which the quality of service will be provided under the current settings;
- research of existing and development of new quality of service mechanisms.

At present the formalization of the quality of service mechanisms in packet-switched networks as a reasonably accurate analytical models is an impossible task. An alternative approach is the simulation of complex systems. Simulation is a research method in which the investigated system is replaced by a model with sufficient accuracy to describe the real system which conducted experiments aiming to the study of this system. This paper presents the analysis, generalization, formalization and simulation method for quality of service mechanisms in data networks with packet switching. The presented method of generalization can be used to build specific simulation models as well as a platform for simulation and research of quality of service mechanisms.

## II. OVERVIEW OF THE QUALITY OF SERVICE MECHANISMS

Let's consider the quality of service mechanisms used in networking equipment. One of the largest manufacturers of communications equipment split provided QoS tools to *Congestion Management* tools and *Congestion Avoidance* tools. Congestion Management tools include FIFO Queueing, PQ, CQ, WFQ, CBWFQ and LLQ. Congestion Avoidance tools include WRED, DWRED, Flow-Based WRED and DiffServ Compliant WRED. FIFO Queueing provides basic store and forward packets. This is the simplest algorithm for processing the queue, which in some cases can be used by default and does not require configuration. PQ allows one to set strict priorities for critical traffic. PQ guarantees the primary processing of the

A. Khobnya, Francisk Skorina Gomel State University, Gomel, Belarus (e-mail: andrey@khobnya.me).  
V. Liauchuk, Francisk Skorina Gomel State University, Gomel, Belarus (e-mail: liauchuk@gmail.com).  
O. Demidenko, Francisk Skorina Gomel State University, Gomel, Belarus (e-mail: demidenko@gsu.by).

most important traffic. Priorities for packages can be flexibly defined depending on ports, source and destination addresses, packet size, etc. CQ reserves a predetermined fraction of the total bandwidth of the network interface for each given type of traffic. If any type of traffic does not use dedicated amount of bandwidth than the other types of traffic use it instead. WFQ splits traffic into multiple streams based on parameters such as source address, destination address, source and destination port, etc. Each traffic stream is of applied priority (or weight). CBWFQ extends the functionality of WFQ providing the ability to configure traffic classes. CBWFQ allows one to select the exact percentage of bandwidth for each class. Up to 64 different classes of traffic can be configured. LLQ allows one to give the absolute priority to one traffic class and adjust the proportion of bandwidth for other classes in CBWFQ. WRED is an implementation of the well-known algorithm RED which uses the IP Precedence to ensure the privilege of traffic with higher priority. DWRED is an extension of WRED and it allows one to set the minimum and maximum thresholds for the packets queue length of different classes of traffic. Flow-Based WRED and DiffServ Compliant WRED is also an extension of the algorithm WRED and it allows management to separate the destruction of packages for different flows and different traffic classes, respectively [1].

Other manufacturers of network hardware and software use similar concepts to build quality of service mechanisms. Also they use the concept of HTB to drop and ensure the priorities of packets [2]. In general the quality of service mechanism in any network classifies incoming packets, distributes them across multiple stores or deletes some of them, then determines the sequence of sending packets from different stores.

### III. ANALYSIS AND GENERALIZATION OF THE QUALITY OF SERVICE MECHANISMS IN PACKET-SWITCHED NETWORKS

After analyzing algorithms of quality of service mechanisms we can distinguish four main components of each QoS mechanism:

- packets queues;
- packet classification algorithm;
- active queue management algorithms;
- scheduling algorithm.

*Packet queue* is the simplest queue with sequential processing. The quality of service mechanism can use one or more packet queues. Generally the differences between the various mechanisms are in the other three components.

*Packet classification algorithms* provide the packet class determination and distribution of packets of different classes across different queues. Let's consider the existing methods for traffic classification used in the quality of service mechanisms.

*Classification by IP Precedence (ToS)* is a classification method that uses a three-bit code from IP Precedence field of Type of Service (ToS) header in IPv4-package. This method allows one to split eight types of traffic based on priority.

*Classification by DSCP* is a classification method that uses a 6-bit code Differentiated Services (DSCP) in 8-bit Differentiated Services Field (Field DS) of IP-packet header. Theoretically the network may have up to 64 different classes of traffic with different values of the DSCP. It gives operators of telecommunications services greater flexibility in the definition of traffic classes.

*Flow-based classification* is a way of traffic classification which can use TOS, DSCP, IP-address of the sender, IP-address of the recipient, the sender port and destination port. It is used to distinguish different traffic flows in the network interface.

*Classification by MPLS QoS* is a traffic classification method to be applied to in cases when MPLS protocol is used. It uses a three-bit field Traffic Class which allows one to separate the

eight traffic classes. Generally this method of classification is used in conjunction with the others. It is often used in next generation networks.

*Deep Packet Inspection (DPI)* is a classification method based on the analysis of package contents which allows one to separate traffic of different application level protocols and applications.

Summarizing the above mentioned facts, let's identify the following main features of traffic classification algorithms:

- usage of special marker fields of IP-packet header in order to identify the class of traffic (ToS, DSCP, MPLS QoS);
- usage of sender IP-address, receiver IP-address, sender port and receiver port;
- usage of information about the application level protocols and applications.

These properties need to be considered in constructing simulation models of the quality of service mechanisms.

The next important component of the quality of service mechanism algorithms is active queue management (AQM). Active queue management (AQM) is the class of algorithms that control the destruction of packet in network interface in case of an overflow or near overflow state of device internal buffer. It is necessary for reducing a load to a network. Let's consider the existing algorithms for active queue management used in the quality of service mechanisms.

*Tail Drop* is a simple algorithm for active queue management used to control the overflow. Unlike more complex algorithms Tail Drop does not differentiate the traffic and handles all of the packets in an identical way. When the size of the buffer queues packets reaches the maximum value, Tail Drop starts to discard all newly arriving packets. Admission of new packages will resume when the buffer has enough space for receiving incoming packets.

*Random Early Detection (RED)* is a more complex algorithm of active queue management designed to prevent network congestion. RED is focused on addressing the major drawbacks of the algorithm Tail Drop: uneven buffer allocation for different traffic flows and global synchronization of TCP-connections which leads to uneven pulsating network load. RED takes into account the average value of the queue size and drops packets based on statistical probabilities. If the buffer is almost empty then it takes all arriving packets. As the size of the queue increases and the probability of dropping, the incoming packet is raised as well. When the buffer is filled, the probability reaches one and, therefore, all arriving packets are dropped.

RED algorithm allocates resources more effectively than Tail Drop because of its being less sensitive to the negative effects of bursty traffic which itself occupies a small fraction of the bandwidth and does not allow it to block the activity of the other TCP-connections. More packets are sent from the source, then it is more likely that a particular packet from a given source will be discarded, since probability is proportional to the number of events in the data queue. Thus RED can be a problem solving of global synchronization of TCP-connections.

*Weighted random early detection (WRED)* is an extension of algorithm RED which enables one to set several different queue thresholds depending on the type of traffic. Thus the probability for packets is different for various classes. For example, the interface can have a low threshold for low priority traffic. When filling in the queue until this threshold, all packets with lower priority will be discarded. Thus WRED protects a high priority traffic in the same queue.

*Adaptive random early detection (ARED)* is an extension of algorithm RED which allows one to take the original algorithm more or less "aggressive" to the packets depending on the value of the average queue length. If the average queue length packet oscillates around the minimum threshold, the risk of dropping the packet increases. Conversely, if the average queue length is around the maximum threshold, the algorithm works more conservatively, i.e. probability of destroying the package is reduced.

*Robust random early detection (RRED)* is an extension algorithm RED designed to protect against network attacks such LDoS (Low-rate Denial-of-Service is denial of service due to its low speed). The basic idea is to identify and filter packet attacks before classical algorithm RED or its modifications will be applied to packets. LDoS attack packets are identified and discarded by this algorithm. As a rule, the sender TCP-packet starts to pause before sending a new packet if it detects a loss. Therefore, if the source sends a packet within a short period of time after the loss of a package sent to them, this sender can be the source of the attack. This idea is the basic one for the identification algorithm packages LDoS attacks inside RRED [3].

*RED with Preferential Dropping (RED-PD)* is an extension algorithm RED which uses the story of the destruction of broadband packages in order to identify flows.

*Random Exponential Marking (REM)* is an active queue management algorithm similar to RED, but it differs from its certain measures load and function of the probability of destroying packets. The dependence of the probability function of the load in REM is close to logarithmic. The function measures of the load depends on the difference between the total rate of incoming traffic and upstream bandwidth, as well as on the difference between the current queue length and a predetermined threshold. REM algorithm aims to provide high load and low losses.

*Blue* is an active queue management algorithm which differs from the RED algorithm that requires no setup by the network administrator. It is able to automatically learn and adapt. Blue maintains the value of the probability  $p$  and discards any incoming packet with a given probability. Whenever the queue overflows,  $p$  is increased by a small constant  $p_d$ , and whenever the queue is empty,  $p$  decreases by  $p_i < p_d$ . If the composition of the traffic does not change, then the  $p$ -value slowly converges to the optimum and makes full use of the communication channel.

*Stochastic fair Blue (SFB)* is an extension algorithm Blue which allows one to separate different traffic flows and to maintain a variety of probability values for each stream. SFB provides uniform distribution of the buffer capacity for each flow of traffic.

*Resilient Stochastic Fair Blue (RSFB)* is an extension of algorithm SFB which is designed to protect against attacks, such as spoofing DDoS. This algorithm improves on the algorithm with SFB flow separation.

*Sontrrolled delay (CoDel)* is an algorithm for active queue management designed to overcome the effect of excessive buffering (buffer bloat) on network interfaces by setting limits on the delay of arrival of the package. The algorithm works within certain intervals. Initial interval lasts for 100 ms. Time delay is calculated for each queued packet which is equal to the timeout of packet in the queue. Minimum time delay is saved. When the last packet is deleted from the queue interval providing that the lowest delay for the current time interval is more than 5 ms, this single packet is discarded, and the interval used for the next group of packets is reduced. If the lowest delay for the current interval is less than 5 msec, the packet is sent and the duration of the interval is reset to 100 ms. The interval is reduced in accordance with the inverse square root of the number of consecutive slots where the packages have been destroyed due to excessive delay values.

*Active Virtual Queue (AVQ)* is an active queue management algorithm based on the use of the virtual queue the capacity of which changes dynamically based on the rate of arrival of new packets and the smoothing parameter. It happens because the destruction of the packages should be more aggressive when the load exceeds the required interface and less aggressive when loading is below the required [4].

*PI Controller* is an active queue management algorithm based on the algorithm of proportional-integrator controller.

Summarizing the information discussed above, let's identify that active queue management algorithms are characterized by the following features:

- the ability of customization (most algorithms have customizable settings);
  - management of only one queue;
  - usage of the data about queue length and buffer capacity;
  - usage of data about source, destination and other fields of the packet header (e.g., WRED identifies different traffic classes, SFB supports multiple traffic flows, etc.);
  - storing and maintaining the internal state (e.g., Blue stores and maintains probability parameter, RED-PD stores history of packets, etc.);
  - great number of algorithms involve the destruction of the package before putting it in place.
- However, for example, CoDel destroys packets when they are removed from the queue.

These properties need to be considered in constructing simulation models of the quality of service mechanisms. Also note that if the quality of service mechanism uses several queues for packets, then different algorithms of active management can be used for different queues.

Another important component of QoS mechanisms is the scheduling algorithm. The algorithms of this class control the sequence of packets to be sent. Let's consider the existing scheduling algorithms used to send packets designed to ensure the quality of service mechanisms.

*First In First Out (FIFO)* is the simplest scheduling algorithm for sequentially processing a single storage working in the principle of "first come - first out". When using this algorithm, traffic prioritization is not performed, i.e. all packets have the same permissions.

*Priority Queuing (PQ)* is the scheduling algorithm which works with absolute priority queues. It provides an absolute priority of a certain class of packets over other packet classes. As a rule, the implementation of this algorithm assumes the use of four queues:

- high;
- medium;
- normal;
- low.

Processing of packages from different queues performs sequentially. Prior to complete processing of all packets from a queue with a higher priority processing of a low-priority packet queue is not performed. In the presence of the incoming packet intensive real-time high-priority traffic the outer channel is monopolized by packages of this type of traffic, i.e. a high priority queue in this case will never be exhausted and all other traffic will be blocked. This is a major drawback of this scheduling algorithm.

*Round-robin* is the scheduling algorithm works by iterating by a circular loop. Queues are processed sequentially in each iteration and packet from each queue is sent.

*Weighted round-robin (WRR)* is an extension of round-robin scheduling algorithm which allows one to set different weights for different queues. WRR per iteration sends the queued packets directly proportional to the weight and inversely proportional to the average size of the last packet in the queue.

*Weighted Fair Queuing (WFQ)* is a scheduling algorithm which provides per-flow bandwidth share depending on its weight. This algorithm is a generalization of the algorithm fair scheduling (Fair Queuing - FQ). If a channel at a rate  $R$  is used for  $N$  streams, than the processing speed of each of them is  $R/N$  if Fair Queuing is used.

Unlike FQ, Weighted Fair Queuing allows to set up the fraction for each stream. Virtual starting time of  $S(k, i)$  processing and virtual time of the end of  $F(k, i)$  processing is assigned to each incoming packet  $p(k, i)$ , where  $k$  is the ordinal number of the packet, and  $i$  is the number of queue. Starting and finishing time is calculated as follows:



$$\begin{aligned}
S(k, i) &= \max(F(k-1, i), V(a(k, i))), \\
F(k, i) &= S(k, i) + L(k, i) / r(i), \\
F(0, i) &= 0,
\end{aligned} \tag{1}$$

where  $a(k, i)$  is the packet arrival time,  $L(k, i)$  is the size of the packet,  $V(t)$  is the virtual time function which is defined as multiplication of argument and reverted sum of active flows weights in current moment. The algorithm chooses the packet with the smallest virtual time of the end of the processing and sends it.

*Low Latency Queuing (LLQ)* is an queuing discipline with low latency. LLQ is an extension of WFQ algorithm with additional priority queue. Additional queue in LLQ enables handling of the most delay-sensitive traffic.

Summarizing the facts discussed above, let's identify that the scheduling algorithms are characterized by the following features:

- the ability of customization (most algorithms have customizable settings);
- usage of data about state of all queues in QoS mechanism;
- storage and maintaining internal state (e.g. WFQ stores and maintain parameters of the virtual time of arrival and sending packets).

#### IV. FORMALIZATION AND BUILDING SIMULATION MODELS OF QUALITY OF SERVICE MECHANISMS

From the point of view of the theory of queuing networks, the quality of service mechanism can be represented by a component to accumulate *transacts* before *service device* which is the outgoing interface of the network node. Packages are represented as service requests (or *transacts*). Each transact must contain the information necessary to simulate the operation of the quality of service mechanism and for the calculation of QoS characteristics: a virtual model time of packet sending, source port, destination port, IP-addresses of sender and recipient, the packet size, protocol type used by application layer, fields ToS, DSCP, MPLS QoS, and other information.

For the simulation of QoS mechanism it is necessary to describe its reaction on the occurrence of two discrete events: the new incoming packet arrives and the next outgoing packet can be sent.

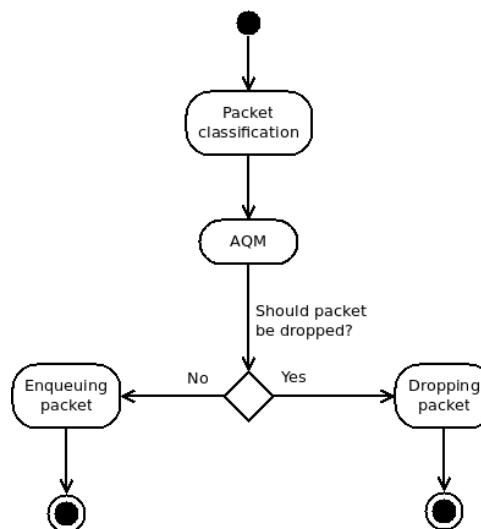


Fig. 1 Activity diagram of the reaction of the quality of service mechanism on the arrival of a new packet

Using the analysis above the reaction of the quality of service mechanism on the arrival of a new packet time can be represented as a UML-activity diagram indicated in Fig. 1. First, packet classification is performed to determine the queue to which it can be placed. Then mechanism of active queue management (AMQ) is applied to the selected queue to determine whether the packet must be dropped. The packet is then queued or dropped.

Using the above analysis, the reaction of the quality of service mechanism, when an event of the next outgoing packet can be sent, can be represented as a UML-activity diagram indicated in Fig. 2. The first step is performed by scheduling algorithm to determine the queue from which the next packet must be sent. Then the second part of the algorithm of active queue management is applied to determine whether the packet must be destroyed or not. The packet is removed from the queue. Then the packet is sent or deleted.

Algorithm of packet classification can be defined by function  $f_c(P, S_c)$ , where  $P$  is a tuple of parameters of packet, and  $S_c$  is the current state of classification component memory. This function can have values from set of all possible pairs  $(n, S'_c)$ , where  $n$  is an ordinal number of queue which packet must be placed to, and  $S'_c$  is a new state of classification component memory. Algorithms of active queue management can be defined by functions  $f_a(P, q, S)$ , where  $P$  is a tuple of parameters of packet,  $q$  is a tuple of parameters of queue state, and  $S$  is the current state of active queue management component. This functions can have values from set of all possible pairs  $(r, S')$ , where  $r$  is the result of algorithm work which can be one of the boolean values, a  $S'$  is a new state of active queue management component memory. Scheduling algorithm can be defined by function  $f_s(Q, S_s)$ , where  $Q$  is a tuple of queues state values, and  $S_s$  is the current state of scheduling component memory. This function can have values from set of all possible pairs  $(n, S'_s)$ , where  $n$  is the ordinal number of queue which packet must be sent from, and  $S'_s$  is a new state of scheduling component memory state. Thus the quality of service mechanism can be defined by parameters  $n, f_c(P, S_c), f_{ai1}(P, q_1, S_1), \dots, f_{ain}(P, q_n, S_n), f_{ao1}(P, q_1, S_1), \dots, f_{aon}(P, q_n, S_n), f_s(Q, S_s)$ , where  $n$  is a number of queues,  $f_c(P, S_c)$  is a function of classification algorithm,  $f_{ai1}(P, q_1, S_1), \dots, f_{ain}(P, q_n, S_n)$  are functions of AMQ algorithms for each queue from 1 to  $n$  that will be applied to incoming packets,  $f_{ao1}(P, q_1, S_1), \dots, f_{aon}(P, q_n, S_n)$  are functions of AMQ algorithms for each queue from 1 to  $n$  that will be applied to outgoing packets, and  $f_s(Q, S_s)$  is scheduling algorithm function.

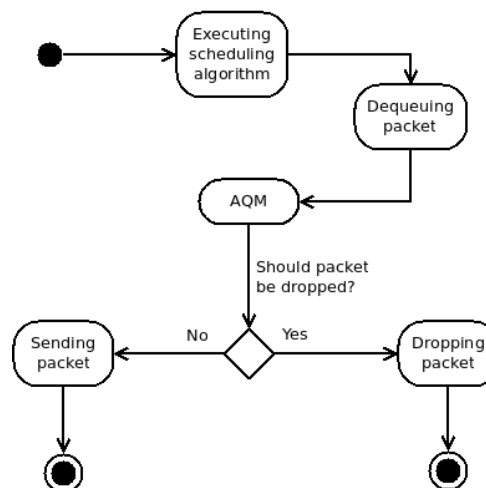


Fig. 2 Activity diagram of the reaction of quality of service mechanism on the next outgoing packet which can be sent

Using the concepts of object-oriented programming, the functions described above can be represented by the classes that implement the corresponding interfaces which contain a single method. This approach is used for building the problem-oriented tool for simulation automation

of next generation networks. Functions of classification algorithms are represented by a class that implements the interface `Classification` which contains a single method `classify(Packet packet)` which returns the number of the queue to the packet to be placed. The function of active queue management algorithms is represented by a class which implements an interface `ActiveQueueManagement` containing a single method `shouldBeDropped(Packet packet, Queue queue)` and which returns a boolean value. The functions of scheduling algorithms are represented by a class which implements an interface `Scheduling` containing a single method `schedule(List<Queue> queue)` and which returns the ordinal number of the queue which packet must be sent from. The toolkit contains the implementation for modeling many existing algorithms of quality of service mechanisms and allows the user to define its own custom algorithms using programming languages Java, JavaScript, Jython, JRuby, Scala, Groovy, Clojure and some others.

For example, the implementation of a classification algorithm that uses only field MPLS QoS can be defined using Clojure language as follows:

```
(gen-class
  :name "MplsQosClassification"
  :implements [Classification]
  :constructors {[[] []]}
  :prefix "mpls-qos-")

(defn mpls-qos-classify [this packet]
  (if (< (.getMplsTrafficClass packet) 2)
    (.getMplsTrafficClass packet)
    (if (< (.getMplsTrafficClass packet) 4)
      2
      (if (< (.getMplsTrafficClass packet) 6)
        3
        4))))
```

Implementation of the modification of AVQ active queue management algorithm for packet arrival can be defined using Jython language as an example as follows:

```
class AvqArrivalManagement (ActiveQueueManagement):
    def __init__(self, alpha, gamma, queue):
        self.alpha = alpha
        self.gamma = gamma
        self.virtual_capacity = queue.getCapacity()
        self.busy_virtual_capacity = 0
        self.last_time = -1

    def shouldBeDropped(self, packet, queue):
        rate = packet.getSize() / (packet.getArrivalTime() - self.last_time)
        self.virtual_capacity = self.alpha
        * (gamma*queue.getCapacity() - rate)
        self.last_time = packet.getArrivalTime()
        if self.busy_virtual_capacity + packet.getSize() <
self.virtual_capacity:
            self.busy_virtual_capacity += packet.getSize()
            return False
        else:
            return True
```

The simulation system calls user-defined algorithms in sequence represented by activity diagrams in Fig. 1 and 2 upon the occurrence of the relevant event. The system simulates traffic generation of different types for the simulation of packet arrival events [5, 6].

This approach allows to describe the quality of service algorithms using high-level languages abstracting the details of the implementation of the simulation system. Decomposition mechanisms for QoS on four types of components allow one to create simulation models of new mechanisms by combining existing components in the system and/or redefining some of them.

## V. CONCLUSION

This paper presents the analysis, generalization, formalization and simulation method for the quality of service mechanisms in data networks with packet switching. The proposed method is used to develop tool for automated simulation of next generation networks which allows to simulate and explore the quality of service mechanisms.

## REFERENCES

- [1] Cisco IOS Quality of Service Solutions Configuration Guide. [http://www.cisco.com/c/en/us/td/docs/ios/12\\_2/qos/configuration/guide/fqos\\_c/qcfintro.html](http://www.cisco.com/c/en/us/td/docs/ios/12_2/qos/configuration/guide/fqos_c/qcfintro.html). Accessed August 2015.
- [2] Traffic Control HOWTO. <http://www.tldp.org/HOWTO/Traffic-Control-HOWTO/components.html>. Accessed August 2015.
- [3] C. Zhang, J. Yin, Z. Cai, W. Chen, "RRED: Robust RED algorithm to counter low-rate denial-of-service attacks", *IEEE Communications Letters*, № 14 (5), 2010, pp. 489-491.
- [4] S. S. Kunniyur, R. Srikant, "An adaptive virtual queue algorithm for active queue management", *Networking, IEEE/ACM Transactions on*, №2, vol. 12, 2004, pp. 286-299.
- [5] V. Kulichenko, O. Demidenko, P. Chechat, "An approach to identify channel capacity of LAN on TCP/IP, ICMP and UDP protocols", *Problems of physics, mathematics and technics*, 2014, № 4 (21), pp. 1-3.
- [6] O. Demidenko, A. Khobnya, "Conceptual model for the generation of VoIP traffic in the NGN", *Francisk Scorina Gomel State University Proceedings*, 2014, №6(87), pp. 117-122.



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Fuzzy model for inventory control under uncertainty

Helena Zhivitskaya, Tanya Safronava

**Abstract**— Modern enterprise faces the huge stream of rapidly changing information. Thereby competitive advantage is fast response of changing external environment by tactical decision making. Decision making problems concern a lot of present science. Many decisions are making under uncertainty or risk. It becomes necessary from faithful deterministic representation goes to sphere of associative fuzzy thinking. These steps give you guidelines for developing decision support system with neuro-fuzzy.

**Keywords**— Inventory control, fuzzy set, decision support system, linguistic variables, uncertainty.

## I. INTRODUCTION

A successful business relies on many factors, one of which is a reliable inventory management system. Inventory control is interesting not just for manufacturing enterprises but for telecommunication company. Huge material supplies freeze cash assets.

A major challenge in supply chain inventory optimization is handling uncertainty, as not all the data required for making decisions are available with certainty at the time of making the decision. This problem of design, analysis, and optimization under uncertainty is central to decision support systems. It becomes necessary from faithful deterministic representation goes to sphere of associative fuzzy thinking.

In decision making we deal with linguistic, not numerical, variables. The concept of linguistic variables is an alternative approach to modeling human thinking – an approach that, in an approximate manner, serves to summarize information and express it in terms of fuzzy sets instead of crisp numbers.

## II. CREATING FUZZY MODEL

Society and industry are becoming knowledge-oriented and relying on different experts' decision – making ability to solve problems. Computer-based systems are capable of understanding the information being processed and can make a decision based on it, whereas the traditional computer systems do not know or understand the data/information they process. Fuzzy logic, a flexible machine-learning technique, is an attempt at mimicking the logic of human thoughts. Human logic is flexible and less rigid when compared to crisp logic [1].

Fuzzy sets were originally introduced in 1965. [2] As was pointed out by Zadeh, conventional techniques for system analysis are intrinsically unsuited for dealing with humanistic systems, whose behavior is strongly influenced by human judgment, perception, and emotions. Fuzzy set theory provides a systematic calculus to deal with such information linguistically. And it performs numerical computation by using linguistic labels stipulated by membership function. Moreover, a selection of fuzzy if-then rules forms the key component of fuzzy inference system that can effectively model human expertise in a specific application. [3]

The fuzzy inference system (fuzzy model) is a popular computing framework based on the concept of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning. It has found successful applications in a wide variety of fields, such as automatic control, data classification, decision

---

H. Zhivitskaya, Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus (e-mail: jivitskaya@bsuir.by).

T. Safronava, Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus (e-mail: tanyshka-mog@yandex.ru).



making, expert systems, time series prediction, robotics, and pattern recognition. Typical structure of fuzzy model can be shown on figure 2.

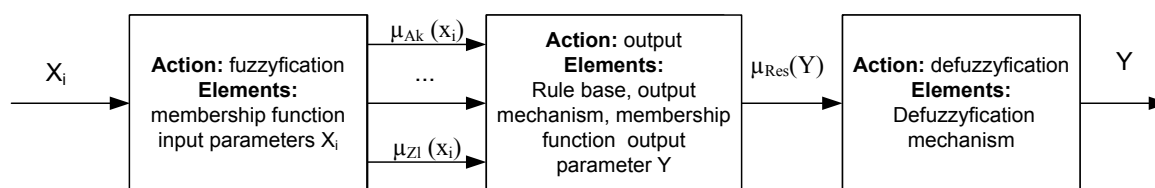


Fig. 1 Typical structure of fuzzy model

Block «*fuzzification*» calculates the membership degree of clear numerical values  $X_i$  fuzzy sets  $A_k, \dots, Z_l$ . Fuzzification unit must have access to a well-defined membership function inputs.

Block «*output*» at the entrance gets a membership degree and the output calculates the resulting so-called membership function of the output value of the model. It consists of rule base, output mechanism, membership function, output parameter.

Block «*defuzzification*» on the basis of the resulting membership function calculates the precise numerical value of the output parameter  $Y$ , which is the result for the output of numerical values  $X_i$ . [4, 10]

### III. CONSTRUCTION OF KNOWLEDGE BASE

A procedure embeds fuzzy logic in the form of fuzzy rules containing linguistic terms with multiple values. There are fuzzy membership functions that map the linguistic values to crisp values, which then can be understood by machines, which can then interpret and execute commands. [1]

Creating a knowledge base follows the five-steps:

- 1) Identify the inputs and their ranges and name them.
- 2) Identify the output and its ranges and name it.
- 3) Create the fuzzy membership function for each input and output.
- 4) Translate the interaction of the inputs and outputs into IF–Then rules.
- 5) Decide defuzzified output. [5]

We will analyze inventory management system for a telecommunications company. [6, 7, 8] At the present time our company uses two technologies with absolute different types of cable: fiber-optic and copper. Fiber-optic cable has priority importance. Fiber-optic cable has different capacitance and divided to main cable, distribution cable and subscriber's cable.

Figure 2 shows structure of capacitance of main cable. The most popular capacitances are 8 and 16. The more the capacitance of main cable the more cost of project. The cable with high capacitance has low demand. As result this type of cable will stay at storehouse. At the same time some projects cannot be implemented without such cable. Time of purchase is once in three month.

For clarity, General Cable demand consists of cable demand calculated by project needs, emergency reserve, demand for scheduled repairs and unexpected demand.

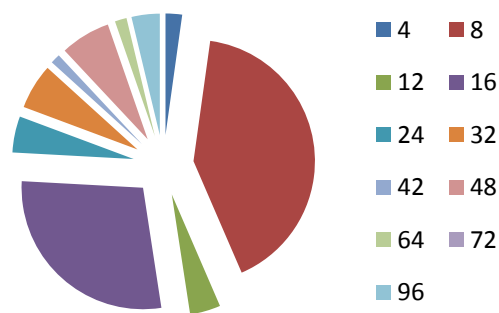


Fig. 2 Structure of capacitance of main

First three demands are crisp values which can be properly calculated. Unexpected demand depends on design error and additional requirements. Design error is linguistic variable. Terms {high, middle, low}.

Design error depends on project complexity. If project has big quantity of consumers for crossing xPON technology then it uses cables with different capacities and different length of the cable. The main design error is inaccurate length of the cable. If the length of the cable is smaller than in reality it led to rise in the cost of project (putting additional cable box) or increase period of construction (pull the cable with right length). If the length of the cable is bigger than in reality it led to rise in the cost of project (waste of the cable). Project author must realize project quickly as there a lot of consumers for crossing xPON technology.

Rule 1: If project quantity is big and time is limited then design error is high.

Rule 2: If project quantity is small and time is limited then design error is middle.

Rule 3: If project quantity is small and time is unlimited then design error is low.

Rule 4: If project quantity is big and time is unlimited then design error is middle.

TABLE I  
DEPENDENCY MISTAKES QUANTITY ON PROJECT QUANTITY PER WEEK

Project quantity	2	4	6	10	12	14	16	18	20
Number mistakes	0	1	4	5	7	9	11	13	14

#### IV. TAKAGI SUGENO

The Sugeno fuzzy model (also known as the TSK fuzzy model) was proposed by Takagi, Sugeno, and Kang in an effort to develop a systematic approach to generate fuzzy rules from a given input-output data set.

Typical view rule of Takagi Sugeno model are

IF  $(x_1=A_{11})$  and  $(x_2=A_{21})$  and ...  $(x_p=A_{p1})$

Else  $(x_1=A_{12})$  and  $(x_2=A_{22})$  and ...  $(x_p=A_{p12})$

Else ...

...

Then  $(y=b_{i0}+b_{i1}x_1+...+b_{ip}x_{pi})$ ,

where  $i$  - number of rule

#### V. SELF-TUNING OF PARAMETERS FUZZY MODEL

The best results can be achieved when structure of optimization model is combined with the optimization of its parameters.

Self-tuning algorithm:

- 1) Determination of the basic system model.
- 2) Tuning the basic model by measurement of input and output parameters.
- 3) Checking the accuracy of the basic model.
- 4) Error detection of the basic model.
- 5) Identifying maximum and minimum errors of the basic model.
- 6) Construction the model error.
- 7) Tuning the parameters of the membership functions of the model error.
- 8) Adding error model to the basic model.
- 9) Step three.

## VI. RESULTS

Using rules for self-tuning algorithm we receive results of error model which are presented in table 2. Average squared error is 0.052.

TABLE II  
RESULTS OF ERROR MODEL

x1	10	13	15	10	15	17
x2	1	1	2	2	3	3
e0	0,32	0,1	-0,1	0,2	-0,23	-0,4
eM0	0,3	-0,041	-0,182	0,192	-0,41	-0,3
e1	0	0,141	0,082	0,008	0,18	-0,1

Neuro-fuzzy network, representing the model constructed on maximum error of the method is similar to the type of neural network RBF [8, 10].

## VII. CONCLUSIONS

In this paper was shown approach creating fuzzy model for inventory management system under uncertainty. The main steps consist of description of the company on the outside and inside. Then using this knowledge you can develop decision support system with modern strategies of inventory control.

## REFERENCES

- [1] R. Arvind Akerkar and P. Srinivas Sajja, "Knowledge-based systems", Jones and Barlet Publishers, LLC, 2010.
- [2] F. Martin McNeil and E. Thro, "Fuzzy logic: a practical approach", Academic Press, Inc., 1994.
- [3] J. Jyh-Shing Roger, Chuen-Tsai Sun, and Eji Mizutani, "Neuro-fuzzy and soft computing: a computation approach to learning and machine intelligence, prentice-Hall Inc., 1997.
- [4] A. Leonenkov, "Fuzzy design in MATLAB and FuzzyTech", St. Petersburg, BHV- Petersburg, 2005.
- [5] L.A. Zadeh, "Fuzzy Sets". Information and control, pp. 338-383, 1965.
- [6] S. Axsäter, "Inventory Control", 2nd ed., Springer Science+ Business Media, LLC, New York, 2006
- [7] D. Waters, "Logistics. An Introduction to Supply Chain Management", PALGRAVE MACMILLAN, New York, 2003.
- [8] M. Muller, "Essentials of Inventory Management", AMACOM, New York, 2003.
- [9] B. Liu, "Theory and Practice of Uncertain Programming", 2nd ed., Moscow: LBZ, 2005.
- [10] A. Piegat, "Fuzzy modeling and control", Physica-Verlag Heidelberg, 2001
- [11] S. Haykin, "Neural Networks. A Comprehensive Foundation", 2nd ed., New Jersey: Prentice Hall, 1998



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Integrated Approach to Improve Reliability of Neural Network

Mary Tyagunova, Tanya Golub, Andrey Proskura

**Abstract**— This article describes complex approach to improve the reliability of neural networks. As a basis it is taken Kohonen neural network, which solves the problem of clustering graphics sets of different signals, ensuring optimum reliability and quality performance. The main test results are given.

**Keywords**— neural network, reliability, SOM Kohonen.

## I. INTRODUCTION

At the present stage of development and applications of neural networks actual direction is to work with graphical data.

Selection images as an input data are related with the study of perspective directions of Kohonen neural network development. Namely, their narrow specialized direction is to work with images that, at this stage of neural networks development is an urgent task. Currently, the development of processing technologies different images, which have different size and quality is very important and relevant. Therefore work with images, not with numerical data is very important, despite the complexity of the problem since the task of clustering antecedent to recognition, i.e. it is an important preparatory step in the recognition process. An important feature of using Kohonen networks is the technology (method) unsupervised learning that significantly affects to the results of using this approach.

Taking into account that created neural network should provide the high level quality of the task performance, there is a question about the feasibility of improving the reliability of its functioning, which in the future will have a direct impact on the quality of its work and the results of clustering. Thus, the chosen theme is relevant.

Analysis methods for improving the reliability are described in [1] in sufficient detail, so attention is paid that it is distinguish functional and parametric [2] reliability. Each method has its advantages and disadvantages. In this paper we propose a comprehensive approach to improve the neural network reliability.

## II. AN INTEGRATED APPROACH TO IMPROVE THE NEURAL NETWORK RELIABILITY

It takes at the basis that the weights of all the neurons and their thresholds are random variables that obey the normal law of probability distributions with known parameters. Based on this knowledge and tailored the identified positive aspects of approaches and improve functional parametric reliability, conducted the study it allocates the following recommendations for an integrated approach to improve the neural network reliability:

1. An increase of the functional reliability of the entire neural network is more rational to introduce a static redundancy, as it has several advantages over other types of redundancy. The important advantage of static redundancy is its flexibility and lack of the need to develop special software for the detection, localization and bug fixes, error correction, without interruption in the operation.

M.Yu. Tyagunova, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine (e-mail: mary@yurich.org.ua).  
T.V. Golub, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine (e-mail: tery86@mail.ru).  
A.S. Proskura, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

2. At increase parametric reliability should be guided by the fact that when an equal number of neurons, the structure of the neural network with a large number of layers has a much lower probability of failure-free operation than the structure with less number; also reliability increases with the growth of quantity (number) of neurons in the first layer of the multilayer neural network, i.e. with the introduction of static redundancy. And if it is introduced into the second and subsequent layers of neural network, reliability is significantly reduced.

Summarizing the above recommendations, it should be noted that the increase in the parametric reliability should be accompanied by an increase in the functional reliability, i.e. use an integrated approach. In turn, increase the parametric reliability by reducing the number of layers and increasing the number of neurons in the first layer by the introducing a static redundancy.

To test the integrated approach to improving the reliability has been implemented the Kohonen neural network, which solves the problem of clustering graphics sets of different signals, ensuring optimum reliability and functioning quality. The importance of addressing this problem is described in more detail in [3].

### III. DEVELOPMENT OF THE NEURAL NETWORK

For the check on the practice of the integrated approach has been designed self-adjusting Kohonen neural network and conducted its test. Fig. 1 shows the control window of created neural network, which is the final stage of the simulation. In this window, are given the opportunity to make adjustments to the configuration and further operation of the created neural network. Among them are: re-training (Train Again), change the number of neurons in layers (Adjust Network Size), import a large volumes of data (Import Large Data Set), Network Testing (Test Network), change the type and the kind of the incoming data (Inputs).

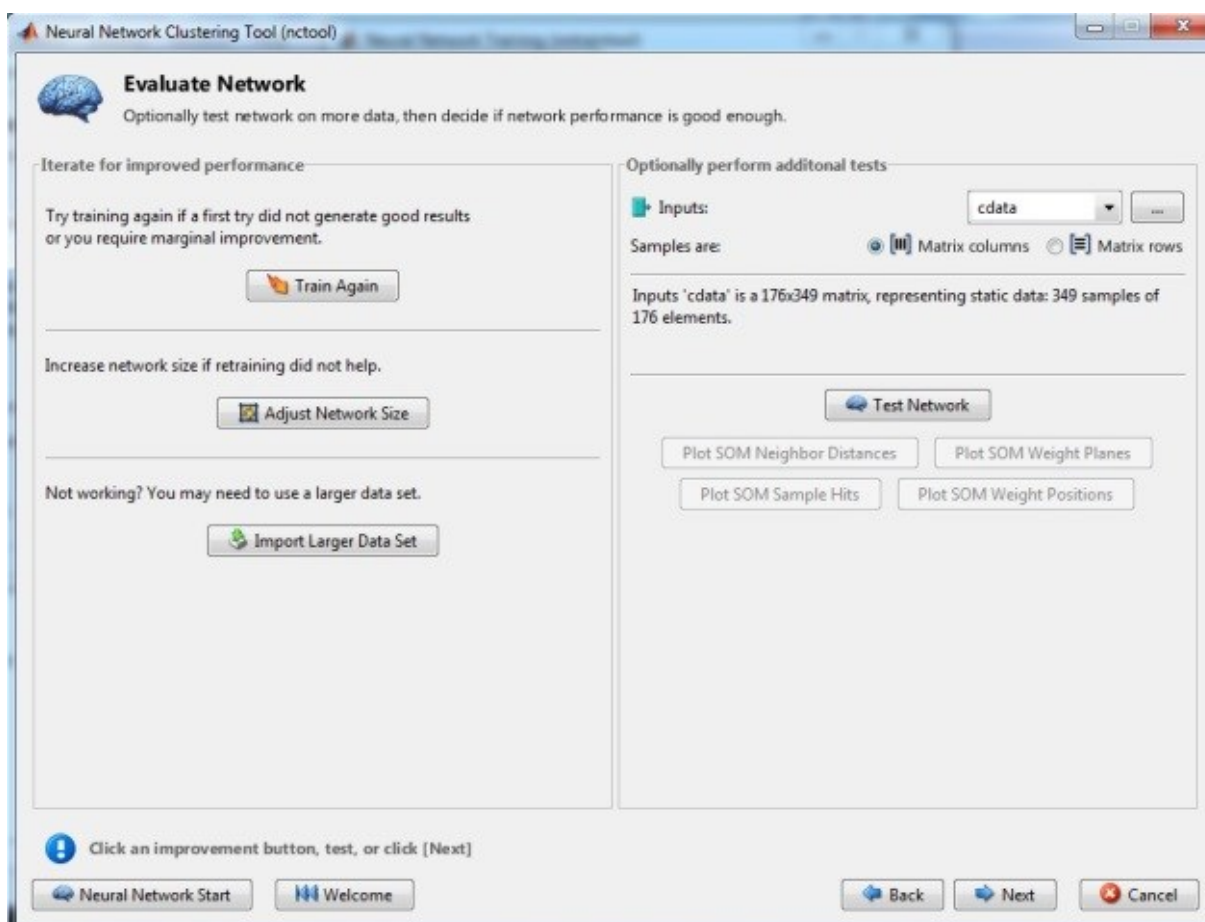


Fig. 1 Window management of created neural network



After testing the network it is opportunity to see results of the created neural network, which are displayed with the following options:

- plot Self Organization Maps (SOM) Neighbor Distances;
- plot SOM Weight Planes;
- plot SOM Sample Hits;
- plot SOM Weight Positions.

#### A. Input Data

The data used as an input data for teaching, training and testing the developed neural networks have the form of images with several signal periods which have different mathematical and physical models, as well as arguments, functions and parameters.

At various stages of the design used different sets of data (signal images), which included the types of signals such as modulated, harmonic, sawtooth. The neural network has distributed corresponding images sets to the relevant clusters at certain stages of its creation. Respectively used in the design process and improve the reliability of the neural network training and test sets contained images of various kinds of signals for more correct and independent results of research and work of the neural network.

As an important feature of using Kohonen networks is a method of learning without a teacher, which significantly affects to the results when using this approach, the results of training, testing, and clustering process directly depends only on the structure of the input data.

Different sets of data (signal image) generated at various stages in the applied mathematical package MatCad15.

#### B. Training Set

The training set consists of 14 similar images of signals which vary harmonically. Some of them are influenced by noise. Some images of the data set are presented in Fig. 2.

Signals from this training set are sine and cosine. They have different amplitude, frequency and presents respective different number of periods.

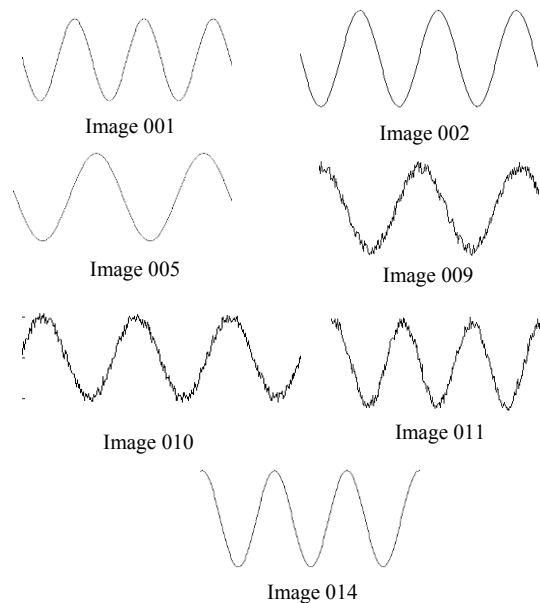


Fig. 2 Training data set

#### C. Test Set

The test signal set includes 12 images. At the generation signals in the system MathCad15 was created 2 obvious signal groups on 3 images in each, further suggesting that the created neural network distributes their according to expected clusters. The remaining 6 images of signals are

generated with more significant differences in the parameters and arguments, as well as noise levels. It is expected that the system distributes them in 2 clusters on 3 in each. The following fig. 3 shows a complete test set of images with the expected distribution them of clusters on the test.

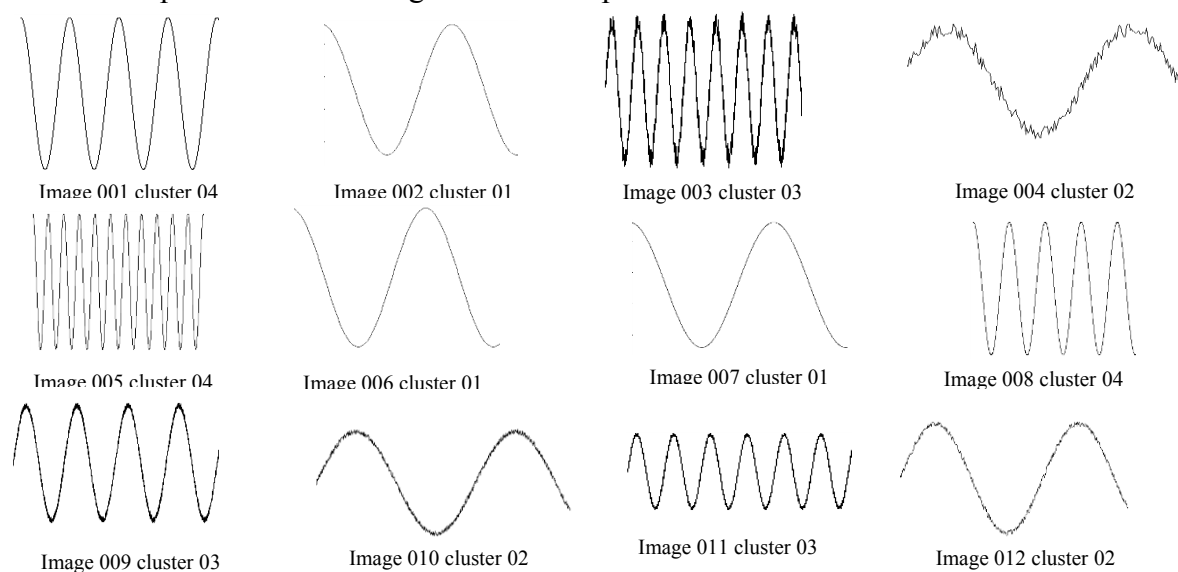


Fig. 3 Test data set

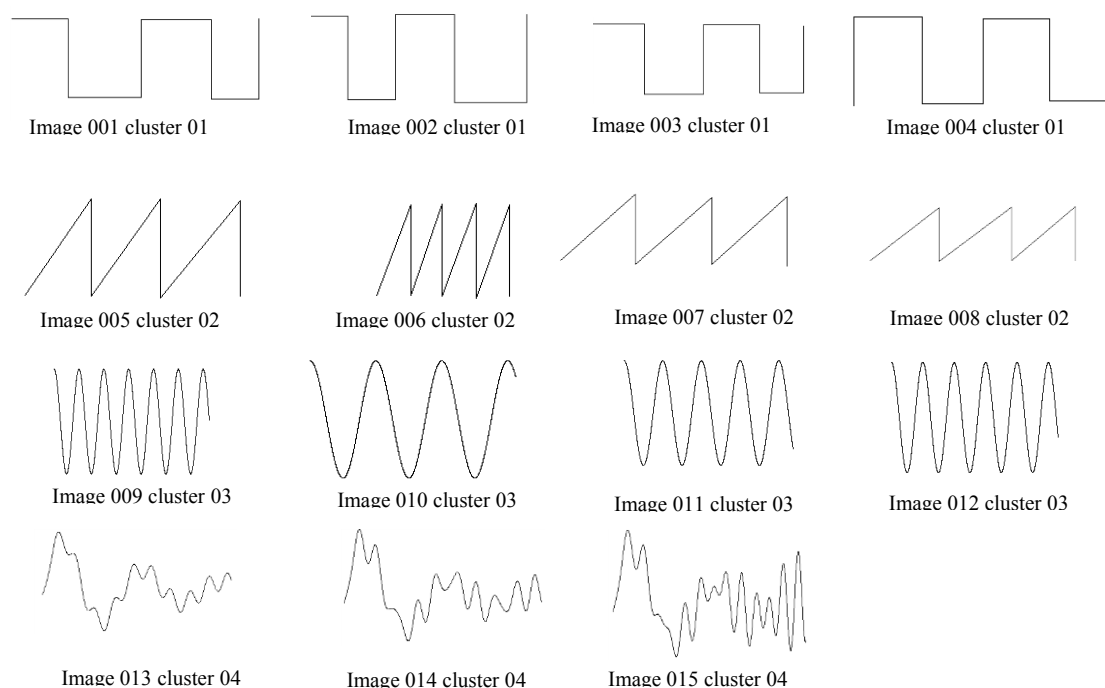


Fig. 4 The data set for clustering

The testing process will be considered successful if the neural network will distribute these 12 images on a 4 clusters, given the recommendations that have been developed in the generation of test image signals, namely in each respective cluster of three corresponding images:

- cluster 01 – test images: 002, 006, 007;
- cluster 02 – test images: 004, 010, 012;
- cluster 03 – test images: 003, 009, 011;
- cluster 04 – test images: 001, 005, 008.

#### D. Set for Clustering

To carry out the process of clustering it was generated set of images, including 4 types of signals: rectangular, sawtooth, harmonic and complex structure signal. It is assumed that the neural network will distribute signals to 4 clusters, and to the 4th cluster will include complex signals, the image of which for the given neural network will be "unknown" and the network will not to compare them with previous three types of signals, and allocates them to the new cluster. Fig. 4 shows set of 15 signals.

The testing process will be considered successful if the neural network will distribute these 15 images on a 4 cluster, given the recommendations that have been developed in the generation of signal images:

- cluster 01 – images: 001, 002, 003, 004;
- cluster 02 – images: 005, 006, 007, 008;
- cluster 03 – images: 009, 010, 011, 012;
- cluster 04 – images: 013, 014, 015.

### IV. THE ARCHITECTURE OF THE NEURAL NETWORK

#### A. The Structure of the Neural Network at the Training Stage

In the training stage of development of the neural network on its inputs was supplied training set, which consisted of 14 images (Fig. 2). The first layer of the neural network consisted of 20 neurons and in the second layer were four neurons.

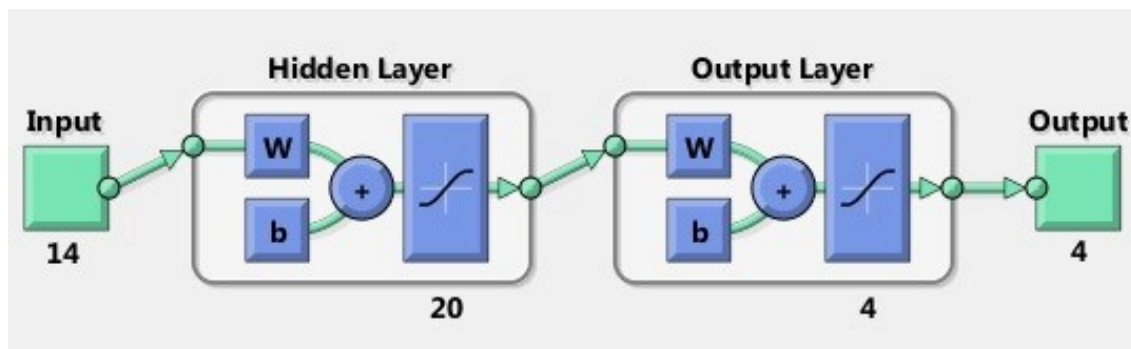


Fig. 5 The architecture of the neural network at the training stage

At the output, respectively, signal images distributed into 4 clusters. This quantitative distribution of neurons in two layers due to the number of input data and amount of output data and has had an enormous impact on the primary structure of the neural network. Fig. 5 shows the primary structure of the neural network.

#### B. The Structure of the Neural Network at the Testing Stage

During the testing phase of the neural network at its inputs it was fed the test set of signal images comprising 12 elements (Fig. 3). The number of neurons in first and second layers remains the same as in the training stage and is equal to 20 and 4, respectively. The number of clusters is 4. Number of inputs is 12 images. The structure of the neural network at this stage is as follows (Fig. 6).

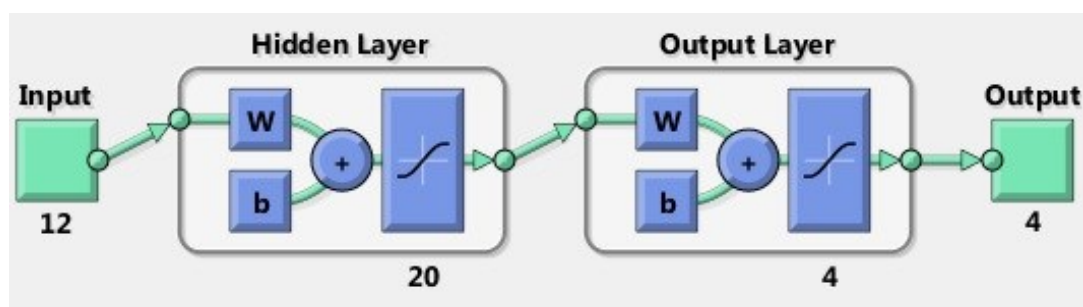


Fig. 6 The architecture of the neural network at the testing stage

### C. The Structure of the Neural Network at the Stage of Clustering

After processing the results of the neural network and increase its reliability, by adding neurons to the first layer, it changed the structure of the neural network. The number of neurons in the first (input) layer increased to 25 neurons. In the second (output) has remained the same - 4.

The input of the neural network received 15 elements of the cluster set signal images (Fig. 4). The structure of the neural network at this stage of implementation is shown in the fig. 7.

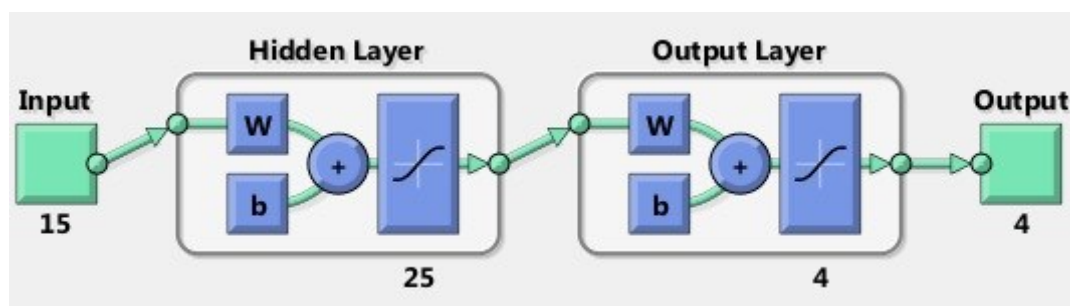


Fig. 7 Architecture of reliable neural network at the stage of clustering

## V. OUTPUT DATA

Output data, which were obtained as a result of the testing process and process of clustering are fully confirmed serviceability and correctness of the established neural network. It confirmed by the expected distribution of the signal images relevant to intended result clusters in steps (processes) of testing and clustering.

Data are shown in graphs (images) of the structure and number of clusters, as well as responses at the command prompt MatLab.

### A. The Test Results of the Neural Network

The test data set includes 12 images with signals of a harmonic type (Fig. 3), which were generated in the applied mathematical package MathCad15. When it created a set of signals a half of them were exposed to noise. An expected distribution of clusters is as follows:

- cluster 01 test images: 002, 006, 007;
- cluster 02 test images: 004, 010, 012;
- cluster 03 test images: 003, 009, 011;
- cluster 04 test images: 001, 005, 008.

Results of distributions obtained during testing are presented in Fig. 8 – 9 .

Fig. 8 shows that the neural network for the distribution of a set identified 4 clusters, each of which has put on 3 objects (images), which corresponded to the anticipated results.

Fig. 9 shows a response of the neural network as function comprising two parameters: the cluster number (h) and the image number (p), which included to this cluster.

Based on the results shown in Fig. 8 - 9 and the expected results in the preparation of the test data set, you can draw the following conclusion: the neural network implemented clustering of the test data set in accordance with the proposed distribution.

Having regard the fact that the proposed results and the obtained results when testing the neural network are identical, it is worth noting that the results of the test confirm the operation and readiness the neural network to implement further clustering process.

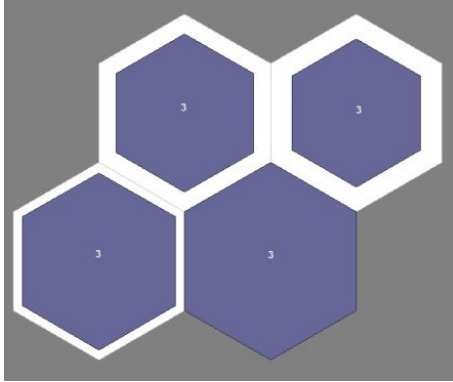


Fig. 8 The graph of images distribution in clusters

```
fx >> y=sim(h,p) %Опос сета
y=(1,002)
(1,006)
(1,007)
(2,004)
(2,010)
(2,012)
(3,003)
(3,009)
(3,011)
(4,001)
(4,005)
(4,008)
```

Fig. 9 Numerical test results presented in the command window

### B. The Results of Clustering Signals

The data set for clustering process include 15 signal images (Fig. 4). It includes 4 types of signals: rectangular, sawtooth, harmonic and signals with complex structures that were generated in the applied mathematical package MathCad15. When it created a set of signals a half of them were exposed to noise. An expected distribution of clusters is as follows:

- cluster 01 images: 001, 002, 003, 004;
- cluster 02 images: 005, 006, 007, 008;
- cluster 03 images: 009, 010, 011, 012;
- cluster 04 images: 013, 014, 015.

The results of distributions obtained during the process of clustering, are presented in Fig. 10, 11.

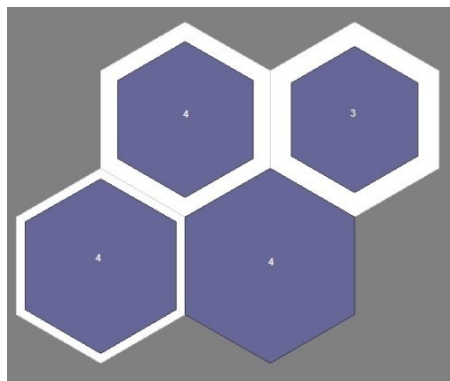


Fig. 10 The graph of images distribution in clusters

```
fx >> y=sim(h,p) %Опос сета
y=(1,001)
y=(1,002)
y=(1,003)
y=(1,004)
y=(2,005)
y=(2,007)
y=(2,008)
y=(3,009)
y=(3,010)
y=(3,011)
y=(3,012)
y=(4,006)
y=(4,013)
y=(4,014)
y=(4,015)
```

Fig. 11 Numerical results of clustering represented in the command window

Fig. 10 shows that the neural network for the set distribution identified 4 clusters, in 3 of which are placed on 4 objects (images), and the rest 3 objects that does not quite match the intended results.

Fig. 11 shows the response of the neural network as a function of having two parameters: the cluster number (h) and the image number (p), included in this cluster.

Based on the results shown in Fig. 10, 11 and expected results in the preparation of the clustered



data set it can conclude the following: the neural network implemented clustering of data set with an error regarding the proposed distribution. The error lies in the definition of image 006 to the cluster 04. As the element 006 is the image of sawtooth signal, the neural network was supposed to define it to cluster 02.

Taking into account the good results of clustering signals and made a small mistake, it is advisable to increase the reliability of the developed neural network, thereby to improve the accuracy and quality of its operation (clusterization).

### C. Results of Clustering Signals by the Neural Network with High Reliability

To ensure the correctness and adequacy of the conclusions about the quality of clustering when comparing results of the distribution to clusters of data processed by upgraded and primary neural networks, it is advisable to use the same set of signal images which includes four types of signals: rectangular, sawtooth, harmonic and signals with complex structure. The total number of images is 15 (Fig. 4). An expected distribution of clusters is following:

- cluster 01 images: 001, 002, 003, 004;
- cluster 02 images: 005, 006, 007, 008;
- cluster 03 images: 009, 010, 011, 012;
- cluster 04 images: 013, 014, 015.

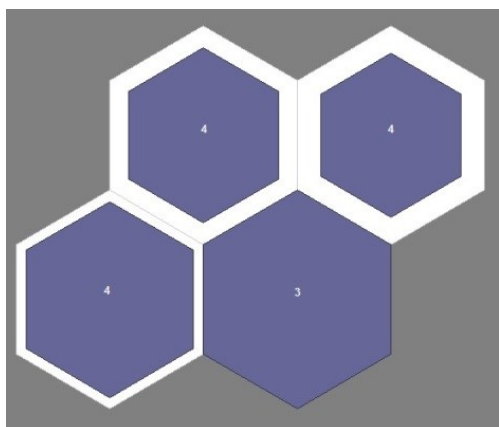


Fig. 12 The graph of images distribution in clusters

```
fx >> y=sim(h,p) %Onpoc cere
y=
(1,001)
(1,002)
(1,003)
(1,004)
(2,005)
(2,006)
(2,007)
(2,008)
(3,009)
(3,010)
(3,011)
(3,012)
(4,013)
(4,014)
(4,015)
```

Fig. 13 Numerical results of clustering represented in the command window

Results of the distribution obtained by carrying out the clustering process are shown in Fig. 12, 13.

Fig. 12 shows that the neural network with high reliability for the set distribution identified 4 clusters, in 3 of which are placed on 4 objects (images), and to the last cluster it distributed 3 objects that corresponded to the intended results.

Fig. 13 shows the response of the neural network as a function of having two parameters: the cluster number (h) and the image number (p), included to this cluster.

Based on the results shown in Fig. 12, 13 and expected results in the preparation of the clustered data set it can conclude the following: neural network implemented clustering of data set without error, with respect to the alleged distribution. Upgraded neural network with higher reliability determined image 006 to the cluster 02, which contained the ramps. Fig. 13 displays images which presence at the sawtooth cluster (cluster 2).

Considering that the estimated and the obtained results by the clustering are identical, it is worth noting that the neural network with high reliability successfully complete the task, thus confirmed its efficiency and readiness for implementation of the quality distribution of similar signal images into classes.

## VI. CONCLUSIONS

The studies proposed a complex approach to improve the reliability of the neural network, which has allowed to provide for the established network high standards of efficiency and performance. Using an integrated approach of classification process of the input data stream it was no error, while using the traditional approach has been fixed misclassification object. It leads to make the conclusion about the effectiveness of the proposed method.

As the software development environment of the neural network used the application of mathematical package Matlab 7.11.0.584 (R2010). To generate test sets of images of various signals used Applied mathematical package MathCad 15.0 (15.0.0.436 [006,041,742]). Software ran by the operating system Windows 7.

## REFERENCES

- [1] A. Galushkin, "The theory of neural networks", M.: IPRZHR, 2000, P. 416.
- [2] I. Potapov, "Mathematical models and optimization techniques of functional reliability of artificial neural networks." The thesis for the degree of candidate Technical Sciences. Tomsk. 2004, p. 175 electronic resources: <http://www.dissercat.com/content/matematicheskie-modeli-i-metody-optimizatsii-funktsionalnoi-nadezhnosti-iskusstvennykh-neiro#ixzz3arRdOCBf>
- [3] T. Golub, M. Tyagunova, "Using neural networks to determine the type of modulation," Annual Scientific Conference of students, teachers, researchers, graduate students and young scientists "Science Week 2013", 15-19 April 2013, Zaporozhye: ZNTU, 2013, p. 225.



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# The Set of Data Transformations for Dimensionality Reduction in Technical and Biomedical Diagnosis

Sergey A. Subbotin

**Abstract**—The set of rapid transformations from the original multi-dimensional space into one-dimensional axis was firstly proposed. They provide a solution of the feature extraction and feature selection problems. The set of criteria allow to compare transformations to the generalized axis is proposed.

**Keywords**—Sample, feature, locality-preserving transformation, dimensionality reduction.

## I. INTRODUCTION

The computational diagnostics is widely used in technology and medicine, where the expert knowledge is missing or insufficient. The theoretical basis of computational diagnostics is the theory of pattern recognition.

The urgent problem in computational diagnosis and pattern recognition is a model complexity reduction and a model construction speed increasing. It is caused by that the samples frequently characterized by a big number of features or contain a big number of instances (precedents). So to reduce the model complexity and to speed-up it's training process we need to reduce the data dimensionality [1, 2].

The data dimensionality reduction usually performed by a feature selection [1]. It needs, as a rule, an exhaustive search of possible feature combination with the least number of features that provides an acceptable accuracy. When the number of features and instances are big such technique require a lot of time. As an alternative to this way the sample selection [3] can be used. Unfortunately, it involves necessity of computation and using a distance matrix in N-dimensional feature space, that causes a combinatorial complexity of large data processing.

The way to solve these problems is the using of transformation from the multidimensional space of initial features to a one-dimensional axis for data dimensionality reduction [2, 3].

There are various methods of transformation for data dimension reduction (Principal Component Analysis, Semidefinite Embedding, Multifactor Dimensionality Reduction, Non-linear Dimensionality Reduction, Partial Least Squares, Independent Component Analysis) [2–11], which, however, require the calculation of distances between instances or feature correlation coefficients and for a large-scale problem they are hardly applicable in practice due to big requirements of time and computer memory in the process of determining the transformation parameters and in the process of transformation execution. This situation is additionally compounded by that the number of known transformations and their modifications is very big and there are no any formal criteria to analyze their quality, as well as to select the best available transformation for a particular task [3].

The purpose of this work is the development of rapid transformations from a multidimensional feature space to a one-dimensional axis, that can be used for a sample selection and the development of criteria for the transformation selection to use in a particular problem solving.

S. A. Subbotin, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine (e-mail: subbotin@zntu.edu.ua)

## II. PROBLEM STATEMENT

Suppose we have an initial (original) sample  $X = \langle x, y \rangle$  the set of  $S$  precedents describing dependence  $y(x)$ ,  $x = \{x^s\}$ ,  $y = \{y^s\}$ ,  $s = 1, 2, \dots, S$ , characterized by a set of  $N$  input features  $\{x_j\}$ ,  $j = 1, 2, \dots, N$ , where  $j$  is the number of feature, and an output feature  $y$ . Each  $s$ -th precedent can be represented as  $\langle x^s, y^s \rangle$ ,  $x^s = \{x_j^s\}$ , where  $x_j^s$  is the value of  $j$ -th input feature and  $y^s$  is the value of output feature for the  $s$ -th precedent (instance) of the sample,  $y^s \in \{1, 2, \dots, K\}$ , where  $K$  is the number of classes,  $K > 1$ .

Then the problem of the sample  $X$  dimensionality reduction can be formally represented as follows: find a transformation  $H: X \rightarrow I$ , which for each instance  $x^s = \{x_j^s\}$  determine the coordinate  $F$  on the generalized axis  $I$  and thus provides a mapping of instances of different classes to the different intervals of the generalized axis.

Since, as a rule, known transformations do not guarantee an exact solution of this problem, further problem arises of designing of indicators to quantify the quality of the transformation and to compare the results of the various transformations between themselves to choose the best transformation of the set.

## III. TRANSFORMATIONS TO THE GENERALIZED AXIS

For large-scale problems it is advisable to ensure the creation of such transformations, which would allow the mapping of individual instances without loading of whole initial sample, as well as taking into account the feature informativity in the process of transforming and to provide a generalization of data.

To ensure the generalization of close located data points (instances) we propose to replace feature values to numbers of feature value interval. For this we need previously to discretize the features by partitioning them into intervals of values.

To partitioning the features into intervals the number of interval (term), which hits the  $s$ -th instance on the  $j$ -th feature is proposed to determine as (1):

$$\hat{x}_j^s = \begin{cases} \text{round}\left(1 + \frac{x_j^s - x_j^{\min}}{\theta_j}\right), \theta_j > 0; \\ 1, \theta_j = 0, \end{cases} \quad (1)$$

where

$$\theta_j = \frac{x_j^{\max} - x_j^{\min}}{k_j},$$

$$k_j = \begin{cases} K, K > \text{round}(\ln S), K < \sqrt[N]{S}; \\ \max\{2, \text{round}(\sqrt[N]{S})\}, K > \text{round}(\ln S), K > \sqrt[N]{S}; \\ \max\{2, \text{round}(\ln S)\}, K < \text{round}(\ln S) < \sqrt[N]{S}; \\ K, \text{round}(\ln S) \leq K, K < \sqrt[N]{S}; \\ \max\{2, \text{round}(\sqrt[N]{S})\}, \text{round}(\ln S) \leq K, K \geq \sqrt[N]{S}, \end{cases}$$

$x_j^{\min}$ ,  $x_j^{\max}$  are the minimum and maximum values of  $j$ -th feature, respectively.

To ensure the generalization of close located data points (instances) we propose to replace feature values to numbers of feature value interval. For this we need previously to discretize the features by partitioning them into intervals of values.

For the mapping of instances from the original multi-dimensional feature space to the one-dimensional generalized axis is suggested to use the following transformations.

Transformation 1 (see Fig.1). For each number of interval of  $j$ -th feature get it binary representation (binary numbers padded with zeros from the left to  $c_j$  – the number of digits in  $k_j$ ). Set the coordinate of  $s$ -th instance on the generalized axis  $F=0$ , set the position (bit) number of generalized axis coordinate  $p = 1$ . Going by the feature numbers  $j$   $s$  in descending order of

their rank and by the group of digits in the interval number  $c = 1, 2, \dots, c_j$  perform in a cycle: if  $p \leq d$ , where  $d$  is a number of bits in a computer bit grid, then record at  $p$ -th position (with numbering from left) of the binary representation of a generalized feature  $F^s$  the  $c$ -th position value (with numbering from left) of interval number, in which the  $s$ -th instance hit on the  $j$ -th feature, and set:  $p=p+1$ . As a result, we will obtain a generalized axis coordinate of instance with the implicit ranking and selection of features.

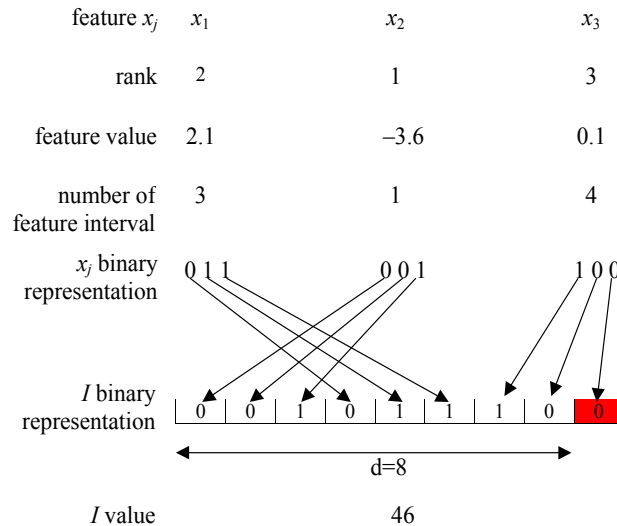


Fig. 1 Illustration of first transformation

Transformation 2 (see Fig. 2). It is an alternative format of constructing a generalized feature for transformation 1. If the total number of bits to represent interval numbers of all features  $c_j k_j N$  does not exceed the number of bits in a binary bit grid  $d$  when the values  $c_j$  are equal for all features: for each interval number of  $j$ -th feature obtain its binary representation (binary numbers padded with zeros from the left to  $c_j$  – the number of digits in  $k_j$ ) set the coordinate of  $s$ -th instance on the generalized axis  $F^s = 0$ , set the position number of coordinate on a generalized axis  $p = 1$ ; looking in a cycle on a group of digits in the interval number  $c = 1, 2, \dots, c_j$  and on feature numbers  $j$  in the descending order of their ranks: put to the  $p$ -th bit position (numbering from the left) of the binary representation of the generalized feature  $F^s$  the  $c$ -th bit (numbering from the left) of interval number, in which the  $s$ -th instance hits on the  $j$ -th feature and set:  $p=p+1$ . As a result, we obtain the generalized axis coordinate with implicit ranking of features.

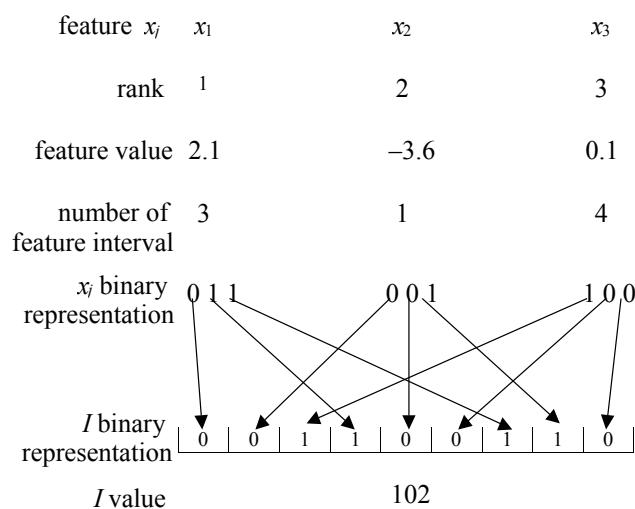


Fig. 2 Illustration of second transformation

Transformation 3. The generalized feature formed on the basis of locality-preserving hashing [12–15]. The initial feature space is divided into  $2^k$  equal hypercubes, each of which identified by the key  $F$  of a  $k$  bit length, where  $k$  is a number of feature partitions. After the  $i$ -th partition the initial feature space split to  $2^i$   $N$ -dimensional cubes, wherein the  $i$ -th partition is carried out on the  $j$ -th dimension:  $j = i \bmod N$ . At the  $i$ -th partition if hypercube located in the top half of the partitioned range, then set to one the  $i$ -th bit of its key, and otherwise set the  $i$ -th bit of its key to zero (set to one the bit in the  $i$ -th position of  $k$ -bit identifier, extended by zeros from the left, if the length is less than  $k$ ). The key  $F$  algorithmically can be generated as follows: set:  $F = 0$ ,  $x_j^{\min'} = x_j^{\min}$ ,  $x_j^{\max'} = x_j^{\max}$ , then for  $i = 1, 2, \dots, k$  do: set:  $j = i \bmod N$ ,  $x_j^{\text{mid}} = (x_j^{\min'} + x_j^{\max'})/2$ ,  $F = 2F$ ; if  $x_j^s > x_j^{\text{mid}}$ , then set:  $x_j^{\min'} = x_j^{\text{mid}}$ ,  $F = F + 1$ , else set:  $x_j^{\max'} = x_j^{\text{mid}}$ .

Transformation 4. The above-described transformations provide mapping to the discrete generalized axis. If the total number of bits to represent numbers of all feature intervals exceeds the number of bits in a bit grid of computer, it is possible to use a transformation to the generalized real axis with partial information loose: add to the real coordinate on generalized feature  $F$  the  $c$ -th bit (numbering from the left) of interval number, in which the  $s$ -th instance hits on the  $j$ -th feature (2):

$$I^s = \sum_{j=1}^N \frac{w_j \hat{x}_j^s}{r_j k_j}, \quad (2)$$

where

$$w_j = \frac{1}{k_j} \sum_{k=1}^{k_j} w_{j,k}, \quad w_{j,k} = \left\{ \frac{\max_{q=1,2,\dots,K} \{S_{j,k}^q\}}{S_{j,k}} \middle| S_{j,k} > 0 \right\}, \quad S_{j,k} = \sum_{q=1}^K S_{j,k}^q,$$

$S_{j,k}^q$  is a number of instances of  $q$ -th class located in the  $k$ -th interval of  $j$ -th feature,  $r_j$  is a rank of  $j$ -th feature (the number of  $j$ -th feature in decreasing order of individual feature importance).

Transformation 5. Define the distance from the  $s$ -th instance to the unit vector in the normalized coordinate system (3):

$$d^s = \sqrt{\sum_{j=1}^N (\hat{x}_j^s - 1)^2}, \quad (3)$$

and the angle between the instance as a vector and the unit vector (4):

$$\varphi^s = \arccos \left( \frac{\sum_{j=1}^N \hat{x}_j^s}{\sqrt{N \sum_{j=1}^N (\hat{x}_j^s)^2}} \right). \quad (4)$$

Thus we map the  $s$ -th instance from the  $N$ -dimensional space into two-dimensional space. Next for coordinates of  $s$ -th instance in formed two-dimensional space by analogy with the first transformation obtain coordinate of  $s$ -th instance on the generalized axis  $F$ .

Transformation 6. Generate  $Q$  support vectors – the centers of pseudo-clusters  $C^q = \{C_j^q\}$ ,  $q=1, 2, \dots, Q$ ,  $K \leq Q \ll S$ ,  $j = 1, 2, \dots, N$ . In the simplest case their coordinates can be set as random taking into account dimensionality and feature scales ( $x_j^{\min} \leq C_j^q \leq x_j^{\max}$ ), or by setting  $Q=K$  to determine the center of each its class (5):



$$C_j^q = \frac{1}{S^q} \sum_{s=1}^S \{x_j^s | y^s = q\}, \quad (5)$$

where  $j = 1, 2, \dots, N$ ,  $q=1, 2, \dots, K$ .

After this calculate the clusters based on their proximity and position in feature space relative to the smallest feature values:

- find the distance from the cluster centers to the point with the lowest feature values (6):

$$R_{\min}(C^q) = \sum_{j=1}^N (C_j^q - x_j^{\min})^2; \quad (6)$$

- find the distance between the cluster centers (7):

$$R(C^q, C^p) = \sum_{j=1}^N (C_j^q - C_j^p)^2; \quad (7)$$

- find the center of cluster closest to the point with the lowest feature values (8):

$$q = \arg \min_{g=1,2,\dots,Q} \{R_{\min}(C^g)\}; \quad (8)$$

– set this center as the current, set a new number of current cluster  $t=1$ , put current cluster in the set of centers with a new index ( $C^* = C^* \cup C^{*1}$ ,  $C^{*1} = C^q$ ) and delete it from the set of centers without a new index ( $C = C / C^q$ );

– while exist at least one cluster without a new index (i.e.  $C \neq \emptyset$ ) perform: among the remaining clusters without a new index in  $C$  find the closest cluster to the current cluster (9):

$$p = \arg \min_{\substack{g=1,2,\dots,Q, \\ C^g \in C}} \{R(C^q, C^g)\}, \quad (9)$$

then increase  $t = t+1$ , put the current cluster to the set of centers with a new index ( $C^* = C^* \cup C^{*t}$ ,  $C^{*t} = C^p$ ) and remove it from the set of centers without a new index ( $C = C / C^p$ ).

As a result we will receive  $C^*$  – a set of cluster centers with numbers corresponding to their proximity to the point with the lowest values of features, and also allowing to determine qualitatively the proximity of the cluster centers.

Further for each instance of the initial sample  $x^s$ ,  $s=1, \dots, S$  do:

- define the distance from it to each cluster center,  $q=1,2,\dots,Q$ , by (10):

$$R(x^s, C^{*q}) = \sum_{j=1}^N (x_j^s - C_j^{*q})^2; \quad (10)$$

- find the index of the nearest cluster center (11):

$$p = \arg \min_{q=1,2,\dots,Q} \{R(x^s, C^{*q})\}; \quad (11)$$

– find the angle between the vectors  $x^s$  and  $C^{*p}$  relative to the point with the lowest feature values (12):

$$\varphi = \arccos \frac{\sum_{j=1}^N (x_j^s - x_j^{\min})(C_j^{p*} - x_j^{\min})}{\sqrt{\sum_{j=1}^N (x_j^s - x_j^{\min})^2} \sqrt{\sum_{j=1}^N (C_j^{p*} - x_j^{\min})^2}}; \quad (12)$$

– assign the  $s$ -th instance with the coordinate on the generalized axis (13):

$$I^s = p + \frac{\varphi}{\pi}. \quad (13)$$

#### IV. CHARACTERISTICS OF TRANSFORMATIONS TO THE GENERALIZED AXIS

The introduced above transformations are encouraged to use the following characteristics of instance mapping process:

- $t^s$  is the time of transforming of one instance from the original feature space to the generalized axis for the sequential computations;
- $m^s$  is the computer memory volume used by the transformation method for processing one instance;
- $\lambda$  is the number of adjustable parameters of transformation needed for its implementation;
- $t$  is the time of calculation of transformation parameters based on the training sample;
- $m$  the computer memory volume used to calculate the transformation parameters on the basis of training sample.

Situations where several instances have equal coordinates may occur in the original and in the synthesized feature spaces. Such situations are called collisions. Under the collision point we will understand the point in the feature space, in which there is a collision.

The collision is quite admissible and even desirable in problems of automatic classification on condition that all instances located at the point of collision belongs to the same class. However, if the instances located at the point of collision, belongs to different classes, the used feature set does not provide a good separability of instances.

Denote the set of points of collision  $\{g_v\}$ ,  $v=1, 2, \dots, V$ , where  $g_v$  is a set of instances belonging to a  $v$ -th point of collision,  $V$  is the number of points of collision, which obviously can not exceed  $0,5S$ .

To estimate the quality of the results of considered transformations we propose to use the following indicators.

The number of points of collisions in which instances belongs to different classes, after the transformation of the sample to the generalized axis can be defined as (14):

$$E_{<I,y>}^* = \sum_{v=1}^V \{1 | \exists s, p=1,2,\dots,S, p \neq s : I^s \in g_v, I^p \in g_v, I^s = I^p, y^s \neq y^p\} \quad (14)$$

This indicator in the best case will be equal to zero when there is no collision points, and in the worst case it maximum value will not exceed  $0,5S$ .

The probability estimation (frequency) of the collision points in which instances belonging to different classes, after the transformation of the sample to the generalized axis can be expressed by (15):

$$P_{<I,y>}^* = \frac{2E_{<I,y>}^*}{S}. \quad (15)$$

The corrected number of points of collisions in which instances belong to different classes, after the transformation of the sample to the generalized axis is defined as (16):

$$E_{<I,y>}^{*'} = E_{<I,y>}^* - E_{<x,y>}^*, \quad (16)$$

where  $E_{<x,y>}^*$  – is the number of collision points in which the instances belongs to different classes in the initial sample (17):

$$E_{<x,y>}^* = \sum_{v=1}^V \left\{ 1 \left| \begin{array}{l} \exists s, p = 1, 2, \dots, S, p \neq s : x^s \in g_v, x^p \in g_v, \\ y^s \neq y^p, \forall j = 1, 2, \dots, N : x_j^s = x_j^p \end{array} \right. \right\}. \quad (17)$$

The indicator  $E_{<I,y>}^{*'}$  more accurately characterizes the quality of transformation to the generalized axis because it eliminates the errors present in the sample. In the best case it will be equal to zero when there is no collisions, and in the worst case it maximum value will not exceed  $0,5S$ .

The corrected probability estimation (frequency) of the collision points in which the instances belong to different classes after the sample transformation to the generalized axis can be obtained by (18):

$$P_{<I,y>}^{*'} = \frac{2E_{<I,y>}^{*'}}{S}. \quad (18)$$

The total number of instances in the collision points in which the instances belong to different classes after the sample transformation to the generalized axis is suggested to calculate by (19):

$$E_{<I,y>}^{\Sigma} = \sum_{v=1}^V \left\{ |g_v| \left| \begin{array}{l} \exists s, p = 1, 2, \dots, S, p \neq s : I^s \in g_v, \\ I^p \in g_v, I^s = I^p, y^s \neq y^p \end{array} \right. \right\}. \quad (19)$$

The more will be value of this indicator, the worse separability of instances on the generalized axis. In the best case it will be equal to zero and in the worst case it will not exceed the number of instances in the sample  $S$ .

The probability estimation of instance hitting to the collision point in which instances belong to different classes after the sample transformation to the generalized axis can be obtained by (20):

$$P_{<I,y>}^{\Sigma} = \frac{E_{<I,y>}^{\Sigma}}{S}. \quad (20)$$

The total number of instances in the collision points of the initial sample in which the instances belong to different classes it is proposed to define as (21):

$$E_{<x,y>}^{\Sigma} = \sum_{v=1}^V \left\{ |g_v| \left| \begin{array}{l} \exists s, p = 1, 2, \dots, S, p \neq s : x^s \in g_v, x^p \in g_v, \\ y^s \neq y^p, \forall j = 1, 2, \dots, N : x_j^s = x_j^p \end{array} \right. \right\}. \quad (21)$$

The more will be the value of this indicator, the worse the separability of instances of the initial sample will be. In the best case it will be equal to zero and in the worst case it not exceed the number of instances in the sample  $S$ .

The probability estimation of instance collision in the sample in which instances belong to different classes can be obtained from (22):

$$P_{<x,y>}^{\Sigma} = \frac{E_{<x,y>}^{\Sigma}}{S}. \quad (22)$$

The number of pairwise collision of instances of different classes after the sample transformation to the generalized axis is proposed to determine by (23):

$$E_{<I,y>} = \sum_{s=1}^S \sum_{p=s+1}^S \{y^s \neq y^p \mid I^s = I^p\}. \quad (23)$$

In the best case, this indicator is zero when there is no any collision, and in the worst case its value will not exceed  $S(S-1)$ .

The probability estimation (frequency) of pairwise collision of instances of different classes after the sample transformation to the generalized axis can be calculated by (24):

$$P_{<I,y>} = \frac{E_{<I,y>}}{S(S-1)}. \quad (24)$$

The corrected number of pairwise collision of instances of different classes after the transformation of training and (or) test sample to the generalized axis is proposed to determine as (25):

$$E'_{<I,y>} = E_{<I,y>} - E_{<x,y>}, \quad (25)$$

where  $E_{<x,y>}$  is a number of pairwise collision of instances of different classes in the original sample (26):

$$E_{<x,y>} = \sum_{s=1}^S \sum_{p=s+1}^S \{y^s \neq y^p \mid \forall j = 1, 2, \dots, N : x_j^s = x_j^p\}. \quad (26)$$

This indicator  $E'_{<I,y>}$  in comparison with the previous indicator more accurately characterizes the quality of the transformation to the generalized axis, because it eliminates the errors present in the original sample. In the best case, it would be equal to zero when there is no any collision, and at worst case, its maximum value will not exceed  $S(S-1)$ .

The corrected probability estimation (frequency) of pairwise collisions of instances of different classes after sample transformation to the generalized axis can be defined by (27):

$$P'_{<I,y>} = \frac{E'_{<I,y>}}{S(S-1)}. \quad (27)$$

The average number of clusters per class on a generalized axis can be calculated by (28)

$$\bar{k} = \frac{k}{K}, \quad (28)$$

where  $k$  is a number of clusters of different classes on a generalized axis.

To determine  $k$ , we need order the instances  $\langle I^s, y^s \rangle$  in ascending order on the generalized axis. Then, looking from left to right, we need to identify clusters – the intervals of one-dimensional axis, all instances of each of which belong to only one class.

The less will be the number of such clusters, the simply is partition of generalized axis.

In the best case when the classes are compact, i.e.  $k = K$ , this indicator is equal to one.

The more will be value of this indicator, the worse the separability of instances will be on the generalized axis.

In the worst case where each instance falls into a single cluster its value will be  $\bar{k} = S / K$ .

The minimum distance between instances of different classes on the generalized axis is offered to determine by (29):

$$R'_{\min} = \min_{\substack{s=1, \dots, S; \\ p=s+1, \dots, S}} \{ |I^s - I^p| \mid y^s \neq y^p \}. \quad (29)$$

The more will be value of this ratio, the better classes will be separated on the generalized axis.

The maximum distance between instances of one class on the generalized axis is offered to determine by (30):

$$R'_{\max} = \max_{\substack{s=1, \dots, S; \\ p=s+1, \dots, S}} \{ |I^s - I^p| \mid y^s = y^p \}. \quad (30)$$

The less will be this indicator value, the more compact instances of each class will be positioned on the generalized axes.

The average ratio of distances on the generalized axis and in the original feature space is proposed to calculate by (31):

$$\bar{\Delta} = \frac{R_{\max}}{0,5S(S-1)R_{\max}^*} \sum_{s=1}^S \sum_{p=s+1}^S \left\{ \frac{|I^s - I^p|}{R(x^s, x^p)} \mid R(x^s, x^p) > 0 \right\}, \quad (31)$$

where

$$R_{\max}^* = \max_{\substack{s=1, \dots, S; \\ p=s+1, \dots, S}} \{ |I^s - I^p| \}, \quad R_{\max} = \max_{\substack{s=1, \dots, S; \\ p=s+1, \dots, S}} \{ R(x^s, x^p) \}, \quad R(x^s, x^p) = \sqrt{\sum_{j=1}^N (x_j^s - x_j^p)^2}.$$

The more will be value of this indicator, the better on the average the transformation on the generalized axis reflects location of instances in the original space and features the better separability of instances on the generalized axis;

Average of the relative distance products on the generalized axis and in the original feature space is defined as (32):

$$\Delta = \frac{\sum_{s=1}^S \sum_{p=s+1}^S |I^s - I^p| R(x^s, x^p)}{0,5S(S-1)R_{\max}^* R_{\max}}. \quad (32)$$

This indicator will vary from zero to one: The more will be its value, the better on the average the transformation on the generalized axis reflects the location of instances in the original feature space.

The indicator of generalized axis feasibility of establishing we define as (33):

$$G = \frac{\min_{j=1,2,\dots,N} \{k_{<x_j,y>}\}}{k}. \quad (33)$$

where  $k_{<x_j,y>}$  is the number of intervals of different classes on the axis of feature  $x_j$ .

This indicator in the best case will be equal to  $S/K$ , and in the worst case will be equal to  $K/S$ . If this indicator will be greater than one, the use of the generalized axis will be feasible, otherwise it can be replaced with the original feature, characterized by the smallest number of intervals of different classes.

## V. THE COMPARISON CRITERIA OF GENERALIZED AXIS TRANSFORMATIONS

On the basis of the indicators characterizing the basic properties of on the generalized axis transformations introduced in the previous section it is possible to determine the criteria for comparison, the criteria for performing and criteria for evaluating the quality of the results of transformations.

The criteria for evaluation of the transformation process is proposed to define as the following:

– the combined criterion of the minimum of time and memory on the instance transformation:  $F_1 = t^s m^s \rightarrow \min$ ;

– the combined criterion of the minimum of time and memory to determine the transformation parameters for the training sample:  $F_2 = \lambda t m \rightarrow \min$ ;

– the integral criterion:  $F_3 = t^s m^s + \lambda S^{-1} t m \rightarrow \min$ .

The criteria for evaluating the quality of results of transformations:

– the criterion of the minimum of probability of instance group collisions:  $F_4 = P'_{<I,y>} \rightarrow \min$ ;

– the criterion of the minimum of probability of instance pair collisions:  $F_5 = P^{*'}_{<I,y>} \rightarrow \min$ ;

– the combined criterion of the minimum probability of pair and group collisions:  $F_6 = 0.5(P'_{<I,y>} + P^{*'}_{<I,y>}) \rightarrow \min$ ;

– the maximum of class compactness-separability:  $F_7 = \bar{k} \rightarrow \min$ ;

– the integral criterion of minimum of collisions-compactness-separability of classes calculated by (34):

$$F_8 = \frac{\bar{k}}{2} (P'_{<I,y>} + P^{*'}_{<I,y>}) \rightarrow \min, \bar{k} > 0; \quad (34)$$

– the integral criterion of minimum of collision-maximum of compactness-separability of classes and maximum of average of relative distance products on the generalized axis and in the original feature space (35):

$$F_9 = \frac{\bar{k} (P'_{<I,y>} + P^{*'}_{<I,y>})}{1 + \Delta e^{-\bar{\Delta}+1}} \rightarrow \min. \quad (35)$$

– the integral criterion of the minimum of collisions-maximum of a generalized axis establishing feasibility-compactness-separability of classes and the maximum of average of relative distances products on the generalized axis and in the original feature space (36):

$$F_{10} = \frac{(P'_{<I,y>} + P^{*'}_{<I,y>})}{G + \Delta e^{-\bar{\Delta}+1}} \rightarrow \min. \quad (36)$$

## VI. EXPERIMENTS AND RESULTS

The proposed transformations to the generalized axis have been implemented as software and experimentally studied in practical problem solving of technical and medical diagnosis, and of automatic classification [3, 16] (see Table 1). The conducted experiments confirmed the efficiency and the practical suitability of the developed mathematical tools. The experiments

TABLE I  
THE CHARACTERISTICS OF INITIAL DATA SAMPLES AND THE FRAGMENT OF THE EXPERIMENTAL RESULTS TO STUDY  
THE TRANSFORMATIONS ON GENERALIZED AXIS

Characteristics	Task			
	Gas-turbine air-engine blade diagnosis	Chronic obstructive bronchitis diagnosis	Agricultural plant recognition on the remote sensing data	Fisher Iris classification
$S$	32	205	3226	150
$N$	513	28	256	4
$K$	2	2	3	3
Best transformation: number	2	1	4	1
$E_{<I,y>}^*$	1	0	0	1
$P_{<I,y>}^*$	0.0625	0	0	0.013333
$E_{<I,y>}$	1	0	0	9
$P_{<I,y>}$	0.0010081	0	0	0.00040268
$E_{<x,y>}^*$	0	0	0	0
$E_{<I,y>}^{**}$	1	0	0	1
$E_{<x,y>}$	0	0	0	0
$E_{<I,y>}'$	1	0	0	9
$E_{<x,y>}^\Sigma$	0	0	0	0
$P_{<x,y>}^\Sigma$	0	0	0	0
$E_{<I,y>}^\Sigma$	2	0	0	10
$P_{<I,y>}^\Sigma$	0.0625	0	0	0.066667
$k$	15	72	1773	11
$R_{\max}$	70.114	918.99	7.5551	7.0852
$R_{\max}^*$	$1.5228 \cdot 10^9$	$1.6927 \cdot 10^9$	3.3217	$1.5126 \cdot 10^9$
$R_{\min}'$	0	1056	$1.1437 \cdot 10^{-7}$	0
$R_{\max}'$	$1.5228 \cdot 10^9$	$1.6833 \cdot 10^9$	3.318	609746944
$\overline{\Delta}$	0.82534	0.69432	0.37993	0.88456
$\Delta$	0.095099	0.098493	0.028541	0.18511
$G$	0.2	0.69444	0.56007	1
$t^s$	0.00097501	0.00053269	$3.8686 \cdot 10^{-5}$	0.000416
$m^s$	4309	262.83	2057.9	82.133
$t$	0.2184	0.093601	4.524	0.1092
$m$	333484	109644	6764132	20340
$\lambda$	1539	84	1024	12
$F_1$	4.2013	0.14001	0.079613	0.034168
$F_2$	$1.1209 \cdot 10^8$	$8.6207 \cdot 10^5$	$3.1336 \cdot 10^{10}$	$0.26654 \cdot 10^5$
$F_3$	$3.5028 \cdot 10^6$	4205.4	$9.7134 \cdot 10^6$	177.73
$F_4$	0.0010081	0	0	0.00040268
$F_5$	0.0625	0	0	0.013333
$F_6$	0.031754	0	0	0.006868
$F_7$	7.5	36	591	3.6667
$F_8$	0.23816	0	0	0.025183
$F_9$	0.42786	0	0	0.041701
$F_{10}$	0.20274	0	0	0.011373



have shown that the proposed transformations allow to significantly reduce the data sample dimensionality.

The Table 1 shows the best transformations for different practical tasks according to proposed criteria.

The proposed transformations can be recommended for use in the construction of diagnostic and recognizing models by precedents, as well as for the formation of the training samples from the source samples of large number of instances.

## VII. CONCLUSION

The urgent problem of the development of mathematical support for data dimensionality reduction was solved in the paper.

The scientific novelty of results consists in that the set of rapid transformations from the original multi-dimensional space into one-dimensional axis was firstly proposed. The proposed transformations provide a solution both the problem of constructing of artificial features (feature extraction), and the problem selection of the most significant features (feature selection).

- the set of rapid transformations from the original multi-dimensional space into one-dimensional axis was firstly proposed. It is based on the principles of hashing and provides taking into account the instance locations in the feature space with respect to the class centers of gravity, and also allows to determine and to take into account the feature weights and thereby implicitly solves the problem of feature selection. Thus, the proposed transformations provide a solution both the problem of constructing of artificial features (feature extraction), and the problem selection of the most significant features (feature selection);

- the complex of indicators characterizing the properties of transformations from multidimensional space to generalized axis was firstly proposed. On the basis of the proposed indicators the set of criteria is defined. It facilitates comparison and selection of the best transformations and results of their work at diagnosis and recognition problems solving by precedents.

The practical significance of obtained results is that:

- the software realizing proposed transformations and indicators characterizing their properties was developed. Its usage allows to automate the data dimensionality reduction and analysis of its results;

- the experimental investigation of the proposed transformations and the indicators characterizing them was conducted at practical problem solving. The results of research allow to recommend the proposed transformations for use in practice for diagnosis and pattern recognition problem solving.

The prospects of further utilization of the results obtained in this work consists in the possibility of their use for automation of training and testing samples formation (data dimensionality reduction through decrease of the number of precedents by the most important precedents extraction).

## ACKNOWLEDGMENT

The work is partially supported by European Commission Tempus program project "Centers of Excellence for young REsearchers (CERES)" (544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

## REFERENCES

- [1] R. Jensen, and Q. Shen, *Computational intelligence and feature selection: rough and fuzzy approaches*, Hoboken: John Wiley & Sons, 2008.
- [2] O. V. Babak, "The solution of some problems of data processing on the basis of the general generalized variable" ["Reshenie nekotorykh zadach obrabotki dannykh na osnove metoda generalnoy obobshchennoy peremennoy"], *Problemy upravleniya i informatiki*, 2002, no. 6, pp. 79–91. (in Russian)
- [3] S. A. Subbotin, An. A. Oleinik, E. A. Gofman et al., *Intelligent information technologies of automated diagnostic and pattern recognition systems design : monograph [Intellektualnye informatsionnye tekhnologii proektirovaniya avtomatizirovannykh sistem diagnostirovaniya i raspoznavaniya obrazov : monografiya]*. Ed. by S. A. Subbotin. – Kharkov: Company "SMIT", 2012.. (in Russian)
- [4] T. W. Lee, *Independent component analysis: theory and applications*, Berlin: Springer, 2010.
- [5] J. A. Lee, and M. Verleysen, *Nonlinear dimensionality reduction*, New York: Springer, 2007.
- [6] C. Padraig, "Dimension reduction," Technical report UCD-CSI-2007-7, Dublin: University College Dublin, 2007.
- [7] Y. Jiang, R. Zhang, G. Liu et al., "Multifactor dimensionality reduction for detecting haplotype-haplotype interaction," in *Proc. Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, 2009, pp. 241–245.
- [8] B. Kulis, A. C. Surendran, and J. C. Platt, "Fast low-rank semidefinite programming for embedding and clustering," in *Proc. Eleventh International Conference on Artificial Intelligence and Statistics*, San Juan, 2007, vol. 2, pp. 235–242.
- [9] O. V. Babak, and A. E. Tatarinov, "An approach to solving classification problems under incomplete information," *Cybernetics and Systems Analysis*, 2005, vol. 41, no. 6, pp 891–897.
- [10] V. I. Vasilev, A. I. Shevchenko, and S. N. Esh, *Principle of reductions it the problems of dependence detection: monograph [Printsip reduksii v zadachakh obnaruzheniya zakonomernostey : monografiya]*, Donetsk: Nauka i osvita, 2009. (in Russian)
- [11] S. Yu, "Feature selection and classifier ensembles: a study on hyperspectral remote sensing data," Proefschrift ... doctor in de wetenschappen. Antwerpen: Universitaire Instelling Antwerpen, 2003.
- [12] J. Jianqiu, J. Li, Sh. Yany, B. Zhang et al., "Super-bit locality-sensitive hashing," in *Advances in Neural Information Processing Systems*, P. Bartlett et al. Eds. 2012, vol. 25, pp. 108–116.
- [13] A. Andoni, and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Communications of the ACM*, 2008, vol. 51, no. 1, pp. 117–122.
- [14] X. Yang and Y. Hu, "A scalable index architecture for supporting multi-dimensional range queries in peer-to-peer networks," in *Proc. International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom-2006)*, Atlanta, 2006, pp. 1–10.
- [15] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Nearest neighbor methods in learning and vision: theory and practice*, T. Darrell, P. Indyk, G. Shakhnarovich Eds. Massachusetts: MIT Press, 2006, pp. 55–67.
- [16] R.A. Fisher, Iris Data Set. Available on-line: <https://archive.ics.uci.edu/ml/datasets/Iris>



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Risk-Oriented Decision Making During Integrated Investment Management under Uncertainty

Valerii Lovkin

**Abstract**—The model of decision-making during integrated investment management under uncertainty, which enables to distribute capital between real and financial investment, is stated. The problem solution method, which allows supporting decisions in distribution and monitoring stage, is proposed. The model and its solution method enable to manage real investments and security portfolio simultaneously allocating all available resources.

**Keywords**—investment management, decision making, real investment, financial investment, unsuccess risk.

## I. INTRODUCTION

Investment decision making is an important process because its results directly impact on development and financial condition of investment activities subject and on the economy of the all country.

Investments include assets, specific deposits, shares, stocks and other securities, technologies, machinery and equipment, credits and all other types of property and property rights, intellectual values etc. But according to the investment goal there are two types of investment: real and financial (security portfolio) investments. Real investment belongs to short-term processes and financial investment can be short-term or long-term.

It is necessary to take into account that financial capital making and turnover are conditional on real capital functioning. So it is necessary for investor to distribute capital among real and portfolio investment to manage these two types of investment simultaneously to increase investment effectiveness. This approach enables to manage investment risks using diversity.

A lot of works are devoted to real investment and financial investment separately.

In [1] Project Definition Rating Index (PDRI) was proposed for project state analysis in the pre-project planning phase by the Construction Industry Institute. The PDRI is used to collect information of the pre-project planning practices in the industrial and building industry. The PDRI for industrial projects is a weighted matrix with 70 scope definition elements grouped into 15 categories and further grouped into three main sections. In [2] the PDRI for building projects was proposed, it has 64 scope definition elements and 11 categories grouped into three sections.

In [3]-[5] dependence between actual project costs and duration deviation from planned values and PDRI is investigated using regression model. In [5] artificial neural networks were proposed for prediction the dependence and the results of artificial neural networks usage were compared with regression model. In [6] artificial neural networks (perceptrons) and their ensembles were investigated for project success modeling.

A lot of classic papers (portfolio theory) by H. M. Markowitz, W. F. Sharpe, J. Tobin were devoted to the security portfolio management problem. In [7] the approach to management of the security portfolio, based on application of the theory of fuzzy sets, was presented. In [8] the approach to the security portfolio management based on D-scores of Russman was modified.

The purpose of this paper is to integrate two main types of investment into the complex investment management problem.

## II. THE MODEL OF DECISION-MAKING DURING INTEGRATED INVESTMENT MANAGEMENT UNDER UNCERTAINTY

Taking into account dependence between real and financial investment the investment decision-making problem should be determined as an allocation of the capital among real investments and security portfolio in such a way that capital is invested in real investment if the investment of the capital, which is required for investment project execution during the necessary investment period of time, is more effective with an allowance for risk than the investment of the same capital for the same period of time in investment (security) portfolio.

Besides that the short-term investment decision-making problem rises as a result of the long-term problem solving. The short-term problem is caused by a real investment fund utilizing way: the capital is utilized during a full investment period, not only in its beginning. Therefore temporary free capital can be used for portfolio investment for short periods of time. The investor should make a decision if it is effective with an allowance for risk to invest temporary free capital in portfolio during the period of its availability.

Investment efficiency with an allowance for risk is estimated by experts based on efficiency and risk measure of investment process. These real and portfolio investment measures can be compared because of the equal capital and investment period.

Decision making during integrated investment management should be done in both short and long periods of time:

- in long period of time capital should be allocated between real investments and security portfolio investments;
- in short period of time capital should be allocated between security portfolio investments and standard allocation of temporarily free resources (for example, deposit).

Let  $Q$  be total number of investment directions:  $Q$  real investment projects and  $Q$  alternative security portfolio investments.  $T$  is a maximum real investment duration (considering all  $i = 1 \dots Q$  projects).

$U = (U_1, U_2, \dots, U_T)$  is a vector, which elements determine current control action in point of time  $t = 1, 2, \dots, T$ . Each element of this vector is a vector  $U_t$ , which can be determined as:

$$U_t = (U_1^t, \dots, U_Q^t, U_{Q+1}^t, \dots, U_{2Q}^t),$$

where  $U_1^t, \dots, U_Q^t$  – control action on every  $i = 1 \dots Q$  real investment project in point of time  $t$  ( $U_1^t, \dots, U_Q^t = \{0, 1\}$ ). Instead of a real investment project a standard allocation of temporarily free resources can be considered, but these free resources should be used in a real investment in a period of time;

$U_{Q+1}^t, \dots, U_{2Q}^t$  – control action on every  $i = Q+1 \dots 2Q$  security portfolio investment in point of time  $t$  ( $U_{Q+1}^t, \dots, U_{2Q}^t = \{0, 1\}$ ).

The model of decision-making during integrated investment management under uncertainty consists in finding such a control strategy  $U$  from feasible solution set  $\Omega$  with initial state  $x_0$ , which maximizes profit derived from real and financial investments during period of time  $T$  and minimizes unsuccess risk of the investment process. The investment process is characterized by object state vector  $x$  considering uncertainty  $\varpi$ . The model can be mathematically stated as:

$$M \sum_{t=0}^T B(U_t, x_t, \varpi) \rightarrow \max_{U \in \Omega}, \quad (1)$$

$$\max_{t=0 \dots T} M \{R(U_t, x_t, \varpi)\} \rightarrow \min_{U \in \Omega}, \quad (2)$$

$$\Omega: x_{t+1} = x_t + f(\hat{x}_t, U_t) + w_t, \quad (3)$$

$$U_i^t \oplus U_{Q+i}^t = 1 \quad (4)$$

$$U_1^t, \dots, U_{2Q}^t = \{0, 1\}, \quad (5)$$

$$S_i^r = S_i^f, \quad (6)$$

$$T_i^r = T_i^f \quad (7)$$

where  $B(U_t, x_t, \varpi)$  – profit from investment, which is characterized by control action  $U_t$ , control object state  $x_t$  and uncertainty  $\varpi$ , in point of time  $t$ ;

$R(U_t, x_t, \varpi)$  – risk of investment, which is characterized by control action  $U_t$ , control object state  $x_t$  and uncertainty  $\varpi$ , in point of time  $t$ ;

$M\{\dots\}$  – mathematical expectation;

$\hat{x}_t$  – predicted investment state: real investment costs, real investment duration, unsuccessful risk of real and financial investment, financial investment profitability.

Equations (1)-(2) define optimization problem. Equations (3)-(7) define model constraints.

Control action defines if it is necessary to invest the same capital in real project or security portfolio for the same period of time (only one variant for each capital amount  $i = 1 \dots Q$ ). This condition is presented by (4)-(7).

The presented model of decision-making during integrated investment management under uncertainty is a dynamic programming problem model and belongs to the problem class of optimal stochastic control in finite horizon.

Uncertainty  $\varpi$  is defined by vector of random disturbances which impact on the system. Vector  $\varpi$  consists of the following elements:

- $\varpi^1$  – vector of random disturbances provided by predicted ( $\xi_i^1$ ) and actual real investment unsuccessful risk value deviation;
- $\varpi^2$  – vector of random disturbances provided by predicted ( $\xi_i^2$ ) and actual investment costs value deviation;
- $\varpi^3$  – vector of random disturbances provided by predicted ( $\xi_i^3$ ) and actual investment duration value deviation;
- $\varpi^4$  – vector of random disturbances provided by predicted ( $\xi_i^4$ ) and actual financial investment unsuccessful risk value deviation;
- $\varpi^5$  – vector of random disturbances provided by predicted ( $\xi_i^5$ ) and actual financial investment profitability value deviation;
- $\varpi^6$  – vector of random disturbances provided by inaccuracy in expert scores ( $\xi_i^6$ ) of real investment projects priority.

It is necessary to predict the following investment metrics to solve the stated problem: real and financial investment unsuccessful risk, actual real investment costs and duration deviation from planned values, financial investment profitability.

### III. THE METHOD OF INTEGRATED INVESTMENT MANAGEMENT DECISION MAKING

The method of integrated investment management decision making is intended for the problem solving defined by the model (1)-(7). This method uses an iterative scheme which consists of two components intended for optimal control action synthesis: prediction of the investment metrics and dynamic optimization of decisions based on prediction results.

At the first stage all projects available for real investment and all securities available on the stock market should be defined. Then preliminary project selection should be done. The total costs of the selected projects should not be more than available capital. The selected projects

should be the most competitive of all available projects. For this purpose the PDRI of all available projects should be estimated by the group of experts. The estimations of individual experts should be adjusted by the Dephi method.

If there is large amount of available projects, experts can be divided into the groups by their speciality. Projects for these groups should be selected using clusterization algorithms. Density tracking self-organizing maps should be used for cluster analysis. The process of selection should be done based on iterative procedure. On every iteration the number of projects should be decreased. This process is similar to genetic algorithms [9].

The fuzzy analytic hierarchy process (FAHP) [10] is proposed to be used for project estimation and selection. At every stage  $N_i$  projects should be selected.  $N_0$  projects should be selected using FAHP and other  $N_i - N_0$  projects should be selected using roulette selection mechanism based on principles of genetic algorithms [9].

At the second stage unsuccess risk, expected duration and costs deviation from planned values should be estimated, using methods of real investment unsuccess risk prognostication and prognostication of actual real investment results deviation from planned real investment results [11].

Projects are considered as successful if their actual costs and duration don't exceed planned values:

$$CI = \frac{AC - PC}{PC} \leq 0, \quad (8)$$

$$DI = \frac{AD - PD}{PD} \leq 0, \quad (9)$$

where  $AC$  and  $AD$  – actual costs and duration of project;

$PD$  and  $PC$  – planned costs and duration of project;

$CI$ ,  $DI$  – costs and duration index.

Unsuccess risk is unsuccess project class belonging probability. Method of real investment unsuccess risk prognostication [11] predicts this probability based on data which defines project characteristics and artificial neural networks ensembles using AdaBoost method.

The method of prognostication of actual real investment results deviation from planned real investment results [11] is based on data that defines project characteristics, Kohonen self organizing maps, neural networks and NeuroEvolutionary Algorithm.

The modified investment portfolio management method based on D-scores of Russman [8] should be used for portfolio optimization to determine security portfolios that should be used as alternative for real investment. This method allows to manage investments more efficiently on unstable stock exchange. In the modified method genetic algorithm is used for the portfolio optimization problem solving: genes in chromosome encode portfolio structure. Yield increase range and system motion path to the goal are used for estimation using the fuzzy group method of data handling.

To use the modified investment portfolio management method based on D-scores of Russman for the method of integrated investment management decision making it is necessary to estimate planned profitability. The Delphi method should be used for this purpose.

After all metrics are predicted it is necessary to evaluate control actions based on dynamic programming method, allocating all available capital in real and financial investment. Using main criterion method optimization problem (1)-(2) can be transformed to optimization problem (1) with constraint:

$$DI = \frac{AD - PD}{PD} \leq 0, \quad (10)$$

So the method of integrated investment management decision making solves the model of decision-making during integrated investment management under uncertainty as optimization problem (1) with constraints (3)-(7), (10).

After this stage all projects and security portfolios should be monitored during period of time  $T$ . If there are free temporary available resources it is necessary to allocate these resources using model (1) with constraints (3)-(7), (10).

#### IV. EXPERIMENTAL INVESTIGATION OF INTEGRATED INVESTMENT MANAGEMENT

In the previous papers [8], [9] the efficiency of the proposed methods usage individually for real investment and security portfolio management was investigated.

For experimental investigation of the presented approach historical data were used.

These data are collected from building and industrial project execution and are used as a real investment data source. The real investment data consist of PDRI, planned project duration, planned project costs, unsuccess class, actual project duration and actual project costs.

The other part of the data consists of the Ukrainian stock exchange trade performance. These data characterize security prices on the Ukrainian stock exchange during last 10 years.

The decision support system for investment management process was developed using programming language C++.

During experimental investigation of integrated investment management approach the integrated investment management decision making method was compared with individual real and financial investment management when capital is initially separated for these two types of investment. Investment period was set equal to 4 years. The proposed approach allowed to increase profitability for 14,3 %. Using this method 55 % of capital was allocated in financial investment and 45 % in real investment.

#### V. CONCLUSION

In this paper risk-oriented decision making approach to integrated investment management was presented. Using this approach real and financial investments can be managed simultaneously.

The model of decision making during integrated investment management under uncertainty was presented. This model can be used for allocating capital between projects, security portfolios and standard allocation of temporary available resources which should be used for financial investment. The model is a dynamic programming problem model and belongs to the problem class of optimal stochastic control in finite horizon.

The method of integrated investment management decision making which is intended for solving the problem defined by the described model was presented. This method uses main criterion method for solving the model problem. It is based on the method of prognostication of actual real investment results deviation from planned real investment results, on the method of real investment unsuccess risk prognostication and on the modified investment portfolio management method based on D-scores of Russman.

Application efficiency of the proposed approach is confirmed by the implementation results.

#### REFERENCES

- [1] G. E. Gibson, P. R. Dumont, "Project definition rating index (PDRI): Construction Industry Institute research report", Austin: UTA, 1996, 95 p.
- [2] C.-S. Cho, G. Edward Gibson Jr., "Building project scope definition using project definition rating index", *Journal of Architectural Engineering*, 2001, Vol. 7, No. 4, pp. 115–125.
- [3] P.-A. Ubach de Fuentes, "Validation of the project definition rating index (PDRI) for MIT building projects", M.S. Thesis, Massachusetts Institute of Technology, 2004, 95 p.
- [4] Y.-R. Wang, "Applying the PDRI in project risk management", Ph.D. Thesis, Austin, TX, 2002, 268 p.



- [5] Y.-R. Wang, G. Edward Gibson Jr., “A study of preproject planning and project success using ANN and regression models”, The 25th International Symposium on Automation and Robotics in Construction, Vilnius: Vilnius Gediminas Technical University, 2008, pp. 688–695.
- [6] Y.-R. Wang, Y.-J. Chen, C. F. Jeffrey Huang, “Applying neural network ensemble concepts for modelling project success”, 26th International Symposium on Automation and Robotics in Construction, Austin: University of Texas at Austin, 2009, pp. 199–204.
- [7] Y. Zaychenko, M. Esfandiyarfard, “Optimization of the investment portfolio in the conditions of uncertainty”, International Journal “Information Technologies and Knowledge”, 2008, № 3, Volume 2, pp. 225–233.
- [8] V. Liovkin, V. Dubrovin, “Investment portfolio management process”, Computer Science and Information Technologies: Materials of the VIth International Scientific and Technical Conference CSIT’2011, Lviv: Publishing House Vezha&Co, 2011, pp. 218–221.
- [9] “The practical handbook of genetic algorithms: applications”, Second edition, edited by Lance D. Chambers, Boca Raton, Florida: Chapman & Hall, CRC, 2000, 501 p.
- [10] T. L. Saaty, L. G. Vargas, “Economic, political, social and technological applications with benefits, opportunities, costs and risks,” Springer, 2011, 288 p.
- [11] V. Liovkin, “Methods and tools for investment decision-making under uncertainty”, The thesis for the candidate degree in technical sciences on the speciality 01.05.04 – System Analysis and Optimal Decision Theory, Kharkiv: Kharkiv National University of Radio Electronics of Ministry of Education and Science of Ukraine, 2013, 23 p.



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Advanced logic cell FPGA

Ruslan Vikhorev, Sergey Tyurin

**Abstract**— Look-up table (LUT) realizing system from  $m$  of logic functions – DC-LUT is presented in this paper. The structure of the disjunctions block of constituents is improved considering the requirement that only one diagram way shall be activated. The carried out modelling in the circuit design system NI Multisim 10 by National Instruments Electronics Workbench Group confirms the operability of the proposed improved technical decision for the disjunctions block of DC-LUT constituents.

**Keywords**— Boolean functions, FPGA, PDNF, Look-up table (LUT).

## I. INTRODUCTION

FPGA PLD Logic Cell, often called LUT (Look Up Table – truth table), is based on a multiplexer that represents a tree of elementary multiplexers 2-1 on the basis of the transmitting MOS transistors, the input of which are configured by the so-called reconfigurable cells of static memory SRAM [1]. Any functions, including constants, may be implemented by uploading the truth tables values of  $n$  variables functions into a configuration memory SRAM.

A typical cell consists of a 4-input LUT (RAM address inputs), but the modern up-to-date FPGAs have the complicated, configurable 6,7-input LUTs. There are reports that the 8-input LUTs exist [4].

## II. DC-LUT LOGIC ELEMENT

As known, the drains and sources of the CMIS transistors used in LUT with the same topological characteristics are actually equivalent. A transistor circuit, inverse to the LUT structure, can be created by “making a U-turn” [5,6] - Fig.1.

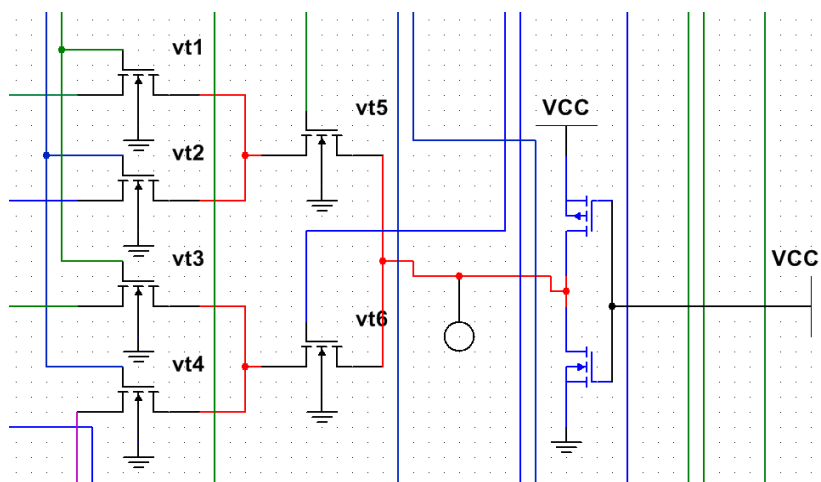


Fig. 1 Transistor circuit – DC –LUT 2 decoder.

By doing so, in accordance with design rules for the circuits from the transmitting transistors, an alternative line should be created for each output inverter to convert its output into “1”.

R. Vikhorev Perm National Research Polytechnic University, Russia (e-mail: vihrusvla@gmail.com).

S. Tyurin, Perm National Research Polytechnic University, Russia (e-mail: tyurinserggeo@yandex.ru).

Therefore, a dual to conjunction expression describing the corresponding line in the inverse tree should be used – Fig.2.

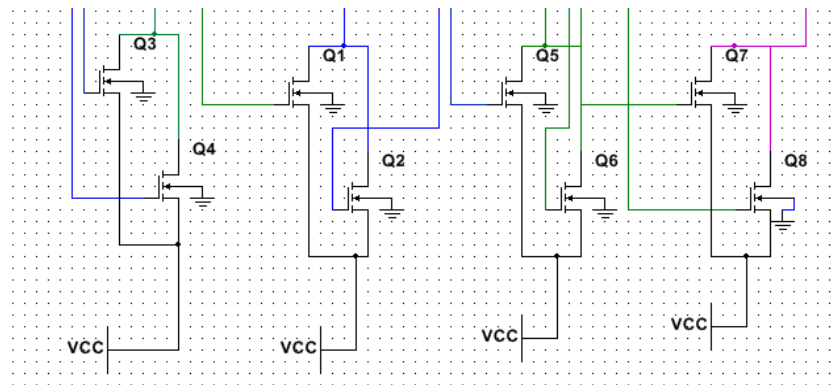


Fig. 2 DC LUT2 zero bit with alternative line.

We got a decoder with  $z$  output functions as follows (without regard to alternative lines):

$$\bar{z}_i = \& x_j^{\sigma_j^{2^n}}, i = 0 \dots 2^n - 1, j = 1 \dots n.$$

Then, by uniting the corresponding outputs from  $2^n$  ones by “OR” [5,6], we implement the  $n$ -digit logic functions system on the basis of the principal disjunctive normal forms (PDNF).

The connections of the outputs 0,1,2,2 can be programmed with the corresponding OR elements in order to implement the logic functions systems. It can be programmed similarly to the interconnects programming [1] - Fig.3. Thus, the disjunctions blocks of the logic functions constituents may be implemented in a manner shown in Fig.3.

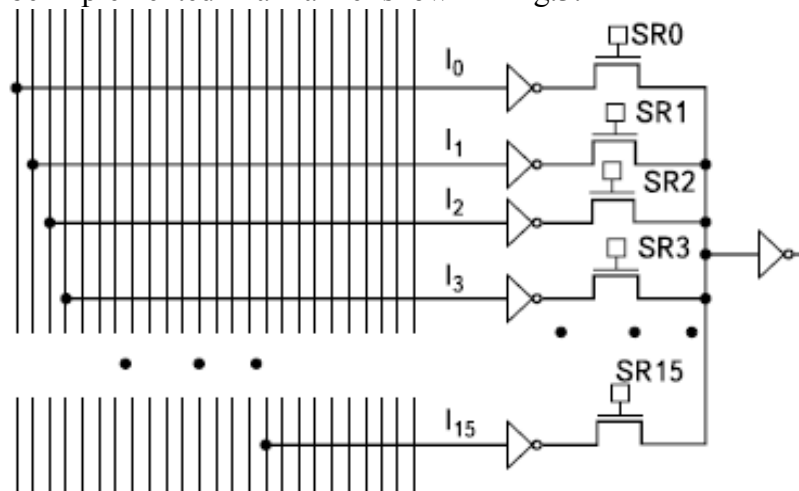


Fig. 3 Programmed interconnects for 16 directions (SR0...15 configuration memory SRAM).

### III. IMPROVING DISJUNCTIONS BLOCK CONFIGURATION METHOD

The former suggested disjunctions block of the logic function constituents [5,6] have looked as shown in Fig.4.

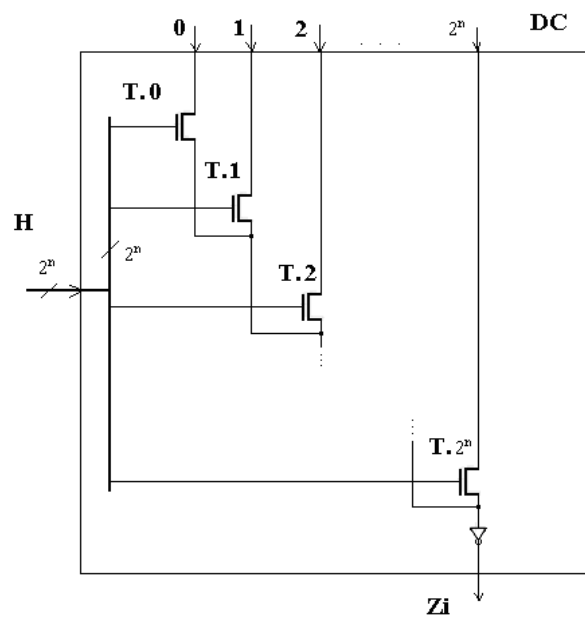


Fig. 4 Initial implementation of disjunctions block of the logic function constituents.

It has been supposed that configuration H for implementing required disjunction of constituents from the modified tree of DC transistors was carried out with the help of the configuration memory [1,8,9,10], not shown in Fig.5. However such implementation does not meet the requirements [7], because more than one MOS transistor can be activated simultaneously – the configuration arrives at gates, whereas the signals from the modified tree of DC transistors arrive at the sources. That is why, the disjunctions block diagram should be changed and look like it shown in Fig.5.

In the diagram of Fig.5 the signals from the modified tree of DC transistors activate the only gate of the only MOS transistor.

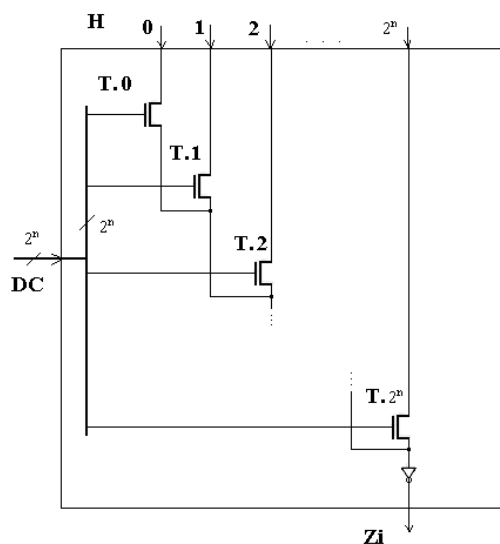


Fig. 5 Improved disjunctions block of the logic function constituents.

#### IV. MODELLING LOGIC ELEMENT FOR IMPLEMENTING LOGIC FUNCTIONS WITH IMPROVED DISJUNCTIONS BLOCK OF THE LOGIC FUNCTION CONSTITUENTS

Let model a logic element for implementing logic functions with improved disjunctions block of the logic function constituents for  $n=2$  (Fig.6) in the circuit design system NI Multisim 10 by National Instruments Electronics Workbench Group. In Fig.7 the keys X1, X2 model the

variables. In Fig. 8 The keys S1- S4 model the configuration H.

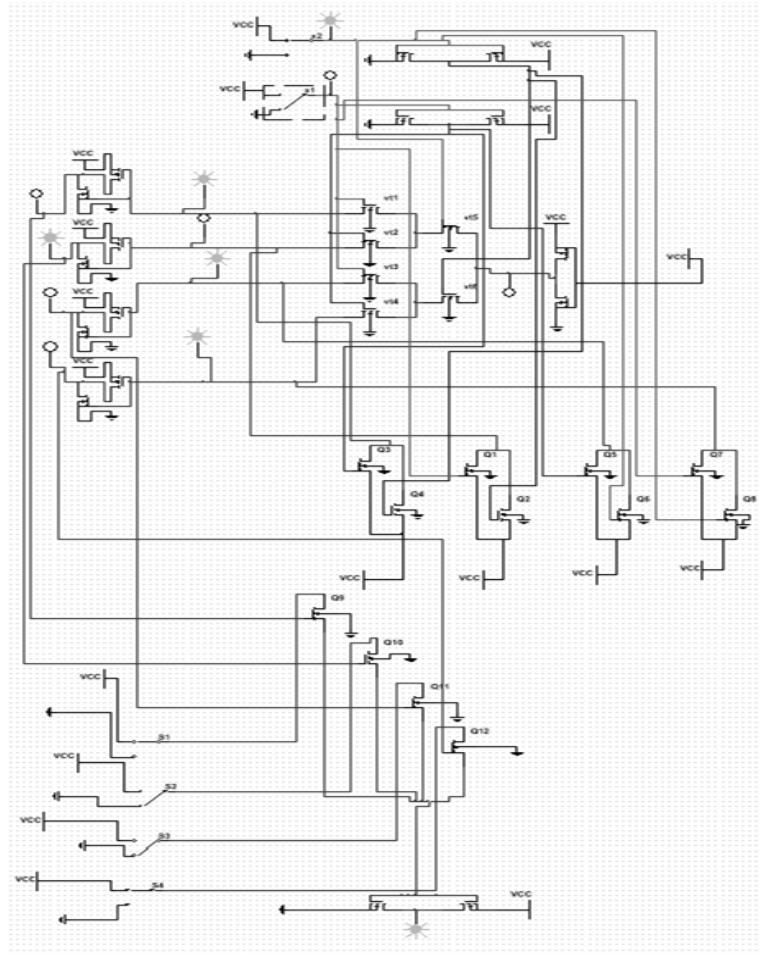


Fig. 6 Logic element for implementing logic functions with improved disjunctions block of the logic function constituents for  $n=2$  with one improved disjunctions block of the logic function constituents.

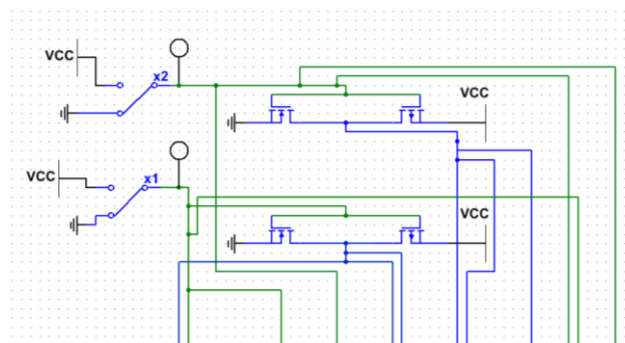


Fig. 7 The keys X1, X2.

The gates of disjunctions block transistors are controlled by invertors installed in the outputs of the detailed (regard to the standard LUT) transistors tree. At the same time, the alternative lines provide the transmission of the logic units when the tree is not activated. Fig.7 shows implementation of the function “Excluding OR” – a LED lights in the OR block output, because the variables X1,X2 have different values.

The carried out modelling confirms the operability of the proposed improved technical decision for the disjunctions block of DC-LUT constituents, an application for a patent is applied.

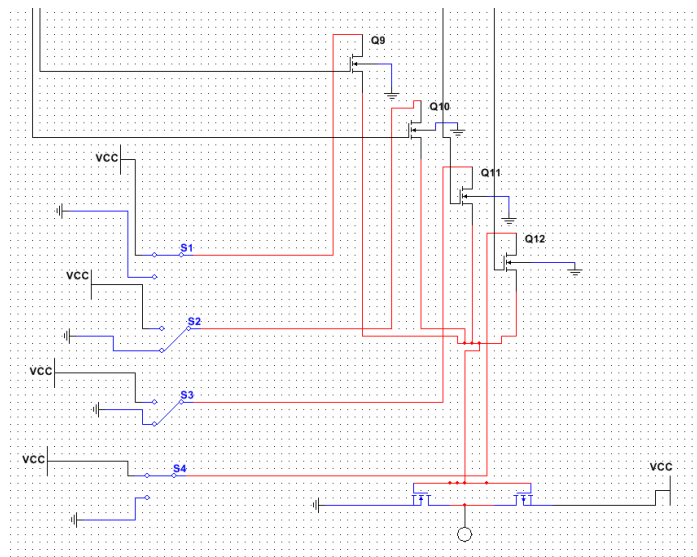


Fig. 8 One improved disjunctions block of the logic function constituents.

## V. CONCLUSIONS

Thus, the structure of the proposed logic element of DC LUT decoder is modified for implementing the systems of the functions in the FPGA PLD, which is based on the diagram of the MOS transistors tree. The structure of the disjunctions block of constituents is improved considering the requirement that only one diagram way shall be activated.

In the changed diagram, the signals from the modified tree of DC transistors arrive at the gates of the disjunctions block MOS transistor, whereas the configuration data from the configuration memory arrives at the sources of these transistors.

The carried out modelling in the circuit design system NI Multisim 10 by National Instruments Electronics Workbench Group confirms the operability of the proposed improved technical decision for the disjunctions block of DC-LUT constituents.

## REFERENCES

- [1] Tsibin, S. Programmable switching FPGA: a view from the inside. [http://www.kit-e.ru/articles/plis/2010\\_11\\_56.php](http://www.kit-e.ru/articles/plis/2010_11_56.php)
- [2] Samkova E. Stratix IV vs Virtex-5. Competition is not finished. <http://www.russianelectronics.ru/leader-r/review/2189/doc/46576/>
- [3] Aria II GX. Altera. <http://icgamma.ru/linecard/altera/arria2gx/>
- [4] An Ultra-Low-Energy, Variation-Tolerant FPGA Architecture Using Component-Speci\_c Mapping <http://thesis.library.caltech.edu/7226/>
- [5] Tyurin, S.F. Logic elements for implementing functions systems in FPGA. Design and technology of electronic instrumentation. 2013. № 4. pp. 33-37.
- [6] Tyurin, S.F., Gorodilov A. Yu., Gromov O.A., Suleymanov A.A. System logic implemetation in FPGA. Bulletin of Perm University. Mathematics. Mechanics. Computer Science. 2013. № 4. pp. 85-90.
- [7] Jeffrey D. Ullman Computational Aspects of VLSI. Translation into Russian.: A.V. Neumann. Edited by P.P.Parkhomenko. - M.: Radio and communication, 1990. - 480 p. ISBN 5-256-00253-8.
- [8] 8. Tyurin S.F., Grekov A.V., Gromov O.A. The principle of recovery logic FPGA for critical applications by adapting (3) P. 328–332. DOI: 10.5829/idosi.wasj.2013.26.03.13474.
- [9] 9. Tyurin S.F., Gromov O.A. A residual basis search algorithm of fault-tolerant programmable logic integrated circuits // Russian Electrical Engineering. – 2013. – 84 (11). – P. 647–651. DOI: 10.3103/S1068371213110163
- [10] A.N.Kamenskikh, S.F.Tyurin. Advanced Approach to Development of Energy-Aware and Naturally Reliable Computing Systems. Proceeding of the 2015 IEEE North West Russia Section Young researches in electrical and electronic engineering conference (2015 EIConRusNW). pp67-69

# Automated decision support system for GTU tests process

Olena Shytikova, Galyna Tabunshchyk

**Abstract**—The paper deals with the problems of reliability and safety of the gas turbine units for terrestrial usage (GTU). Developed by the author models of GTU workflows, documents workflows and mathematical models of uncertainty in GTU tests process were put in the DSS. In the article presents main requirements for DSS, its architecture, the data storage structure, subsystems for information administration, monitoring, data visualization and analysis.

**Keywords**—automated decision support system, gas turbine unit for terrestrial usage, tests process.

## I. INTRODUCTION

The contemporary operation concept of gas turbine units for terrestrial usage (GTU) is determined by customer requirements towards qualitative reduces of their life cycle costs with simulations performance of the reliability increase. For the confirmation of the products reliability are used numerous tests.

In practice all the reliability characteristics are determined by statistical data processing, which is quite time-consuming, and therefore must be properly organized. And with the number of issued GTU the process of registration and storage of large amounts of statistical data only becomes more complicated.

One of the functions of the GTU Automatic Control System (ACS) is the formation of the back-up information about the parameters and status of GTU during functioning, including information tests collected.

However, in addition to the back-up files with the parameters GTU created by ACS, there are formed a numerous additional data, which is necessary to analyze, organize and ensure its safety during GTU life-cycle. Today, it's all stored printed on paper reports. In addition, data processing and analysis take a long time.

The transition from the huge paper archives of data towards automated decision support systems (DSS) can as significantly reduce the amount of information, but also accelerate the access process to the necessary sources, and reduce the processing time and improve the quality of the results, excluding the human factor from the process.

## II. PROBLEM DEFINITION

There are various means of lifecycle automation, such as the enterprise management system and ACS of workflows and documents flow, etc. But they did not take into account the features of the GTU tests process. Therefore, the development of the software that would take into account all the features of the GTU tests process is actual task [1-3].

The object domain analysis has allowed to design as the models of GTU test process workflows [4], which made it possible to determine the basic format and structure of electronic documents, as well a formalized model of GTU tests process (1) [5].

O. Shytikova, Software Tools Department, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine (e-mail: helenshitikova@gmail.com).

G. Tabunshchyk, Software Tools Department, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine (e-mail: galina.tabunshchik@gmail.com).



$$M_H = \langle T, O, R, X, Y, Z, U, D \rangle, \quad (1)$$

where  $T$  – the set of GTU test process stages;

$O$  – the set of objects;

$R$  – the set of the links between objects;

$X, Y$  – the set of input and output parameters;

$Z$  – the set of objects states that vary according to the condition  $u$  in the set  $U$ ;

$U$  – the set of conditions to change the objects states;

$D$  – the set of operations performed at model time  $t$  on the object  $o \in O$ , located in the state  $z \in Z$  until a condition  $u \in U$ . It is defined as a set of ordered elements

$$d = \langle t, o, z, \tilde{X}, u, M'_R \rangle,$$

where  $\tilde{X}$  – the subset of  $X$ , which includes input parameters  $x$ , necessary to perform operations  $d$ ;

$M'_R$  – risk models of GTU test process.

To improve the reliability and safety of GTU have been developed GTU risk models (Fig. 1) [6].

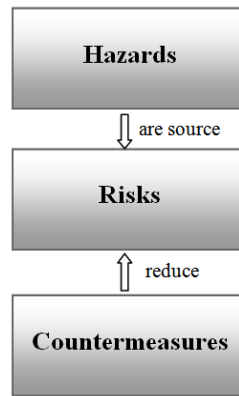


Fig. 1 GTU risk macro-model

Risk models of test process (2) [7], the method of uncertainty control in resource-limited settings [8] and the method for choice of alternative countermeasures [9] were also obtained.

$$M_R = H(D, A), \quad (2)$$

where  $M_R$  – hypercube of data from a lots of room, which correspond to the set measurements of  $D$  and a set of attributes  $A$ ;

$D = \{d_1, d_2, \dots, d_i\}$  – hypercube measurements, which correspond to the components of risk models (e.g. for GTU  $d_1$  - "Hazards",  $d_2$  - "Risks",  $d_3$  - "Countermeasures");

$A = A_{d1} \cup A_{d2} \cup \dots \cup A_{di}$ ;

$A_{di} = \{a_1^i, a_2^i, \dots, a_g^i\}$ , where  $i = 1, 2, \dots$

These earlier models and methods will be used as a basis to develop DSS of GTU test process.

The work purpose is to design the DSS of GTU tests process, which is aimed collect and register the tests results information. It include carry out the effective storage and regulated access rights to it, and provides the analysis, in particular the risk assessment for GTU and for its test process. This requires the development of DSS architecture, storage structure, subsystem for information administration, subsystems for data visualization, monitoring and analysis

### III. SOFTWARE ARCHITECTURE

DSS architecture was designed with the following requirements:

- loading of the GTU tests data by a network directly from ACS;
- support of the centralized data archive with hierarchical storage system;
- storage of the tests results during all life cycle of GTU;
- analyzing of the possible extraordinary situations;
- differentially access to the data for various users groups (on GTU, on test, kinds of works);
- the information visualization about the GTU parameters;
- the cartographical representation of operating GTU.

As seen in Fig. 2, the DSS architecture developed using the structure "client - server", since the system must support multiplayer mode, as well, because the test objects are widely distributed geographically.

Data from ACS to database server is communicated by a protected radio channel, so for remote access to all geographically distributed GTU, the amount of which can reach up to several hundreds.

Fig. 3 is a DSS diagram of the components for the GTU test process.

The data storage system include:

- of DB for storage the necessary information from the server;
- of file structure which uses a uniform data storage format of language marking XML from the client.

At loading of data from other formats are used converters.

Also it was designed the database structure, and to provide a cross-platform database interaction - corresponding WCF (Windows Communication Foundation) services.

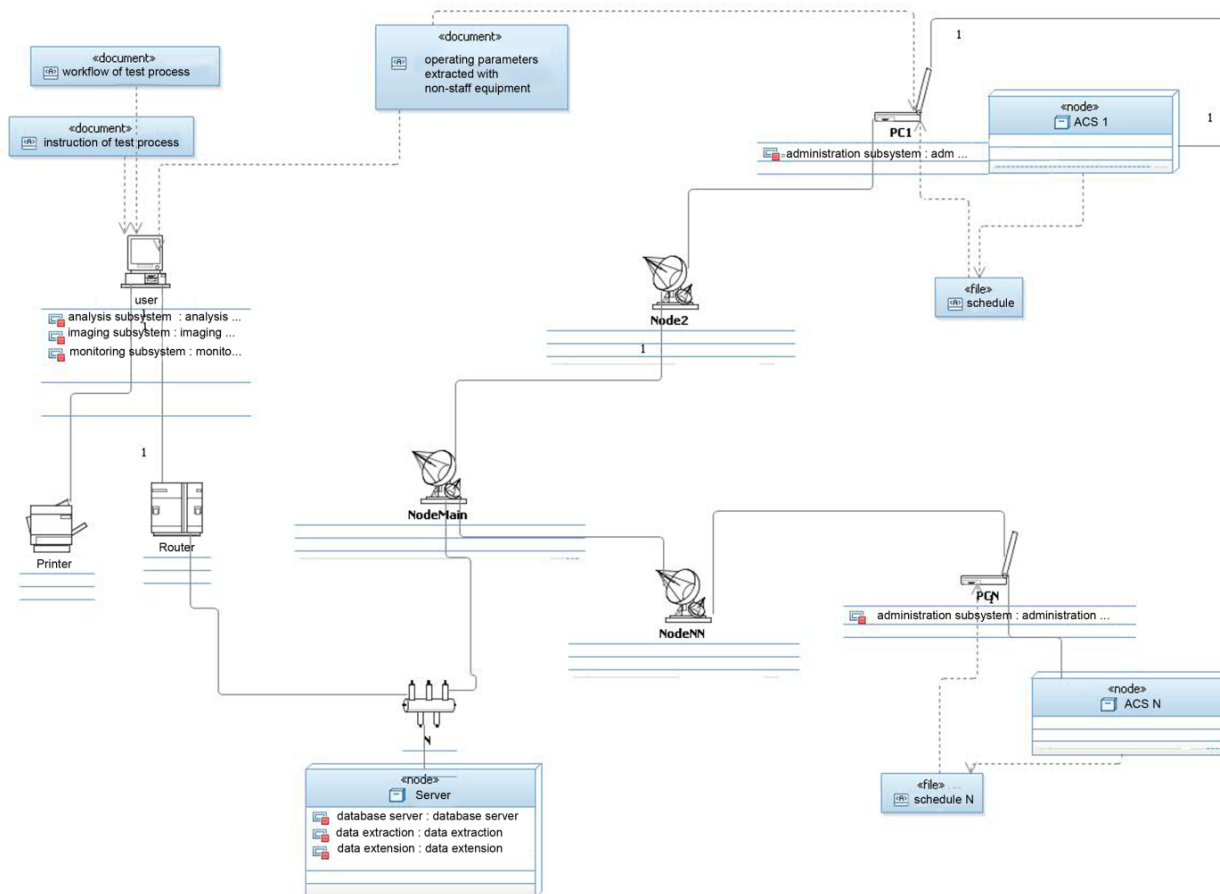


Fig. 2 DSS architecture for GTU test process

#### IV. DEVELOPMENT OF SUBSYSTEMS DSS

##### A. Subsystem for information administration

Was developed subsystem for information administration that allows:

- create a new file or open an existing project for each GTU, which be stored throughout the life cycle;
- differentially access to the data for various users groups;
- to export / import data in other applications;
- provide access to all other subsystems: monitoring, data visualization, analysis, etc .;
- generate reports on user requirements with the possibility of printing and other.

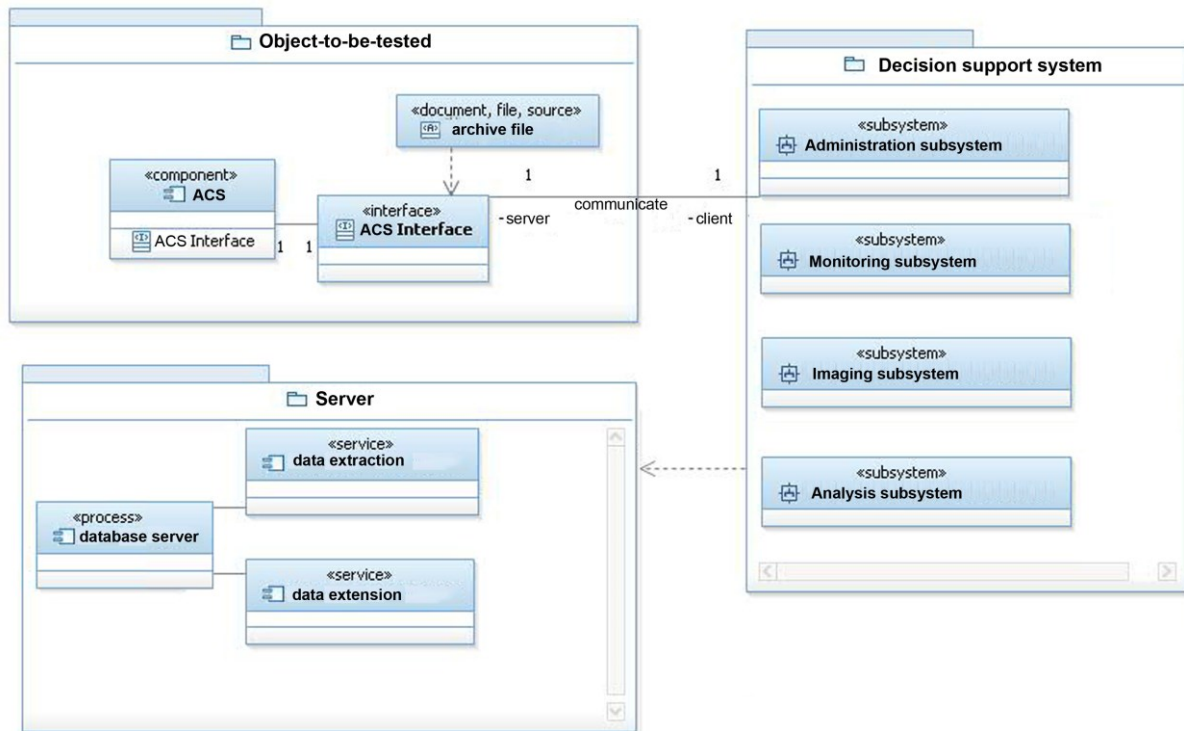


Fig. 3 DSS diagram of the components for the GTU test process

##### B. Monitoring subsystem

The basic idea is information integration of GTU tests process and ACS. Information integration can be carried out provided that the monitoring subsystems will operate not with traditional documents, and with the formalized documentary information models describing GTU test process.

The monitoring subsystem will boot ACS archive files containing the parameters and status of GTU during operation, including during test. It uses a specially designed converters in storage format XML.

The monitoring subsystem also provides data entry contingency equipment to download data collected manually. Loading is done with an access point PC connected to ACS, or user PC with a specially organized forms of requests.

##### C. Subsystem for data visualization

Based on the developed GTU test process model, as well as the electronic documents structure [4], it was designed subsystem for data visualization. This subsystem provides the correspondence between the ongoing processes and their visual presentation, focused the attention operator on the essential processes and the states with the possibility of up to the required level of detail.

Developed subsystem for data visualization allows visualizing information about the GTU parameters, as well as cartographic representation of all units in operation. Interface latter is shown in Fig. 4.

#### D. Analysis subsystem

In the analysis of reliability and safety of technical systems important place takes a risk-based methodology. Therefore, the developed analysis subsystem must ensure the risk assessment GTU test process, based on previously developed GTU risk models and test process [6-7], as well as the method of managing uncertainty in resource-limited settings [8] and the method for choice of alternative countermeasures [9].

Analysis subsystem performs the following functions:

- enables the identification of the research object, the assets representing the value of this object, as well as the dangers and risks that threaten;
- for the correct risk identification need to analyze the possible contingencies, so the analysis subsystem allows organization of user criteria samples for comparative analysis of parameters and operating conditions, the search for identity denials, organizes summary reports;
- not all risks are the same threat, so should be highlighted the most critical of them, so a subsystem allows a qualitative risk ranking by a group of experts;
- checked the consistency of expert opinion;
- for the most critical risks allow the identification of countermeasures if the alternatives has a mechanism of choice;
- produces a final report containing recommendations to the list of the most critical risks and the possible countermeasures that can prevent or minimize these risks.

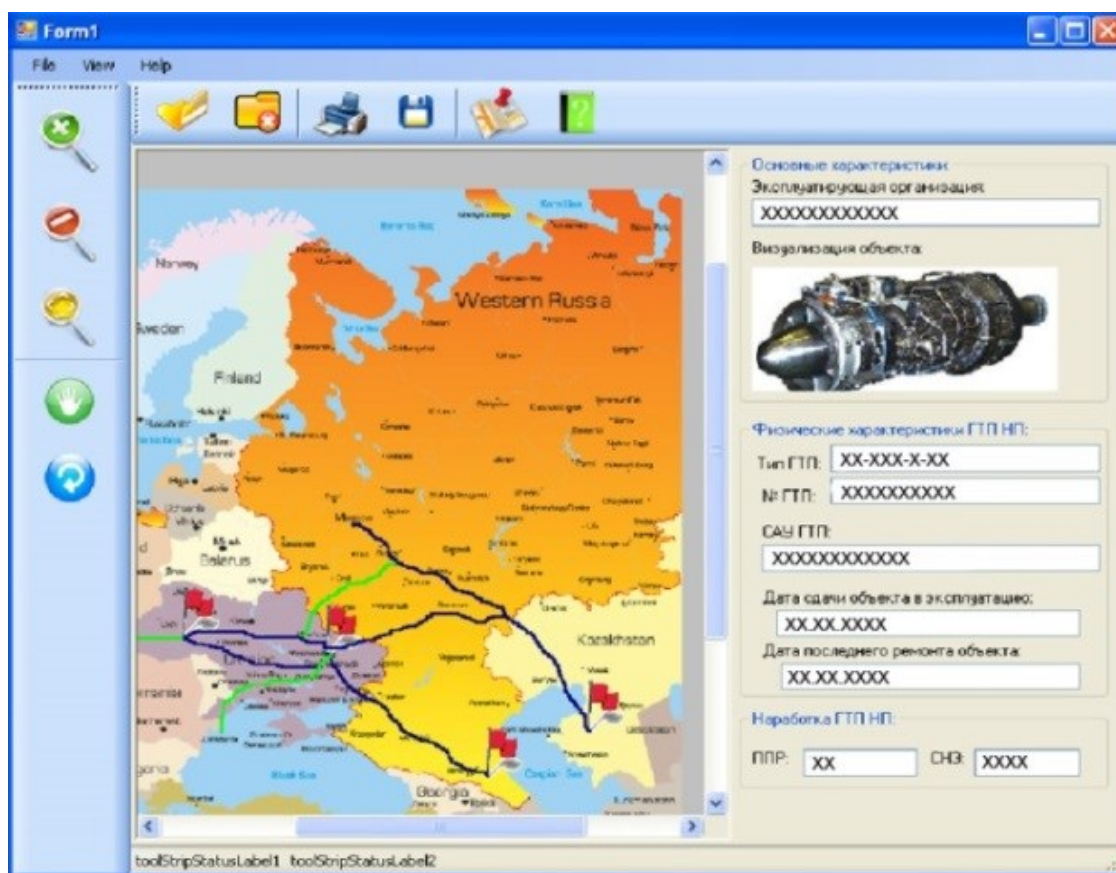


Fig. 4 Interface of Subsystem for data visualization

## V. CONCLUSION

The DSS for GTU test process were developed by the authors. This allows to improve the manage process of test, by increasing accuracy of data collection, as well as by automate the processes of data storage and analyses.

The main requirements for DSS were designed, developed DSS architecture, the data storage structure, subsystems for information administration, monitoring, data visualization and analysis.

The developed DSS for GTU test process are passing  $\beta$ -testing at Motor Sich JSC, Ukraine. In the experiment data collections involved five power plants and one gas pumping stations.

The implementation of this DSS allows to improve the quality of controlled information, and to reduce the risks of incorrect information related to the human factor. Analysis subsystem of DSS allows to assess risks for the timely implementation of effective measures to improve the consumer properties of GTU and increase its reliability and safety. At the same time, the storage of this information throughout the GTU lifecycle ensure continuity of experience and formed a training base for young specialists of the enterprise.

The prospects for further development of this work are to optimize the process of collection and transmission of information from geographically dispersed GTU with regard to limitations on available resources.

## ACKNOWLEDGMENT

We would like to thank and acknowledge Valeriy Morozov, Volodymyr Mytin and Viktor Raznoshinsky at the Power Generating Units and Gas Pumping Stations Department of Motor Sich JSC, Ukraine for their assistance and help.

## REFERENCES

- [1] Zelenkov Ju.A. Kompleksnaja avtomatizacija ispytanij gazoturbinyh dvigatelej. Chast' 1: Upravlenie stendom i sbor dannyh // Vestnik UGATU. Serija Upravlenie, vychislitel'naja tehnika i informatika Publ., 2011, part 15, vol. 2(42), pp. 119-125. (In Russian).
- [2] Zelenkov Ju.A. Kompleksnaja avtomatizacija ispytanij gazoturbinyh dvigatelej. Chast' 2: Hranenie i obrabotka dannyh // Vestnik UGATU. Serija Upravlenie, vychislitel'naja tehnika i informatika Publ., 2011, part 15, vol. 2(42), pp. 126-131. (In Russian).
- [3] Pogorelov G.I. Informacionnaja podderzhka upravlenija zhiznennym ciklom ispytanij GTD na osnove CALS-tehnologij // Vestnik UGATU. Serija Upravlenie, vychislitel'naja tehnika i informatika Publ., 2007, part 9, vol. 4(22), pp. 57-63. (In Russian).
- [4] Shitikova E.V., Tabunshchik G.V. Information model of test process of gas turbine unit for terrestrial usage. // Radioelektronika, informatyka, upravlinnja Publ., 2013, vol. 1(28), pp. 101-108. (In Russian).
- [5] Shitikova E.V., Tabunshchik G.V. GTU tests process model. // Informatyka. Kul'tura. Tehnika : III ukrai'ns'ko-nimec'ka konferencii', Odesa, 23 -24 kvitnja 2015 r. : tezy dopovidej, ONPU Publ., 2015, pp. 53-54. (In Russian).
- [6] Shitikova E.V., Tabunshchik G.V. Risk analysis of gas turbines units for terrestrial usage. // Visnyk dvygunobuduvannja Publ., 2012, vol. 1, pp. 54-59. (In Russian).
- [7] Shitikova E.V., Tabunshchik G.V. Risk model of the gas turbine units for terrestrial usage testing process // Naukovi praci Donec'kogo nacional'nogo tehnic'nogo universytetu, serija: "Problemy modeljuvannja ta avtomatyzacii' proektuvannja" Publ., 2012, vol. 1(10)-2(11), pp.79-89. (In Russian).
- [8] Shitikova E.V., Tabunshchik G.V. Method of managing uncertainty in resource-limited settings // Radioelektronika, informatyka, upravlinnja Publ., 2015, vol. 2 (33), pp. 87-95. (In Russian).
- [9] Shitikova E.V., Tabunshchik G.V. The method for choice of alternative countermeasures in risk analysis of complex technical systems // Systemnyj analiz. Informatyka. Upravlinnja (SAIU-2013) : IV Mizhnar. nauk.-prakt. konf., 13–16 bereznja 2013 r., m. Zaporizhzhja : materialy konf, KPU Publ., 2013, pp. 282-283. (In Russian).



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Service for Vulnerabilities Analysis and Security Assessment of Open Source Systems

Anastasiia Strielkina, Artem Tetskyi, Bogdan Selin, Oleksandr Solovyov, Dmitriy Uzun

**Abstract**—The article describes the service for vulnerabilities analysis and security assessment of open source information systems. The article deals with determination of information security threats and vulnerabilities and actuality of protection sensitive information. The service that will help to improvement information security from cyber-attacks has been considered. The most valuable components of web-application penetration testing are shown. Also it is spoken in details about SQL-injections, brute-force, potential damage assessment model and device for anonymization. Operation principles of this device are described. The article can be used by information security specialists, network administrators and owners of web-resources and wireless networks.

**Keywords**— anonymization, brute force, open source, penetration testing, damage assessment model.

## I. INTRODUCTION

For today informatization is one of the priority development directions of all economic sectors. Almost each organization, commercial or governmental, has its own website and introduces all kinds of online services. Personal information of customers and employees, financial information and economic activity data are stored in electronic form. Therefore the task of ensuring security of web applications is becoming more important every year. Unfortunately, the developers of corporative information systems do not always follow the safety requirements [1] - due to lack of the necessary experience, or simply focusing on the development of other system purposes. The annual researches indicate that a large number of web applications contain high risk vulnerabilities, which may cause financial or reputational damage [2]. Exactly an attack against web applications is often the first step on the large companies' networks hacking and a publication of owner discrediting information on the official web sites serves as a weapon in the information war. All these suggest that the international community is in dire need of qualified information security specialists at the present time.

## II. TOPICALITY AND PROBLEMS

The rapid growth and widespread use of electronic data processing and electronic business conducted through the Internet, along with numerous occurrences of international terrorism, fueled the need for better methods of protecting the computers and the information they store, process and transmit. And for today there is a rapid growth in the number of Internet resources, practically each organization has its own website. The functionality of sites varies, ranging from business card sites and ending with online stores, auctions, etc. Due to various content management systems, website creation does not require knowledge of programming languages and, especially, the knowledge of information security aspects. Therefore, the following problems occur:

- 1) Resource owners do not care about the security of their sites. Most of the owners begin to

A. Strielkina, National Aerospace University, Kharkov, Ukraine (e-mail: strelkina.anastasiya@hotmail.com).  
 A. Tetskyi, National Aerospace University, Kharkov, Ukraine (e-mail: artem.tetskiy@csn.khai.edu)  
 B. Selin, National Aerospace University, Kharkov, Ukraine (e-mail: selin.bv@gmail.com).  
 O. Solovyov, National Aerospace University, Kharkov, Ukraine (e-mail: extsand@gmail.com).  
 D. Uzun, National Aerospace University, Kharkov, Ukraine (e-mail dmitriy.d.uzun@gmail.com).



pay attention to security after attack on their site. If a site is, for example, an online store, the consequences of such an attack may carry substantial material damage. Other good example is PlayStation user data loss due to PlayStation Network service disabling in 2011. This hack has brought the loss to the company and caused the damage to users [1].

- 2) Not all developers have a sufficient knowledge level in the information security field. Topical example is the development of various plugins and modules for open source content management systems. Before the community or users will find a vulnerability in a component, that component will be already installed on some web-sites. Perhaps vulnerability will be closed in this component future releases, but there is a problem again - the resources owners do not follow the used software upgrading in a timely manner.

One of the main threats is cyber-attack on the target system.

Targeted cyber-attacks on resources often lead to the following consequences:

- the loss of important information can harm the competitiveness in the market (industrial and commercial espionage);
- the vulnerabilities presence in the system predetermines the potential financial loss (a denial of service, loss of profit, commercial espionage, fraud);
- loss of productivity.

### III. Condition of the Problem and Existing Solutions

Thus, to prevent emergence/reduce risk of the aforementioned consequences is necessary to organize comprehensive protection against cyber-attacks.

There is a special testing to detect existing vulnerabilities in the studied system - penetration testing. This testing is a simulation of malicious attacks.

Conducting the penetration testing procedure is necessary to provide to user information about the real state of information security in the system.

Penetration tests are usually performed using manual or automated technologies to systematically compromise servers, endpoints, web applications, wireless networks, network devices, mobile devices and other potential points of exposure.

Providing of services for conducting penetration testing is not new. There are tools, methodologies and standards for penetration testing. Most of the necessary tools were assembled in the operating system Kali Linux [3], [4]. Kali Linux is a Debian Linux based Penetration Testing arsenal used by security professionals (and others) to perform security assessments. Kali offers a range of toolsets customized for identifying and exploiting vulnerabilities in systems.

It is also developed and actively applied testing methodology described in the OWASP Testing Guide. And Penetration Testing Execution Standard describes all the stages of testing.

Most of web applications testing methodologies are based on OWASP (Fig. 1) [5], [6].

Thereby the procedure for penetration testing should follow the steps described below [7].

- 1) Research information about the target system. Computers that can be accessed over the internet must have an official IP address. Freely accessible databases provide information about the IP address blocks assigned to an organization.
- 2) Scan target systems for services on offer. An attempt is made to conduct a port scan of the computer(s) being tested, open ports being indicative of the applications assigned to them.
- 3) Identify systems and applications. The names and version of operating systems and applications in the target systems can be identified by "fingerprinting".
- 4) Researching vulnerabilities. Information about vulnerabilities of specific operating systems and applications can be researched efficiently using the information gathered.
- 5) Exploiting vulnerabilities. Detected vulnerabilities can be used to obtain unauthorized

access to the system or to prepare further attacks.

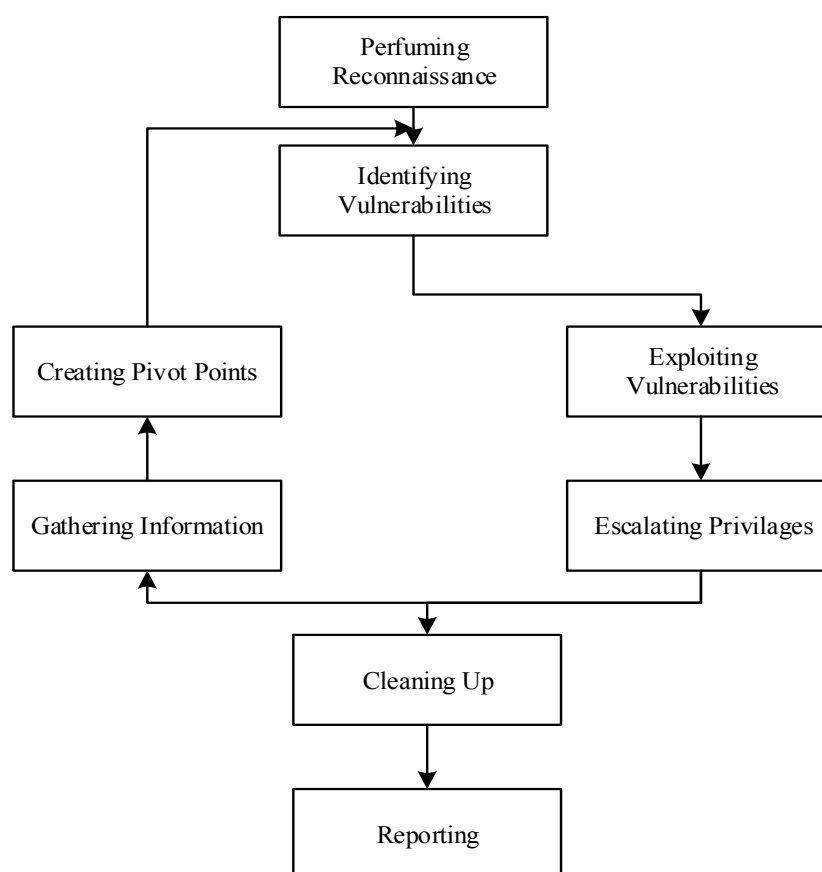


Fig. 1. OWASP testing methodology

Client goals that can be attained by penetration testing can be divided into four categories:

- 1) Improving security of technical systems.
- 2) Identifying vulnerabilities.
- 3) Having IT security confirmed by an external third party.
- 4) Improving security of organizational and personnel infrastructure.

The result of an IT penetration test should therefore be more than just a list of existing vulnerabilities; ideally it should also suggest specific solutions for their elimination.

So penetration verification is a reliable method for information security, which confirms the breaking possibility not only theoretically, but also empirically. Penetration testing can provide valuable insight into company's risk exposure, including resistance to real world-style attacks, level of sophistication required to compromise systems, countermeasures that mitigate risk, attack detection and incident response etc.

There are companies that provide penetration testing services in Ukraine for today:

- “Active Audit Agency”;
- “FireEye”;
- “AMI”;
- “Berezha Security”.

The main drawback of these companies is the price for provided services.

So it is necessary to have a service that provides penetration testers with an acceptable price / performance ratio.



#### IV. OBTAINED RESULTS

Service for vulnerabilities analysis and security assessment of open source systems is a comprehensive solution in the field of cyber security that will help to eliminate security breaches (identify and prevent cyber-attacks at every stage of invasion), protect valuable information and ensure data confidentiality with the help of technology, intelligence, knowledge and skills of the project team.

The service is based on open source software and does not contain proprietary software.

The implementation of actions is carried out both from a position of an insider and external potential attacker and involves active using of security vulnerabilities.

##### A. Web-Vulnerabilities

According [8] to describe web-application penetration testing can be used Markov processes.

In general, the description of the web-application penetration testing as graph is shown on Fig. 2

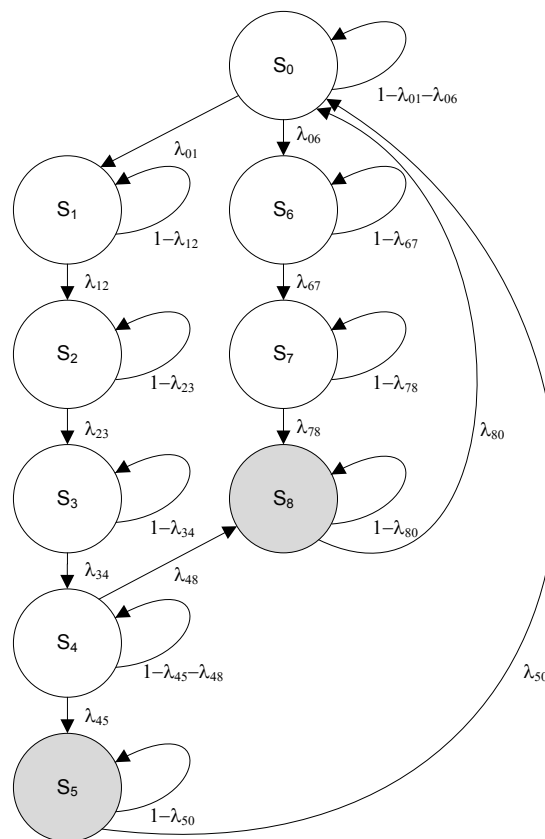


Fig. 1 State graph for web-applications testing

Meanings of the states graph that is shown on Figure 1 are:

- S<sub>0</sub> – initial state without any concrete knowledge;
- S<sub>1</sub> – content management system and its version are known;
- S<sub>2</sub> – all vulnerabilities identifiers are known;
- S<sub>3</sub> – critical vulnerabilities identifiers are known;
- S<sub>4</sub> – exploits for critical vulnerabilities are exist;
- S<sub>5</sub> – vulnerability exploitation (excluding SQL-injection);
- S<sub>6</sub> – paths to data processing scripts are known;
- S<sub>7</sub> – the scripts without input data sanitization are found;
- S<sub>8</sub> – SQL-injection exploitation.

Table I shows the transition probabilities for the graph shown on figure 1.

TABLE I  
THE TRANSITION PROBABILITIES FOR THE GRAPH

<i>from</i>	<i>to</i>	$S_0$	$S_1$	$S_2$	$S_3$	$S_4$	$S_5$	$S_6$	$S_7$	$S_8$
$S_0$		$1-\lambda_{01}-\lambda_{06}$	$\lambda_{01}$					$\lambda_{06}$		
$S_1$			$1-\lambda_{12}$	$\lambda_{12}$						
$S_2$				$1-\lambda_{23}$	$\lambda_{23}$					
$S_3$					$1-\lambda_{34}$	$\lambda_{34}$				
$S_4$						$1-\lambda_{45}-\lambda_{48}$	$\lambda_{45}$			$\lambda_{48}$
$S_5$		$\lambda_{50}$					$1-\lambda_{50}$			
$S_6$								$1-\lambda_{67}$	$\lambda_{67}$	
$S_7$									$1-\lambda_{78}$	$\lambda_{78}$
$S_8$		$\lambda_{80}$								$1-\lambda_{80}$

In Table I  $\lambda_{ij}$  is a probability of transition from state  $S_i$  to state  $S_j$ .

For example, it is possible to go into state  $S_8$  by two ways:

- 1) By finding vulnerable page with CVE-ID from the state  $S_4$ ;
- 2) By testing all pages from the state  $S_7$ .

This graph may be expanded with new vulnerabilities by adding new columns for appropriate vulnerabilities.

Also during the research the methods of getting the database contents using SQL-injections have been identified experimentally. Fig. 3 shows these methods.

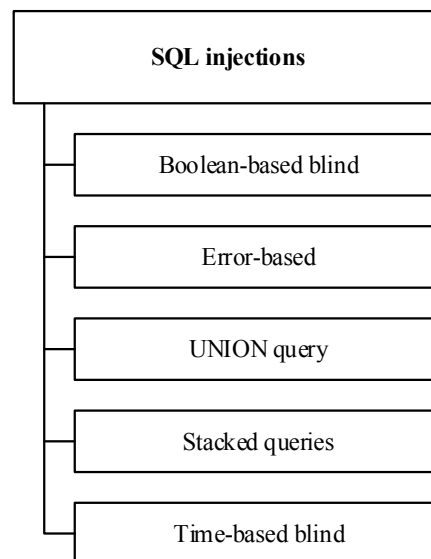


Fig. 3. Varieties of SQL-injections

- 1) Boolean-based blind SQL injection.

The method is based on the selection of the code table symbol number by adding structures using operators AND/OR. A feature of this method is that the information obtained from the database is not displayed. Binary search algorithm is used, so to get a single character an average of 7 database queries should be executed. For the Unicode characters it need more requests because code range becomes wider. Therefore, due to low speed and high load on the server, this method is not suitable for obtaining large amounts of data.

- 2) Error-based SQL injection.

This method is based on the fact that the contents of the cell is included in the error text

generated by the database server as a result of an incorrect request. With a single query it is possible to get the contents of a single cell. This method is possible only when the script displays an error returned by the database server. Using MySQL database in PHP the following construction exists:

```
mysql_query("...") or die( mysql_error() );
```

It is important to prevent error messages displaying in the browser.

### 3) UNION query SQL injection.

This method is classic and the easiest for understanding injection of SQL-code. The principle is to combine two SELECT queries with the operator UNION. The peculiarity is that the number of columns in queries should be the same. If the result of the query is processed in a cycle, then one query can get contents of several columns in the table (depending on the number of columns in the first query).

### 4) Stacked queries SQL injection.

The principle of this method is using multiple queries to the database, separated by semicolons. It is the most dangerous injection type because requests may occur to insert/update/delete records. Therefore, using of stacked queries in PHP + MySQL is prohibited for safety reasons.

### 5) Time-based blind SQL injection.

This method is a modification of the first method, because similar principle to obtain information is used. In the «Boolean-based blind» method a sign of fulfillment of the condition is the withdrawal of the correct result, in this method such feature is the implementation of a time delay in a database query..

By comparing the time of query with the correct parameter with the time of a query with a parameter that contains the SQL-code, it can be concluded that the condition was performed in the modified SQL-query. Delay is produced using the SLEEP(). The presence of these delays is the additional drawback of this method.

As a result of research the methods of getting contents of the database were identified, these methods are confirmed by the sources [9]. The SQL-injection should not only compromise the database, but also gain full access to the server (at the confluence of certain circumstances). To prevent this kind of attack it is strongly recommended to avoid the simple concatenation of parameters in queries, instead using the appropriate drivers (for example, PDO in PHP to work with MySQL).

During the research penetration testing has been conducted of such large web-sites:

- 1) Internet-shop "Tehnomaster".
- 2) Internet-shop "Mobilluck".
- 3) The conference "Dessert" site.
- 4) National Aerospace University "KhAI" site.
- 5) Computer systems and networks department (NAU "KhAI") site.

A vulnerability was detected on one of the sites via which was held SQL-injection, which led to the the database compromising, that stores information about employees. Also in the database stored the password in an unencrypted form, so it was accessed to the admin panel.

### *B. Hardware and Software System for Anonymization*

One of the main components of penetration tester secure work is the anonymity of his activities. For this purpose it is necessary to have networking device-anonymizer that allows access to the Internet regardless of reception and transmission device settings and furthermore provides anonymous access to the Internet. This device may be useful not only to penetration tester, but also for secure Internet surfing, safe data transferring and to bypass censorship in places where Internet access is limited.

This device is based on microcomputer, additional standardized equipment (i.e. case for microcomputer, microSD memory card, Wi-Fi adapter, power source, Ethernet cable), open source operating system and a set of additional services (i.e. DNS, Gateway, Router, DHCP client, ISC DHCP server, Hostapd, service Tor and etc.).

After configuring the microcomputer, additional equipment and the operating system components (services) can be obtained anonymous access to the Internet.

This device encrypts all web traffic, not just the traffic that arrives from one program. That is, there is no need to download additional software or set up accounts, register. It is only necessary to connect the device to the Ethernet interface and user can connect to an anonymous Wi-Fi network.

Anonymizer operating principle can be divided into levels (Fig. 4). The first level is the network gateway, which is needed to convert the protocols among different network model levels. This layer functions as router - receives a signal from the network cable, converts into the protocol that is required and sends it to the recipient address. Also there is the Domain Name System (DNS) on the first level which is responsible for converting network nodes domain names into IP addresses. The second and third levels are Dynamic Host Configuration Protocol (DHCP), responsible for IP addresses distribution among clients, which are connected to the access point. The fourth level is a service (daemon) Hostapd, which is designed for server access points, security configuration and authentication that is for conversion the microcomputer to the access point. Thus, at this stage, devices can connect to the configured access point. The fourth level is the service Tor that is responsible for traffic anonymization that passes through the access point by using a technology "onion routing" that encrypts the data and sends a lot of time through a large number of servers, each of which removes one layer of encryption. The sixth level is responsible for connecting end users to the access point and further access to the Internet through the configured access point on the microcomputer.

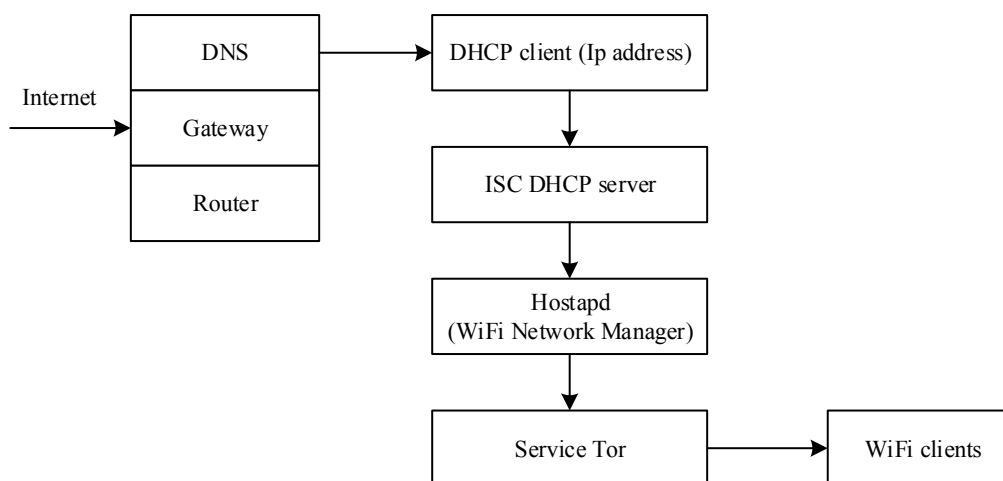


Fig. 4 Anonymizer operating principle

During the implementation of access to the Internet resources could leak user personal data (location, other personal information), and the using of the above described device allows to hide all the data.

### C. Distributed Brute Force

For today one of the most common vulnerabilities is a possibility of making automated selection of the credentials and passwords (Brute Force Attack). A disadvantage that allows exploiting this vulnerability exposed almost 70% of the systems that ranks the second place in

the ranking of vulnerabilities after the "Cross-Site Scripting» (Cross-site Scripting, XSS) by rating of Positive Technologies [10].

An exhaustive search or so-called "brute force" is so common among hackers because it is suitable for attacks on almost any system because it can be used against practically any existing encryption method. More often, this method is used in cases when other weaknesses cannot be found in the system because despite the fact that brute force gives an absolute guarantee to solve the problem, it requires enormous time and (or) computing resources. This is due to the fact that the complexity of exhaustive search depends on the number of all possible solutions and may require exponential time operation. For example, using the alphabet of Latin letters and numbers of one register and speed of selection of 100 000 passwords per second on selection passphrase of 6 characters 6 hours, and the phrase from 8 characters - 11 months are required. Here are examples those shows the time required for password guessing using a processor i7 3840QM 4x3.8 GHz ~ 4700 pmk/s:

- 8-digit password:  $\frac{10^8}{4700 \cdot 3600} = 5.91$  hours;
- 8-character password:  $\frac{26^8}{4700 \cdot 3600 \cdot 24} = 514$  days;
- 10-digit password:  $\frac{10^{10}}{4700 \cdot 3600 \cdot 24} = 24.6$  days;
- 10-character password:  $\frac{26^{10}}{4700 \cdot 3600 \cdot 365 \cdot 24} = 952$  years.

The calculation was performed according to the formula:

$$T = \frac{d^c}{p \cdot t}, \quad (1)$$

where  $T$  – the result, time required for guessing a password;  $d$  – number of characters in the dictionary;  $c$  – number of characters in the password;  $p$  – speed of computing which gives the device;  $t$  – the number of seconds in an hour, hours in the day and so on (for a more convenient time display).

These numbers can instill that this method is ineffective, especially in order to ensure the best protection can be used alphabet of more characters (for example, Latin letters of both registers, numbers and special characters). However, human nature is such that the users (even privileged) too lazy to remember a complex password consisting of more than 8 characters, especially when it is not meaningful phrase and a random set of letters and numbers. According to the analysis conducted by Positive Technologies [11], about 52% of users of Russian companies is used as a password only numbers, only 18% use the Latin lowercase letters and the same percent of people use Latin letters in lowercase + numbers. As for the length of a passphrase, only about 15% of users have more than 8 characters, 25% of use 8 characters long, and about 50% of the users' passphrase have 6-7 characters or less. This statistic is not much more different for network or web resources for administrators. In addition, frequently used passwords are recording in the dictionaries, attacks which are much more effective.

Another argument convincing of the danger of brute force attacks is increasing at huge rates computing power and bandwidth of the network connection. One of the methods to increase the efficiency of the exhaustive search is solutions becoming more common based on the computing power of the GPU (GPGPU). The newest graphics processor allows increasing the speed of password guessing tenfold. Using (1) and speed of password selection on GPGPU AMD 7990 ~ 220 kh/s (kh = 1000 pmk):

- 8-digit password:  $\frac{10^8}{220000 \cdot 60} = 7.5$  minutes;
- 8-character password:  $\frac{26^8}{220000 \cdot 3600} = 10.9$  days;

- 10-digit password:  $\frac{10^{10}}{220000 \cdot 3600} = 12.6$  hours;
- 10-character password:  $\frac{26^8}{220000 \cdot 3600 \cdot 24 \cdot 365} = 20.3$  years.

The difference is obvious, but even the use of this method requires a huge time-consuming, if the passphrase is quite complicated. Therefore, to improve the efficiency of brute force attacks often use distributed computing: in this case the set of keys splits into subsets that are processed parallel on multiple machines. Attackers can create a cluster from the n-th number of GPGPU and (or) CPU and perform calculations on it. Fig. 5 shows this graphically.

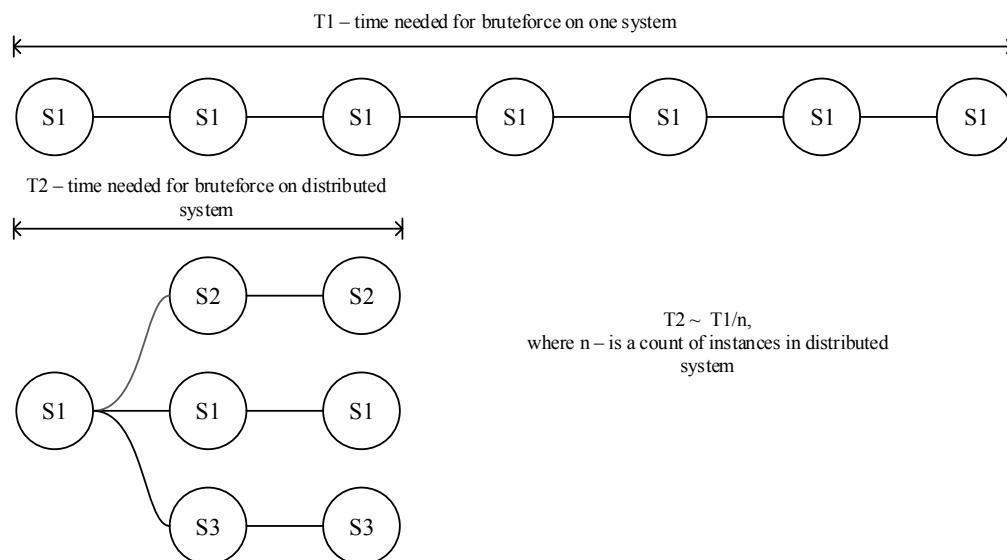


Fig. 5 Simple graphical view of brute force time decreasing with distributed system using

In this case, time required to compute the passphrase will decrease, respectively, as many times as many instances is used in a distributed system (on the condition that they have the same computing power).

Using (1) approximate time of the key calculation using the cluster of 40 AMD 7990 ~ 10000 kh/s is:

- 8-digit password:  $\frac{10^8}{10000000 \cdot 60} = 10$  seconds;
- 8-character password:  $\frac{26^8}{10000000 \cdot 3600} = 5.8$  hours;
- 10-digit password:  $\frac{10^{10}}{10000000 \cdot 3600} = 16$  minutes;
- 10-character password:  $\frac{26^{10}}{10000000 \cdot 3600 \cdot 24 \cdot 365} = 163$  days.

According to this principle hackers often create a botnets, a network of infected unsuspecting PCs users, by means of which produces a distributed brute force attack on web resources. This type of attack is most advantageous for the attacker, both in terms of efficiency, as the botnet can total tens of thousands of infected machines that give a huge amount of computing power, and in terms of financial costs - no need to establish and maintain computing cluster.

Similarly lately, after the spread of cloud computing a rent of needed for brute force attacks computing resources become more popular. Since the essence of cloud easily scalable them quite easy to use and profitable - 200 entities Amazon EC2 ~ 10000 kh/s and using (1):

- 8-digit password:  $\frac{10^8}{10000000 \cdot 60} = 10$  seconds;
- 8-character password:  $\frac{26^8}{10000000 \cdot 3600} = 5.8$  hours;

– 10-digit password:  $\frac{10^{10}}{10000000 \cdot 3600} = 16$  minutes;

– 10-character password:  $\frac{26^{10}}{10000000 \cdot 3600 \cdot 24 \cdot 365} = 163$  days.

There is even online service WPACracker, based on 400 entities Amazon EC2, which provides services to hack WPA-key by a downloaded "file handshake".

All of these attacks can be applied by hackers to penetrate into the corporate wireless networks to steal confidential information or for unauthorized access to other password-protected infrastructures.

Described service provides services of testing to penetrate the enterprise wireless network in order to verify a cryptographic strength of passphrase, professionalism and integrity of the enterprise information network administrator, fulfillment security standards of employees, etc.

The software that used for testing is the free open source software, such as pyrit, which specializes in automated brute force attacks on wireless networks and allows integrating into a cluster any number of computers of any power, whether they support or not GPGPU technology (from this depends only on increase in power) through the Internet. This allows saving money by combining together the computing resources available to the team of developers and not having to buy additional equipment. As one more software means serves also open source software hascat, which is based solely on the use of distributed computing on clusters of GPGPU. This software is used to recover passwords of the intercepted hash functions and can be used in the expansion of service and acquisition of additional processing power, as it is well amenable to parallelization on a large number of clusters of GPGPU.

#### D. Potential Damage Assessment Model

It is possible to use the methods of game theory to assess the potential damage [12].

The basic concepts of game theory are:

- the game is a mathematical model of conflict;
- the parties involved in the conflict are called players;
- the move is the player choice one of the courses of action;
- the player strategy is a set of rules that determines the behavior of the player at the move;
- the goal of game theory is to develop methods to determine the optimal strategy for each player [13].

In the present study are considered two players - the attacker and the security administrator. The relationship between these players is determined as the payoff matrix (Table II). In drawing up the matrix of the game can be assumed that the attacker aims to inflict the greatest possible

TABLE II  
TABLE OF THE MATRIX GAME

	$y_1$	$y_2$	$y_3$	...	$y_m$
$x_1$	$a_{11}$	$a_{12}$	$a_{13}$	...	$a_{1m}$
$x_2$	$a_{21}$	$a_{22}$	$a_{23}$	...	$a_{2m}$
$x_3$	$a_{31}$	$a_{32}$	$a_{33}$	...	$a_{3m}$
...	...	...	...	...	...
$x_n$	$a_{n1}$	$a_{n2}$	$a_{n3}$	...	$a_{nm}$

damage to the attacked computer system. The purpose of the security administrator in the matrix game is to allow the attacker to inflict the least damage at the lowest cost of the means of protection.

As the attacker strategy we mean string  $x_i (i = \overline{1, n})$  of the matrix, and as the security administrator strategy – columns  $y_j (j = \overline{1, m})$ . To the attacker strategy can be attributed various types of attacks on the computer

system, and to the strategy administrator – protection means of computer information. The parameter  $a_{ij}$  is a result of the game. As this parameter can be regarded, for example, the annual loss to all combinations variants of  $x_i (i = \overline{1, n})$  and  $y_j (j = \overline{1, m})$ . For this is necessary to compare each attack with each protection method and to determine the damage that can be suffered in this case. Purchase, installation and use of protective equipment can require additional costs, which should also contribute to the damage in the calculations.

Assume also are known the following parameters:  $n$  – the number of attacks types;  $m$  – the number of protection means;  $S_j (j = \overline{0, m})$  – the cost of protection means;  $D$  – the value of alleged damage;  $p_{ij}^{(p)} (i = \overline{1, n}, j = \overline{1, m})$  – the probability of attack  $x_i$  reflection using protection means  $y_j$ , i.e. probability of protection against attack;  $p_{ij}^{(a)}$  – the probability of the attack  $x_i$ ;  $p_{ij}^{(d)}$  – the probability of damage during an attack  $x_i$  using protection means  $y_j$ .

The condition for the effective protection is a rule: the cost of protection means should be less than the cost of the losses incurred in the successful implementation of attack.

Since the probability of using at least one protection means equals 1, draw up the following inequality:

$$1 \times S_j \leq p_{ij}^{(d)} \times D, \quad i = \overline{1, n}, j = \overline{1, m}. \quad (2)$$

Represent the probability of damage  $p_{ij}^{(d)}$  as probability of protection  $p_{ij}^{(p)}$  and the probability of the attack  $p_{ij}^{(a)}$ :

$$p_{ij}^{(d)} = (1 - p_{ij}^{(p)}) \times p_{ij}^{(a)}, \quad i = \overline{1, n}, j = \overline{1, m}. \quad (3)$$

Substituting (3) into (2), we obtain:

$$S_j \leq (1 - p_{ij}^{(p)}) \times p_{ij}^{(a)} \times D, \quad i = \overline{1, n}, j = \overline{1, m}. \quad (4)$$

Dividing both sides of (4) to the expression in the right-hand side of this inequality, we obtain:

$$\frac{S_j}{(1 - p_{ij}^{(p)}) \times p_{ij}^{(a)} \times D} \leq 1, \quad i = \overline{1, n}, j = \overline{1, m}. \quad (5)$$

Denote in (5) left-hand side as  $\lambda_{ij}$  and call as a coefficient of effective protection:

$$\lambda_{ij} = \frac{S_j}{(1 - p_{ij}^{(p)}) \times p_{ij}^{(a)} \times D}, \quad i = \overline{1, n}, j = \overline{1, m}. \quad (6)$$

According to (5) the condition of the effective protection is the ratio:

$$\lambda_{ij} \leq 1, \quad i = \overline{1, n}, j = \overline{1, m}. \quad (7)$$

This condition is necessary to use with the Wald's maximin model [14]-[16]. That is, the matrix of game is complemented by a string, each element of which has a minimum value of the gain in the strategy. The optimum for this criterion considered to be that strategy when



choosing minimum value is the maximum payoff. The chosen strategy in this way eliminates the risk.

Thus, after developing the game matrix and analyzing its values, it is possible to assess in advance the cost of each decision for the protection of computer information and choose the most effective variants to ensure security for the entire range of attacks.

## V. CONCLUSION

The article describes security problems because of cyber-threats, the service for vulnerability analysis and security assessment of open source systems.

The proposed service has some advantages over other services that provide penetration testing services. This service can offer a lower price in the implementation of testing services at the required quality, as it is possible due to lower costs of doing business by a small group of individual entrepreneurs. Also the use of software with open source has a positive effect on the cost price.

The article gives examples of hacking techniques, which are based on exhaustive search on (brute force) and SQL-injections. Also state graph for web-applications testing was presented. And the device for anonymization operating principle was described. In addition, were established interconnections and proposed a quantitative method for assessing potential damage after implementation of attack.

Ways of further investigations are improvement of testing methods, review of application source code to detect backdoors, improvement of anonymization device using such techniques as I2P and VPN, conducting simulation experiments and analysis of results obtained to develop recommendations to prevent potential damage and attacks.

## REFERENCES

- [1] "PlayStation Network Restoration Begins", PlayStation Network / PSN News. United Kingdom: Sony. 2011-05-17. Retrieved 2011-10-20.
- [2] *2015 Trustwave Global Security Report*, Trustwave Holdings, Inc, 2015.
- [3] J. Muniz, A. Lakhani, "Web Penetration Testing with Kali Linux". Packt Publishing Ltd, Birmingham B3 2PB, UK, 2013.
- [4] R. W. Beggs, "Mastering Kali Linux for Advanced Penetration Testing". Packt Publishing Ltd, Birmingham B3 2PB, UK, 2014.
- [5] *OWASP Application Security Verification Standard*, 2014.
- [6] *OWASP TESTING GUIDE v3.0*, 2013.
- [7] *Study: A Penetration Testing Model*, Federal Office for Information Security, Germany.
- [8] S. Abraham, S. Nair, *Cyber Security Analytics: A Stochastic Model for Security Quantification Using Absorbing Markov Chains*, Journal of Communications Vol. 9, No. 12, December 2014.
- [9] J. Clarke, *SQL Injection Attacks and Defense*, Justine Clarke, Syngress Publishing, Inc., 576 p., 2009.
- [10] *Statistics of web application vulnerabilities*, Positive Technologies, 2013.
- [11] *Analysis of the problems of password protection in Russian companies*, Positive Technologies, 2013.
- [12] K. Hausken, "Probabilistic Risk Analysis and Game Theory. Risk Analysis", Vol 22, No1, 2002.
- [13] G. Owen, "Game Theory", 2nd edition, Academic Press, 1982.
- [14] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses", The Annals of Mathematics, 10(4), 299-326, 1939.
- [15] A. Wald. "Statistical decision functions which minimize the maximum risk", The Annals of Mathematics, 46(2), 265-280, 1945.
- [16] A. Wald, "Statistical Decision Functions", John Wiley, NY, 1950.



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Francisk Scorina Gomel State University in implementation of Tempus project «Eastern partnership in pedagogical innovations in inclusive education (INOVEST)

Sergei A. Khakhomov, Uladzislau A. Beizerau

**Abstract**—The article deals with the development of integration of Belarussian system of education into European educational space. Implementation of international programs, held by the higher educational establishments of the Republic of Belarus is discussed. The role of universities in European integration is critically analyzed. As the example of a long-scale European project, Tempus-INOVEST (Eastern Partnership in Pedagogical Innovations in Inclusive Education) project is described and analyzed. The role of F. Scorina Gomel State University in developing European integration processes on post-Soviet space is discussed.

**Keywords**—European educational space, European Union, integration, ERASMUS, Higher educational institutions, inclusive education, INOVEST, integration processes, TEMPUS, universities.

## I. INTRODUCTION

Today the most wildly spread and dominating tendencies in the World educational space are globalization, internationalization, commercialization and regionalization. One of the major driving factors of globalization and internationalization in education is the factor of competitiveness in the educational market. Now to be competitive, the education system has to be global, provide all types of educational services demanded by consumers and economy. In turn, institutions of higher education have to be global and international to increase competitiveness.

Most often met definition of internationalization in higher education was given by Altbach and Knight as "The process of integration of higher education institution into the world educational space including training, scientific researches and other functions which are carried out by higher educational institutions in interaction with foreign partners". It is possible to offer another following definition of internationalization in education: Internationalization of higher educational institution is a process of integration of establishments, and also their compound (students, teachers, resources) into global (world) educational space. This definition allows to describe the process of internationalization more precisely [1].

Thus, internationalization of universities assumes activity in various directions (educational, scientific, cultural, innovative, etc.). For example, educational direction is realized through training of foreign students, undergraduate and graduate students at universities, realization of joint educational programs, opening of joint educational institutions, and also an exchange of undergraduate, graduate students and teachers. The scientific direction is realized through joint scientific programs, research, carrying out joint conferences, seminars, functioning of joint scientific laboratories, etc.

S. A. Khakhomov F. Scorina Gomel State University, Gomel, Belarus (e-mail: khakh@gsu.by)  
U.A. Beizerau F. Scorina Gomel State University, Gomel, Belarus (e-mail: bejzerov@gsu.by)

## II. FACTORS OF INTERNATIONALIZATION

The factors which increase the process of internationalization of universities are described in scientific literature. In particular, Spring allocates academic (or educational, assuming global influence of education and scientific research), economic (search for new sources of income and development out of the countries). These two groups of factors explain why universities get benefit from internationalization. Internationalization promotes realization the main educational mission by universities. Even especially local or regional universities which serve the local educational market now are compelled to integrate the international component into the training programs, as the world around becomes more and more connected organism that demands knowledge of it and the ability to put the knowledge into practice in various parts of the planet from students [2].

One of the major driving factors of internationalization in the higher education is the factor of competitiveness of universities in the educational market. Now to be competitive, the university has to be global, i.e. provide all types of educational services demanded by students. There are two reasons. The first is the demand of students and employers for the academic courses, programs and subjects of scientific researches having the international value and actual in any country. The second competition against other institutions of higher education [3].

Along with the domestic students, being internationalized, universities have an opportunity to take foreign students that is positively reflected in the general atmosphere of the university (it is supposed that foreign students, paying fees, are mostly motivated to study). Institutions of higher education with well developed international strategy are rather easily capable to catch the best students and the best teachers, having sufficient experience of recruiting, and also are attractive to students and teachers from around the world. Use of English in teaching is important. Internationalization is also an opportunity to diversify activities of the higher educational institution in the geographical plan. The entry into the world educational markets allows universities to carry out programs, which are not popular among local students. However different countries are at the different level of economic and technological development and are characterized by various demographic situation. Level of demand for educational services in various countries respectively depends on influence of these factors. For universities existence of foreign branches is also a good opportunity to use the labor potential flexibly. In the absence of demand for any programs in one country, there is no need to carry out reduction of teachers. In similar cases they are temporarily transferred to branch where these programs are demanded by potential consumers [4].

It is possible to refer high cost of the carried-out international projects, and also high risks for the higher education institutions opening branches abroad to problems of internationalization of economic character.

Internationalization of higher educational institutions is the developing, multipolar process. As a rule, this process originates from internationalization of structure of students and the academic programs, proceeds in internationalization of professorial and research structure. Apogee of internationalization of higher education institution is opening of foreign branch, or the whole network of branches of this university [5].

The most effective instruments of integration in education are the international projects and programs. And, these projects differ on duration (short-term and long-term), the purposes and scales.

The Republic of Belarus, being in the heart of Europe, is actively involved into the European integration processes in education.

Here we can list the objective factors of the national educational system in the world educational space, which determine the main strategic directions of modernization, are:

- geopolitical (being in the center of Europe; interaction with the European Union with the countries of the CIS and Baltic States; the Union of Belarus and Russia; Eurasian economic community; the forthcoming creation of the strengthening of the processes of globalization in the world; the long isolation of Belarus from the world education space in the last century, etc.);
- socio-cultural (the European nature of the social and cultural spheres; one of the best education systems in the world; a powerful scientific and socio-cultural potential; reform of the education system, taking into account the characteristics of the processes of integration, etc.);
- economic (significant economic potential; the transition to market-based forms of educational organization and commercialization of educational services, etc.);
- proper educational (legal security; principles, priorities for the development of the national educational system; the functioning of educational institutions of various types and forms of ownership, structure and duration of education at all levels is approximately the same to duration of education in European countries; the introduction of a two-tier system of training specialists with higher education; the introduction of credits, etc.).

### III. REPUBLIC OF BELARUS AND INTERNATIONAL PROGRAMS

Since the beginning of the 1990 's in the Republic of Belarus, programs that allow citizens of the country with higher education, to study at UK universities developed. To a large number of options-from receiving the certificate – to master's degree.

The large-scale project is a project of VISBY, introduced by Swedish Institute since 2004 year. The objective of this project, funded by a Swedish government was to strengthen cooperation in the field of higher and secondary education and expand the exchange of students between Sweden, Belarus, Ukraine, and the North-Western and Central regions of Russia. The benefits of this program can benefit students, teachers and scientists from these countries and regions. The Institute offers a number of scholarships for long-term (1-2 years) and short-term (1-2 weeks). Long-term visit, in most cases, involves training master at one of the universities in the country. The applicant must possess sufficient Swedish or English and have a basic education. The main advantage of VISBY is that they can make use of whole groups of students and senior high school students of lyceums and gymnasiums. This is the interchange of the students of the two countries [6].

The value of joint educational projects for modernization and improvement of the education system of our country and the further integration of the national high school (which is still considered to be one of the best in the world) in the world education system should not be underestimated.

Highest stage of international educational projects is the opening of the joint educational institutions. The clearest example of such cooperation was the Franco-Belarusian Management Institute, opened September 1, 1995, on the basis of the educational establishment “F. Scorina Gomel State University”, with the cooperation of Belarusian and French sides. Competence of students formed in the educational space of the two countries, and at the end of the training they receive two diplomas-Belarusian and French. For just 19 years, more than 500 students received diplomas of different levels [7].

Gomel State University actively supports the trends arising from modern educational paradigm. The consequence of which is the recognition of the institution as one of the best universities of CIS in 2014. A certificate of international rating agency «expert RA» was received by GSU resulting from the analysis of the statistics and surveys of the academic and scientific community, employers, students and graduates. GSU holds the 153-th place in the list of higher education institutions of Russia, Ukraine, Kazakhstan, Belarus, Moldova,

Uzbekistan, Azerbaijan and Kyrgyzstan. In the ranking of universities Ranking web of universities GSU takes 4 place in the Republic of Belarus (February, 2014).

Currently, GSU takes an active part in the international project TEMPUS. TEMPUS-programme, funded by the European Union, aimed at supporting the modernisation of higher education in the partner countries of Eastern Europe, Central Asia, the Western Balkans and the Mediterranean. TEMPUS is part of the European political neighbourhood programs the purpose of which is to contribute to the prosperity, stability and security of countries in the regions of Eastern neighbourhood. The program promotes cooperation in the field of higher education between the EU and partner countries and promoting voluntary integration of the higher education systems of the partner countries in the European processes.

Since 1993 the TEMPUS projects has played a significant role in the integration of the academic community of Belarus into the European network of institutions of higher education. Belarus is involved in projects from 1994. According to the Office of the ERASMUS (TEMPUS) in the Republic of Belarus, all higher education institutions participated in 64 projects.

The 2014 concluded with the selection of projects for funding under the sixth competition of TEMPUS IV projects. 171 projects were selected. Among them are 13 projects involving Belarusian institutions of higher education.

#### **IV. GOMEL STATE UNIVERSITY AND TEMPUS PROGRAM**

Gomel State University is one of the leaders on the number of selected projects among Belarusian universities. It should also be noted that the participation of the GSU in implementation of international projects is possible thanks to the close coordination of work with the National Erasmus (Tempus) Office in Belarus, participation in training seminars, meetings with potential partners.

Only results from the latest competition, the University will participate in the two projects "Centres of excellence for young scientists from Belarus, Russia and Ukraine» and «Ecological education for Belarus, Russia and Ukraine.

Currently, the GSU. takes part in 4 Tempus projects: 1. Preparation and professional development for teachers in the field of tourism (Teacher Education and Training in Tourism in Belarus); 2. Eastern partnership in the field of pedagogical innovations in the framework of inclusive education (Eastern Partnership in Pedagogical Innovations in Inclusive Education); 3. Environmental education for Belarus, Russia and Ukraine (Ecological Education for Belarus, Russia and Ukraine); 4. Centers of excellence for young scientists from Belarus, Russia and Ukraine; and one project IANUS II-Inter-Academic Network mobility under Erasmus Mundus programme (Erasmus Mundus), aimed at strengthening cooperation and international relations in the field of higher education and the universities of Belarus, Ukraine, Moldova, Armenia, Azerbaijan and Georgia with the universities of the countries of the European Union.

INOVEST project (Eastern Partnership in Pedagogical Innovations in Inclusive Education) TEMPUS IV began with in October 2012. The goal of the project is the creation by the end of 2015, the methodological and educational framework for the development and implementation of training programmes for teachers in the field of pedagogical innovations in inclusive education, as well as their further development and sustainability within the University structure; creation and launching of a national network of promoters in the field of pedagogical innovations in the inclusive education included for the support and implementation of innovative education in Belarus, Ukraine and Moldova [8].



Fig. 1 INOVEST logo

The project involves universities-partners from 6 countries (Spain, Germany, Sweden, Belarus, Ukraine and Moldova). The international university consortium is composed of 4 Belarusian partners (Ministry of education, Belarussian State Pedagogical University, Minsk City Institute of Education Development, and F.Scorina Gomel State University). Two laboratories were created in each of the partner universities: the Laboratory of research and development of talents and intelligence and Innovative multimedia didactic laboratory [6].



Fig. 2 INOVEST laboratory at F. Scorina Gomel State University

Within the framework of the project, the partner sides are actively implementing special courses for teachers of institutions of secondary education. Experts develop and adapt training programmes for teachers and school principals in the area of pedagogical innovations in inclusive education, based on the latest assistive technologies.

Belarus joined some major international instruments on human rights (UN Declaration on Human Rights, UN Convention on the Rights of Persons with Disabilities, the Convention on the Children Rights) and committed themselves to respect human rights, in particular, to ensure the right to education of children with special educational needs (SEN). In these countries over the past 10 years, reforms have been undertaken in special education, which resulted in a significant number of children with SEN having been transferred into mainstream schools. The

positive public opinion on inclusive education has been created. State regulations aimed at the development of inclusive education have been developed and adopted.

According to operative data, in Belarus there are more than 50.000 children with SEN integrated in secondary schools. However, the organizational and methodological principles of the learning process in secondary schools in Belarus are focused on children with regular development, and does not take into account the peculiarities of the educational-cognitive activity of children with SEN. Non-compliant forms and methods of pedagogical influence on these children can create the preconditions for the formation of their negative attitudes toward school and deviant behavior. Belarus is currently facing several major issues in social protection, which primarily affect children with SEN. Lack of alternative education for children with SEN and lack of adequate support for their families in bringing up such children inevitably reduce the quality of their education. Therefore most of these children graduate from schools without acquiring proper skills and abilities necessary for life and many of them are being deemed “uneducable”.

The successful implementation of inclusive education for children with special needs, requires the solution of many problems at the state level. One of them is the introduction of Pedagogical Innovations in Inclusive Education, based on the latest assistive technologies (AT) and information and communications technologies (ICT).

The concept of innovative education enjoys high priority within international and European policies. The European Council has repeatedly recognized innovation as crucial to Europe’s ability to respond effectively to the challenges and opportunities of globalization. In December 2006, for example, it is noted that “Europe needs a strategic approach aimed at creating an innovation-friendly environment where knowledge is converted into innovative products and services”. In many European countries (including Belarus) innovative education and innovative schools became a reality.

On the other hand, institutions of teachers’ education do not offer training courses for teachers’ re-qualification or retraining teachers in Pedagogical Innovations in Inclusive Education. In consequence, apart from isolated short term or punctual in-service, trainings on individual aspects of innovative education practically do not exist [6].

To tackle these problems the consortium of European universities was created and proposed the project aiming at the development of a holistic concept and a pilot structure for coordinated teachers’ training in Pedagogical Innovations in Inclusive Education by creating correspondent networks and fostering universities' contribution to the development of educational, legal and technological support for the implementation of Pedagogical Innovations in Inclusive Education with a regional-wide impact. The first step in the development of innovative education is training and/or retraining of teachers of pre-university educational institutions to equip them with the knowledge and skills to work with children with SEN and with their parents. As a result, children with SEN will acquire easier skills in: reading, writing, mathematics, ICT and foreign languages in correspondence with EU strategies regarding basic competences.

Therefore the consortium members expressed their will to establish a particular inter-university regional network of promoters for Innovative Education. The consortium members believed that the creation of this network would be an important step towards a knowledge society.

An important role in the project constituted the experience in this field of Germany, Spain and Sweden, where innovative educational systems were created in the frame of the education reform. Another important advantage is the experience of Belarussian HEIs in the provision of professional training for teachers as well as their experience in the initial education of teachers. Thus the project’s idea not only corresponded with the regional development priorities but fit

well into the competencies and development priorities of the partner universities and educational institutes in Belarus and other countries, which have developed a range of activities related to innovative education, including research and studies to various national and international forums.

The previous experience of the consortium members' universities in many national and international projects demonstrated their capacity and expertise to contribute to the project objectives.

The aim of the project was set – to contribute to the development and implementation of Pedagogical Innovations in Inclusive Education policies in Belarus and a better coordination of lifelong learning professional formation in education through the enhancement of the quality and relevance of services of Belorussian universities in the provision of professional teachers' training in Pedagogical Innovations in Inclusive Education.

The project team worked on the basis of 6 work packages: 1. Increased knowledge on Pedagogical Innovations in Inclusive Education through sharing of experiences and visions on the progress of innovative education at national and regional levels (It included following steps: In-depth analysis of the state of implementation of Pedagogical Innovations in Inclusive Education and teachers training needs in Belarus, Ukraine and Moldova; Comparative analysis and benchmarking of best practices in Pedagogical Innovations in Inclusive Education and teachers training in the EU partners' countries; Stakeholders round tables for the discussion and promotion of the methodology of Pedagogical Innovations in Inclusive Education at schools and related teachers training. 2. Publication of resource guides on Pedagogical Innovations in Inclusive Education and professional teachers training (Development and publication of a guide on Pedagogical Innovations in Inclusive Education and requirements for professional teachers training; Development and publication of a road map for the implementation of Pedagogical Innovations in Inclusive Education and lessons learned from the project). 3. Establishment and launch of the new programs on Pedagogical Innovations in Inclusive Education ( Preparation of administrative and technical conditions for the establishment and launch of the INOVEST training programs; Selection of administrative staff and trainers for new training programs; Preparation & implementation of short-term intensive training courses for the staff (training of trainers). 4. New curricula and courses for professional training in Pedagogical Innovations in Inclusive Education developed and implemented (Development and publication of curricula, courses and training materials for professional training of teachers and decision makers in Pedagogical Innovations in Inclusive Education; Development and publication of curricula, courses and training materials for professional training of promoters in Pedagogical Innovations in Inclusive Education; Implementation of continuing training courses in Pedagogical Innovations in Inclusive Education. 5. Creating a regional network of Pedagogical Innovations in Inclusive Education promoters developed (Selection of promoters in Pedagogical Innovations in Inclusive Education and setting up of a regional network; Training of promoters of Pedagogical Innovations in Inclusive Education). 6. Dissemination (Creating and up-date of a dissemination website; Development and publication of dissemination materials; Internal and external seminars and information events; Preparation & Implementation of the final conference).

During the INOVEST project course materials were developed within the following curricula areas: Psychology (3 courses); Pedagogical innovations (6 courses); ICT in inclusive education (5 courses): Introduction to the psychology of human uniqueness. Psychodiagnostics. Psychological tests. Teaching Thinking Skills (training). Innovative pedagogics. Teaching today. Creative and innovative education at school. Pedagogical Innovations in Inclusive Education at School. Pedagogical Innovations in Inclusive Education at HEI. School management of Pedagogical Innovations. ICT for students with SEN. Assistive technologies



in education. Cloud technology in inclusive schools. ICT accessibility features for improvement and disclosure the abilities of students with SEN. (laboratory practice). Digital Laboratory's FourierEdu (laboratory practice).

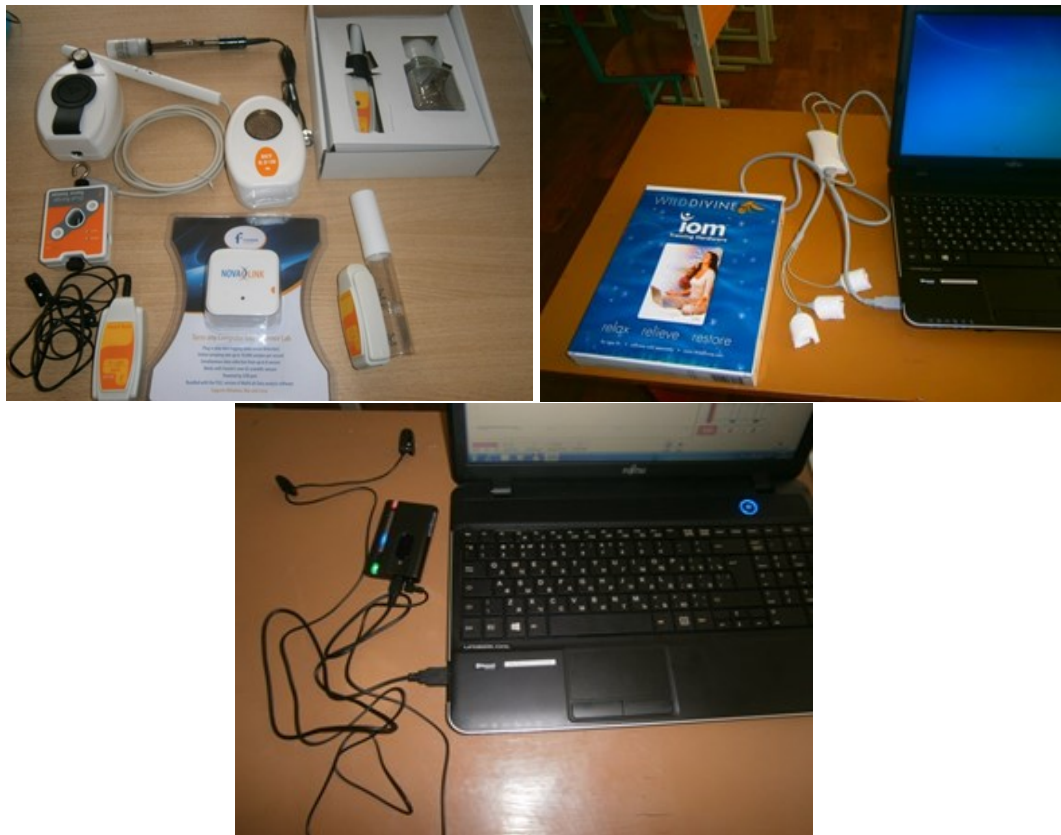


Fig. 3 INOVEST laboratories and equipment

## V. CONCLUSION

During the project the methodological and educational framework for the development and implementation of professional teachers' training programs in Pedagogical Innovations in Inclusive Education by 2015 was created; national network of Pedagogical Innovations in Inclusive Education promoters for the support and implementation of innovative education in Belarus was created. Partner universities do much in dissemination of the results of the project, including creation of web-pages of the project on the university's web-sites [9].

The project resulted in the following outcomes and outputs:

1. Increased knowledge on Pedagogical Innovations in Inclusive Education through sharing of experiences and visions on the progress of innovative education at national and regional levels;
2. Publication of resource guides on Pedagogical Innovations in Inclusive Education and professional teachers training;
3. Establishment and launch of the new programs on Pedagogical Innovations in Inclusive Education;
4. Development and implementation of new curricula and courses for professional training in Pedagogical Innovations in Inclusive Education;
5. Development of a regional network of promoters in Pedagogical Innovations in Inclusive Education.

## REFERENCES

- [1] P. G. Altbach, J. Knight, The internationalization of higher education: Motivations and realities, *Journal of Studies in International Education*, № 11. – pp. 290–305, 2010.
- [2] J. Spring, Globalization of education, NY.: Routledge, 2009.
- [3] Internationalization and trade of higher education: challenges and opportunities, Paris.: OECD, 2004.
- [4] S.B. Sutton, Developing Strategic International Partnerships. Models for Initiating and Sustaining. Innovative Institutional Linkages, NY. : Institute of International Education, 2011.
- [5] S.B. Sutton, Transforming internationalization through partnerships, *International Educator*, 19(1), pp.60–63, 2010.
- [6] S. A. Khakhomov, U.A. Beizerau, V. N. Dvorak, F. Scorina Gomel State University as an active participant of integration of the Belarusian higher school into the European educational space, Proceedings of the international conference "The European Union and Republic of Belarus: cooperation prospects", Minsk.: BSU, 2014, pp. 340–343.
- [7] A.V. Rogachyov, A.V. Gavriluk, M. Shenevua, French-Belarusian institute of management of F. Scorina Gomel State University, Vysheyschaya school, №2, pp.39–44, 2011.
- [8] INOVEST project website, 2015, <http://www.inovest-project.eu/index.php/en/>, accessed March 2015.
- [9] GSU INOVEST project page, 2015, <http://www.gsu.by/mundus/530417.asp>, accessed March 2015.

# Optimization of medical institutions in Ukraine

Anna Trofymenko

**Abstract**— In this paper, there were developed mathematical models and tasks settings of optimal placement of medical institutions in terms of health system restructuring using optimal set partitioning methods. Represented a developed program with the results of optimal partitioning of Dnepropetrovsk region to the service zones (medical districts) by hospitals of the second level.

**Keywords**— optimization, medical institutions, mathematical methods.

## I. PROBLEM FORMULATION

One of the priority direction of development of any country in the world is maintenance the necessary health level of its citizens. Maintenance and improvement the overall health of the population is only possible with effectively functioning health care system. Since 2011 in Ukraine there has being a health sector restructuring. The main purpose of this restructuring is improvement healthcare service. One of the main areas of sector restructuring is the health institutions network optimization through the formation of the hospital districts. Therefore, there is an urgent task of identifying the hospitals optimal location and identifying areas of service by them i. e. finding optimal medical district's zones. Moreover it is important to satisfy people needs in all types of medical care, take into account the distance to the hospitals and theirs capacity powers.

## II. ANALYSIS

The study of the state regulation problems in the health field in Ukraine and challenges of its restructuring as well as its effective operation involved the following scientists: V. Lekhan [1, 2], Y. Radysh [3], O. Schepin [4].

The problem of optimal health institutions allocation with the definition of the boundaries of service areas for the population is a multifaceted and algorithmically complex task. Usually, the problems of optimal sets partitioning are solved for the optimal firms location that produce several types of products and identifying areas where to supply these products in. Herein the total production expenditures and goods transportation are minimized and service areas are determined separately for each product. Typical representatives of continuous problems of optimal set partitioning are infinite dimensional transport problems or more general – infinite dimensional enterprises location problems with simultaneous division of a given region, constantly filled with consumers to the field of consumers, each of which is served by one enterprise in order to minimize transportation and production costs [5].

The difference of the problems of optimal medical institutions allocation lies in the fact that they should be placed, considering simultaneously all services that will be provided by each institution.

The issues of optimal sets partitioning are investigated by such scientists as A. Kiselova [5], L. Lozovskaja [6], N. Shor [7] and others.

Also, the continuous problems of optimal sets partitioning are studied by foreign authors such

as H. Corley, S. Roberts [8], R. Francis [9], J. Moreno, C. Rodriguez, N. Jimenez [10].

However, despite on considerable researches in this area, their thorough analysis showed that in scientific studies there are not deeply examined enough the questions of mathematical device application for deciding actual issues that arise during the healthcare sector restructuring.

### III. THE MAIN PART

The effective functioning of the health system is impacted by many factors such as validity and adaptability of solutions to national realities, interest actions by all levels of government as well as readiness and motivation of managers of health institutions.

One of the priorities of the health sector is the availability of medical services for the people that is confirmed by the Law of Ukraine about Health System. Improving the efficiency of the health system requires its reorganization. During the period of 2011-2013, the Government of Ukraine held a medical reform based on the restructuring of health care in the “pilot regions”, which include Vinnitsa, Dnepropetrovsk, Donetsk regions and Kyiv [11].

Ukrainian government conducts a series of activities for the reorganization of health care system at the primary level, in which polyclinics are reorganized into outpatient clinics of family type that geographically close to the places of accommodation of the population that they serve [12,13]. In order to improve the quality of specialized medical services, there are held events concerning the formation of the secondary (specialized) level of health care and the creation of hospital districts, which involves combining a network of health facilities which will provide medical assistance in case of need of specialized methods of treatment, diagnostics, rehabilitation and use of complex advanced medical technologies. Decisive place of secondary level should occupy general hospitals which provide intensive medical care (primarily emergency medical assistance) since from their activities the lives and health of patients are depended. In developed countries general hospitals make up approximately 90% of all hospitals [2].

Hospital districts may combine health facilities of several rural areas or cities and regions that depends on the density and nature of the population resettlement, takes into account the state of transport communications, logistical and staff potential of hospitals, their profiles and structure of health care services. At the same time it is important to take into account the people needs in health care, the creation of appropriate conditions for the quality care and optimal use of material, financial and human resources. It should be noted that in Europe and the United States there exist hospital districts, that allows to state rationally allocate medical resources and provide more qualitative and prompt qualified medical care.

Currently, medical care in Ukraine is divided on three levels: primary, secondary and tertiary levels.

The primary leveled centers provide primary health care. In order to improve the quality of secondary health care, it is expedient to create hospital districts, which represent a specific organizational and functional union of medical facilities and satisfy current needs of the population in specialized health care.

The need for a hospital districts formation can be explained by the fact that for the providing of quality intensive secondary health care, institution must be fitted with appropriate medical equipment, staffed by doctors and has the ability to make the required number of interventions, which may be provided when this institution serves approximately 200 thousand people [2].

As health care institutions that needed to be placed in each district have different profiles of health care, perform medical procedures of varying complexity and are suggested to service a certain number of people, the optimal reprofiling and location of these centers in existing health care facilities is important, given the existing number of staff physicians and the number of

beds [1]. For this purpose, medical institutions, which will be included in the medical district it is expedient to choose by their capacity criterion, which is calculated as a proportion of doctors and beds in each institution relatively to the total number of doctors and beds of selected health facilities. Accordingly to this, the general hospitals may be located in the institutions with the highest capacity criterion PI (Power Indicator), in institutions with lower PI will be located other kinds of hospitals.

To determine the location of health facilities and the borders of the territories served by them, we use the model of optimal partition of [3].

To create optimal conditions for the functioning of the health sector and ensure an adequate level of health care to the population of the second level, first of all, in a given area need to optimally place the health facilities and identify areas (districts) public service by these institutions so that they do not overlap between other and meet the needs of the population for health care. Cost of service delivery and shipping costs are known. This service costs should be minimal, and the load - and the maximum capacity of all health facilities in the region should meet the needs of the total number of services.

We distinguish some types of tasks the medical institutions of the second level and formulate mathematical models of these tasks as continuous single-commodity special type of optimal partition [5]:

1. Place the second level medical institutions, determining the coordinates of their position and find the limits areas that will be served according to the institutions, taking into account the needs of the medical service at some time. A mathematical model of the problem will be as follows:

Let  $\Omega$  – limited closed set, measurable by Lebesgue that specifies the region to allocate medical institutions. It is need to be divided into  $N$  service areas  $\Omega_1, \dots, \Omega_N$  (which do not intersect with each other) of patients of  $i$ -th medical institution and to place centers  $\tau_1, \dots, \tau_N$  of these zones in the region  $\Omega$ , i.e. find the unknown beforehand coordinates of the centers  $\tau_1, \dots, \tau_N$  so that the total number of patients who served by the  $i$ -th institution and lives in the area  $\Omega_i$  does not exceed the given volumes:

$$\text{mes}(\Omega_i \cap \Omega_k) = 0, \quad i \neq k, \quad i, k = 1, 2, \dots, N, \quad (1)$$

where  $\text{mes}(\cdot)$  – the Lebesgue's measure,

$$\bigcup_{i=1}^N \Omega_i = \Omega, \quad (2)$$

total number of patients who are served by  $\tau_i$ -th hospital and lives in the area  $\Omega_i$  does not exceed the given volumes:

$$\int_0^T \int_{\Omega_i} \rho^j(x, y, t) dx dy dt \leq b_i^j, \quad i = 1, \dots, N, \quad j = 1, 2, \dots, M, \quad t \in [0, T], \quad (3)$$

where:

$\rho^j(x, y, t)$  – the need in the  $j$ -th service at the point  $(x, y)$  at the moment of time  $t \in [0, T]$ ;

$T$  – period of time for which the dynamic of the services is known if it is not known, then  $\rho^j(x, y, t) = \rho^j(x, y)$ ;

$b_i^j$  – maximum number of  $j$ -th service that is provided by the  $i$ -th medical institution;

and thus the total cost functional

$$F(\{\Omega_1, \dots, \Omega_N\}, \{\tau_1, \dots, \tau_N\}) = \sum_{i=1}^N \int_0^T \int_{\Omega_i} \sum_{j=1}^M [c(x, y, \tau_i) + a_i^j] \cdot \rho^j(x, y, t) dx dy dt \quad (4)$$

reached the minimum value.

Here and in the following models:

M – number of services;

N – number of centers;  $c(x, y, \tau_i)$  – function of determining transport costs for one medical visit or patients' transport costs;

$\tau_i = (\tau_i^x, \tau_i^y)$  – the coordinates of the  $i$ -th center;

$a_i^j$  – cost price of providing  $j$ -th service by  $i$ -th medical institution.

The total capacity of all health facilities must meet the second level collective need for all health care services in the region. That is, the condition:

$$R = \int_{\Omega} \sum_{j=1}^M \rho^j(x, y) dx dy \leq \sum_{i=1}^N \sum_{j=1}^M b_i^j, \quad 0 < b_i^j \leq R, \quad i = 1, 2, \dots, N, j = 1, 2, \dots, M, \quad (5)$$

where R – the total demand for health care in the region. That is, the possibility of providing medical services to all hospitals in the region should meet the total needs of the population.

For simplicity, we consider the following problem without taking into account the dynamics of demand for services.

- Determine the optimal limits areas served by existing health facilities. A mathematical model of the problem will be as follows:

Let  $\Omega$  – properly closed sets, Lebesgue measurable by that specifies the region to accommodate medical facilities. You must break it into N service areas  $\Omega_1, \dots, \Omega_N$  patients and the first medical institution that does not intersect with each other, so that the total cost functional

$$F(\Omega_1, \dots, \Omega_N) = \sum_{i=1}^N \int_{\Omega_i} \sum_{j=1}^M [c(x, y, \tau_i) + a_i^j] \cdot \rho^j(x, y) dx dy \quad (6)$$

reached a minimum value conditions (1-3) and (5).

Thus, developed mathematical models and tasks to them in their implementation in software environment can determine the optimal area of people medical service by existing health facilities including total maintenance costs and transportation costs for calls.

Due to the mathematical models and setting objectives for them, created software in C++ language environment Microsoft Visual Studio 2010, which allows to determine the optimal service area of population by existing medical facilities, taking into account the total cost of maintenance and transportation costs for calls. Fig. 2 shows a window with the results of its work.

If necessary, when changes in population dynamics known, the algorithm can be modified without significant complications.

Operations are tested to determine the optimal partitioning of Dnepropetrovsk region in the service area (district) hospitals the second level. On the fig. 1 hospitals provisions of the second level is represented by the points with the following coordinates, each separate service area hospital depicted by its color.

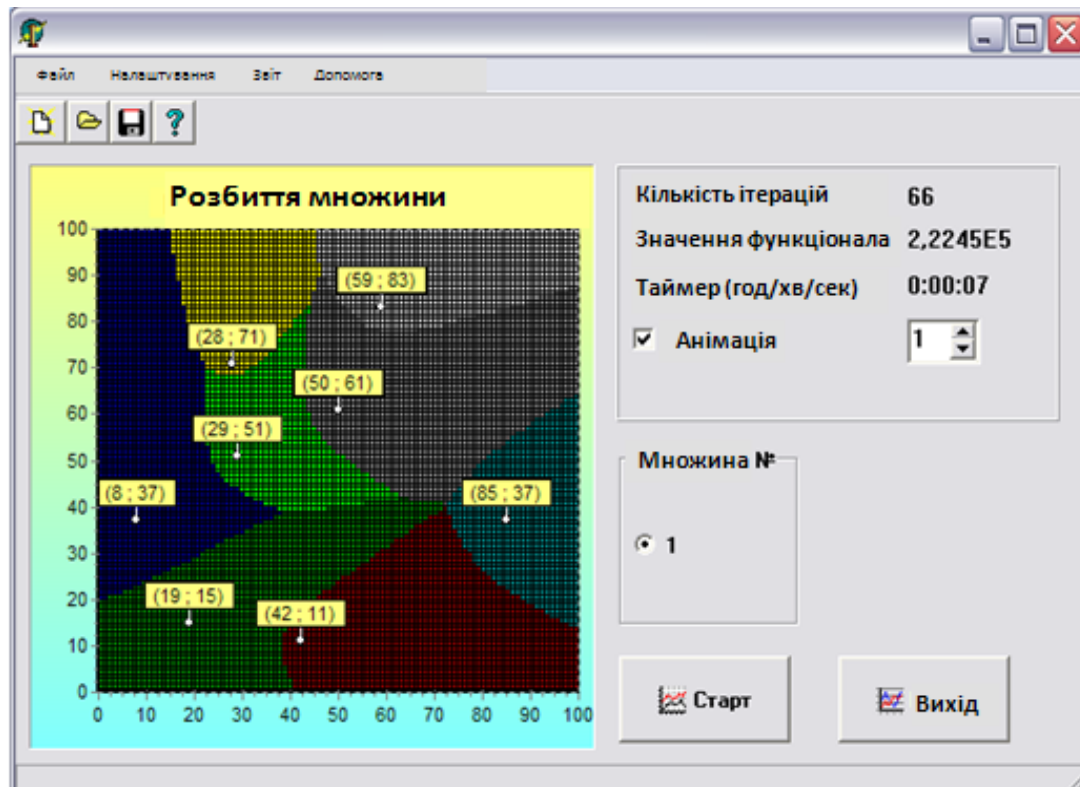


Fig. 1 Optimal service area of the second level hospitals

Source: Developed by the author.

Input data for the medical institutions of the second level are the coordinates of existing health facilities Dnepropetrovsk region with high enough values of the capacity of the hospital PI (Power Indicator, which is calculated as a percentage of the total number of doctors and beds). As the reception centers of the second level hospitals considered: Krivorozhskaya hospital №2, for which  $PI = 0.093$ , coordinates (8, 37), Apostolove central district hospital, where  $PI = 0.07$ , coordinates (19; 15), Verkhnedneprovsk central district hospital,  $PI = 0.068$ , coordinates (28; 71), Krinichansky Central District Hospital,  $PI = 0.065$ , coordinates (29; 51), Marganetsk Central District Hospital,  $PI = 0.073$ , coordinates (42; 11), Dnipropetrovsk city Multidisciplinary Clinical Hospital № 4,  $PI = 0.097$ , coordinates (50; 61), Novomoskovsk Central Hospital,  $PI = 0.72$ , coordinates (58; 83), Intercession Central District Hospital,  $PI = 0.63$ , coordinates (85; 37).

Number of services that could be provided with hospital determined the average load on the doctor per day, which is 15 patients [16]. So №2 of city hospital could provide 2895 services a day, Apostolove Central District Hospital - 2685, Verkhnedneprovsk Central District Hospital - 1215, Krinichansky Central District Hospital - 1395, Marganetsk Central District Hospital - 2685, multi Dnipropetrovsk City Clinical Hospital №4 - 2985, Novomoskovsk Central District Hospital - 1575 and Intercession Central District Hospital - 1425 services per day. To simplify the analysis of the results, the cost of medical services in all institutions will be considered the same.

The optimal location of new second-level medical institutions and identified areas of service are showed on fig. 2. We assume that the cost of service delivery while placing new medical institutions is not changed and is the same cost of the provision of services by existing institutions.

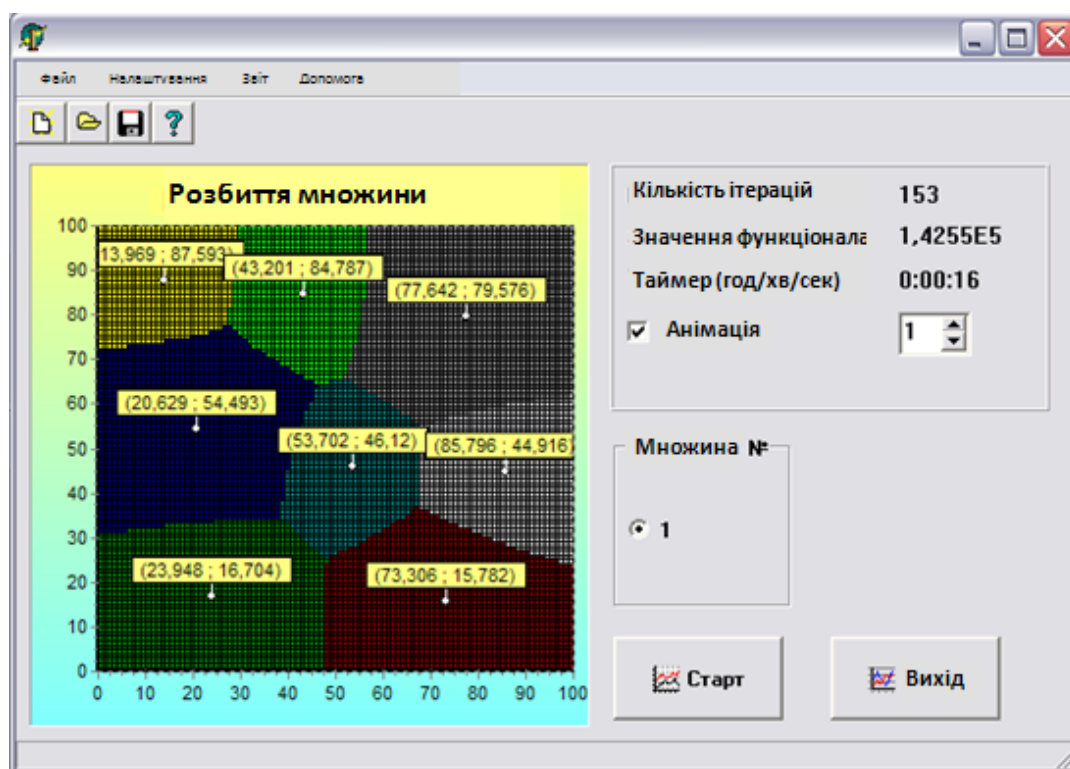


Fig. 2 Optimal location of new hospitals of the second level and their service area

Source: Developed by the author.

Thus, servicing the existing public hospitals makes functional expenses is 222 450 UAH. / Day (Fig. 1) and the placing of new medical institutions makes functional expenses 142 555 UAH./ Day (Fig. 2).

Thus, the optimal placement of new hospitals, the difference in transport costs and costs of medical services is 36% ( $((222450-142550) / 222450) * 100\%$ ), which, in turn, allows us to provide the most efficient high level of required quality health care for the population in the region.

These results do not include the cost of construction. Despite on the fact that the effectiveness of the location of new medical institutions is high enough and the target functional costs much less, in real terms the construction of new hospitals is unprofitable because it requires the transportation of all hospitals to new places.

This approach of the optimal location of new centers can be used in making decisions on the construction of hospitals in the area where the second level health facilities do not exist (for example, in the new residential area) or they do not meet the medical standards and require demolition.

Thus, the division of given territory into spheres of influence zones by existing hospitals on condition of their redevelopment is more appropriate.

Obviously, these results are approximate, because the actual situations which develop such models of optimal partitioning sets are often characterized by some degree of uncertainty due to lack of reliability of information and based on which the selection decision is done.



#### IV. RESULTS

In the article is it discussed the urgent issue of improving the provision of medical care by reorganizing health care on the secondary level accompanied by optimization of the network of health institutions and the formation of hospital districts. Methods of optimal partitioning sets have been used to solve this problem. Implementation in practice developed models of health facilities optimal location problems can reduce transport costs and costs of medical services by finding the optimal location of new health facilities taking into account certain restrictions, allowing more efficient use of available sector resources and provide the appropriate level of medical services.

Thus, the models of the selected types of problems allow place medical institutions of the second level optimally given into account the existing restrictions on the distance to the farthest point of the served territory, the capacity of medical institutions, change the demand for services in time, uncertainty in need of medical care and calls transportation costs.

#### REFERENCES

- [1] Lekhan V. "Main ways of further development of health care in Ukraine", For the Society, Publisher Rajewski, 2005. - P. 74.
- [2] Lekhan V. "Development strategy of the health system: Ukrainian Dimension", [electronic resource], News of medicine and pharmacy. - 2010. - 5 (311). - Access regime: <http://www.mif-ua.com/archive/issue11757/article11795/>.
- [3] Radysh J. "State Department of Health in Ukraine: genesis, problems and reforms: Monograph", Publishing UAPA, 2001. - 360 p.
- [4] Shchepin O. "About state health reform mechanisms", Problems of social hygiene and the history of medicine, № 1, 1997, P. 34-38.
- [5] Kiseleva E. "Continuous problem of optimal partitioning of sets: Theory, Algorithms, Applications: monograph", Naukova Dumka, 2005. - 564 p.
- [6] Lozovskaya L. "Models and methods for solving problems of optimal continuous decomposition sets in incomplete information", The dissertation, - Dnepropetrovsk, 1994. - 149 p.
- [7] Shor N. "Using the r-modification algorithm for finding the minimum polynomial functions", Cybernetics and Systems Analysis, № 4, 1997, P. 28-49.
- [8] Corley H. W., Roberts S.D. "A partitioning problem with applications in regional design", Operat. Res., Vol. 20. №5, 1972, P.1010-1019.
- [9] Francis R. L. "Sufficient conditions for some optimum-property facility design" Operat. Res., V. 15. №3, 1967, P. 448-466.
- [10] Moreno J., Rodriguez C., Jimenez N. "Heuristic cluster algorithm for multiple facility location - allocation problem. - Operations research", Vol.25, №1, 1991, P. 97-107.
- [11] The Cabinet of Ministers of Ukraine "On approval of the establishment of hospital districts in Vinnytsia, Dnipropetrovsk, Donetsk and the city. Kiev" of 24 October 2012 p. № 1113 [electronic resource]. - Access: <http://zakon2.rada.gov.ua/laws/show/1113-2012-%D0%BF>.
- [12] MON Ukraine of 23.02.2000 № 33 "On staffing ratios and standard staffing of health facilities" (amended) [electronic resource]. - Access: [http://www.moz.gov.ua/ua/portal/dn\\_20000223\\_33n.html](http://www.moz.gov.ua/ua/portal/dn_20000223_33n.html).
- [13] The new European policy for health – Health 2020: Vision, values, main directions and approaches: Regional Committee for Europe EUR / RC61 [Electron resource]: according to the World Health Organization / Regional Office for Europe. - Access: <http://www.euro.who.int/ru/health-topics/health-policy/health-2020-the-european-policy-for-health-and-well-being/about-health-2020>.



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Identification of explicit lexicon for opinion mining in the media

Liauchuk Veranika

**Abstract**—The paper deals with the problems of opinion mining in respect to the media discourse. The peculiarities of analytical media genres as the sources for opinion extraction are considered, the opposition and distinction between opinionative and non-opinionative phrases (opinions and facts) are analyzed. The paper presents quantitative survey on opinions in media discourse, discusses opinion lexicon and presents a classification of opinionative phrases in the discourse of mass-media. The results can be used in the algorithms and programs of automatic identification and extraction of opinion.

**Keywords** — explicit means, media discourse, opinion lexicon, opinion mining, opinionative phrases

## I. INTRODUCTION

Opinions fulfill all spheres of human activity and sway the perception of reality, behavior, choices and judgments upon a particular topic. Due to the role of the media as one of the sources to gain knowledge and, accordingly, the influence on a recipient, research of different media genres is becoming more important.

Studies of different points of view, comments and ratings in the media as well as facts and documental information create an opportunity to understand an object or a situation under consideration in a more profound way. Tracking of opinion trends as such and separate opinions in particular is helpful while investigating public opinion tendencies in the media-discourse. It is also noted that opinions (together with evidence and beliefs) record valuable data, and thus, become both a prerequisite and a result of perception [1].

## II. HISTORICAL OVERVIEW OF THE ISSUE

### A. Previous surveys

As far as opinions are considered a result of intellectual activity of a human-being, they became a subject of investigation in different fields (psychology, logics, philosophy, theory of communication, etc.). There is a range of works (Aleksandrova, Wittgenstein, Dmitrovskaya, Zaliznyak, Karagodin, Pavilens) dedicated to the opposition of knowledge and opinion (in this case a particular opinion is regarded as a proposition that requires verification or argumentation unlike propositions of knowledge), a number of surveys dealing with opinionated predicates (Apresyan, Arutunova, Zadvornaya, Ivanova), which provide foundations for further surveys.

### B. Opinion mining and Sentiment analysis

The domain of subjectivity analysis is being elaborated nowadays. It is studied in two main areas – Opinion mining and Sentiment analyses. The first aims at extracting and processing

views in a broad meaning, while the latter intends to determine attitudes. In a way the problems of both are similar, that is why at times these terms are used interchangeably.

The surveys of customer reviews, film reviews or ratings about products or services dominate in both areas (L. Lee, B. Pang, H. Tang, J. Wiebe, E. Riloff), thus, aiming at identification of emotional attitudes. There is also a number of works dedicated to differentiation of opinions and judgments which do not lack an emotional component (Dmitrovskaya, Pazelskaya, etc.). However, as Arutunova suggests “there are no distinct boundaries that divide mental and emotional spheres, will and wishes, perceptions and judgements, knowledge and beliefs in the inner life of a person” [2], emphasizing by this the correlation between opinions and judgments.

### III. THE POTENTIAL OF MEDIA-DISOURSE FOR OPINION MINING

#### A. *Mass-media discourse peculiarities*

The media have a unique capability for creation of cognitive trends and their manipulation. Being affiliated in the process of production of meanings, images and metaphors, in the process of mass information creation (content), the media sway public perception and behavior, control intentions and the way they are pronounced and disseminated.

Marking suggestive influence of the media on a mass recipient, Zheltuhina stresses that “this role is unlikely to change in the nearest centuries as speed, quality and quantity of the information sent and received (newspapers, magazines, radio, television, the Internet) is modifying together with the technological progress. So, the level of the media influence is increasing due to the interpretation of events and transformation of the information transmitted”. [3]

Werlich calls the types of modern texts “matrixes of elements that state texts fixed in linguistic experience of speakers” [4]. Moreover, Brinker notes that “texts represent a conventional example of complex linguistic actions” [5]. In this case a text can be regarded a complex unit with typical contextual (situational), communicative, functional and structural traits.

#### B. *Analytical genres of the media*

Analytical genres of the mass-media attract a particular interest for opinion extraction and study: analytical article, commentary, press or letter overviews, etc. The feature that unites them is orientation towards “interpretation of facts, phenomenon, elaboration of author point-of-view” and at times “a clash of opposite reader opinions in one material, dwelling upon it and revealing a personal judgment” [6]. It is logical to admit, the majority of linguistic means and tools are aimed at implementation of an affecting function “through argumentation, demonstration and reasoning” [7].

### IV. SOURCES FOR THE SURVEY

The survey presented was carried out on the base of articles of informational and analytical genres from the British newspapers “The Guardian”, “The Daily Telegraph”, “The Independent”, “The Daily Mirror” and Belarusian editions “Zvyazda”, “Novy Chas” and “Culture”. Through the method of contextual analysis more than 3000 opinionative phrases were selected from the mentioned texts.

## V. OPINIONATIVE CONTEXTS

### A. *Opinionative phrases*

Wittgenstein distinguished opinions as propositions that need verification against knowledge that does not need such verification. According to Apresyan, “the main difference between opinion and knowledge is in the relation to the verity of a proposition that contains the sense of a phrase” [8]. He also suggests a number of opinion-predicates and fact-predicates. Dmitrovskaya adds the criterion of probability that is implied by a speaker in a phrase [9]. In the surveys of Zaliznyak, opinion are characterized as phrases that have argumentation, comparison as a source of perception, they can have different levels of certitude, and finally, all unverifiable propositions should be considered opinions [10].

### B. *Opinions vs Facts*

Naturally, not all the phrases in media texts contain opinionative markers. It is important to distinguish between opinionative and non-opinionative phrases [11].

In our research we understand “opinion” as a proposition, by which some person shares a particular point of view on some object or situation. The obligatory component in the structure of an opinionative phrase is subjectivity.

There are several options by which opinions are opposed to facts (as verifiable objects, situations that are unbiased empirical knowledge):

- Author interpretations, logical arguments and conclusions to which another author can object. For example, a phrase “Ernest Hemingway received a Nobel prize for the novel “The Old Man and the Sea” is a fact. However, a phrase “The novel “The Old Man and the Sea” by Ernest Hemingway received a Nobel prize because the author managed to show unique experience” is opinion as another author can declare that “[it was] due to the representation of the strength of a human soul” or “[...] thanks to a particular writing style”.
- Emotional component in the structure of a phrase. Then, a phrase “The novel “The Old Man and the Sea” is the best work by Ernest Hemingway” is an opinion since one can oppose to it a similar in structure but a different in contents phrase – “The novel “For Whom the Bells Tolls” is the best work by Ernest Hemingway”.
- Modality: “Ernest Hemingway should have dedicated the book to the fisherman Santiago”.
- Statements about capabilities: “Ernest Hemingway could describe a life in a small fishing village in a thousand pages”.
- Statements about the future: “The novel will always be popular”, etc.

### C. *Opinion phrase*

One of the problems for extracting opinions is to define a minimum unit (a phrase or context) that is enough to perceive a particular opinion. In our survey we define an opinion phrase as a fragment of a text that contains a personal point of view (opinion of an author or another sender of information on a topic discussed) that is enough to perceive the sense of this point.

Such units can vary profoundly in their size:

- a part of a sentence;
- a whole sentence;
- several sentences;
- a passage or even the whole text (especially if an opinion is announced in a heading).

## VI. OPINIONS IN THE MEDIA: QUANTITATIVE PARAMETERS

### A. *Explicit and implicit opinions*

When mining opinions, one deals not only with positive, negative or neutral grades, but rather with the extraction of information, to be more precise: what and how somebody thinks of something. Whatever the purpose is, opinions are results of intellectual activity of a person, of something that is created, developed and preserved in a human's consciousness. That is why the first step is to build opinion lexicon that contains markers of intellectual, logical, etc. activity of a person.

It is important to note that the contents of every phrase (both opinionative and non-opinionative) differs by the level of explicitness. Correspondingly, there are explicit (obvious) and implicit (concealed, implied) statements. An explicit phrase is a statement which sense can be extracted reasoning from a surface structure, meaning of words it consists of, without additional transformations [12]. It is also right that explicit statements "are represented directly in the lexical and syntactical structure of a phrase" [13]. Explicit markers (words and constructions) are instrumental for the extraction, analysis and further classification of opinions [14].

### B. *Quantitative characteristics of opinions for informational and analytical genres of media discourse*

One part of our survey considered the saturation of opinionative phrases in informational and analytical genres of media discourse in general. To fulfill this task, we considered how one fact (an accident) was presented and discussed in both genres in the media both in English and Belarusian (attack on "Charlie Hebdo" offices on the 7<sup>th</sup> of January, 2015). And thus, only 16% of statements in the texts of analytical genres in English and 37% in Belarusian contain exceptionally facts (Table 1). For comparison, the proportion of such statements in the texts of informational genres is 49% and 60% accordingly. Moreover, while in informational genres authors express opinion explicitly rarely (9% of phrases in the English language and 8% in Belarusian) and prefer to share opinions by quoting other participants of communication, in the texts of analytical genres the proportion of the phrases with explicit means by author is much higher (48% in English and 47% in Belarusian).

TABLE I. STATISTICAL DATA ON OPINIONATIVE PHRASES IN THE TEXTS OF INFORMATIONAL AND ANALYTICAL GENRES OF THE MEDIA (REGARDING THE SAME FACT)

English		Belarusian		Option
<i>Information</i>	<i>Analitical</i>	<i>Information</i>	<i>Analytical</i>	
49%	16%	60%	37%	The volume of non-opinionative phrases
51%	84%	40%	63%	The volume of opinionative phrases
59%	9%	66%	8%	The percentage of quotations within opinionative phrases
26%	48%	25%	47%	The percentage of opinionative phrases with explicit markers
15%	43%	9%	45%	The percentage of opinionative phrases with implicit markers

### C. Quantitative characteristics of opinions for the genres “Comment” and “Opinion”

The peculiarity of a media text is in its secondary nature in respect of a founding text (a text itself or a real-life situation). This feature is especially urgent for the media genres “Comment” and “Opinion” since their aim is to represent author’s opinion on some topic. Accordingly, the main peculiarity of mentioned genres is high saturation with opinion phrases and explicit opinionative phrases in particular (Table 2). The rest of a text is a description of a situation discussed or facts.

The survey showed that the majority of opinionative phrases in the media genres “Comment” and “Opinion” are fulfilled with explicit of means of opinion expression. The tendency is similar in both languages. However, in English implicit opinionative phrases are also rather frequent (in comparison with explicit – 43% to 57% accordingly). This means that authors express opinion allegorical with the help of presupposition, irony and other means (for example, the author of an article running about the referendum in Scotland made a hint about a particular status of the region as compared to Wales or Northern Ireland in the following way: “Scottish independence: There’s a kind of magic in our United Kingdom”).

TABLE II. STATISTICAL DATA ON OPINIONATIVE PHRASES OF MEDIA GENRES “COMMENT” AND “OPINION”

Option	English	Belarusian
The percentage of opinionative phrases with explicit markers	57 %	82 %
The percentage of opinionative phrases with implicit markers	43 %	18 %
The percentage of quotations within opinionative phrases	6,5 %	48 %
The percentage of the opinionative phrases of an author (explicit and implicit)	93,5 %	52 %
The proportion of explicit and implicit opinionative phrases of an author	53,2 % to 46,8 %	66 % to 34 %
The percentage of the explicit opinionative phrases of an author	88,4 %	41,5%
The percentage of quotations within explicit opinionative phrases	11,6%	58,5%

As one could notice, the percentage of quotations (a direct means to represent opinion in the media discourse) in English is only 6,5% in the genres concerned, while in the Belarusian language it reaches 48% of all the opinionative contexts. In this respect the survey showed that with these quotations authors tend to provide a detailed opinion of another person (an expert, a participant of a problem situation, etc.) who finally becomes a “co-author” of a text. Taking into consideration this qualitative trait of Belarusian media discourse, the proportion of opinionative phrases of an author makes 85%, which proves the point.

## VII. OPINION LEXICON

Building a comprehensive data base, opinion lexicon, is an important task for creation of an accurate algorithm (and a tracking program) in the framework of opinion mining. Identification of explicit means of opinion expression is easier in comparison with the implicit ones, though still it is efficient due to their frequency on the texts. Thus, it is possible to extract a considerable array of useful information from opinion lexicon.

Below one can find examples of such opinion lexicon identified in the texts of analytical genre of the media (concerning the mentioned above accident).

### A. Words and structures of interpretation

Within explicit means of opinion representation by an author words and structures that help to interpret an event dominate (35% of phrases in English and 63% in Belarusian).

1) Units of logical relations, explanation of patterns (Belarusian variants are omitted).

a) verbs: *explain, attribute (to)*

b) prepositions: *by (doing smth)*

c) colloquial particles

d) conjunctive structures with the meaning of consequence: *if... (then), when... (then), interpretation*

For e.g., “We explained that the artists working at a magazine printed drawings that made two men angry”.

2) Markers with the semantics of reason

a) verbs and nouns: *cause, reason* (for the reason; the reason is...)

b) conjunctions: *because, for/as/since* (in the meaning of “because”), *due to*

For instance, “The reason is that it is a technique of conflict, not a cause”.

3) Descriptions and definitions

a) verbs with the meaning “to represent, to depict, to imply”: *mean, seem, represent, describe, see (as), view (as), symbolize*

b) cliché-structures: *the thing is*

c) prepositions: *to plus an subject* (to somebody something is...)

For e.g., “On the cover of what is meant to be an anti-establishment magazine, this just symbolises – at best – egalitarian bigotry”.

4) Aim

a) verb: *aim at*

b) prepositions: *for*

For e.g., “[...] the assassins’ bullets were aimed squarely at free speech itself”.

5) Result

a) verbs: *result, lead to*

b) nouns: *result*

c) adverb: *generally*

For instance, “As a result, all advice at the time was for America not to universalise its response to 9/11 [...]”.

#### *B. Markers of opinion*

1) A number of units that only mark a phrase as opinion, though do not add any new semantic senses to the context.

a) verbs: *suggest, show, create*

b) structures with the meaning of enumeration: *in one regard*

c) parenthesis: *it must be said, to be fair, by the way, among others*

As an example, “To be fair, that article was simply guilty of the Endtimes hubris that affects us all [...]”.

## 2) Personal characteristics of an author.

There is a range of linguistic means by which the attitude of an author towards the topic is identified. All such contexts contain a pronoun in the first person, singular (I, for me, etc.).

a) words and structures that mark emotions and mood of an author: *I am afraid/appaled/heartbrocken/glad/aware, I like/love/worry, etc.*

For e.g., “I’m glad to see development of detailed efforts in this country at government and local authority level to address extremism”.

By this an author emphasizes important for him details, and a recipient can get some information about a personality of the author. As a rule a clause or a separate sentence combined with the marker by sense contain an opinion on the object blamed or supported by the author.

b) A range of opinion verbs and structures that mark opinion contexts: *suppose, believe, think, find, say, agree, regard, consider, my point is, etc.*

The context becomes more personal as a recipient get information about the opinion of an author directly.

For example, “I think the surviving Charlie Hebdo journalists really had no alternative but to show some image of Muhammad on this week's cover”.

### C. Theories, ideas, recommendations

Authors of the media texts strive not only to present some interpretations, but also suggest some new ideas or pieces of advice considering possible solutions to the problem discussed. Up to 29% of the phrases in English and 18% in the Belarusian language have in its structure explicit opinionative means that mark contexts containing some concepts and recommendations. For example, “We must not allow the assassin's veto”.

The most frequent are:

#### 1) Modals

a) modal verb *can* in the meaning “be able” (“It [terrorism] can kill people and damage property”) and “it is needed, it is vital” (“freedom of speech can only be absolute”).

b) modal verbs *should, must*

c) modal verbs *need, have to, ought to*

d) modal structures: *it's important (to do smth)*

#### 2) Question-answer bodies

3) Nouns: *question, problem, idea, sense*

4) Predicates with positive or negative grades

5) Verbs in imperative mood

### D. Expectations about future

22% of phrases in English and 11% in Belarusian contain words and structures by which authors make predictions and forecasts about the topic or it separate aspects.

1) Verbs in the forms of future tenses

2) Oblique mood

3) Some adverbs: *will, would, might, may, perhaps, probably*



For e.g., “There will now be cries from the security services and parliament for more powers and more surveillance”.

## VIII. CLASSIFICATION OF OPINIONS IN THE MEDIA

Intentional and semantic analysis of the opinionative phrases in the texts of the media discourse shows that there are three significant types of opinions. Each type helps to extract a particular information.

### A. *Opinions-interpretations*

These opinionative phrases contain interpretations, comments, descriptions and characteristics of the discussed topic. For instance, “If much can be done from a legal and contractual side without marriage, then marriage loses all credibility”. Accordingly, a recipient is provided with different approaches, explanations and logical interconnections.

### B. *Opinions-theories*

By the phrases of this type authors share ideas, conclusions, possible solutions to the problem, personal notions and concepts. For example, “Love shouldn't be completely unconditional, but it also shouldn't be a gun to the throat” or “Give the poor creatures a break!” Thus a recipient gets some recommendations and ideas which can be used in real life.

### C. *Opinions-prognosis*

Authors can not only analyze facts, situations and trends, but also build their own prognosis and make predictions concerning the probable outcomes, changes, development of the topic or a situation as such. To illustrate this, “England will pay a high price for such arrant selfishness”.

### D. *Comments upon classification*

1) It should be noticed that sometimes it is impossible to divide a phrase into several separate opinionative phrases. That is why a number of contexts should be considered a mixed type (“But it's not just the national newspapers which should be considered in this debate because the blow will land hardest on local newspapers and micro news sites”).

2) A range of phrases of different types may contain judgments in its structure, while judgment is not the purpose of a phrase but only a means to identify separate elements in the whole phrase.

3) The phrases with markers of opinion (that do not imply additional meanings) should be classified according to the type of an opinion expressed in a clause or a separate sentence combined with the marker by sense).

## IX. CONCLUSIONS

The survey allows making several conclusions:

1) Modern mass media serve as reviews and chronicles of the current event, sources of knowledge about the world, source of analytic data, recommendations, prognosis and are able to serve as opinion polls to a certain extent. The media have a unique capability for creation of cognitive trends and are instrumental for manipulation of public opinion.

2) Both informational and analytical genres of the media are saturated with opinionative phrases, especially, with the ones having explicit means of opinion expression in their structure.

3) To extract data from the media it is important to build a particular opinion lexicon, containing words, structures and markers of opinion. Explicit means are divided into several groups: interpretational (logical relations and explanation, reasoning, definition, aim, result),

markers of opinion (neutral by semantics but provide information about point of view of an author), ideas and recommendations (modals, question-answer bodies, particular predicates, etc.) and expectations (verbs in future tenses, particular adverbs, etc.).

4) Opinionative phrases in the media are classified according to the type of information that can be extracted: interpretations, theories, prognoses. In some cases contexts should be considered a mixed type.

5) The results can be useful (as data base, principles and goals) for the algorithms and programs of automatic identification and extraction of opinion.

## REFERENCES

- [1] O.S. Aleksandrova, "Cognitive status of opinion: abstract of dissertation", Moscow, 2001, p.3 (in Russian).
- [2] N.D. Arutunova, "Cosider" and "see" (to the problem of mixed propositional attitudes)", Moscow, 1989, pp. 9 – 15 (in Russian).
- [3] M.R. Zheltuhina, "Political and mass-media discourses: influence – perception – interpretation", Moscow: MAK Press, 2003, p.49 (in Russian).
- [4] E. Werlich, "Problems of text typology", St.-P., 1984, p.8 (in Russian).
- [5] E.V. Akulova, "Genres of the German routine communication", Saratov, 2001, pp. 66 – 67 (in Russian).
- [6] L.A. Mutovkin, "Printed media. Newspaper genres", unpublished (in Russian).
- [7] M.S. Kardumyan, "Linguistic peculiarities of analytical type of mass-media discourse", Stavropol, 2011, p. 5 (in Russian).
- [8] U.D. Apresyan, "System senses "know" and "consider" in the Russian language", Moscow, 2001, p.7 (in Russian).
- [9] M.A. Dmitrovskaya, "Knowledge and opinion: image of the world, image of a person. Logical analysis of a language", Indrik, 2003, p. 15 (in Russian).
- [10] A.A. Zaliznyak, "Polysemi in a language and the means of its representation", Moscow, 2006, p. 476 (in Russian).
- [11] M. Tsytarau, "Survey on Mining Subjective Data on the Web", unpublished.
- [12] A.N. Baranov, "Hidden (implicit) statement in linguistic expertise of a text", unpublished (in Russian).
- [13] I.M. Kobozeva, "Linguistic semantics", Moscow, 2010, p. 214 (in Russian).
- [14] B. Liu, "Sentiment Analysis and Opinion Mining", California: Morgan & Claypool Publishers, 2012, p. 12.



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Testing methods library for applications with web-based interfaces

T. Bragina, G. Tabunshchyk, D. Moroka

**Abstract** — The purpose of this work is to describe the practical implementation of testing methods library for applications with web-based interfaces, which are used for testing at each stage of the software project life cycle, beginning from the prototype, based on an evolutionary prototype creation.

**Keywords** — testing, web-based interface, metrics, application.

## I. INTRODUCTION

The web-project quality is a complex integrated indicator. Quantitative indicators of quality are defined by the introduction of criteria relating to performance, reliability, operation and many other characteristics. However, the most important quality criteria are reliability and compliance with the functional specifications of the project, or, in other words, compliance with specified requirements of the system. One of the effective tools in this case, is the quality control testing, which should be applied at all stages of the project life cycle. There is the problem to make decisions about choosing the necessary tools and metrics for quality control, because of the large selection offered quality control methods, including testing. Therefore, the objectives were as follows:

- To explore the types of web-projects, and identify their characteristics and specific tools for their testing;
- To examine metrics and quality control methods;
- To implement a comprehensive method of testing a web-based interface.

Most attention should be paid to finding bugs early in the project since it is in the early stages the most important decisions are taken. Moreover, with the development of the project at each stage cost of correcting the defect grows. Nevertheless, the initial level of quality control is minimal and it increases with the promotion of the development process and complete testing starts only at later stages. The result is too late error detection, followed by the road processing system or its individual parts.

## II. PROBLEM DEFINITION

Web-projects have its specific problems. First, data and requirements for web-projects often change, stored in various structured and unstructured formats in the system directory, creating a problem of tracking their relevance and accuracy. Second, tests for web-projects should be designed so that they are ready for possible changes in the technologies requirements and overall architecture. All these factors necessitate a well-organized quality control process, which laid the analysis stages and requirements specification for web-application and confirmed them at all stages of life cycle.

T. Bragina, Zaporizhzhya National Technical University, Ukraine, (e-mail: bragina.zntu@gmail.com).  
G. Tabunshchyk, Zaporizhzhya National Technical University, Ukraine, (e-mail: galina.tabunshchik@gmail.com).  
D. Moroka, Zaporizhzhya National Technical University, Ukraine, (e-mail: hotsauce9292@rambler.ru).

A set of test methods depends on the specific structure and purpose of web-projects, but mostly on the result of personal tester experiences. Therefore, the study of types of web-projects is important in order to determine their characteristics and specific tools for their testing [1].

Research scientists in the field of software quality verification touch various aspects of information systems testing and prove the need for a control strategy and implementation methods of testing in the early information systems stages. However, proposed methods do not take into account the internal structure of the web projects and especially the web interface, and there are no formal methods for assessing the functionality of web projects.

On the basis of this, it has been tasked to implement information technology for functional verification web-oriented systems.

### III. INFORMATION TECHNOLOGY FOR FUNCTIONAL VERIFICATION WEB-ORIENTED SYSTEMS

Information technology for functional verification web-oriented systems is presented in Fig. 1 and consists of the following methods and models [2-6]:

- Web-oriented systems verification model;
- The regression testing method for web-oriented systems [1];
- Model assessment of the uncertainty level for the process of information system development [2, 3];
- Model and method of functionality assessing of web-oriented systems in conditions of uncertainty;
- Estimation time reserves method for the software web-oriented systems development based on uncertainty [1, 4];
- Evaluation model of software web-oriented systems functionality, taking into account losses [5].

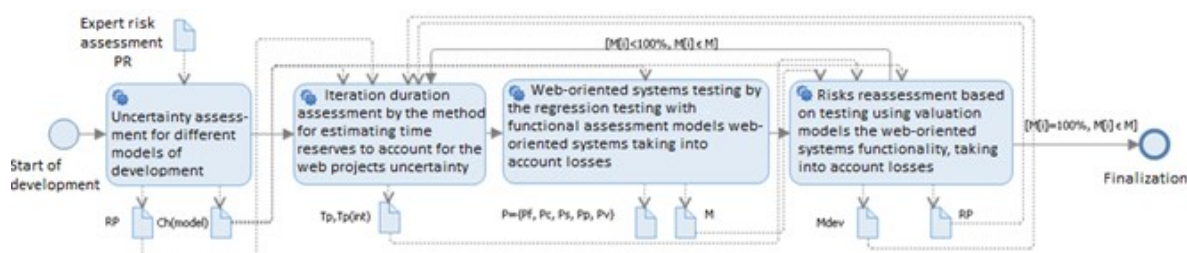


Fig. 1 Information technology for functional verification web-oriented systems

Realized information technology [6] allows estimating the functional suitability of web-oriented systems with the uncertainty of scheduling, and includes the following steps:

*Step 1.* Uncertainty assessment for different models of development:

- Gathering expert estimates the probability of risks occurrence in the development of web-oriented systems;
- Evaluation of the index reduces the efficiency of information system development by the uncertainty level model assessment during the web-oriented systems development, using fuzzy inference;
- Drawing up the initial schedule using the method of time reserves evaluation during developing web-oriented systems for uncertainty.

*Step 2.* Iteration duration assessment by the method of estimating time reserves to account for the web-projects uncertainty.

*Step 3.* Web-oriented systems testing by the regression testing with functional assessment models for web-oriented systems taking into account losses - after the planning stage sprint web-oriented systems implement the tasks assigned to the development of the current sprint, followed by a test of the results. At this stage, automated re-evaluation of input parameters by information are obtained through testing developed in the current sprint artifacts.

*Step 4.* Risk reassessment based on testing, using evaluation models for the web-oriented systems functionality, taking into account losses.

*Step 5.* Completion of the development.

Information technology is implemented as a test methods library that consists of the following modules:

1. "Xenus Report Analyzer" module (for automation of the regression testing method for web-oriented systems, verification model of web-oriented systems, model and method of functionality assessing of web-oriented systems in conditions of uncertainty);
2. "The risk assessment methods library" module (for the automation for the process uncertainty level assessment of information system development and estimation the time reserves method for the software web-oriented systems development based on uncertainty).

The heart of developed testing methods library for applications with web-based interfaces is Xenus Report Analyzer, because to control the quality of web-project manager it is required to process and monitor a large number of quality parameters of the project, including compliance with the developed web-project prototype approved by the customer. We need to use the necessary means testing at each stage of the project life cycle beginning from creating an evolutionary project prototype [1]. Given the large size and complexity of modern web-projects, there is a need for tools that automate and simplify the control of the developed product. On this basis, it was concluded that there is a necessity for implementation of a module for automation of the regression testing method for web-oriented systems.

#### IV. DESCRIBING XENUS REPORT ANALYZER

To assess the compliance of prototype, testing was performed by following metrics: compliance with the number of pages per second and third level of nesting prototype version of the project; number of pages being deeper in the third level of nesting; matching project to the structure prototype. With these metrics, the established test for compliance at every stage is artifact with customer.

Visual C# compiler was selected as the environment development method development environment for complex testing of web-projects, because it significantly reduces the amount of code through an ordered hierarchy of program structures and classes [7].

The Xenus Report Analyzer checks the development of the web-sites basing on the reports of the program Xenu, containing information on the list of accessible web-pages of the site and the links between them [1]. Class ExcelWorker was designed to work with these reports, containing methods for converting a file report by Xenu Microsoft Excel, obtaining the name of the report, downloading the report, a list of pages and their nesting level, a list of all the pages of a unique report.

Tested software showed that the greatest amount of time occupied the Page Map project downloading and determination for nesting pages level, because reports have a large number of rows. Matrix operations forming relationships took the least of all time.

The testing Xenus Report Analyzer is shown in Fig. 2. The left side of the interface matrix shows links between pages, and the right one - all pages of the project and their level of nesting. Here are the calculated metrics conformity assessment and call buttons prototype reference values and conversion metrics at the bottom.

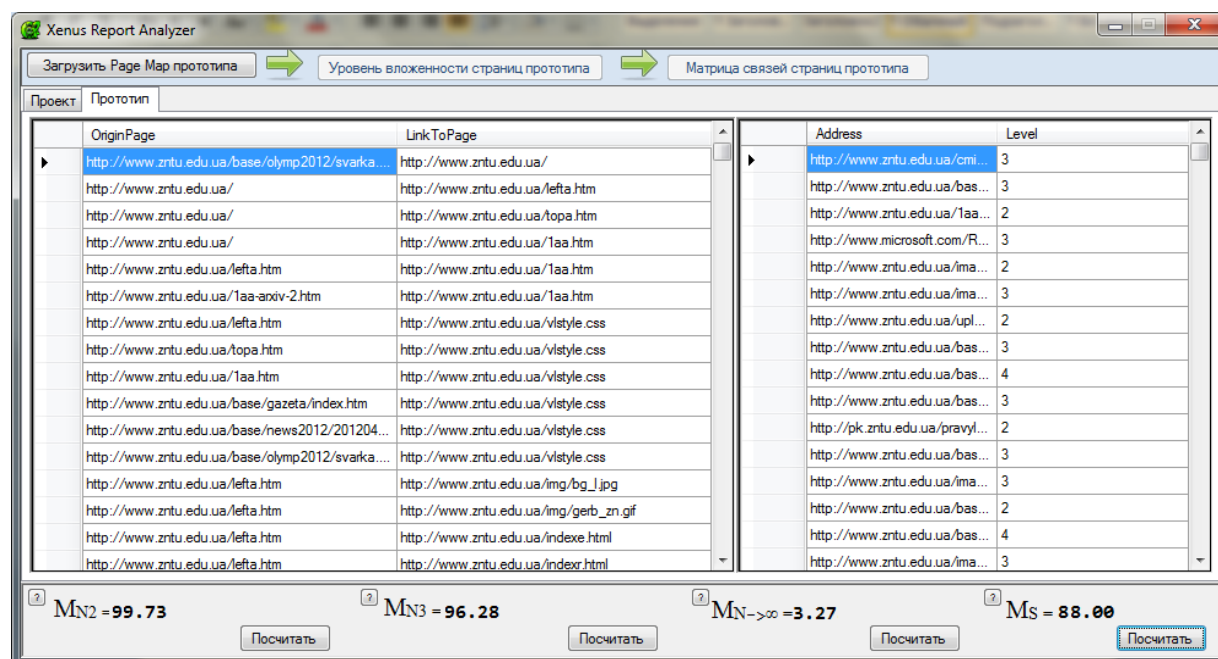


Fig. 2 Testing Xenus Report Analyzer

Value metrics in Fig. 2 shows that the project is in the final stage of development and meets the requirements for the number of pages of different nesting levels.

## V. TESTING XENUS REPORT ANALYZER

To test software four reports were taken. The first two reports contain general information and a map of the project site, and the other two - the prototype. The report contains general information on the project, having 9052 lines, and report on the site map, having 26,949 lines. The report keeps the overall site information prototype contains 429 lines, and report on the site map contains 459 lines.

The tests have shown that the greatest amount of time was occupied by project Page Map downloading, and the determination for the project nesting pages level, because reports have a large number of rows. Matrix operations forming relationships took the least of all time.

TABLE I. TESTING RESULTS

Operation	Approximate time spent on the operation (s)
Download Page Map Project	8.1
Determining the level of the project nesting pages	7.4
The formation of matrix connections between project pages	1.8
Downloading the Page Map prototype	1.7
Determining the prototype nesting pages level	2
Forming links between prototype pages matrix	0.5

## CONCLUSIONS

The library for applications with web-based interface were developed, implementing the developed methods and algorithms of dynamical verification. This library allows to assess the compliance of web-site prototype at each stage of the project life cycle.

The heart of developed library is Xenus Report Analyzer, allowing to gather information and to increase the efficiency of testing by analyzing the structure of the prototype and developed web-applications.

## REFERENCES

- [1] T. Bragina, G. Tabunshchik, "The strategy of testing web - projects", Proceedings of the Donetsk National Technical University. Series "Informatics, Cybernetics and Computer Science" (IKOT 2012). Issue 15 (203), pp. 118-124. (in Russian)
- [2] T. Bragina, G. Tabunshchik, "Comparative analysis of iterative software development models", *Radyoelektronika. Informatics. Management*, Vol. 2, pp. 130 – 139, 2010, (in Russian).
- [3] T. Bragina, G. Tabunshchik, "Analysis of approaches to risk management in software projects with an iterative lifecycle", *Radyoelektronika. Informatics. Management*, Vol. 2, pp. 120-124. 2011 (in Russian).
- [4] T. Bragina, G. Tabunshchik, "Fuzzy analysis of project risk", *Information Processing Systems*. Vol. 3 (93), pp. 15-21, 2011, (in Russian).
- [5] T. Bragina, "Development integration tools", *Information Processing Systems*, Vol. 8 (106), pp. 127-130, 2012, (in Ukrainian).
- [6] T. Bragina, G. Tabunshchik, "Information technology risk-based evaluation of the functionality of web-oriented systems", *Information Processing Systems*. Vol. 2 (118), pp. 245-252. 2014, (in Russian).



Co-funded by the  
Tempus Programme  
of the European Union

This publication is the result of the project implementation:  
TEMPUS CERES: Centers of Excellence for young REsearchers.  
Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES



# Genome-wide association study of prostate cancer

Jurečková Jana, Babušíková Eva, Kliment Ján, Dobrota Dušan

## I. INTRODUCTION

Prostate cancer is the second most frequently diagnosed malignancy and the sixth leading cause of cancer-related death in men. Prostate cancer is multifactorial disease and the major risk factors contributing to its development are age, ethnicity, environmental factors or family history of prostate cancer. Some men with prostate cancer remain asymptomatic with latent prostate carcinoma and their number is still higher than the number of men with clinically detected disease. A better understanding of the genetic and biological mechanisms that determine prostate cancer aggressiveness is needed. The genome-wide association study provides an unbiased screen for genetic loci potentially associated with prostate cancer and may help to identify new candidate genes or targets for intervention.

## II. MATERIALS AND METHODS

Our study population consisted of 48 prostate cancer patients and 46 control subjects. All prostate cancer patients enrolled in this study have Gleason score  $\geq 7$ . Blood samples, collected from all participants, were used for purification of high-quality DNA samples. The Genome-Wide Human SNP Array 6.0 was used for detection of about 900 000 single nucleotide polymorphisms. Statistical analysis of results was performed using PLINK 1.9 software package (<http://pngu.mgh.harvard.edu/purcell/plink/>).

## III. RESULTS

We identified seven single nucleotide polymorphisms to be significantly associated with prostate cancer risk ( $p < 1.0 \times 10^{-8}$ ) (rs12136562, rs41439745, rs9426908, rs12328643, rs9423252, rs1465512, rs2826099). Their association with prostate cancer risk remains significant ( $p < 0,05$ ) also after Bonferroni correction for multiple testing. Three of them are located in the intergenic regions and four of them are located in genes (rs41439745 - LOC101929023, rs9426908 - PRRX1, rs12328643 - REEP1, rs9423252 - ACADSB). The exact functional impact of these polymorphisms is not known. Also direct connection between four genes, in which we detected significant polymorphisms, and prostate cancer is not described in literature. Only PRRX1 is in generally connected with the process of metastases formation, while tumors with high amounts of PRRX1 have better prognosis as they are not able to form metastases.

## IV. CONCLUSIONS

Genome-wide association study represents the important tool for identification of multiple single nucleotide polymorphisms associated with susceptibility to prostate cancer, which could in the future help to predict higher risk of prostate cancer development or worse prognosis in individual patients.

*This work was supported by ICGEB Research Grant CRP/SVK 13-01: „Molecular profiling of cancer: biomarkers of progression, therapeutics and predisposition.”*

J. Jurečková, E. Babušíková, D. Dobrota, Comenius University in Bratislava, Jessenius Faculty of Medicine in Martin, Department of Medical Biochemistry, Martin, Slovakia (e-mail: jana.jurecekova@jfmed.uniba.sk).

J. Kliment, Comenius University in Bratislava, Jessenius Faculty of Medicine in Martin and University Hospital Martin, Department of Urology, Martin, Slovakia.



# A new perspectives in a detection of markers of human diseases in Slovakia

Eva Babusikova, Milos Jesenak, Dusan Dobrota, Jana Jureceková

## I. INTRODUCTION

Nowadays is very popular individualized therapy. This advance is characterized by application of genetic (genome) and molecular and biochemical (RNA, proteins, metabolites) data for exploration disease etiology and pathophysiology. A big challenge is to define persons with a predisposition to any disease and phenotype of a disease as well. Analysis of DNA and single nucleotide polymorphisms is a modern strategy for these aims. Common variable immunodeficiency (CVID) is the most common symptomatic primary immunodeficiency. Symptoms of common variable immunodeficiency are heterogeneous and unspecific therefore its diagnosis and management is challenging. This immune deficiency is considered to be a collection of genetic immune defects with complex inheritance patterns but exact gene modifications are still unknown. Our aim is to find an early marker, to optimize therapy and to increase efficacy of treatment.

## II. METHODS

We decided to use DNA chip array (Genome-Wide Human SNP Array 6.0 – Affymetrix) and we analyzed DNA from patients with common variable immunodeficiency and DNA of healthy persons. We analyzed 906 600 single nucleotide polymorphisms and 946 000 probes for the detection of copy number variation.

## III. RESULTS

We found several significant changes and it is possible that some single nucleotide polymorphisms can be responsible for protection against common variable immunodeficiency or can participate in origin and development of common variable immunodeficiency. We identified 28 221 single nucleotide polymorphisms with  $p$  less than 0.05 and 27 single nucleotide polymorphisms with  $p$  less than  $10^{-5}$  in our small population of patients with common variable immunodeficiency.

## IV. CONCLUSIONS

Common variable immunodeficiency represents a clinical and immunological syndrome that merges various diseases with different genetic roots. It is still problem to diagnose common variable immunodeficiency during childhood and very often patients are without diagnose and an adequate treatment. Using DNA chip array or whole exome sequencing is needed to illuminate the causes of this disease. We found several potential positive and negative markers of common variable immunodeficiency. Additional deeper analysis of determined single nucleotide polymorphisms is necessary.

*This work was supported by VEGA 1/0252/14.*

E. Babušiková, D. Dobrota, J. Jurečeková, Comenius University in Bratislava, Jessenius Faculty of Medicine in Martin, Department of Medical Biochemistry, Martin, Slovakia (e-mail: Eva.Babusikova@jfmed.uniba.sk).  
M. Jesenak, Comenius University in Bratislava, Jessenius Faculty of Medicine in Martin, Pediatrics Clinic, Martin, Slovakia.