by Anhalt University of Applied Sciences and Perm National Research Polytechnic University

Proceedings of the 7th International Conference on Applied Innovations in IT

Volume 7

Issue 1

EDITION
Hochschule Anhalt

Proceedings of the 7th International Conference on Applied Innovations in IT

Volume 7 | Issue 1

Koethen , Germany
6 March 2019

Editors:

Prof. Dr. Eduard Siemens* (editor in chief),
Dr. Bernd Krause*,
Dr. Leonid Mylnikov**

(*Anhalt University of Applied Sciences,
** Perm National Research Polytechnic University)

This volume contains publications of the International Conference on Applied Innova-tions in IT (ICAIIT), which took place in Koethen March 6th 2019. The conference is devoted to problems of applied research in the fields of automation and communications. The research results can be of interest for researchers and development engineers, who deal with theoretical base and the application of the knowledge in the respective areas.

ISBN: 978-3-96057-086-8 (Online)
ISSN: 2199-8876

Additional copies of this publication are available from:
FB6 Anhalt University of Applied Sciences
Postfach 1458
D-06354 Koethen, Germany
Phone: +49 3496 67 2327
Email: eduard.siemens@hs-anhalt.de
Web: http://icaiit.org

Content

Section 1. Communication Technologies

Section 2. Control, Management and Automation

Section 3. Data Analysis and Processing

Evaluation of the Time-Aware Priority Queueing Discipline with Regard to Time-Sensitive Networking in Particular IEEE 802.1Qbv

Manish Kumar, Martin Boehm, Jannis Ohms, Oleksandr Shulha and Olaf Gebauer
Research Group Communication Systems, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302, Wolfenbüttel, Germany
{m.kumar, ma.boehm, jannis.ohms2, o.shulha, ola.gebauer}@ostfalia.de

Keywords: TSN, Industry 4.0, Scheduling, Real-Time, QoS, M2M.

Abstract: Within the evolvement of Industry 4.0, the need for flexible real-time communication technologies emerges. Time-Sensitive Networking enables deterministic communication in IEEE 802 networks. The Time-Sensitive Networking Working Group published a set of standards whose implementation is in progress. IEEE 802.1Qbv standard introduces the concept of Time-Aware Shaping (TAS). TAS enables determinism by dividing traffic in different preconfigured time-slots configured in a Gate-Control-Lists (GCL). Intel recently released a time-aware scheduler based on the IEEE 802.1Qbv standard. This paper investigates Intel's implementation. In order to test the scheduler a testbed is created. The scheduler is configured to filter and group different types of incoming packets in queues. The packets are transmitted in accordance to the configured GCL. Ingress and Egress traffic of the scheduler are analyzed in accordance to the configured time-slots. The results show, that packets which are arriving outside of their respective time-slot are buffered and transmitted in the beginning of their time-slots. It shows that no traffic interfered with the incorrect time-slots.

1 INTRODUCTION

Industry 4.0 is rapidly evolving and raises the need for real-time M2M communication. However, there are already communication technologies for several decades, which provide determinism. PROFINET as an example offers the ability for deterministic networking. One disadvantage of these technologies is the vendor lock-in, which strictly limits the number of supported hardware. Furthermore, reconfigurations of these systems are costly and inefficient.

The Time-Sensitive Networking (TSN) Task Group, established by IEEE, targets these problems by developing standards for flexible deterministic communication in IEEE 802 networks. One of the most important standards for real-time communication besides precise time-synchronization is the scheduling of the network traffic for each device. The IEEE 802.1Qbv (Enhancements for Scheduled Traffic) standard, offers the ability to schedule traffic based on the traffic type [7]. So far, there are theoretical investigations about the scheduler [1].

This paper investigates an implementation of the IEEE 802.1Qbv standard. Chapter 2 gives an overview of the related work. Later, Chapter 3 gives an overview of TSN and its standards. Chapter 4 explains the IEEE 802.1Qbv standard in more detail. Chapter 5 presents a test-setup to test the timing-characteristics of the Ingress and Egress traffic for the scheduler. The discussion of the results takes place in Chapter 6. Chapter 7 wraps up the topic and exposes gaps and questions for future work.

2 RELATED WORK

J. Vila-Carbó et al. [2] show how to use queuing disciplines to provide bandwidth limitation for traffic classes in real-time communications. Their results show, that it is possible to avoid collisions between traffic classes and reduce delays about 30%. This reduction of delay does not provide the determinism required for industrial automation.

Craciunas et al. [3] analyze the algorithmic complexity of automated configuration synthesis for

TSN. Their results show, that the problem is NP complex. This opens the need for efficient heuristics.

In order to provide an auto configuration mechanism, all updates need to be timely synchronized to avoid unwanted network configuration states. This problem is addressed by Mizrahi et al. [4].

In [5], Gutiérrez et al. propose MQPRIO as a queueing discipline for the Linux operating system. Their work focuses on the use of a patched Linux kernel, called Real-time Preemption patch (PREEMPT-RT), for real-time communication in robotic applications.

In order to communicate between a TSN and other communication systems, Böhm et al. [6] propose an architectural design for a gateway which connects TSN with Software-Defined Networking resp. OpenFlow. The gateway forwards packets between the networks while preserving the real-time capabilities of the TSN network. The scheduler examined in this paper is one of the main components described in their requirements.

This paper investigates an implementation of the IEEE 802.1Qbv standard for "Enhancements for Scheduled Traffic" [7]. The functionalities and preciseness of the time-aware scheduler will be examined.

3 TIME SENSITIVE NETWORKING

In 2012, the Audio/Video Bridging Task Group which developed standards for time-synchronized low latency streaming services, renamed themselves to Time-Sensitive Networking Task Group. They focus on real-time communication through IEEE 802 networks and provide a set of standards. Depending on the case of application, multiple standards can be used together.

The combination of time synchronization (IEEE 802.1AS-Rev - Timing and Synchronization for Time-Sensitive Applications) and time-aware traffic shaping (IEEE 802.1Qbv - Enhancements for Scheduled Traffic) enables determinism, by processing different traffic types in their respective time-slots. Guard bands block time before each time slice to prevent conflicts of overlapping packets. Due to the waste of time for each guard band, where no traffic is allowed to be transmitted, they introduced Frame Preemption (IEEE802.1 Qbu - Frame Preemption) [10] to reduce the size of guard bands. This standard enables transmission of frames to be interrupted and later resumed.

TSN also offers the ability to reconfigure all network devices dynamically. As shown in Figure 1, end devices can request the Centralized User Configuration (CUC) for their specific deterministic communication flow including packet size, frequency etc. (IEEE 802.1Qcc - Stream Reservation Protocol (SRP)) [8]. The Centralized Network

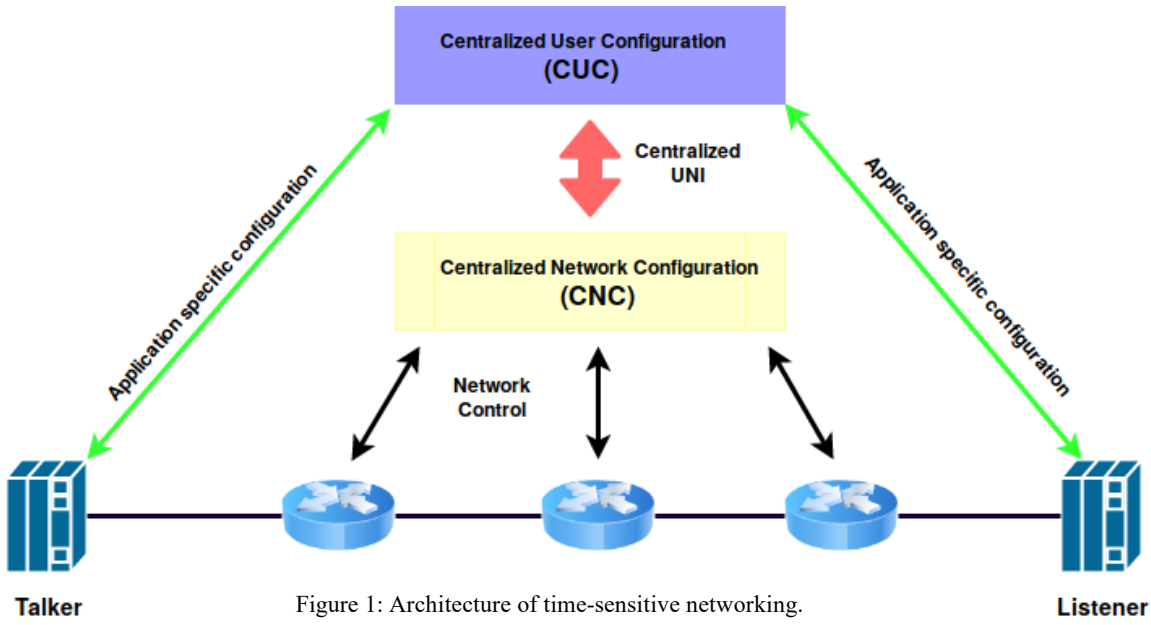

Figure 1: Architecture of time-sensitive networking.

Configuration (CNC) on the other side, which has a global view on the network, calculates configurations including time-slots size, VLAN to map traffic etc. Flow requests can also be rejected if resources are not sufficient. All devices get updated to the new configuration (IEEE 802.1Qcp - YANG Data Model) [9].

The majority of standards have been published. A few standards are not finalized yet. The next chapter introduces the IEEE 802.1Qbv TAS in detail.

4 TIME-AWARE SHAPING

The basic architecture of IEEE 802.1Qbv is visualized in Figure 2. The time-aware shaper consists of 1 to 8 queues. Each queue has a transmission selection algorithm which selects the next packet transmitted from the queue. This can be a queueing disciplines like first-in, first-out (FIFO) or token bucket filter (TBF). The state of a gate can either be open or close, only open gates transmit packets. A time-aware gate opens and closes according to its configured time. The schedule of the gate states is specified in the GCL. Each entry of the list consists of a set of gate states and their duration. The entries are repeated cyclically. Furthermore, the standard requires another parameter called base-time which specifies when the execution of the GCL starts. This parameter is used to assure that each device in a TSN network starts their schedule at the same time to avoid delays.

One open source implementation, developed by Intel, for the IEEE 802.1Qbv is called Time-Aware Priority (TAPRIO). It is a classful queueing discipline for the Linux Kernel. It timely opens and closes gates respective to the current entry of the GCL.

The next chapter shows the test-setup to test the functionality of the scheduler. Test-cases are presented.

5 TEST-SETUP AND TEST-CASES

This chapter introduces a test-setup as well as test-cases for TAPRIO, an implementation of the IEEE 802.1Qbv standard.

The test-setup is visualized in Figure . All systems are based on Ubuntu 18.04. TAPRIO is not part of the mainline Linux kernel. A custom version of the Kernel (GNU/Linux 4.19.0-rc5 x86_64) has been compiled. The bridge is equipped with one Intel i210 Ethernet controller which provides hardware offloading and precise timestamping for the TAPRIO queueing discipline. The i210 controller is used as an egress port. The ingress and egress traffic of the bridge is mirrored by two hardware network taps and recorded in a measurement system using Wireshark. Based on the recorded traffic the bridge delay and the time

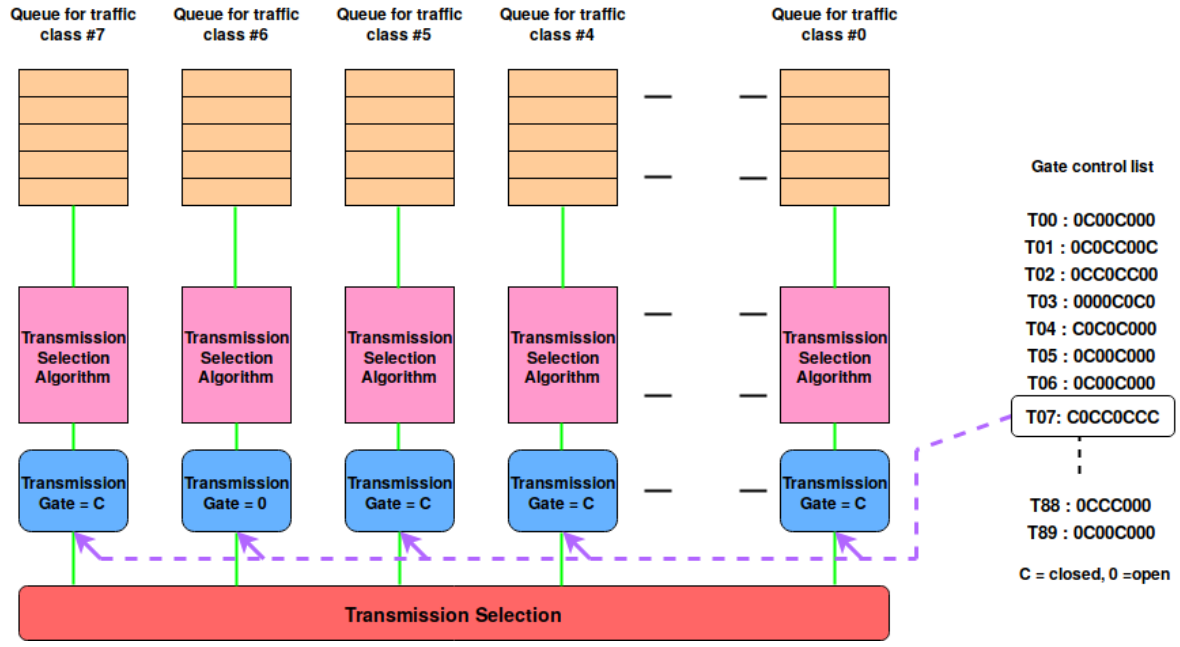

Figure 2: Selection for transmission-gates.

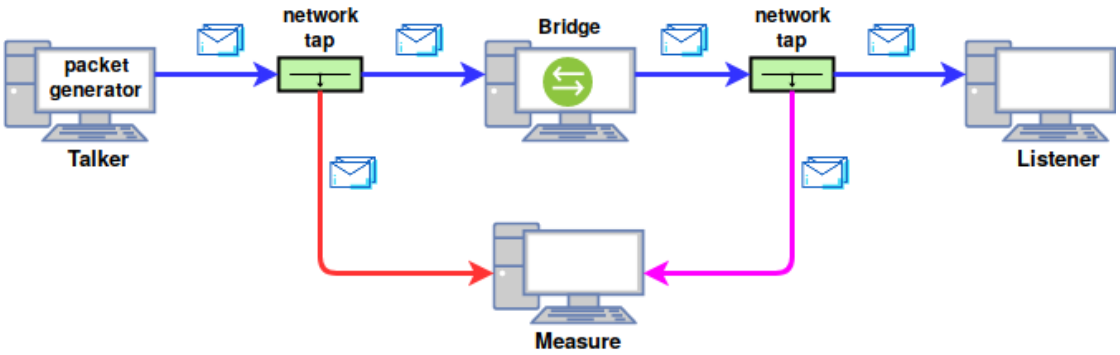

Figure 3: Test-setup for time-aware scheduling.

distribution of ingress/egress traffic is calculated. It has to be noted, that no real-time Linux like Linux RT or Industrial Linux has been used for the setup.

The Talker creates two UDP packets every 10 milliseconds. The packets use two different IP TOS values: 0x00 and 0x08. Each packet contains a unique number as payload for later mapping. The i210 controller supports four transmission queues while the test-cases use three time-slots for three different traffic classes. The configuration of TAPRIO, which is shown in Figure 4, consists of three queues Q_0 - Q_2. Each queue is opened for 30 milliseconds starting with Q_0. Traffic with a TOS value of 0x00 is assigned to Q_2 and traffic with a TOS value of 0x08 is assigned to Q_1. No traffic is assigned to Q_0, to make sure, that no traffic interferes with this queue and the respective slot always remains empty. The base-time parameter, which represents the starting point of the schedule, is configured to be a point in time in the near future.

```
qdisc replace dev eth0 parent root handle 100 taprio
    num_tc 3
    map 2 2 1 2 0 2 2 2 2 2 2 2 2 2 2 2
    queues 1@0 1@1 2@2
    base-time $BASE_TIME
    sched-entry S 01 30000000
    sched-entry S 02 30000000
    sched-entry S 04 30000000
    clockid CLOCK_REALTIME
```

Figure 4: Configuration of the time-aware priority queueing discipline.

The setup helps to determine if the timing characteristics of the GCL parameters of the TAPRIO queueing disciplines works properly. It measures packets before and after they are scheduled to calculate the delay between both points. This shows the distribution of processing plus waiting time for each packet. The captured egress packets show, if they are transmitted in their respective time-slots.

6 DISCUSSION OF THE RESULTS

This chapter discusses the results of the test-cases presented in the previous chapter.

The ingress time of each packet is visualized in Figure 5. Each colored dot represents a packet. Two packets enter the bridge every ten milliseconds with two different TOS fields. The x-axis shows the unique packet number of each packet while the y-axis shows the arrival time of the packets. It shows, that there is a linear rise of unscheduled incoming packets over time.

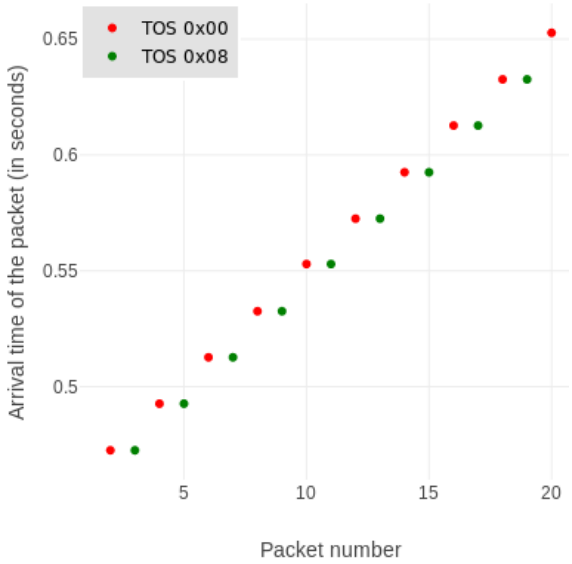

Figure 5: Ingress traffic with two TOS felds over time.

Figure 6 shows the egress time of each packet. The axes are the same as in Figure 5. In the graph, each horizontal line represents the start of a new time-slot. The grey highlighted region represents an unused time-slot.

Both, the green area as well as the red area, always start sending packet immediately after their time-slot starts. The traffic, which is not sent directly in the time-slot, is buffered and sent when the next allocated time-slot starts. In the graph, linear sequences of packets show the transmission of buffered packets. Packets which arrive in their respective time-slot are transferred directly, which can be seen between the horizontal lines. The graph also shows that no packet gets transmitted outside of the appropriate time-slots.

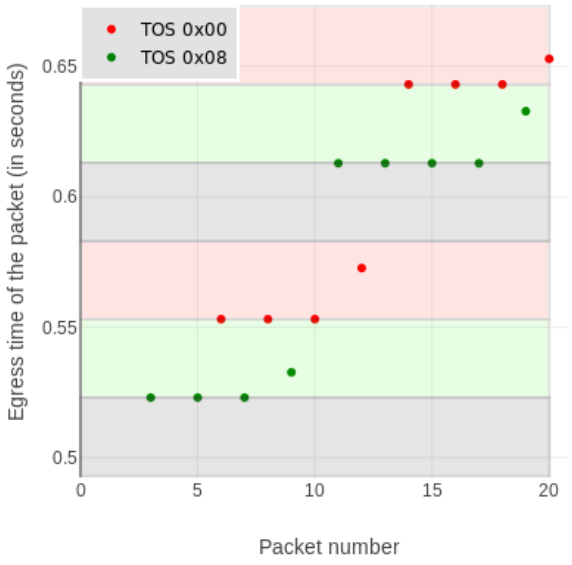

Figure 6: TOS based prioritized egress scheduled traffic with 3 time-slots.

The difference between egress and ingress time is visualized in Figure 7. It shows the amount of time each packet spends in the scheduled bridge. Since there are three slots with 30 milliseconds time-window each, the upper bound delay of the bridge should be maximum 60 milliseconds due to a maximum buffer time of two slots. This assumption can be validated by Figure 7.

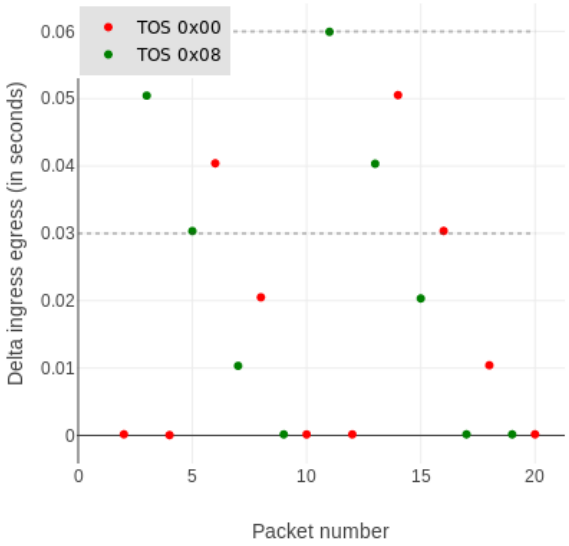

Figure 7: Distribution of the delta (difference between ingress- and egress-time of the traffic).

7 CONCLUSION AND FUTURE WORK

The results presented in this paper show, that the implementation of the IEEE 802.1Qbv scheduler works as expected. This has been validated with a test-setup by analyzing the traffic before and after is has been scheduled. It shows, that different traffic classes never interfere with the wrong time-slots. It also shows that packets, which arrived outside of their time-slots are buffered and processed as soon as their time-slots begin.

There are a few things which should be noted. Real-time traffic should be sent in their respective time-slot to avoid waiting time in the buffer. In this case, the Talker needs to be time-aware too. Due to the usage of a buffer, there is a possibility that the buffer is overfilled with packets. This can be a huge security risk.

Another topic, which is not mentioned in the standard is the assignment of queues for time-synchronization traffic. The impact of scheduled time-synchronization packets should be investigated in the future.

ACKNOWLEDGMENTS

This work was partly funded by the Ministry for Science and Culture of Lower Saxony as a part of the research project SecuRIn (VWZN3224) and the

Federal Ministry for Education and Research within the KMU-innovativ program as a part of MONAT (16KIS0782).

REFERENCES

[1] N. G. Nayak, F. Dürr, and K. Rothermel, "Routing algorithms for IEEE802.1Qbv networks," ACM SIGBED Review, no. 15(3), pp. 13-18, 2018.

[2] J. Vila-Carbó, J. Tur-Masanet, and E. Hernandez-Orallo, "An evaluation of switched ethernet and Linux traffic control for real-time transmission," IEEE International Conference on Emerging Technologies and Factory Automation, 2008, pp. 400-407.

[3] S. S. Craciunas, R. S. Oliver, M. Chmelík, and W. Steiner, "Scheduling real-time communication in IEEE 802.1 Qbv time sensitive networks," In Proceedings of the 24th International Conference on Real-Time Networks and Systems, ACM, 2016, pp. 183-192.

[4] T. Mizrahi, E. Saat, and Y. Moses, "Timed consistent network updates in software-defined networks," IEEE/ACM Transactions on Networking, no. 24(6), pp. 3412-3425, 2016.

[5] C. S. V. Gutiérrez, L. U. S. Juan, I. Z. Ugarte, and V. M. Vilches, "Real-time Linux communications: an evaluation of the Linux communication stack for real-time robotic applications," arXiv preprint arXiv: 1808.10821, 2018.

[6] M. Böhm, J. Ohms, O. Gebauer, and D. Wermser, "Architectural design of a TSN to SDN gateway in the context of industry 4.0," 23. VDE/ITG Fachtagung Mobilkommunikation - Technologien und Anwendungen, 2018.

[7] IEEE 802.1Qbv, IEEE Standard for Local and metropolitan area networks - Bridges and Bridged Networks - Amendment 25: Enhancements for Scheduled Traffic, 2016.

[8] IEEE 802.1Qcc, IEEE Draft Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks Amendment: Stream Reservation Protocol (SRP) Enhancements and Performance Improvements, 2017.

[9] IEEE 802.1Qcp, IEEE Standard for Local and metropolitan area networks - Bridges and Bridged Networks - Amendment 30: YANG Data Model, 2018.

[10] IEEE 802.1Qbu, IEEE Standard for Local and metropolitan area networks - Bridges and Bridged Networks - Amendment 26: Frame Preemption, 2016.

Adopting Minimum Spanning Tree Algorithm for Application-Layer Reliable Mutlicast in Global Mutli-Gigabit Networks

Kirill Karpov[1], Dmitry Kachan[1], Nikolai Mareev[1], Veronika Kirova[1], Dmytro Syzov[1], Eduard Siemens[1] and Viatcheslav Shuvalov[2]

[1]*Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany*
[2]*Department of Transmission of Discrete Data and Metrology,*
Siberian State University of Telecommunications and Information Sciences, Kirova Str. 86, Novosibirsk, Russia
{kirill.karpov, dmitry.kachan, nikolai.mareev, dmytro.syzov, veronika.kirova, eduard.siemens}@hs-anhalt.de,
shvp04@mail.ru

Keywords: Application Layer Multicast, Point-to-Multipoint, RMDT, Cascaded Data Transmission, Minimum Spanning Tree, DCMST, Networking, High Bandwidth.

Abstract: Data transmission over the Wide Area Networks (WAN) is a common practice in nowadays Internet, however, it has its limitations. One of them is that IP multicast data transmission rarely can be applied outside of Local Area Networks (LAN). Due to its vulnerability, multicast traffic is blocked by most Internet Service Providers' (ISP) edge equipment. To overcome this limitation, an Application Layer Multicast (ALM) is proposed, where multicast functionality is implemented on the end-hosts, instead of network equipment. For the application of ALM no changes in the network are needed, what significantly facilitate deployment of multicast services. The key point of this work is to implement ALM for reliable high-speed data transmission over WANs using RMDT transport protocol and Minimum Spanning Tree (MST) algorithm, which shall improve bandwidth utilization and provide a higher data rate for data propagation across multiple sites.

1 INTRODUCTION

Transmission of big data chunks over WANs to multiple sides can be implemented using point-to-point approach – when sender host simply initiates data transmission to several destinations in parallel or one by one in the queue. In the first case, data flows share the same link, at least on sender's last mile, and the TCP protocol doesn't share the network resources evenly [5]. Moreover, higher usage of shared bandwidth in that case means lower bandwidth per individual connection. Another approach will provide entire available bandwidth for the receivers, however, each of them will receive data only in its turn. Both solutions will cause unnecessary use of bandwidth since each data set will be sent separately to each receiver.

Usually, LAN connections and connections on short distances have higher bandwidth, fewer impairments and lower latency. In contrary, WAN connections have cross traffic between the data endpoints, many intermediate network devices which cause additional network impairments and a higher level of latency. Moreover, bandwidth in WAN connections has usually a higher price than in LANS.

Using an MST algorithm, it is possible to employ metrics, which will evaluate connections between involved hosts to create optimal ALM topology for data propagation, which will send data over LANs and short distance connections in parallel, and send data in an ad-hoc manner over WAN links. To get the benefit of multicast service, the ad-hoc connection will not completely receive the data before forwarding it further, instead, it will pass it to the next host alongside with confirmation of successful reception of each consequent data chunk. This allows usage of e.g. file-based video transmission, when all users may start to process the file without waiting till end of data transmission.

To make high-speed data transmissions over WAN possible, the RMDT [2] transport protocol was used, since it satisfies all necessary conditions described above:

- The protocol provides WAN acceleration service, which makes network impairments and latency up to 1 second nearly negligible.
- It can serve up to 10 receivers in parallel within a single session natively – means no fairness issues will be among receivers and available bandwidth will be shared evenly. Moreover, it has a centralized congestion control, which allows the coexistence with the cross traffic in IP WANs.
- RMDT is a pure user-space software library which makes it possible to create network applications capable of forwarding the received data chunks further to the next receiver.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 describes the developed application algorithm. Section 4 is devoted to the experimental setup, describes hardware and software equipment, testing environment, measurement and representation methods for the current research. The results of conducted experiments is presented in Section 5. Finally, conclusion discusses the results, followed by the future work.

2 RELATED WORK

A detailed survey of existing tree-based application layer multicast algorithms has been made by Computing Laboratory, University of Kent [6]. However, the efficiency of observed protocols have been investigated only in terms of tree cost and delay optimization.

S. Banerjee and B. Bhattacharjee [7] have also analyzed various application layer multicast algorithms and determine the fields of applicability for them. They substantiated that tree-first application layer multicast approach is useful for high-bandwidth data transfers, however it is less suited for real-time applications.

The Narada performance study [8] provides several useful performance metrics such as latency, bandwidth, stress, resource usage, etc.

The given paper describes and study performance of application layer multicast in combination with high-bandwidth data transport applications.

3 ALGORITHM DESCRIPTION

The key part of developed application layer multicast system is minimum spanning tree algorithm, which is supposed to construct an optimal tree, based on the chosen metrics. In the given research, RTT is the optimization metric. It has been chosen, because RTT is one of the basic characteristics of a network, which is easy to obtain, unlike the available bandwidth, which might cause undesirable effects to the network operation while being measured.

The given ALM realization uses tree-first approach, therefore the first step of application workflow is to recognize the network environment among all hosts which are involved in transmission process. With the chosen metrics the protocol forms an adjacency matrix. This matrix represents a complete directed weighted graph, where the weights are the values of the chosen metric e.g. RTT, available bandwidth, air distance, etc. The result of MST operation will be an adjacency matrix with zeroed non-optimal paths which represents the optimal spanning tree without loops.

The given tree is the directive map for multipoint transmission application, in this case – Data Clone a point-to-multipoint data copy application based on RMDT.

4 EXPERIMENTAL SETUP

4.1 Testing Environment

As an experimental environment, Amazon AWS has been chosen. It provides virtual infrastructure in selected continents and regions. Cascade network transmission infrastructure based on c5.xlarge virtual instance with an Ubuntu 18.04 operating system, 4 vCPU, 8 Gb RAM, and up to 10 Gbps available network access bandwidth have been choosen.

In order to minimize disk I/O operations overhead of getting data from the disk storage, a RAM disk as data storage has been configured on each host.

The instances are distributed all over the world in the following AWS regions: US West (Oregon), EU (Frankfurt), EU (London), Asia Pacific (Singapore), Canada (Central). Using a geo IP service, it has been found out that the hosts from Canada (Central) region are located in Montreal, and the US West (Oregon) data center is located in Boardman. In each

region 3 c5.xlarge virtual instances have been deployed. The regions have been chosen to get different variations of network conditions, such as long and short distances, international and intercontinental links. The air distances between AWS data center locations are shown in Table 1. The minimum spanning tree obtained based on the air distances between AWS regions is shown on Figure 1.

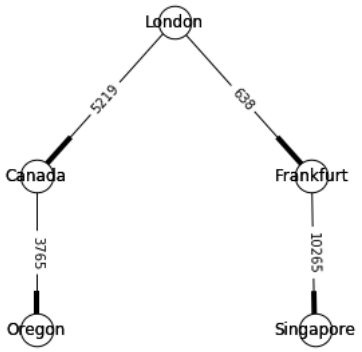

Figure 1: The tree generated by minimum spanning tree algorithm based on the air distance between AWS instance locations.

Table 1: Air distance between virtual instances locations in kilometers.

AWS regions	Oregon	Frank-furt	Lon-don	Singa-pore	Cana-da
Oregon	0	8393	7906	13094	3765
Frankfurt	8393	0	638	10265	5842
London	7906	638	0	10854	5219
Singapore	13094	10265	10854	0	14803
Canada	3765	5842	5219	14803	0

4.2 Software Equipment

For the experiments, the following software and technologies have been used.

1) **Dataclone** – RMDT-based software, which provides point to multipoint data transport functionality [2]. It uses BQL congestion control [3] which is tolerant to big delays and dramatic packet loss rates. In the experiments it will allocate 100 MB of RAM for both send and receive buffers.
2) **Multipoint sender** – a TCP-based application, developed by us to implement point-to-multipoint data transport capability and cascading functionality as Dataclone is doing. It has been created as TCP reference in multipoint field. It uses different threads for simultaneous multi-destination transmission and barrier type of synchronization.

5 EXPERIMENTAL RESULTS

As mentioned in Section 3, the first step of tree-first application layer multicast is the investigation of the given network environment. The result of RTT measurements between the deployed AWS regions is shown in Table 2.

Table 2: Packet RTT delays between virtual instances, in milliseconds.

AWS regions	Ore-gon	Frank-furt	Lon-don	Singa-pore	Canada
Oregon	0	100	141	162	65
Frankfurt	100	0	13	174	100
London	141	13	0	173	87
Singapore	162	174	173	0	219
Canada	65	100	87	219	0

As can be seen from the tables, the RTT values between regions are corresponding to distance metrics, however, the dependency between delay and distance is not linear due to the physical network paths [9] and other factors, such as cross-traffic, configurations and types of intermediate devices, their number, etc.

Based on obtained metrics, the minimum spanning tree for the given set of hosts has been constructed. It is shown on Figure 2.

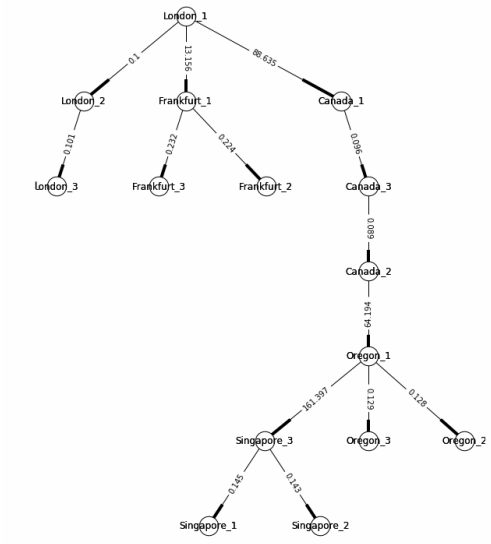

Figure 2: The tree generated by minimum spanning tree algorithm in the AWS network cloud infrastructure using RTT as weights (in milliseconds).

The multipoint TCP realization, which has been described in Section 4.2 produced a constant data rate during all transmission time, which was about 76 Mbps for each edge of the tree.

With Dataclone application as a carrier, the data rates are distributed less uniformly across the transmission tree, in comparison with the TCP multipoint experiment, as shown on Figure 3.

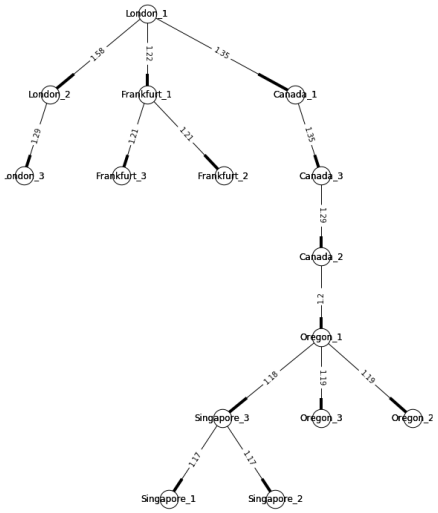

Figure 3: The data transmission tree with average data rates with edges weighted in Gbps.

During the test that lasted 145 seconds, 20 GB of data has been transmitted. The average data rate for the whole tree was 1.14 Gbps, which is 15 times higher than during TCP multipoint experiment. Hereby the average data rate was calculated as the size of the transmitted data chung (20 GB) divided by runtime of the root sender at London_1 and so is slighthly less than the lowest data rate at the graph on figure 3 London_1 is the node with the highest outbound traffic of 4.15 Gbps in total. The average link bandwidth across the edges was 1.28 Gbps.

More detailed result of the experiment is shown on Figure 4. Data sets with rates were processed with Savitzky-Golay Filter to get rid from the outliers.

The plot shows that, starting from 20 seconds, data rates on all long links after London region were stabilized near 1.2 Gbps. This value can be used for stable multipoint streaming.

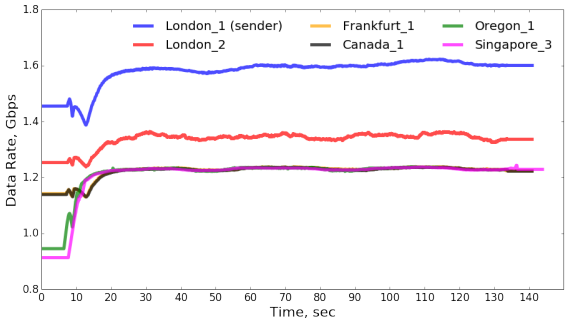

Figure 4: Data rates from the nodes of the tree.

6 CONCLUSIONS

Several conclusions can be drawn from the conducted experiment:

1) The alternative way for data transmission, e.g., via flat, non-hierarchical connection topology requires about 16 Gbps outbound link bandwidth for sender, with 100% network resource utilization. The experimental results show that the hierarchical scenario achieves the same performance with only 5 Gbps of link bandwidth at the maximum loaded sender.
2) RMDT with its BQL congestion control is able to serve in both flat (pure point-to-multipoint) as well as in hierarchical (tree-based) connection topologies, and achieves much higher bandwidth per link utilization in the latter case.
3) Despite the fact that ISP providers have fat pipes, e.g 10 Gbps, it is not always possible to fit into such limit due to variety of obstacles, such as server hardware or virtual configurations limits. However, with WAN acceleration it is possible to overcome most of these limits.
4) RMDT with BQL congestion control shows 15 times higher data rate, than a comparable TCP-based multipoint data transmission realization.

There are a lot of ways to further improve the ALM approach for data transmission. The future steps towards the improvement of the current approach might be the following:

1) To change MST algorithm, which currently does not consider maximum output number limitations of the sender. The alternative to MST could be degree-constrained minimum spanning tree (DCMST) algorithm [10].
2) Using additional metrics for tree nodes. MST algorithm does not consider the performance of

the node and decides to split or to make a chain of nodes only on edges metrics. Thus, a question remains open: what kind of local topology gives the best performance? There is room for investigation of that question in the future.

3) Another weak point of RTT-based MST is that it does not take into account the lower layer infrastructure. Considering this circumstance, it makes sense to build a transition tree, based on available L4 infrastructure.
4) The public network is always changing due to variety of factors, such as cross traffic from other customers, their activity in specific time and date, and so on. Tracking the history of such events and analyzing their effect to network conditions might be helpful for WAN acceleration applications.

ACKNOWLEDGMENTS

This work has been funded by Volkswagen Foundation for trilateral partnership between scholars and scientists from Ukraine, Russia and Germany within the CloudBDT project: “Algorithms and Methods for Big Data Transport in Cloud Environments”.

REFERENCES

[1] V. Kirova, E. Siemens, D. Kachan, O. Vasylenko, and K. Karpov, “Optimization of Probe Train Size for Available Bandwidth Estimation in High-speed Networks,” in MATEC Web of Conferences, vol. 208, p. 02001, 2018.

[2] A. V. Bakharev, E. Siemens, and V. P. Shuvalov, “Analysis of performance issues in point-to-multipoint data transport for big data,” in 2014 12th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE), 2014, pp. 431–441.

[3] N. Mareev, D. Kachan, K. Karpov, D. Syzov, E. Siemens, and Y. Babich, “Efficiency of a PID-based Congestion Control for High-speed IP-networks,” in Titel: Proceedings of the 6th International Conference on Applied Innovations in IT, 2018.

[4] M. Hock, R. Bless, and M. Zitterbart, “Experimental evaluation of BBR congestion control,” in 2017 IEEE 25th International Conference on Network Protocols (ICNP), 2017, pp. 1-10.

[5] R. L. Graham and P. Hell, “On the history of the minimum spanning tree problem,” Annals of the History of Computing, vol. 7, no. 1, pp. 43-57, 1985.

[6] S. Tan, G. Waters, and J. Crawford, “A survey and performance evaluation of scalable tree-based application layer multicast protocols,” 2003.

[7] S. Banerjee and B. Bhattacharjee, “A comparative study of application layer multicast protocols,” Network, vol. 4, no. 3, 2002.

[8] Y. Chu, S. G. Rao, S. Seshan, and H. Zhang, “A case for end system multicast,” IEEE Journal on Selected Areas in Communications, vol. 20, no. 8, pp. 1456-1471, Oct, 2002.

[9] J.-M. Beaufils, “How Do Submarine Networks Web the World?,” Optical Fiber Technology, vol. 6, no. 1, pp. 15-32, Jan, 2000.

[10] S. C. Narula and C. A. Ho, “Degree-constrained minimum spanning tree,” Computers & Operations Research, vol. 7, no. 4, pp. 239-249, 1980.

Custom UDP-Based Transport Protocol Implementation over DPDK

Dmytro Syzov, Dmitry Kachan, Kirill Karpov, Nikolai Mareev and Eduard Siemens
Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany
{dmytro.syzov, dmitry.kachan, kirill.karpov, nikolai.mareev, eduard.siemens}@hs-anhalt.de

Keywords: High-Speed Data Transport, Packet Processing, User Space.

Abstract: As industry of information technologies evolves, demand for high speed data transmission steadily increases. The need in it can be found in variety of different industries – from entertainment (trend for increasing resolution of video-cast for example) to scientific research. However, there are several problems that hinder network application capabilities. One of them is slow packet processing due to significant overheads on system calls for simple network operations. There are hardware solutions, but from the economical point of view, using legacy equimpent is preferable due to high cost of updating network infrastructure. Thus, software solutions to these problems can be preferable. One of them is DPDK toolset which gives the ability to tailor network operations to the application. RMDT is a custom transport protocol aimed at high speed data transmission over lossy networks with high latency. The protocol is built over standard Linux UDP sockets. Thus it is heavily reliant on the networking stack performance. The goal of this work is to improve RMDT performance by means of DPDK in a 10G network and to assess the benefits of such an implementation.

1 INTRODUCTION

The nature of network operations on Linux OS, with overheads on system calls, several memory copies during recv() and send(), results in a low performance in cases of high speed connections. While there are multiple solutions to increasing packet processing rate, for example Receive Packet Steering [1], but they are usually aimed at TCP optimization, or maintaining many low-rate connections. If there is a need in high speed transmission and TCP is not fitting for cases of high latency and lossy network, out-of-the-box options are limited. Their functionality is also often dependent on a specific implementation by the manufacturer, thus making development of a widely applicable network utilities more expensive and harder to maintain.

DPDK [2] toolset aims at boosting packet processing performance by giving developers access to a low-level management of network stack. One of the main benefits is the avoidance of user space to kernel space switches. However it does not provide transmission protocols to use out-of-the-box.

The common bottleneck is receive performance, as in case of standard Linux network operations, packets have to go through multiple memory copy operations and additional management operations necessary for the correct delivery to applications. In case of DPDK, there is an opportunity to tailor these operations specifically to the application. Having more control over timings of various send- and receive- related operations can improve latency, improving performance in use cases such as streaming. Also such control can deliver more precise measurements of round trip time, which consequently can improve behavior of congestion control as standard kernel method can introduce fluctuations in the overall time of an operation.

This work attempts to adapt internal structure of the RMDT [3] protocol to DPDK library and to asses the benefits of DPDK over standard Linux approach. At this stage, the goal is to create a simple RMDT over DPDK implementation to test the possibility of improving its' perormance with DPDK. As the main measure of the efficiency in our tests we are using the achievable data rate. Comparison between synthetic packet generation tests and RMDT tests can show the difference in ratio of time spent on network operations to time spent on custom protocol functionality, allowing an assessment of the necessity to improve the implementation. Thus a simple test of a clean send and receive is to be performed as well.

2 RELATED WORK

As DPDK is a generally applicable network development kit, there is a large amount of projects implementing DPDK for a variety of goals. These include using DPDK to build a light-weight TCP/IP stack to achieve better efficiency with resource limited systems [4] as presented in a paper by R. Rajesh et al., building a high performance software router [5] as presented in a paper by Z. Li. M. Miao et al. developed and tested a self-tuning packet I/O aimed at dynamic optimization of data rate and latency by controlling a batch size in a high throughput network system [6]. As can be seen, an improvement in networking operations is in demand by different types of applications.

3 TESTBED DESCRIPTION

All tests have been performed in 10 GE Laboratory of Future Internet Lab Anhalt [3]. The core element here is the WAN emulator Netropy 10G [7] that can be used to create an emulation of WAN links with various impairments like packet losees, delay, reordering etc. It collects data regarding data passed through it and is used in this work to assess the resulting performance.

Servers, which are used in tests have following characteristics:

- Kernel: 4.15.0-45-lowlatency.
- NIC: 82599ES 10-Gigabit SFI/SFP+ by Intel Corporation.
- Memory: 64 GB DDR3.
- CPU: 2xIntel Xeon E5-2643 v4, 3.40GHz.

Software consists of two RMDT builds and two synthetic tests with pure packet generation and reception. Builds are for standard Linux networking stack and DPDK respectively. For an interface to UDP over DPDK, an already existing software was used – F-Stack [8].

4 SOFTWARE DESIGN

On Figure 1 the flow chart of a basic DPDK receiver functionality test is presented.

Here, the overall loop includes a basic, F-Stack provided, polling interface, which is derived from DPDK's own polling mechanisms. Apart from basic functionality, necessary for receiving packets via DPDK, additional checks are added to assure that data is received correctly. This functionality is put in the "Corruption check" box. A test for a sender is the same, but without polling for EPOLLIN.

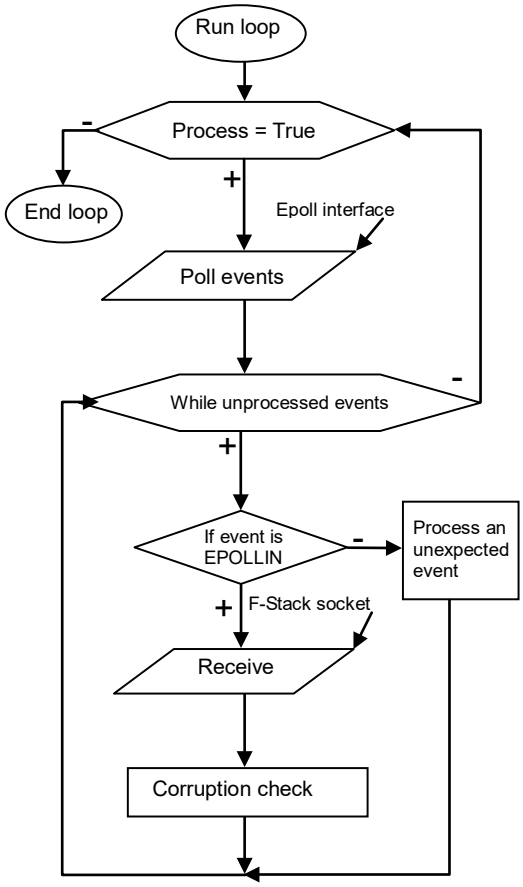

Figure 1: DPDK receive loop.

On Figure 2 the basic RMDT structure for the receiver side is presented.

Here, "Receive handler" is tasked with receive and some basic processing for both user data and service packets. Rest of the protocol functionality is put into "Transport control functionality" box. That includes tasks regarding sending service packets. However, sender functionality is not the aim of this work as it does not bottleneck RMDTs' overall performance in a point-to-point configuration with MTU of 1500 bytes (Ethernet standard [9]) and F-Stack is not optimized for the send process. Both parts work concurrently with the memory buffer and all of the stack is controlled by a master thread which provides protocol interface to an application. It shall be noted that to perform network operations a context switch from user space to kernel space has to be performed, which is one of the contention points.

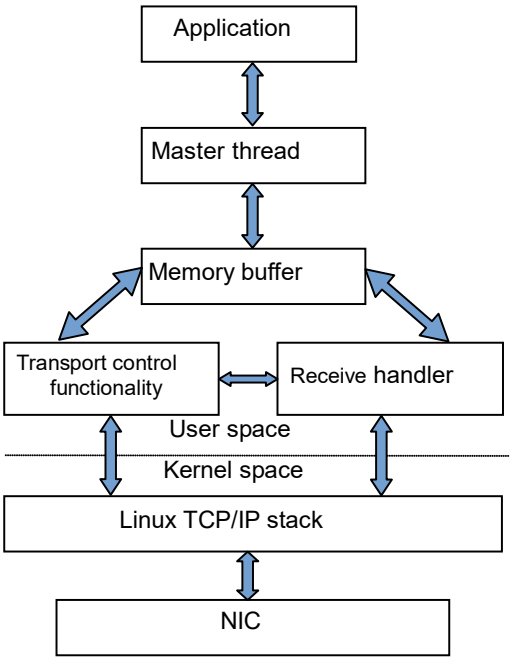

Figure 2: Simplified RMDT structure.

In order to implement F-Stack into RMDT protocol some changes to the protocols' networking subsystem have to be made. F-Stack requires a separate loop function to be run on a dedicated CPU core and that function has to be static. Thus, due to OOP structure of RMDT, all functionality regarding receiving and sending packets has to be moved to a separate thread that is not a direct part of any class in RMDT stack (a global static function).

In the modified structure, additional blocks for send and receive loops represent separate threads which have to run on dedicated cores and perform receive polling and sending via DPDK (Figure 3). These threads are separated from the overall RMDT structure and transmit received data via Single-Producer/Single-Consumer queues, while threads that were handling network operations previously are now polling said queues. Receive/send loop flow is similar to the one presented in figure 1, but with addition of interprocess communication after receiving or before sendig of each batch of packets. Here, switch to kernel space is not needed as DPDK works fully in user space.

5 TEST RESULTS

Firstly, basic DPDK tests have been performed without RMDT to assess the capabilities of hardware while working with DPDK. At this stage both pure

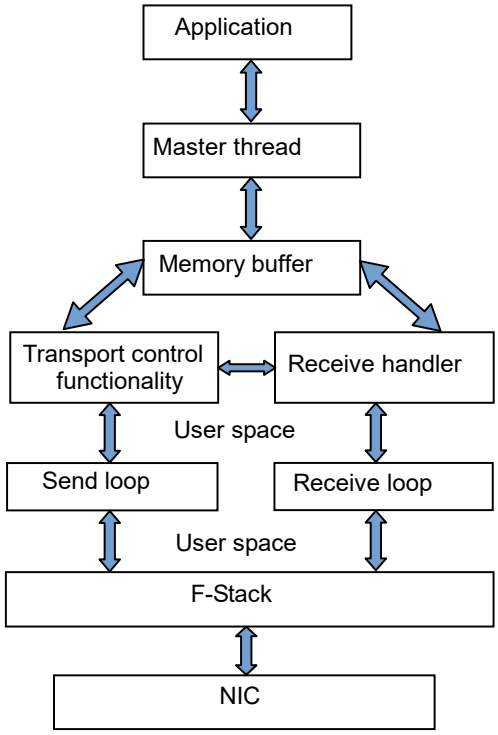

Figure 3: RMDT over DPDK structure.

send and pure receive in a case of point-to-point data transmission have been tested. In the simplest configuration, presented in the previous section, sender was able to achieve up to 6 Gbps, while receiver was capable of achieving maximum link capacity of 10 Gbps (unlike standard TCP/IP, which bottlenecked at receive). However, during testing, certain fluctuations in performance were noticed with sending data, when the rate would drop to 5.2 Gbps or less frequently vary between 5.2 and 6 Gbps. Possible reason for this could be an additional memory copy operations in F-Stack sending interface. The exact cause for such behavior was not studied in this work. Further tests with RMDT were performed only for a DPDK-based receiver. Sender used the standard Linux TCP/IP stack as in multithreaded configuration it was able to achieve 10 Gbps rates, unlike the F-Stack/DPDK test.

Subsequent tests with RMDT were performed – at first with a standard TCP/IP stack to compare it with a DPDK-based RMDT. Standard RMDT showed datarate of 6 Gbps. The bottleneck in such configuration is the receiver as in a test in point-to-multipoint configuration with two receivers, 10 Gbps datarate was achieved. In a test with DPDK-based RMDT, peack achieved datarate was 8 Gbps, althought it was observed to behave inconsistently, sometimes dropping to 6 Gbps. This behavior can be

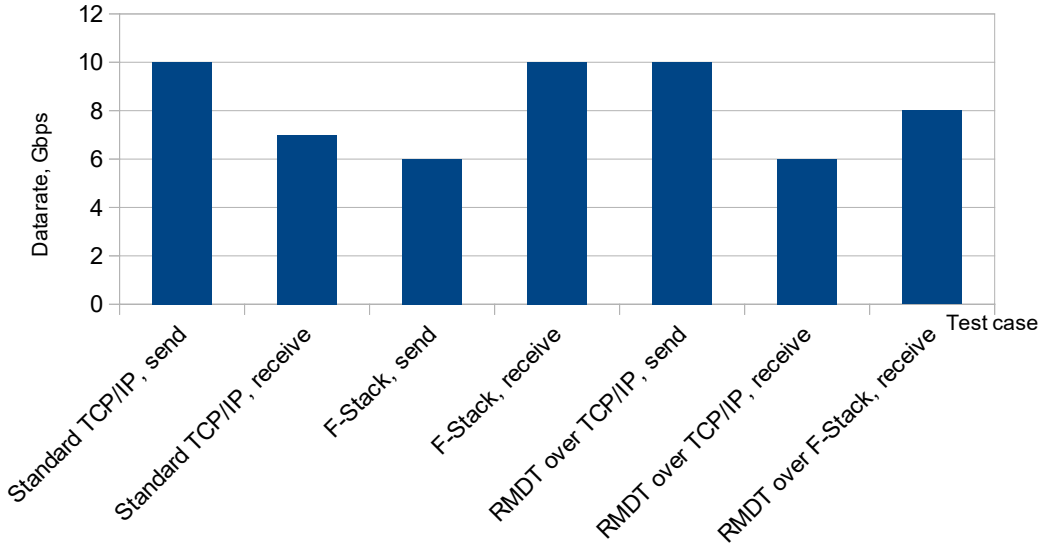

Figure 4: Test results.

explained by unoptimized inter-process communication between F-Stack loop and main RMDT threads. This can be observed by comparison of a clean DPDK test. One of the main reasons are additional memory copy operations. However, even in an unoptimized state, the increase in performance can be seen. A summary of the test results can be seen on Figure 4.

6 CONCLUSIONS

The demand for tools providing high speed data transmission grows, thus leading to development of new SDKs that revise outdated approaches to network applications as for example DPDK/F-Stack does. In this work an attempt to modify a custom UDP-based transport protocol to use DPDK capabilities was made with a goal of increasing performance in a 10G network in a point-to-point configuration with 1500 bytes MTU.

Tests showed an increase in performance in copmarison to standard Linux TCP/IP stack, but full link utilization was not achieved due to the fact that current RMDT structure does not yet fully use DPDK capabilities.

7 FUTURE WORK

In order to continue tests with RMDT over DPDK, significant changes have to be made to the protocol's structure. In particular, better memory management should be implemented. With an improved version additional tests in a 10G and 40G network could be made.

Another possible continuation of this work is developing and testing transport-related applications that could use DPDK functionality for better performance. Network probing algorithms, for example, might improve with lower latency and more stable measurements.

ACKNOWLEDGMENTS

This work has been funded by Volkswagen Foundation for trilateral partnership between scholars and scientists from Ukraine, Russia and Germany within the project CloudBDT: Algorithms and Methods for Big Data Transport in Cloud Environments.

REFERENCES

[1] "Scaling in the Linux Networking Stack", kernel.org, 2018 [Online]. Available: https://www.kernel.org/doc/Documentation/networking/scaling.txt, Accessed on: Dec 01, 2018.

[2] "Data plane development kit", dpdk.org, 2018 [Online]. Available: https://www.dpdk.org/about/, Accessed on: Dec 01, 2018.

[3] "Big Data Transmission | F I L A", fila-lab.de, 2018 [Online]. Available: https://fila-lab.de/index.php/our-work/big-data-transmission/, Accessed on: Dec 01, 2018.

[4] R. Rajesh, K. B. Ramia, and M. Kulkarni, "Integration of LwIP stack over Intel (R) DPDK for high throughput packet delivery to applications," in 2014 Fifth International Symposium on Electronic System Design, 2014, pp. 130-134.

[5] Z. Li, "HPSRouter: A high performance software router based on DPDK," in 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 503-506.

[6] M. Miao, W. Cheng, F. Ren, and J. Xie, "Smart batching: A load-sensitive self-tuning packet I/O using dynamic batch sizing," in 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016, pp. 726-733.

[7] "Apposite Technologies Netropy WAN Emulators", Apposite Technologies.

[8] "F-Stack | High Performance Network Framework Based On DPDK", f-stack.org, 2018 [Online]. Available: http://www.f-stack.org/, Accessed on: Dec 01, 2018.

[9] C. Hornig, "A standard for the transmission of IP datagrams over ethernet networks," 1984.

Efficiency of BQL Congestion Control under High Bandwidth-Delay Product Network Conditions

Nikolai Mareev, Dmitry Kachan, Kirill Karpov, Dmytro Syzov and Eduard Siemens
Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany, {nikolai.mareev, dmitry.kachan, kirill.karpov, dmytro.syzov, eduard.siemens}@hs-anhalt.de

Keywords: Long Fat Networks, Transport Protocol, IP Networks, High-Speed Network, Congestion Control.

Abstract: BQL congestion control is aimed to utilize full available bottleneck bandwidth while keeping bottleneck buffer queue load on some low level to prevent it from producing avoidable additional delays or delay jitter. In this paper, an intermediate result of research in delay-based congestion control is presented. Using RMDT protocol we have evaluated its performance under high bandwidth delay product network conditions and compared it with TCP BBR using the iperf utility. High bottleneck bandwidth utilization in a wide area of delay/bandwidth/loss conditions have been reached. Some performance issues of BBR in some cases has also been observed and investigated.

1 INTRODUCTION

Congestion control algorithms take a significant role in the efficiency of data transport over IP networks.

In general, there are three main challenges for a modern congestion control algorithm: high bottleneck bandwidth utilization, resource sharing, and low influence on network buffers. The most common congestion control type in a network is loss-based congestion control. Using packet losses as a congestion indicator leads to performance degradation in lossy networks, and additional delays caused by the bottleneck queue load. Available buffer space of the bottleneck queue buffers significantly increased last time, what makes this additional delay significant. These consequences show the need of new congestion control algorithms with lower influence on the network and with non-congested packet loss tolerance.

One such solution is a BQL (Bottleneck Queue Level) congestion control developed in the course of the CloudBDT and BitBooster projects at the Future Internet Lab Anhalt. Mentioned projects operate with a Reliable Multi-Destination Transport protocol RMDT [1], [2]. It is a delay-based congestion control solution with packet loss tolerance and low influence on the network infrastructure.

The aim of this paper is to present intermediate results the actual advances in research on a delay-based congestion control. For this, a series of tests of the efficiency of data transport using the developed BQL algorithms have been performed.

Results of such a solution in high bandwidth delay product network conditions in comparison to TCP BBR [4] (Bottleneck Bandwidth and Round-trip propagation time) have been shown. TCP BBR - is besides BQL another modern congestion based congestion control solution with similar aims as BQL and it is the closest solution to the proposed algorithm.

The content of this paper is organized as follows: In section 2 a brief observation of data transport issues over high bandwidth delay product network is provided. Section 3 describes the testbed network. In section 4 test scenarios and test results are provided. Section 5 includes conclusions over experiments and further work.

2 RELATED WORK

The first test results of a BQL congestion control were presented in the 6th International Conference on Applied Innovations in IT [3], (ICAIIT 2018). The main idea of this solution is to use a modified PID (Proportional – integral – derivative) controller to keep link always slightly congested with the aim to reach full bottleneck bandwidth utilization. The most significant states of an algorithm in the current implementation are the Gain state (to quickly reach

bottleneck bandwidth limit) and the Control state (to keep bottleneck slightly congested). In [3] first performance tests have been provided, which show a fair resource sharing capability, full bottleneck bandwidth utilization and the overall structure of the BQL algorithm.

In this paper, new results after some development period of the algorithm are presented. Hereby, the stability and efficiency of congestion control have been significantly increased. Transport delays (delay between action and reaction in terms of a controller) generally caused by RTT in the current version of the algorithm have now much less destructive influence on the performance of control, what allows to keep a necessary number of bytes in a bottleneck queue buffer more precisely.

Most changes during development were done in control state of an algorithm. The current version of RMDT allows using more accurate delay metrics of network congestion which lead to higher performance in long pipes. Paired modified PID controllers now can stabilize throughput on full utilized bottleneck bandwidth under extreme conditions of up to 1000 ms of RTT delay with nearly 200 KB of memory usage by network device queue buffer. It is a benefit in the context of usage in networks with high throughput and tiny buffers. Paired controllers provide more precise control in cases with network delay jitter what leads to higher performance in noisy networks.

Another modern congestion-based congestion control solution is TCP BBR which provides high bottleneck bandwidth utilization in a wide area of conditions while keeping buffer load on some low level. The mechanism of keeping high bandwidth utilization of this solution is a bandwidth probing what leads to RTT jitter and rate losses caused by congestion control which is increasing with the growth of network delay.

Performance degradation of TCP BBR during resource sharing in 1Gbps and 10Gbps links is presented in [5]. Here shown that small bottleneck buffers can lead to packet losses caused by congestion during resource sharing and unfair coexistence of TCP BBR with other congestion control, especially in cases with diverse flow round trip time. In [6] a cyclic performance drop of TCP BBR was observed. However, BBR shows higher performance in cellular networks [7] in comparison to other congestion control algorithms. In [8] a detailed analysis of TCP BBR algorithm behaviour is presented. In this work performance degradation of BBR in cases with shallow buffers caused by overestimating the bottleneck capacity has been observed. In these cases, BBR cannot recognize that the network is congested what leads to datarates higher than available bandwidth and so to massive packet losses.

It is worth to mention that in many other cases, BBR congestion control algorithm can reach high bottleneck bandwidth utilization along with keeping low mean bottleneck buffer load level and nearly fair coexistence with other flows.

3 TESTBED NETWORK

The testbed network topology for our investigations is presented in Figure 1.

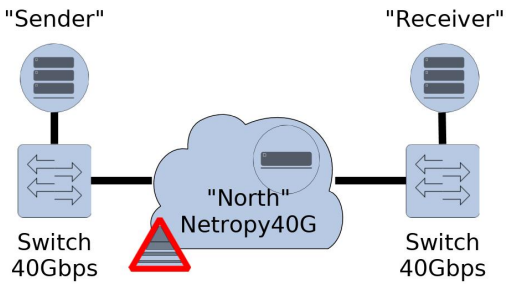

Figure 1: Test network setup.

WAN emulator Netropy 40G is the core element here, it can be used to emulate WAN links with up to 40 Gbps throughput and up to 1000s delay and to collect different statistics such as datarate and bottleneck buffer load level. Both servers run in Ubuntu 16.04 (kernel: GNU/Linux 4.15.0-45-lowlatency x86_64) and are equipped with Intel(R) Xeon(R) CPU E5-2643 v4 3.40GHz, 64GB of RAM and 40000baseSR4/Full supported link modes on Emulex Corporation OneConnect NIC.

The first bunch of test aimed to evaluate behavior of BQL under different round trip time conditions and comparing its performance with TCP BBR. The second bunch of tests is aimed to demonstrate the performance of BQL in a wide range of round trip time / packet loss rate / bottleneck bandwidth conditions. Mean datarate mentioned in these tests refers to the amount of transmitted data divided by time at the sender elapsed to transmit it. All tests have been performed in 40 GE Laboratory of Future Internet Lab Anhalt (FILA).

4 EXPERIMENTAL RESULTS

Results of the first bunch of tests are presented in Figures from 2 to 4. Each of these tests were evaluated over testbed network with next parameters of Netropy link: BBW (bottleneck bandwidth) = 1 Gbps; RTT = {0, 100, 500, 1000} ms; Queue buffer size = 80 MB, drop tail queuing algorithm. TCP BBR flows were executed with iperf utility.

On Figure 2 differences in behavior between BBR and BQL are shown.

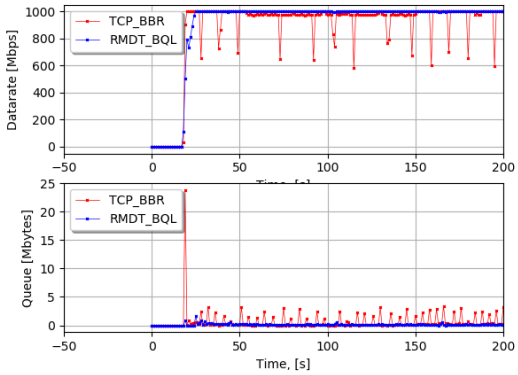

Figure 2: BBW 1 Gbps, base RTT 100 ms.

Both algorithms reach maximum available bandwidth. 75 Gigabytes were transmitted in approximately 10 min. Mean bottleneck buffer load levels during transmission were: 158 KB occupied by BQL and 705 KB occupied by BBR. Mean datarate during transmission were: 994.6 Mbps by BQL and 964.2 Mbps by BBR. The most significant difference between these two flows is buffer jitter. For BBR it can reach up to 4 MB while buffer jitter caused by BQL is less than 200 KB during the Control state period.

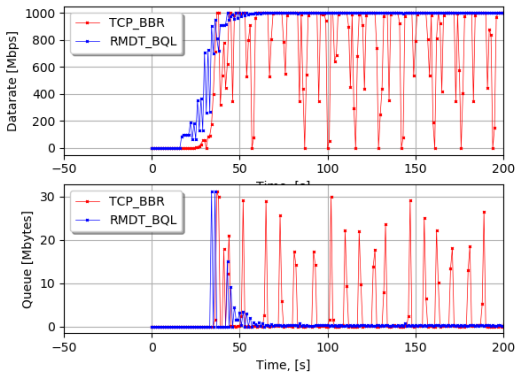

Figure 3: BBW 1 Gbps, base RTT 1000 ms.

Figure 3 demonstrates a more severe difference between these solutions: active bandwidth probing by BBR leads to significant rate decay during transmission and buffer jitter up to 30 MB. Mean bottleneck buffer load level during transmission was: 0.387 MB occupied by BQL and 3.633 MB occupied by BBR. Mean datarate during transmission was: 972.9 Mbps by BQL and 821.6 Mbps by BBR.

In Figure 4 mean datarates and buffer load level caused by these algorithms in different network delays are shown.

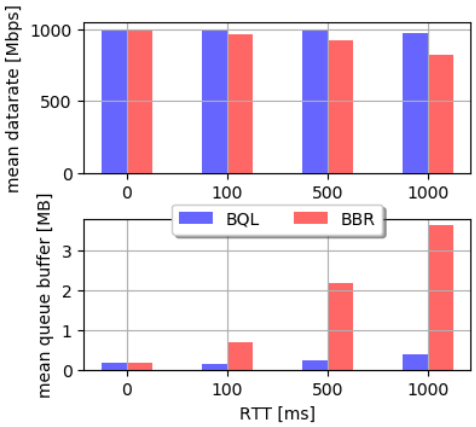

Figure 4: BBW 1 Gbps, 75 GB, summary results.

On Figures 5 and 6 a comparison between algorithms under packet loss conditions is presented. Both BBR and BQL do not use packet losses as congestion indicator what allows reaching high bandwidth even in presence of significant PLR (Packet Loss Rate). During 1 Gbps tests here in each case, a 60 Gb of data have been transmitted.

Figure 5 reveals the growing performance difference between BBR and BQL with increasing of RTT delay. Nevertheless, both algorithms provide high performance in such cases.

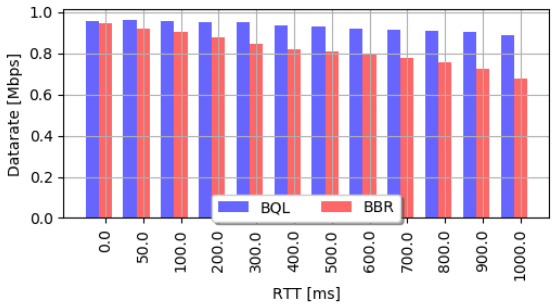

Figure 5: BBW 1 Gbps, PLR 1 %, 60 GB.

Figure 6 demonstrates tests bunch under 10 Gbps bottleneck bandwidth, 0.7 % packet loss rate and variety round trip time delays.

TCP stack of both server and a receiver was tuned up to its maximum but it turned out that it is not enough for such test conditions. It can be seen that RMDT under these conditions has a key advantage – a user-space protocol buffers and faster lost packets processing algorithm.

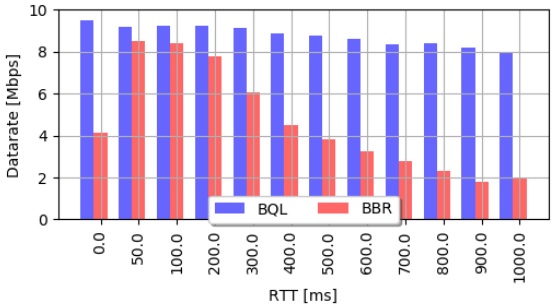

Figure 6: BBW 10 Gbps, PLR 0.7 %, 200 GB.

On 0 ms RTT case (in fact there present small network delay of appx. 150 µs) BBR showed significant performance drop. BQL as the congestion control algorithms in RMDT keeps on reaching high performance. During 10 Gbps tests in each case, a 200 GB of data has been transmitted.

5 CONCLUSIONS AND FURTHER WORK

In this article performance investigations of BQL in high bandwidth delay product network has been provided. With the main aims of BQL – to be scalable to different link cases, this solution can provide high bottleneck bandwidth utilization in wide conditions area. The raising RTTs does not have a significant influence on its control performance. Bottleneck buffer load level during all tests was kept on a low level, the mean value of bottleneck buffer load level during all 1 Gbps tests was nearly 250 KB. Packet losses do not have a significant effect on congestion control performance. Comparison with TCP BBR under the same conditions is also provided. This solution can provide high performance in many cases. However, performance degradation in high bandwidth delay product conditions was observed.

BQL congestion control algorithm is under active development. One of the main aims of the next work is an adjustable resource sharing algorithm - providing a mechanism of fair / low priority / aggressive coexistence of BQL with loss-based and delay-based common TCP congestion control algorithms. Boosting performance in wireless networks in common and in wi-fi networks in particular.

ACKNOWLEDGMENTS

This work has been funded by Volkswagen Foundation for trilateral partnership between scholars and scientists from Ukraine, Russia and Germany within the project CloudBDT: Algorithms and Methods for Big Data Transport in Cloud Environments.

REFERENCES

[1] E. Siemens, D. Syzov, D. Kachan High-speed UDP Data Transmission with Multithreading and Automatic Resource Allocation Proc. of: the 4th International Conference on Applied Innovations in IT, (ICAIIT 2016), pp. 51-56, Koethen, 2016.

[2] S. Maksymov, D. Kachan, E. Siemens Connection Establishment Algorithm for Multi-destination Protocol Proc. of: the 2nd International Conference on Applied Innovations in IT, (ICAIIT 2014), pp. 57-60, Koethen, 2014.

[3] N. Mareev, D. Kachan, K. Karpov, D. Syzov, E. Siemens, and Y. Babich, “Efficiency of a PID-based Congestion Control for High-speed IP-networks,” Proc. of: the 6th International Conference on Applied Innovations in IT, (ICAIIT 2018), pp. 1-5, Koethen, 2018.

[4] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and Van Jacobson, “BBR: congestion-based congestion control,” Communications of the ACM, vol. 60, no. 2, pp. 58-66, Jan. 2017.

[5] M. Hock, R. Bless, and M. Zitterbart, “Experimental evaluation of BBR congestion control,” in 2017 IEEE 25th International Conference on Network Protocols (ICNP), 2017, pp. 1-10.

[6] K. Miyazawa, K. Sasaki, N. Oda, and S. Yamaguchi, “Cyclic Performance Fluctuation of TCP BBR,” in 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 2018, vol. 01, pp. 811-812.

[7] E. Atxutegi, F. Liberal, H. K. Haile, K. Grinnemo, A. Brunstrom, and A. Arvidsson, “On the Use of TCP BBR in Cellular Networks,” IEEE Communications Magazine, vol. 56, no. 3, pp. 172–179, Mar, 2018.

[8] D. Scholz, B. Jaeger, L. Schwaighofer, D. Raumer, F. Geyer, and G. Carle, “Towards a Deeper Understanding of TCP BBR Congestion Control,” IFIP Networking 2018, pp. 109-117, Zurich, Switzerland, May, 2018.

Hardware Implementation of IP Packet Filtering in FPGA

Ana Cholakoska, Danijela Efnusheva and Marija Kalendar
Computer Science and Engineering Department, Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Karpos II bb, PO Box 574, 1000 Skopje, Macedonia
{acholak, marijaka, danijela}@feit.ukim.edu.mk

Keywords: FPGA, IP Header Fields Extracting, IP Packet Filtering, Network IDS Systems.

Abstract: In the present rapid expansion of the number of computers and devices connected to the Internet, one of the top three issues that need to be addressed is the network security. The greater the number of connected users and devices, the attempts to invade privacy and data of connected users becomes more and more tempting to hostile users. Thus, network intrusion detection systems become more and more necessary and present in any network enabling Internet connections. This paper addresses the network security issues by implementing NIDS style hardware implementation for filtering network packets intended for faster packet processing and filtering. The hardware is based on several NIDS rules that can be programmed in the system's memory, thus enabling modularity and flexibility. The designed hardware modules are described in VHDL and implemented in a Virtex7 VC709 FPGA board. The results are discussed and analyzed in the paper and are presenting good foundation for further improvement.

1 INTRODUCTION

Many concepts of security measures for computer communication networks have been developed over the years. Consequently, it has been shown that some of them are more effective when it comes to the resilience to various network intrusions and attacks in comparison to other known systems. Network Intrusion Detection Systems (NIDS) allow greater control over the traffic generated in the network while applying several mechanisms and rules for filtering known and sometimes predicting unknown types of network attacks according to anomalies detected in the monitored network traffic.

Naturally, these kind of IDS systems are generally software defined, and are still vulnerable to unknown or novel types of attacks. Nevertheless, the software defined NIDS systems are quite flexible, modular and easily upgradeable. Despite the flexibility, the main potential liability of these software based NIDS systems is their inability to handle and process the continuous and daily increasing quantities of network traffic.

Consequently, the concept of filtering network packets in this paper has been based on an existing software system for protection against unauthorized intrusions. Namely, SNORT – a network IDS system, is well known for its ever evolving architecture and the vast collection of rules for detecting unwanted network traffic. Precisely those rules are taken as the basis for the hardware implementation and the packet filtering tests.

Despite the software solutions, several specialized hardware solutions intended for packet filtering have also been proposed, in order to bring additional speed to the process of filtering. Regarding hardware network packet processing, one of the most popular and vastly used solutions are the Network processors (NPs) [1]. In general, they represent devices specially tailored to perform various network processing operations: header parsing, bit-field manipulation, pattern matching, table look-ups, and data movement, [2]. Similarly, one of the more renowned and studied architectures of network packet processing is the NetFPGA architecture [4]. NPs are usually used in different types of network equipment, including routers, switches, IDS or firewalls, [3]. Accordingly, NPs spend significant part of processor cycles on packet header parsing, especially when the packet header fields are non byte- or word-aligned. Improving the number of processor cycles needed for packet header parsing has been addressed in our previous work [8], enabling a single-cycle memory access to these non byte- or word- aligned header fields. The simulation results and the flexibility of the proposed solution

were investigated utilizing a reconfigurable hardware platform Virtex7 VC709 FPGA.

Resuming this previous work and building on, after the network packet headers have been parsed and accessed in memory, this paper investigates the possibilities for implementing software IDS packet filters in hardware. Consequently, the primarily goal of this paper is to augment software defined IDS systems by implementing and simulating network packet filtering modules in hardware, using the faster hardware resources of an FPGA board while retaining the flexibility of the software based rules. Such hardware/software co-design would bring speed, as well as flexibility while implementing and applying the rules for network packets filtering.

The rest of this paper is organized as follows: Section II gives an overview of the state of the art in the area presenting different network processing and filtering hardware solutions. Section III describes the design of the rule filtering hardware intended for increased security and layouts the benefits of the hardware design and the flexibility due to the programmability of the FPGA system. Section IV presents the additional hardware module for extracting/ writing ip header fields from/to memory in a single-cycle access, especially for non byte- or word- aligned packet header fields. Section IV presents simulations and synthesis results from the FPGA implementation of the IP packet filtering hardware module in VHDL. Section V concludes the paper, outlining the benefits of the proposed IP packet filtering module.

2 STATE OF THE ART

Contemporary technology advances increase the pace of rapid expansion of the number of computers and devices connected to the Internet on daily basis. As a result, one of the highest priority issues that need to be considered in this enormous network is the network security. The greater the number of connected users and devices, the attempts to invade privacy and data of connected users becomes more and more tempting to hostile users. Thus, Network Intrusion Detection Systems (NIDS) become more and more necessary in any network connected to the Internet, and are taking the lead in the battle against intruders.

In order to enhance the security NIDS have to inspect incoming network packets looking for unwanted and hostile traffic. This is foremost done via various software platforms (e.g. Snort), but the major increase in daily network traffic imposes a great challenge for software platforms. Therefore, many researchers have already turned to hardware designs for many network issues, including security.

One of the first topics for hardware designed processing is aimed to network packets header parsing, which is a prerequisite for the packet filtering operations.

Many researchers have been working in both areas since they are interconnected. Namely, the process of identifying and extracting fields from a packet header and doing it in hardware for faster processing is being addressed in many works [7], [9]. In most cases NPs are used to perform fast packet processing where the IP header is being processed, by analyzing, parsing and modifying its content, [3]. NPs might include some specialized hardware units to perform classification of packets, lookup and pattern matching, queue management and traffic control which on the other side can be used for the purpose of packet filtering for security reasons. In recent research, NP software is getting closer to the NP hardware, such as in [10] where part of the packet processing tasks such as classification or security are offloaded to application-specific coprocessors. Other proposals make big use of FPGA technology for packet parsing, enabling implementation of pipeline architectures and thus achieving high-speed network stream processing, [12]. Actually, the reconfigurable FPGA boards can be used to design flexible multiprocessing systems that adjust themselves to the current packet traffic protocols, which in turn makes them very suitable for packet filtering regarding security.

FPGA technology is widely used in combination with NIDS and packet filtering for security purposes. For example, the authors in [15] propose a modular approach for grouping homogeneous traffic, and then splitting it for filtering in different specialized hardware blocks, each supporting a (smaller) rule set tailored for the specific traffic category. The rule sets are based on the well known Snort NIDS. The paper concludes that the exploitation of traffic classification and load statistics may bring significant savings in the design of HW NIDS. Similarly, [18] introduces a packet pre-filtering approach, based on the observation that very rarely a single incoming packet fully or partially matches more than a few tens of IDS rules. Finally, packet pre-filtering prevents matching at least 99% of the SNORT rules per packet, thus minimizing processing time and improving the scalability of the system.

Other reconfigurable FPGA approaches include [16] and [19] who propose modular and programmable hardware accepting configuration changes (for the rules) in real time, while processing the important packet header fields and/or payload.

Finally, [20] investigate an approach suitable for current and future high speed networks of 100 Gb/s and more, by trimming the traffic that needs to be filtered by using FPGA-based packet filters. The goal is achieved by implementing a new "network grammar" for specifying protocols and filtering rules for continuous stream of data. The grammar compiles directly to Verilog code for packet filters. The new concept was tested on two proof-of-concept designs: a DNS filter and a simple firewall.

3 RULE FILTERING HARDWARE FOR INCREASED SECURITY

Considering network IP packets processing, usually the main operation is packet header parsing, mostly for packet routing purposes. Nevertheless, since the header fields of the packet have been already extracted, further processing is possible and very useful for the aim of traffic engineering, traffic shaping, network security.

This paper is taking into account further packet processing regarding security, i.e. applying specific IDS software rules on received network packets for the purpose of filtering unwanted packets from the network. The IP header that is received at the input of the previous module [8], is usually immediately written in memory without previously going through any inspection. Consequently, to be able to enforce greater security, in this architecture, a block for filtering IP headers is being implemented previous to writing the packet into memory. The filtering block uses rules previously defined by the system administrator (in the considering case, predefined rules from the Snort IDS for checking IP headers). The corresponding rules have been programmed into memory on a particular location.

Figure 1 presents a proposed scheme for an implementation of such IP packets filter that enables applying appropriate rules, therefore filtering the network packets. Following the packet parsing module, a selected IP header field is being input in this proposed module (whether a TTL, protocol etc.) and is subsequently compared to the rule in memory. If these two fields match, then the Alert signal outputs a high level signal, which in turn allows the appropriate field to be written in memory. The appropriate location is calculated through the BaseAddress and MemoryOffset information. On the contrary, if these fields do not match, the package is rejected and not written in memory. This is signalled again through the Alert signal that outputs a low level signal.

Initially, the Alert signal may not be present, however this kind of signalling proves very useful to the previous module keeping the headers (or the entire) IP package. The value of the Alert line would signal the IP packet parser module that in fact there is at least one field of the header that has not passed the filter, enabling it to make a rejection of the entire packet header and delete all packet fields.

Figure 2 presents a simulation conducted with the HDL programming package VIVADO showing the signal layout for an unsuccessful rule matching procedure. If the incoming IP packet header field does not match the expected value in the filter field, the Alert signal outputs a low level signal value (0 value), indicating the incoming packet field is invalid.

Figure 3 presents the contrary situation where the incoming IP packet header field matches the rule, and the Alert signal outputs a high level signal value (1 value) , thus indicating that the IP packet header field passes the security check and can be saved in memory.

The presented hardware solution enables faster hardware comparison of the IP packet headers against the programmed rules compared to the software approach. Moreover, programming the filtering rules in memory imposes great flexibility and possibility for implementing and testing new future rules for new and emerging threats. The presented hardware module is a proof of concept solution requiring further investigation in the area of hardware rule optimization, as well as field comparison optimization.

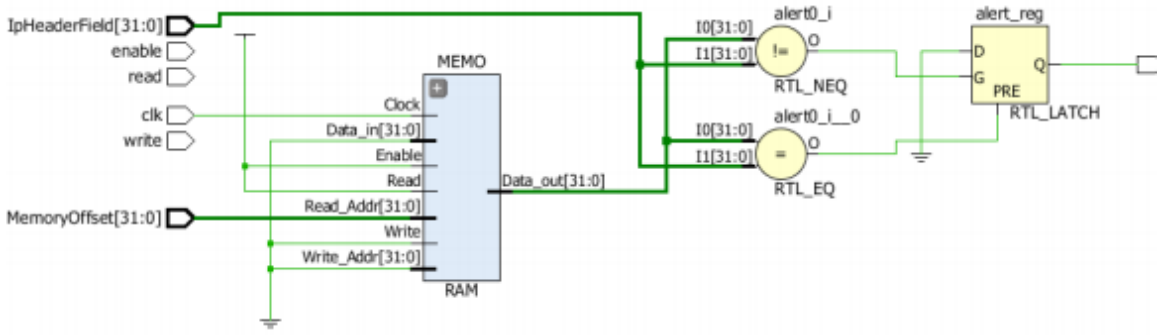

Figure 1: Scheme of the implementation of an IP filter.

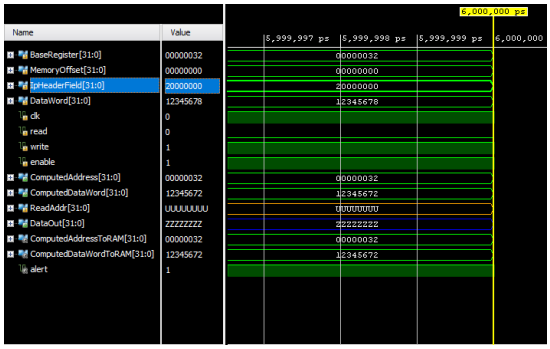

Figure 2: Field filtering - non matching.

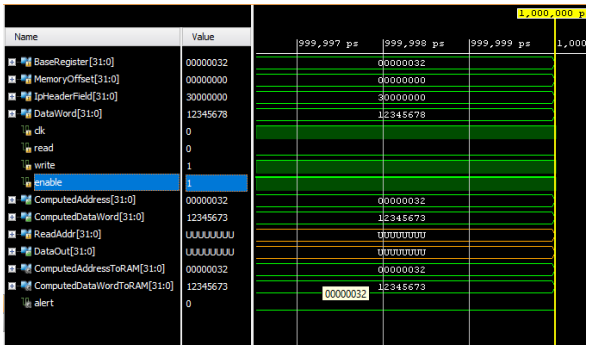

Figure 3: Field filtering – matching.

4 HARDWARE FOR WRITING IP HEADER FIELDS IN MEMORY

Besides the concept of IP header fields filtering, this paper details and proposes specialized hardware modules for IP packet header fields extraction before processing, and subsequently IP packet header fields rewriting after processing. These modules help the direct extraction of the needed IP packet field in a single memory access operation, further improving the speed for information transfer from/to the processor as proposed in our previous work [8]. The IP packet fields extracted to the processor are then used in all other processing (filtering) operations. In order to return the corresponding field to the proper memory location (proper location in the), several reversed operations have to be performed. This section details these operations.

After the processing/filtering of the specified IP header field, the field should be returned to the correct position in the complete IP packet header in memory. The signals lines augmenting this process are BaseRegister, MemoryOffset and DataWord, as well as the IpHeaderField containing the shifted address to the processed IP header field. The BaseRegister carries the initial IP packet address into the memory, MemoryOffset tells which header field is being considered (order number). Than this value goes through a Look-up table where the order numbers of the fields are paired with one of the six words on the from the IP header. Every header field is mapped with the word that it is contained in. For example the field “Time to Live” is field 7 (it starts with 0) and it is mapped with the word 2 represented in 32 bits (it starts again with a count of 0). The values from BaseRegister and MemoryOffset are summarized and thus the address where the new value should be written is calculated.

The order number of the word containing the IP header field under consideration, arrives to the decoder, which in turn decides which of the FieldLogic fields to activate. IpHeaderField and DataWord are also present enabling the assembly of the IP header word with the selected order number. At this point, having the needed data and the address in memory, the content of the IP header word containing the processed IP header field is written in memory.

To illustrate the process a selected example will be explained taking into account the “Time to Live” field. “Time to Live” field is the seventh header field in order and it is contained in the second word (counting from 0). Let the value of this second word

be DataWord = 12345678 (hexadecimal). The hardware would extract only the TTL field and shift its value to the left, resulting in IpHeaderField = 78000000. "BaseRegister" represents the first address of the IP packet header in memory, and for ease of the example calculation it can be freely chosen. So let's assume BaseRegister = 00000000. The initial value of MemoryOffset = 00000007. Then, by searching through the Look-up table, the value 00000002 will be output and stored in the MemoryOffset field (the seventh field is in the second word of the header). Now let's assume that the FieldLogic module initiates some change in the IpHeaderField resulting in a new field value IpHeaderField=77000000. The correct memory address where this new value has to be rewritten is calculated from the existing values:

BaseRegister + MemoryOffset =
= 00000000 + 00000002 = 00000002

thus marking that the field should be rewritten in the second word of the header. Finally, the correct data from the DataWord and the new IpHeaderField need to be combined and written back to the correct place.

The FieldLogic module is responsible for correctly combining the DataWord and the IpHeaderField using the order number of the IP header field (7 in the example). Consequently, the seventh part of the FieldLogic module would be activated, enabling shifting and adding logic to finally calculate the new value of the DataWord (incorporating the IpHeaderField at the last two "digits"). Finally the new and correct DataWord exiting the FieldLogic module can be written at the correct address calculated previously.

Since the finest details regarding the IP header fields extraction and rewriting are "hidden" in the FieldLogic module, Figure 4 and Figure 5 give details regarding the hardware realization of the FieldLogic parts designed for the Version IP header field and the Header Checksum IP header field, just for the purpose of illustration. As it can be seen from these figures, the number and position of the bits in the IP packet header influences the complexity of the hardware intended for IP header field extraction/writing. Version IP header field is 4 bits long, and positioned at the end of word 1. Hence, several operations of shifting, multiplexing and additions are needed for the operations. On the other hand, the Header Checksum IP header field is 16 bits long and is byte aligned in word 3 of the IP packet header. Consequently, the needed hardware is less complex. This is true for all the other IP packet fields with similar observations.

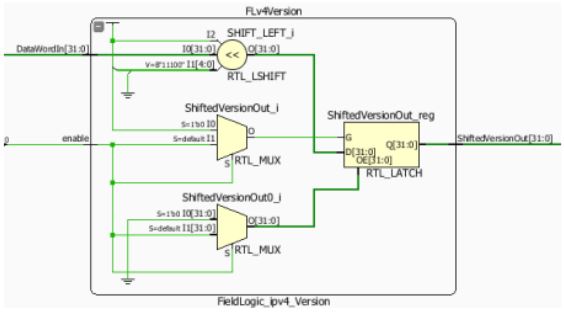

Figure 4: Field_logic_ipv4 Version design.

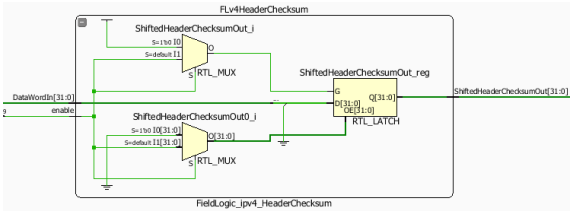

Figure 5: Field_logic_ipv4_HeaderChecksum design.

The last field of the IPv4 header is the SourceAddress field (the sender of the packet). In this case, the field has a size of a one word, or 32 bits. Therefore, since this field can be transferred completely to memory, and on the right position, no additional selection is performed for certain bits.

As presented, specialized hardware is needed for each IP packet header field, and special FieldLogic parts have been designed for all IP packet header fields for the IPv4 and IPv6 packet headers.

Regarding IPv6 packet header, the fields: Version, Traffic Class, Flow Label, Payload Length, Next Header, Hop Limit are mapped similarly as the IPv4 fields. Interesting cases with IPv6 are the Source and Destination Address fields taking 128 bits each.

5 USE OF RESOURCES

completing the functional simulation of the filter and parser of the IP headers, comes the FPGA synthesis and the implementation of the device itself.

The results of the synthesis presented in Figure 6 indicate that the proposed filter for IP network packets can be implemented on in Virtex7 VC709 evaluation platform by utilizing less than 0.01% of the slice registers and 0.16% of the slice LUT resources, which represents less than 1% of the possible FPGA resources. As it is obvious from the

low FPGA resource's utilization, the initial IP header filtering logic design can be further extended and then implemented in the same Virtex 7 VC 709 FPGA board.

This means that as further research, the proposed IP packet filter can be expanded by incorporating more packet header fields and/or packet payloads, as well as additional and more complex packet header filters, all implemented on the same FPGA platform. From the ability to reconfigure the FPGA device, it can be concluded that this kind of module can be very easily adapted to work with other protocols, which in turn indicates great flexibility and low cost.

1. Slice Logic

Site Type	Used	Fixed	Available	Util%
Slice LUTs*	689	0	433200	0.16
LUT as Logic	689	0	433200	0.16
LUT as Memory	0	0	174200	0.00
Slice Registers	36	0	866400	<0.01
Register as Flip Flop	0	0	866400	0.00
Register as Latch	36	0	866400	<0.01
F7 Muxes	0	0	216600	0.00
F8 Muxes	0	0	108300	0.00

Figure 6: Used resources of Virtex 7.

6 CONCLUSIONS

This paper addressed a very current and vastly spread issue of network security that affects more and more large institutions and companies every day. Nowadays, this topic of research is placed on the top three priority places and great efforts are being made for it to be improved. Building from here, the idea for a combined software/hardware solution for better network protection from unauthorized intrusions.

Firstly, the concept of filtering network packets has been based on some of the existing software systems for protection against unauthorized intrusions. As the basic software solution, one of the widely used open source systems has been chosen. Namely, SNORT – a network IDS system, is well known for its ever evolving architecture and the vast collection of rules for detecting unwanted network traffic. Exactly those rules are input as the basis for hardware implementation.

In order to be able to offer a complete hardware and software solution, building on top of Snort, a VHDL hardware design was implemented and tested. The hardware design encompasses a packet filter based on hardware implementation of Snort rules, as well as a hardware accelerator for IP packet header fields extraction and rewriting.

The hardware design was implemented on the Virtex 7 VC709 FPGA board, thus proving the functionality and envisioning the future possibilities for improvement. The realization of the module for network packet filtering, presented solid results showing that the proof of concept filter can be implemented by using only <0.01% of the slice registers and as little as 0.16% from slice LUT resources, which represents less than 1% of the possible FPGA resources in total.

The results obtained in this manner indicate the great flexibility and low cost of the module, as well as the possibility for its expansion towards filtering different types of packages, protocols, packet behaviour, as well as adding additional selection filters.

As for the future development of this module, one of the possibilities is to design an additional module for separating the header from the packet, and enabling parallel processing of both, regarding network processing and filtering for increased security reasons.

REFERENCES

[1] B. Wheeler, "A new era of network processing," LinleyGroup Bob Wheeler's White paper, 2013.

[2] P.C. Lekkas, "Network Processors: Architectures, Protocols and Platforms," McGraw-Hill Professional, 2013.

[3] R. Giladi, "Network Processors - Architecture, Programming and Implementation", Ben-Gurion University of the Negev and EZchip Technologies Ltd., 2008.

[4] J. Naous, G. Gibb, S. Bolouki, N. McKeown, "NetFPGA: reusable router architecture for experimental research", in Sigcomm Presto Workshop, 2008.

[5] B. Doud, "Accelerating the data plane with the Tile-mx manycore processor", in Linley Data Center Conference, 2015.

[6] J. M. P. Cardoso, M. Hubner, "Reconfigurable Computing: From FPGAs to Hardware/Software Codesign", Springer-Verlag, 2011.

[7] G. Gibb, G. Varghese, M. Horowitz, N. McKeown, "Design principles for packet parsers", In ACM/IEEE Symposium on Architectures for Networking and Communications Systems, 2013, pp. 13–24.

[8] D. Efnusheva, A. Tentov, A. Cholakoska, M. Kalendar, "FPGA Implementation of IP Packet Header Parsing Hardware", In Proc. of the 5th International Conference on Applied Innovations in IT, (ICAIIT), 2017, pp. 33-41.

[9] J. Kořenek, "Hardware acceleration in computer networks". In 16th International Symposium on

Design and Diagnostics of Electronic Circuits Systems, 2013.

[10] L. Kekely, V. Puš, J. Kořenek, "Software Defined Monitoring of application protocols", In IEEE Conference on Computer Communications, 2014, pp. 1725–1733.

[11] R. Bolla, R. Bruschi, C. Lombardo, F. Podda, "OpenFlow in the Small: A Flexible and Efficient Network Acceleration Framework for Multi-Core System", In IEEE Transactions on Network and Service Management, 2014, pp. 390-404.

[12] V. Puš, L. Kekely, J. Kořenek, "Design methodology of configurable high performance packet parser for FPGA", In 17th International Symposium on Design and Diagnostics of Electronic Circuits Systems, 2014, pp. 189-194.

[13] M. Attig, G. Brebner, "400 Gb/s Programmable Packet Parsing on a Single FPGA", In Seventh ACM/IEEE Symposium on Architectures for Networking and Communications Systems, 2011, pp. 12-23.

[14] G. Brebner, W. Jiang, "High-Speed Packet Processing using Reconfigurable Computing", In IEEE Micro, vol. 34, no. 1, 2014, pp. 8-18.

[15] S. Pontarelli, G. Bianchi, S. Teofil, "Traffic-aware Design of a High Speed FPGA Network Intrusion Detection System". In IEEE Transactions on Computers, Vol. 62, Issue: 11, 2013, pp. 2322 - 2334.

[16] R. Ajami, A. Dinh, "Embedded Network Firewall on FPGA", In Proc. of 8th 2011 International Conference on Information Technology: New Generations, 2011.

[17] S. Yusuf, W. Luk, M.K.N. Szeto, W. Osborne, "UNITE: Uniform hardware-based Network Intrusion deTection EngineS". In Proc. of ARC 2006: Reconfigurable Computing: Architectures and Applications, 2006, pp 389-400.

[18] I. Sourdis, V. Dimopoulos, D. Pnevmatikatos, S. Vassiliadis, "Packet Pre-filtering for Network Intrusion Detection". In Proc. of ANCS'06, 2006.

[19] A. Wicaksana, A. Sasongko, "Fast and Reconfigurable Packet Classification Engine in FPGA-Based Firewall", In Proc. of 2011 International Conference on Electrical Engineering and Informatics, 2011.

[20] J.F. Zazo, S. Lopez-Buedo, G. Sutter, J. Aracil, "Automated synthesis of FPGA-based packet filters for 100 Gbps network monitoring applications", In Proc. of 2016 International Conference on ReConFigurable Computing and FPGAs (ReConFig), 2016.

Passive Perimeter Security Systems Based on Optical Fibers of G 652 Standard

Alexey Yurchenko[1], Ali Mekhtiyev[1,2], Yelena Neshina[1,2], Aliya Alkina[1,2] and Vyacheslav Yugay[2]

[1]*Engineering School of Nondestructive Testing and Safety, Tomsk Polytechnic University, Lenina avenue, 30, Tomsk, Russia*

[2]*Faculty of Power Engineering, Automation and Telecommunications, Karaganda State Technical University, Mira blvd, 56, Karaganda, Kazakhstan*

niipp@inbox.ru, barton.kz@mail.ru,1_neg@mail.ru, alika_1308@mail.ru, slawa_v@mail.ru

Keywords: System, Protection, Sensor, Optical Fiber.

Abstract: This article deals with the problem of ensuring protection of limited access objects and other objects of state significance from unauthorized access. There is given the analysis of systems already developed by Russian and foreign scientists. A passive perimeter security system is proposed for consideration the main element of which is an optical fiber. The measurement principle is based on controlling the magnitude of the additional dissipation losses under mechanical action measured in dB. There have been carried out field experiments using the proposed security system. In conclusion there are describes the results of the study using a reflectometer.

1 INTRODUCTION

Ensuring the protection of limited access and hazardous objects of state importance or just private territories that occupy large areas from unauthorized access, in contrast to local objects, requires large expenditures and complex communication to build a perimeter security system and monitoring. This fact increases significantly the cost of security systems. Today a lot of security systems of different technical levels and costs have been developed that are based on different principles: infrared, vibro-acoustic, magnetometric, capacitive, seismic and other types of systems [1-7]. Perimeter security systems monitor continuously the area of space along the protected boundary by some active physical parameter (physical field). When it is violated and the parameters go beyond the permissible, an alarm is triggered that is sent to the information collecting and processing system. All the systems can be divided into active and passive. The former are more expensive and can be detected by the intruder before they are triggered. Such devices require supplying electrical power, as well as a communication line for transmitting signals or a wireless data system over the air [1-4]. Special tools can be used that bring them out of action. Passive systems are less costly, and in contrast to the active ones are secretive. For example, scientists from Novosibirsk developed a passive perimeter security system using seismic acoustic sensors (geophones) [8-14]. Passive systems control changes in the physical field of the environment and the soil oscillations, the parameters of which are generally random, but there is no radiation of energy into the surrounding space, which complicates their detection.

These systems have many advantages, but they also have disadvantages. For example, signal processing requires complex algorithms and devices, since otherwise false alarms and inaccurate parameters may be found in detecting the intruder. These systems provide absolute secrecy, since their principle of operation is passive, seismic sensors and connecting wires are usually immersed in the ground. When spreading over large areas or boundaries, this method is very complex and costly, since first of all there takes place attenuation of the electrical signal, and there are difficulties in their interaction with each other in the group based on the data obtained.

Based on the above-said, we propose to consider a passive perimeter security system that has all the advantages of a system with seismic acoustic sensors (geophones), but differs in the principle of operation, while not being inferior in technical parameters of work and less complex in configuration. An

important difference is its lower cost. The basis for developing the system was the work aimed at monitoring and measuring in real time the deformation parameters of building objects (foundations, pipelines, bridge construction elements, etc.). Similar developments are underway in the field of mining, to monitor and control the deformation of mine workings, to protect personnel from the sudden collapse of the mine [15–16]. These systems are based on the use of optical fiber in communication systems for information transfer. The measurement principle is based on controlling the magnitude of the additional dissipation losses under mechanical action measured in dB. During mechanical action on the optical fiber, the energy dissipation indicators of the light electromagnetic wave mode passing through the optical fiber change. Considerable work has been done in this field, a number of experiments have been carried out and the original results obtained. Using an optical fiber, it is possible to measure a variety of electrical and non-electrical parameters in parallel with fairly high accuracy [11]. The annual reducing of the cost of optical fiber in the market and increasing its consumer properties, for example, in terms of transparency windows, make it very attractive for using in perimeter security systems. Today one km of a single-mode optical fiber can be bought for about $9, which makes it out of competition with a copper couple that is used for communication with seismic sensors (geophones), since the cost of copper wire in the market is very high. Electromagnetic interference does not affect an optical fiber. Therefore, the use of optical fibers to build passive systems for protecting the perimeters and borders of various objects is an extremely promising trend.

The main idea of the work is connected with the use of telecommunication optical fibers of the G 652 standard as a sensitive sensor capable of identifying mechanical effects. In this case an optical fiber is used as a sensor and a guiding system for transmitting information. All the obtained measurements in the form of a modified optical signal are processed by a microprocessor device, after which it is possible to identify impacts and to determine the distance to the point of the alleged violation of the protected perimeter.

2 METHODOLOGY OF CARRYING OUT FIELD STUDIES

The system development requires a series of studies based on the systemic approach. It is necessary to carry out computer simulations and field experiments to study the processes associated with the light propagation in an optical fiber and caused by mechanical strain deformations. The estimation of losses in the object under study is associated with developing deterministic models that reflect the physical essence of phenomena and contain a description of the mechanisms of the elementary processes taking place in them. The structural diagram of the developed system is presented in Figure 1.

The object of study is represented by an optical single-mode fiber of the G.652 standard that has low losses in the region of the hydroxyl peak (1383 nm), which makes it possible to use CWDM technologies more widely during transmission.

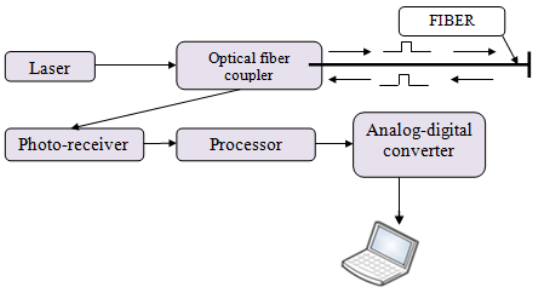

Figure 1: Structural diagram of the system developed.

The system we developed was compiled and tested. It consists of a laser, an optical coupler, a fiber, a photodetector, a microprocessor, and a laptop with software. We use a semiconductor laser as a radiation source. An optical receiving module is mounted at the output of the fiber, which makes it possible to estimate the loss value with accuracy of 10-3 dB. It is known that during mechanical deformations or vibrations in an optical fiber, the conditions for the light propagation or its internal reflection (dissipation) change, as a result of which the phase and spatial characteristics of the beam at the cable output undergo changes. The resulting changes will be recorded by the photodetector and processed by the signal analyzer. Preliminary experiments show that the light at the output of the optical fiber has a “spectrum” structure, this is an irregular system of light and dark spots, and under a mechanical effect on the optical fiber the “spectrum”

structure changes (Figure 2). To fix these changes it is necessary to use space-sensitive photo-receivers.

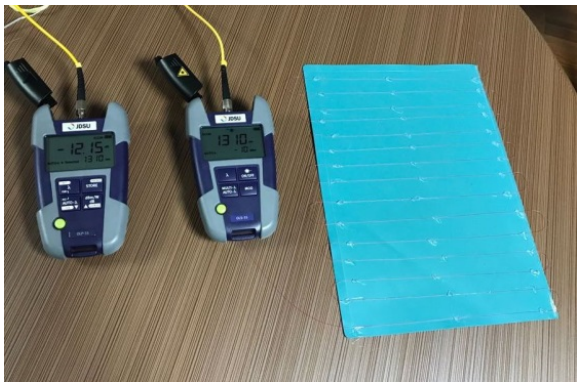

Figure 2: A laboratory test bench for practical approbation of the study theoretical results.

Our task is to develop a security system with the length of one zone up to 100 km with accuracy of detecting an intrusion site up to ±10 m. The depth of the optical cable is from 5 to 50 cm, the width of the sensitive zone is up to 4-5 meters. Under the effect of mechanical vibrations the optical fiber cables give a response in the frequency range from 1 Hz to 100 kHz; in the future for the security system it is planned to limit the frequency range from 100 Hz to 10 kHz.

The optical cable is attached to the plastic grid to increase the system sensitivity and the probability of detecting an intruder or group of violators on the ground, who can walk or run. We use the method of the correlation processing of signals from two fiber-optic cables, which allows isolating the signals of real intrusion from their background to filter out the noise. Three systems will be involved in the studies.

The first one is based on the use of the micro-stress method of mechanical action on an optical fiber and will use two optical fibers in which the laser beam passes (Figure 3). At the end of the zone we perform the interference of both beams in a special optical module. If a mechanical effect is exerted on the cable, the nature of radiation propagation in both fibers changes, and the dynamics of the interference pattern in the optical module allows registering an invasion.

The second system is based on the use of the classical Mach-Zander interferometer (Figure 4). The sensor has already three optical fibers. Two fibers are used as sensing elements, a laser beam operating in continuous mode is fed through them, and the third (output) fiber is used to transmit signals to the system analyzer from the terminal optical module. The radiation source is located in the analyzer unit, from which laser radiation is passed in the passive fiber to the initial module. In this module radiation is split into two beams that are fed to two sensitive fibers.

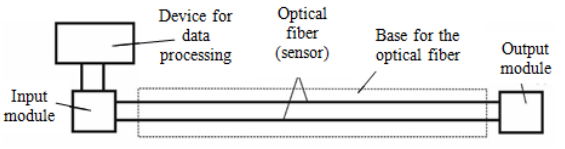

Figure 3: Structural diagram of the method based on the use of micro-stresses of mechanical action on the optical fiber.

In the terminal module both beams interfere. If both arms of this interferometer are in the unperturbed state, then the interference pattern in the terminal module remains unchanged. At this, the signal transmitted from the terminal module via the output optical fiber to the analyzer does not have a variable component. With the cable deformations or vibrations the optical path difference in the sensitive fibers (i.e., the interferometer arms) changes and the terminal module registers the variable component of the signal transmitting it to the analyzer. A specific feature of this interference system is that it determines the relative time delay of the recorded signals in both arms of the interferometer. This allows determining the location of the system invasion with accuracy of several meters.

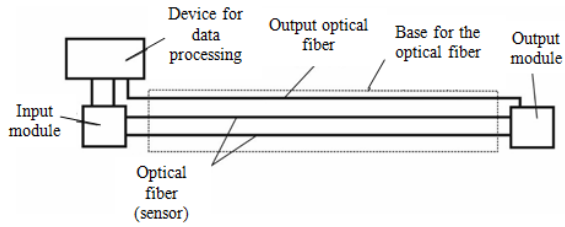

Figure 4: Structural diagram of the interference system.

The third system is based on the use of the method of coherent optical reflectometry with the time resolution (Figure 5). To the controller there are connected optical light guides that pass a laser beam. There takes place the known dissipation effect and part of optical radiation is reflected back from various irregularities. When the fiber optic light guide is subjected to mechanical stress (vibration), an alarm signal can be registered by the parameters of reflected optical radiation to trigger the system. The effectiveness of the system is significantly increased if regular refractive index irregularities with a spatial period comparable to the laser

radiation wavelength are specially developed in the fiber. It is necessary to develop the conditions for Bragg dissipation. This method will allow determining the location of the invasion based on the calculation of the reflected signal delay time. It is possible to establish accurately the location of the invasion with an error of up to 10 meters.

To implement this system, it is necessary to lay not less than two fiber-optic light guides using the underground method to the depth of about 7-10 cm along the protected perimeter. The fibers must be attached to a plastic grid to increase the system accuracy and sensitivity. The correlation processing of signals from two fiber optic cables allows filtering out interference signals (noise of rain, traffic, etc.) and highlighting the signals of real intrusion on their background. This system can be used to protect and to monitor the integrity of pipelines. It is possible to configure the system using a closed loop, when both ends are connected to electronic units. When a sensor breaks, the system switches to the operation mode with two separate beams signaling the location of the cable break. At this, the operation is maintained throughout the entire perimeter. Separate conductors of standard fiber communication cables with optical losses of approximately 0.3 dB/km at the wavelength of 1550 nm are used as sensitive elements.

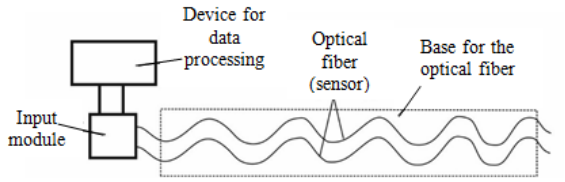

Figure 5: Structural diagram of the system based on the use of coherent optical reflectometry.

3 THE NUMERICAL SIMULATION RESULTS

Using a laboratory test bench, experiments were carried out to determine the loss of optical fiber at various pressures.

The numerical study of the VOD system model was carried out using the Wolframalpha program that is an interactive system for processing the results of experiments focused on working with the data files [8].

The boundary conditions are as follows: the energy of pressure on the fiber is from 0 to 15 Nm, the interval of the step is 2.4 Nm, total 7 steps, the temperature in the laboratory is 23 °C. The movement along the axes until pressure is applied: OX = 0m; OY = 0m; OZ = 0m. As a result of automated data approximation there were obtained the single-factor mathematical models.

Optical fibers with the wavelength of 1310 and 1550 Nm were studies. A plot of the optical fiber with the wavelength of 1310 Nm loss with the incremental pressure increase is presented in Figure 6.

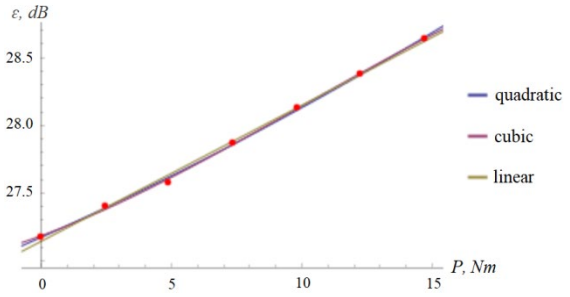

Figure 6: Loss values in the optical fiber with the wavelength of 1310 Nm with the step increment of pressure.

When performing an automatic approximation, the following results were obtained:

1) $0.000124408P^3 + 0.00373935P^2 + 0.0707259P + 27.1854 = \varepsilon$ is the third degree approximation (cubic);

2) $0.100308P + 27.1445 = \varepsilon$ the approximation is linear;

3) $0.000994195P^2 + 0.0856824P + 27.1744 = \varepsilon$ is the second degree approximation (quadratic).

Since the best mathematical model is considered to be the model with the lowest value of the AIC (Akaike Information Criterion), the dependence of the loss values in an optical fiber is better represented by a quadratic approximation.

A plot of the optical fiber with the wavelength of 1550 nm with step-by-step pressure increment loss is presented in Figure 7.

Evaluating the results we can conclude that the loss values in the optical fiber dependence is better represented by the quadratic approximation: $0.00195471P^2 + 0.256042P + 24.1281 = \varepsilon$.

To determine the distance to the place of violation of the perimeter security, the YOKOGAWA AQ1200 OTDR reflectometer was used. On the traces (Figure 8) it is clearly seen in which part of the optical fiber the loss changes.

On the trace it is shown that in the range of 0.411273-0.43224 km the return loss of the optical

fiber is 0.334 dB, which indicates that pressure on the optical fiber in this interval is above the norm.

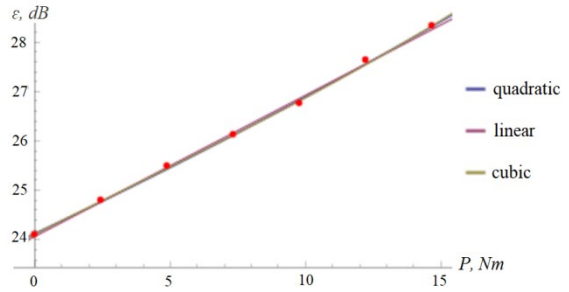

Figure 7: Loss values in the optical fiber with the wavelength of 1550 Nm with the step increment of pressure.

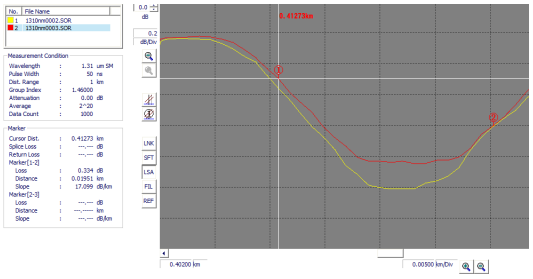

Figure 8: A fiber trace.

Based on our analytical and practical studies, we can draw a number of conclusions on the advantages and disadvantages of these systems. The advantages include the difficulty in detecting fiber-optic sensors (visually or technically), since the sensors are not susceptible to electromagnetic and radio frequency interference, a significant area of the protected perimeter is up to 60...100 km long with the intrusion detection accuracy of up to several meters. The disadvantages include a high cost of the equipment, complexity of setup, a number of preparatory measures for trenching, the need to fulfill certain conditions that ensure reliable system operation with optimal performance, the system planning and mounting the sensors. There is a probability that the system may give a false alarm, as it is sensitive to soil vibrations and seismic signals caused by passing nearby vehicles, large trees, railways, objects under construction. If these factors are present, then the sensors must be placed in shallow trenches filled with small gravel within the "forbidden" zone between two parallel fences, which allows partial isolating the sensors from the effects of these ground seismic effects. It is not recommended to mount the sensors directly into the ground, since a change in soil compaction and its movement can change significantly the system sensitivity and reduce the probability of an intruder being detected. When the sensors are laid directly in the grass, their sensitivity is greatly reduced, an intruder can be detected only when it strikes a fiber-optic light guide. A trench in which fiber-optic sensors are laid must be provided with a drainage system to drain water. The presence of water can lead to freezing at low temperatures (in winter), which will cause decreasing the system sensitivity. Soil erosion can cause the exposure of underground sensors or their immersion to the depth, and this will reduce the system sensitivity or make it completely inoperable.

4 CONCLUSIONS

As a result of the study the experiments were carried out with an optical fiber with the wavelength of 1310 Nm and 1550 Nm using a reflectometer to determine the location of violating the integrity of the protected perimeter. Using the reflectometer it was found that in the tested range of 0.411273-0.43224 km, the return losses of the optical fiber were 0.334 dB, which indicates that pressure on the optical fiber in this interval was higher than normal. In the future, a hardware-software complex will be developed that will make it possible to estimate automatically the obtained parameters of the protection section and to fix violations when the trace curve changes from the specified reference one. The signal will be transmitted automatically to the operator.

REFERENCES

[1] J. C. Juarez and H. F. Taylor. "Field test of a distributed fiber-optic intrusion sensor system for long perimeters," Applied Optics, vol. 46, no.11, pp. 1968-1971, 2007.

[2] Sh.-Ch. Huang and H. Lin, "Counting signal processing and counting level normalization techniques of polarization-insensitive fiber-optic Michelson interferometric sensors," Applied Optics, vol. 45, no.35, pp. 8832–8838, 2006.

[3] H. M. Hashemian, C. L. Black, and J. P. Farmer. "Assessment of fiber optic pressure sensors," United States, N., 1995, doi:10.2172/71391, [Online].

[4] Jonas H. Osório, et al., "Simplifying the Design of Microstructured Optical Fibre Pressure Sensors," Scientific Reports, 7, 2017.

[5] S. Poeggel, et al., Optical Fibre Pressure Sensors in Medical Applications, Sensors, no. 15(7), pp. 17115-17148, 2015.

[6] G. Hayashi, M. Cristiano, B. Cordeiro, A. Marcos, R. Franco and F. Sircilli, Numerical and Experimental Studies for a High Pressure Photonic Crystal Fiber Based Sensor Juliano, AIP Conference Proceedings 1055, pp. 133-136, 2008; doi: 10.1063/1.3002521. [Online] Available: https://doi.org/10.1063/1.3002521.

[7] F. Urban et al., Design of a Pressure Sensor Based on Optical Fiber Bragg Grating Lateral Deformation, Sensors (Basel), no. 10 (12), pp. 11212-11225, 2010, doi: 10.3390/s101211212 [Online].

[8] B. S. Vvedensky, "Волоконно-оптические сенсоры в системах охраны периметра," Мир и безопасность, no. 4-5, 2006.

[9] .V. Polyakov, M. A. Ksenofontov. "Frequency fiber-optical alarm system," International Conference on Laser, Applications and Technologies (LAT-2007), Minsk, June, p. 93, 2007.

[10] A. L. Markhakshinov, A. A. Spector, "Estimation of the trajectory of a person's movement in a local area in a seismic protection system," Coll. of scientific works of NSTU, no. 1 (59), pp. 59-64, 2010.

[11] D. O. Sokolova, A. A. Spector, "Classification of moving objects by spectral signs of seismic signals," Autometry, no. 5, pp. 112-119, 2012.

[12] D. O. Sokolova, A. A. Spector, "Nonparametric detection of seismically active objects with continuous impact on the ground," Scientific Herald of NSTU, no. 4, pp. 20-28, 2012.

[13] D. O. Sokolova, A. A. Spector, "Classification of moving objects based on spectral features of seismic signals," Optoelectronics, Instrumentation and Data Processing, no. 5, pp. 522-528, 2012.

[14] A. L. Markhakshinov, A. A. Spector, "Evaluation of the local characteristics of the movement of an object in a seismic protection system," Avtometriya, no. 5 (45), pp. 48-53, 2009.

[15] A. L. Markhakshinov, "Evaluation of the characteristics of human movement in the seismic protection system," Proceedings of the All-Russian Scientific Research Conference of Young Scientists "Science. Technology. Innovation", Novosibirsk, part 2, 2009, pp. 111-112.

[16] A. V. Yurchenko, A. D. Mekhtiev, N. I. Gorlov, A. A. Kovtun, Research of the Additional Losses Occurring in Optical Fiber at its Multiple Bends in the Range Waves 1310nm, 1550nm and 1625nm Long, Journal of Physics, Conference Series, pp. 27-31, Jul, 2015, doi:10:1088/1742-6596/671/1/012001. [Online] Available: http://dx.doi.org/10.1088/1742-6596/671/1/012001.

[17] A. Yurchenko, A. Mekhtiyev, A. Alkina, F. Bulatbayev, Y. Neshina. The Issues of Development of Fiberoptic Sensors for Measuring Pressure with Improved Metrological and Operational Characteristics. VII Scientific Conference "Information-Measuring Equipment and Technologies", MATEC Web of Conferences 79, p. 01085, 2016, doi: 10.1051/01085/matecconf/ 201679001085, [Online].

Development of Exercise Designing Module for Computer Training Complex

Filipp Shkliaev and Rustam Fayzrakhmanov
Department of Information Technologies and Automated Systems, Perm National Research Polytechnic University, 29 Komsomolsky prospekt, Perm, Russia
fishkliaev@gmail.com, fayzrakhmanov@gmail.com

Keywords: Training Simulation Complex, Ontological Modelling, Exercise Designing.

Abstract: The development of exercises for computer training complexes is a time-consuming process, which includes many factors that must be considered. In the training complex, it can be simplified by a special module that takes into account the characteristics of the subject area and simplifies the development of the exercise. However, at the moment such a module has not been submitted. The aim of the article was to develop an exercise designing module that takes into account the possibility of interaction between elements of the simulator and compiles the exercise scenario. The ontological model was developed for the selection of consistent elements; finite-state machine was comprised the exercise scenario. The novelty of developed module was the integration of 3D-scene creating tool and an exercise scenario designer, based on the ontology of simulator' subject area. The developed module was tested by designing an exercise scenario included in the examination program of the National Commission for Certification of Crane Operators (NCCCO).

1 INTRODUCTION

Computer training complexes (CTC) are designed to train personnel and develop the necessary skills for the effective, high-quality and safe work. An instructor is responsible for a personnel training at CTC. His tasks include personal work with the trainees, tracking their learning progress and designing a training program which consists of individual exercises. Exercises are necessary for developing professional skills. Each exercise in the CTC consists of a 3D-scene and a scenario of actions that the student must perform in order to complete the exercise successfully. When developing a 3D-scene for an exercise, it is worth considering that not all of its elements can interact with each other. For example, grab can not be used for container moving. Therefore, developing and combining all the components of an exercise is a time-consuming process. Therefore, the actual task is to develop a module, which would take into account the parameters of objects on the 3D-scene, the possibility of their interaction with each other and simplified the process of creating an exercise scenario. This module will be useful for both the instructor and the developer, as it reduces a time for exercise designing and allows to be flexibly customizing.

Some authors have described the various parts of the exercise using ontology and finite-state machine. The authors of the [1], [2] described possibilities of ontological modelling used in simulation systems, citing as an example CTC.

In the [3], [4] the authors developed a system for calculating the trajectory of the weapons shells flight, based on the ontological approach. The ontological model contained equations, parameters of elements and entire structure of system. To calculate the trajectory, it was necessary to convert the ontology into the MATSIX project for using in MATLAB.

In [5] and [6], ontology described geometric-temporal concepts of trajectory and motion. The authors presented the path as a set of key points consisted of its position and timestamp. The model was used to compile routes using data from various sensors. In [7], control points, in addition to timestamp and position, were characterized by speed and acceleration. However, these models worked with already prepared data, without defining the requirements for the exercise.

In [8], a finite-state machine of a water pump was developed. The nodes of model were the states, and links consisted conditions for transition from one state to another. However, the authors of the article have not described the process of designing a model. Hence, the problem of developing a system for designing of exercises can't be considered solved.

Therefore, the purpose of this article was to develop a module for the designing of an exercise which takes into account the possibility of interaction between the elements of the simulator and allows to compile the exercise scenario.

2 EXERCISE DESIGNING ALGORITHM

As an exercise in this article, we understood the complex of objects involved in the learning process and the scenario of the exercise.

The module was developed as a part of CTC, described in [9], which has a control system for the formation of sensorimotor skills [10]. Control system automates learning process by defining of exercises for the student, depending on his skill level. A set of necessary exercises need to be designed by the developed module.

For the exercise designing, we need to identify objects, involving in a learning process. As a result of the analysis of various teaching programs and training standards [11], typical objects were united in classes and presented in Table 1.

Table 1: Learning process parts and classes, related to them.

Part	Class
Operating machine	Crane
	Lifting device
Cargo properties	Cargo
	Quantity of cargo
	Starting place
	Destination place
Workers	Banksman
	Signaller
	Slinger
Environmental conditions	Weather conditions
	Time of day
	Speed and direction of wind

Every object in a simulator belongs to some class. When designing an exercise, a user should choose from all objects of a class a specific one.

However, it is worth noting that not all elements of the subject area can interact with each other. For example, a grab can only interact with bulk loads, and a 40-foot container cannot be connected to a hook. In order to take into account the restrictions when setting the parameters of a 3D-scene, it was necessary to develop an ontology that takes into account the components of the simulator and its relationships. The developed model was used as the basis for generating a 3D exercise scene.

The second step of the exercise designing is to set up a scenario which includes a set of control points and the conditions for their passage. In addition to the correct sequence of actions, the scenario can also contain student's erroneous actions and the response of the simulator to it.

Based on the material studied, the exercises designing algorithm was compiled. It included such steps as:

1) Selection of components involved in the exercise, based on the description of the subject area;

2) Arrangement of selected components on the 3D-scene and generation of start and end control points of cargo position;

3) If necessary, the placement of additional control points and setting conditions of their passage.

3 THE ARCHITECTURE OF THE EXERCISE DESIGNING MODULE

Figure 1 shows the architecture of module that was developed based on the exercises designing algorithm.

The Ontology of simulator is an ontology that contains data about all simulator elements, which can be used in an exercise designing and relationships between them. The choice of ontology is explained by the simplicity of organization and obtaining of necessary information. In addition, a reasoner is able to infer new logical consequences, that helps in complex queries.

The Scene manager module output is 3D-scene in a simulator. There were 3 ways to get a 3D-scene as a result: 1) Creating a new scene from basic components manually; 2) Generating new scene by setting some input parameters; 3) Loading a scene, that was saved before.

The easiest to instructor way to prepare a 3D-scene is automatically generating it or to load saved scene. In case of scene generation, module used the elements selected in the interface and information

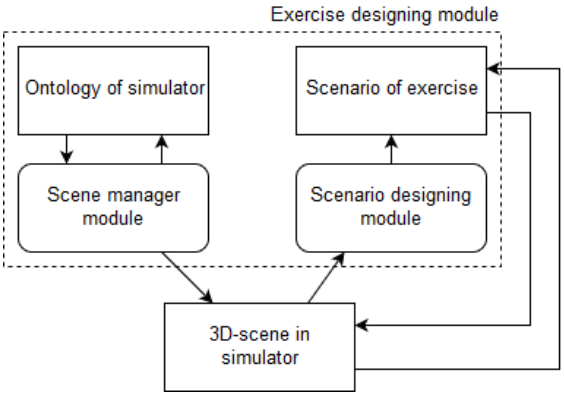

Figure 1: Architecture of an exercise designing module.

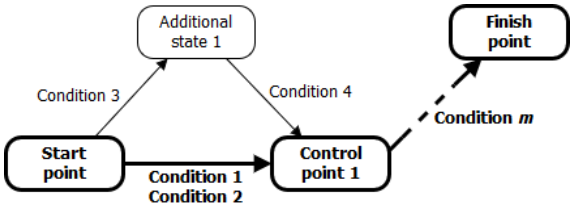

Figure 2: Example of exercise scenario scheme.

about them from the ontology and then converts it into a 3D-scene, which can be improved later by the developer or instructor. SPARQL queries were used by the interface for interaction with ontology of simulator.

The purpose of the Scenario designing module was to create a sequence of control points and set conditions for their passage. The result of the module work was a finite-state machine, represented an Exercise scenario. The nodes of the finite-state machine are control points, and the links contains conditions for the transition from one point to another.

Figure 2 shows the scheme of an exercise scenario, where bold-style blocks and arrows composed the path, which cargo should pass to finish an exercise. In addition, additional states can be configured that take into account the possible erroneous behavior of the student, examples of ones presented below:

- Out from permissible borders;
- Touched the ground;
- Cable break, etc.

Conditions can track the following parameters:

- Reaching the defined position;
- Cargo collision with a ground or other scene objects;
- Scene object position and rotation;
- Cables length and tension;
- Cargo speed and acceleration.

Scenario designing module generates start and finish only, based on position of starting and destination places, defined in Scene manager module. The other states and conditions can be added manually.

The result of the exercise design module was generated 3D-scene and a scenario for the exercise.

4 MODEL TESTING

An exercise designing module for the AnyCrane CTC [12] was developed.

To test the developed module, a full cycle of creating an examination exercise from the NCCCO certification standard [10] for tower crane operators was designed. The 3D-scene contained a crane, a cargo and poles that limit the route. At the beginning of the exercise, the cargo was at the start point; the hook and the cargo were connected by slings. The task was to move the cargo to the finish point.

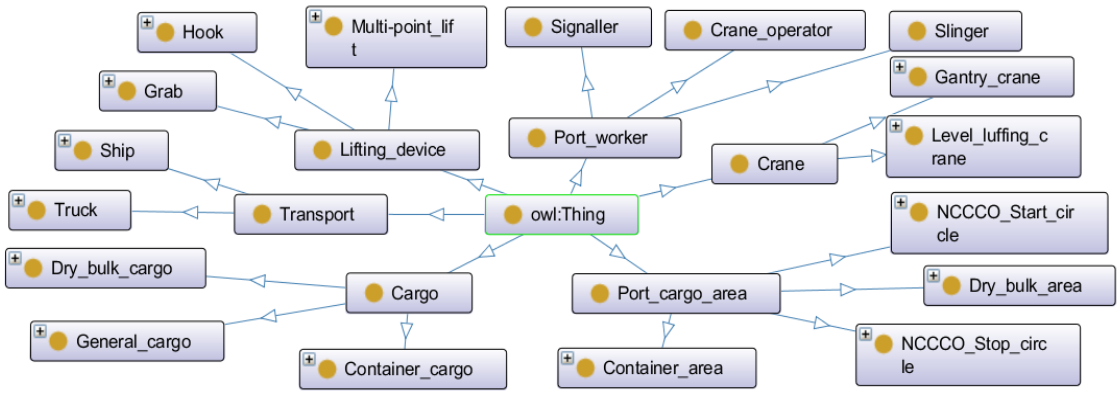

Figure 3: Class hierarchy of ontology model.

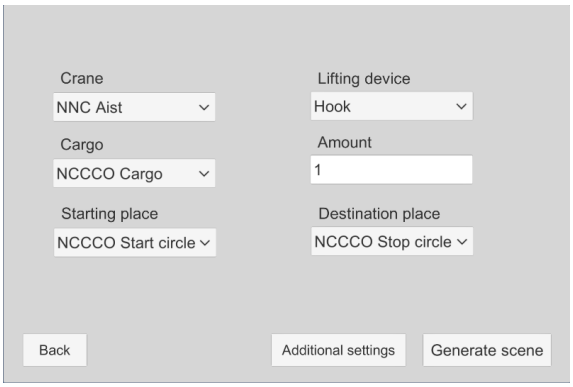

Figure 4: Scene options selection in the Scene manager module.

The sequence of the exercise is defined as follows:

1) Raise the hook with a cargo at least 10 feet above the ground to avoid obstacles and personnel.
2) Move the cargo from the start circle to the end circle.
3) Once the cargo has reached the end circle, place it in such a way that the cargo contacts the ground inside the circle and remains there.
4) When the cargo is contacted the ground surface inside the finish circle, cargo is not allowed to rise above the ground.
5) The examiner will give a stop signal when the cargo is under control.

In accordance with the subject area of the CTC, an ontology was developed. It includes a structural model of classes, instances of simulator objects with its parameters, and relationships between ontology objects. Figure 3 shows the 2-level class hierarchy of the ontology model.

For example, a query can retrieve a connection device, which can be used if a user has selected CMM Aist crane and brick pallet as a cargo. The SPARQL query, which was sent to the ontological model is presented below.

```
SELECT ?LiftDev
WHERE
{
:CMM_Aist :canBeConnectedWith
?LiftDev .

?LiftDev :canTransports
:Brick_pallet
}
```

The query result is:

```
:Hook
```

Figure 4 demonstrates the interface that allows user to select elements of the exercise. Every parameter can be chosen from the dropdown list, which was formed by obtaining query result. The following parameters were selected: NNC Aist, Hook, NCCCO Cargo, 1, NCCCO Start circle, NCCCO Stop circle. The scene was edited manually: poles, the start and stop circles were added.

Using the scenario development module, checkpoints and conditions for their passage were added in accordance with the sequence of the exercise. The first control point is placed above the starting position in accordance with step 1. Intermediate points were set at the corners of the route. The last point was determined automatically, in the stop circle.

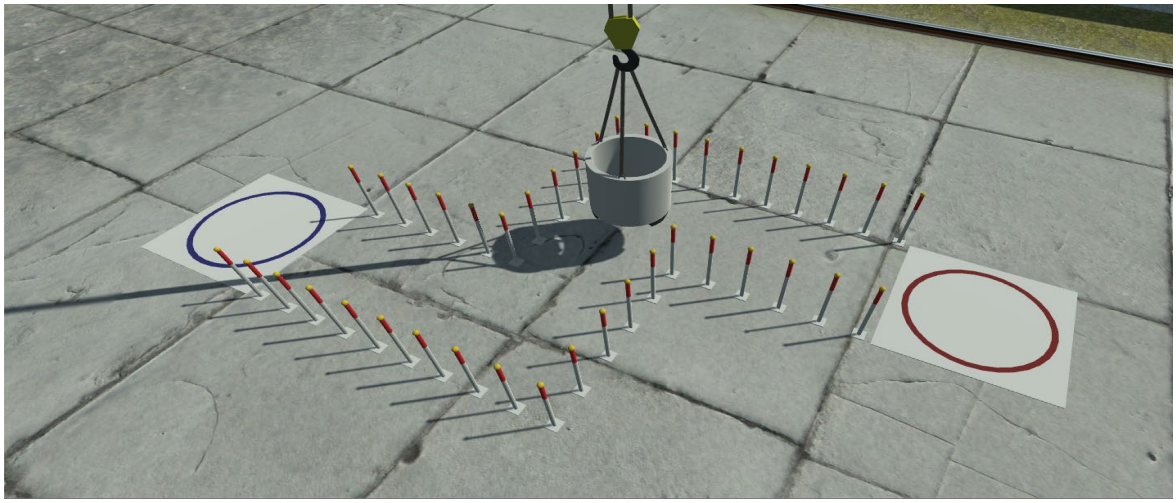

Figure 5: Exercise performing process in the 3D-scene of AnyCrane CTC.

Then the main exercise scenario was compiled by defining conditions for passing each control point. The main scenario of the exercise was extended with additional branches, taking into account the touch of the poles and the ground surface.

Figure 5 shows the exercise process. A banksman can be placed on the 3D-scene. It will signal the target position and conditions for passing the point.

5 CONCLUSIONS

As a result of the work, the original technique and the software module for the exercise designing were developed. Also, the module was integrated into the AnyCrane TSC. The module was tested by building an examination scenario for the "crane operator" NCCCO qualification.

The novelty of the developed system was the integration of 3D-scene creating tools and an exercise scenario designing based on the ontology of simulator' subject area. The use of the ontological model made it possible to determine the consistent elements of the 3D-scene and compile the exercise scenario.

The module can also be used in another simulator' subject area, such as technological process or medical operations. A subject area ontology of the CTC should be composed.

However, the developed module has limitations. The first is the inability to set different conditions for control points for different cargos. The second is the need to write additional code to register events occurring in alternative ways of the exercise scenario.

ACKNOWLEDGMENTS

The reported study was funded by RFBR according to the research project № 18-38-00835.

REFERENCES

[1] P. Benjamin, M. Patki, R. Mayer, "Using ontologies for simulation modeling," Proceedings of the 38th conference on Winter simulation, 2006, pp. 1151-1159.

[2] E. Holohan, M. Melia, D. McMullen and C. Pahl, "The generation of e-learning exercise problems from subject ontologies," Advanced Learning Technologies on Sixth International Conference, 2006, pp. 967-969.

[3] U. Durak, H. Oğuztüzün, C. Köksal, and Ö. Özdikiş, "Towards interoperable and composable trajectory simulations: an ontology-based approach," Journal of Simulation 5, no. 3, pp. 217-229, 2011.

[4] U. Durak, S. Güler, H. Oğuztüzün and S. K. İder, "An exercise in ontology driven trajectory simulation with MATLAB SIMULINK (R)," Proceedings of the 21th European Conference on Modelling and Simulation (ECMS), 2007, pp. 1-6.

[5] M. Manaa and A. Jalel, "Ontology-based trajectory data warehouse conceptual model," International Conference on Big Data Analytics and Knowledge Discovery, pp. 329-342, 2016.

[6] M. Manaa and A. Jalel, "Ontology-based modeling and querying of trajectory data", Data & Knowledge Engineering 111, pp. 58-72, 2017.

[7] T. P. Nogueira, R. B. Braga, H. Martin, "An ontology-based approach to represent trajectory characteristics," Fifth International Conference Computing for Geospatial Research and Application, 2014, pp. 102-107.

[8] N. Walkinshaw, R. Taylor, J. Derrick, "Inferring extended finite state machine models from software executions," Empirical Software Engineering, vol. 21, no. 3, pp. 811-853, 2016.

[9] R. Fayzrakhmanov, I. Polevshchikov, A. Khabibulin "Computer Simulation Complex for Training Operators of Handling Processes," Proceedings of International Conference on Applied Innovation in IT, vol. 5, pp. 81-86.

[10] R. Fayzrakhmanov, I. Polevshchikov, A. Polyakov, "Computer-aided Control of Sensorimotor Skills Development in Operators of Manufacturing Installations", Proceedings of International Conference on Applied Innovation in IT, vol. 6, pp. 59-65.

[11] The National Commission for the Certification of Crane Operators (NCCCO), [Online], Availible: http://www.nccco.org/home.

[12] R A. Fayzrakhmanov, I. Polevshchikov, A. Khabibulin, F. Shkliaev, R. R. Fayzrakhmanov, "ANYCRANE: Towards a better Port Crane Simulator for Training Operators," 15th International industrial simulation conference, 2017, pp. 85-87.

Method of Data Dimensionality Reduction in Brain-Computer Interface Systems

Rustam Fayzrakhmanov and Roman Bakunov
Department of Information Technologies and Automated Systems, Perm National Research Polytechnic University, 29 Komsomolsky prospekt, Perm, Russia
fayzrakhmanov@gmail.com, bakunov_roman@mail.ru

Keywords: Brain-Computer Interface, Information-Measuring System, Data Dimensionality Reduction, Linear Discriminant Analysis.

Abstract: The article is devoted to the problems of performance increasing of information-measuring and control systems based on brain-computer interface technology (BCI). BCI is a technology that allows communication between the brain and the external environment only on the basis of processing of the electroencephalogram (EEG). The functioning of the BCI system can be represented as a cycle. In each iteration, the EEG signal is measured and preprocessed, the characteristic features are extracted, the classification is implemented and the control action corresponding to the recognized command of the operator is generated. For the functioning of BCI systems in real-time mode it is necessary to solve the problem of the processed data dimensionality reduction (without losing significant information). The article describes the author's algorithm which is designed to use for that purpose. The algorithm is based on the use of digital signal processing and cluster analysis. Also, the results of experimental testing of the approach are described in the article. The experiments showed that proposed approach allows to significantly reduce the time required to perform operations of data dimensionality reduction. In addition, it's using has not negative affect on the clustering quality of processed sets of signals. It is experimentally confirmed that the developed algorithm effectively works in conjunction with the linear discriminant analysis (LDA), acting as a preprocessor for the LDA. At the same time, the speed of such bundle is much higher than speed of LDA without the preprocessor.

1 INTRODUCTION

Brain-computer interface (BCI) technology is based on the measurement of user EEG signals and the recognition of conscious brain electrical activity using digital signal processing techniques.

In BCI systems, electroencephalogram (EEG) signals analysis is often performed in the frequency domain. In this case, the set of processed signals can be represented as a set of points in N-dimensional space, where N is the number of allocated spectral components. Often this value can be equal to 32, 64, 128, etc. A large amount of coordinates in the processed vectors generates a number of problems.

Firstly, it is impossible to visualize them on a plane or in three-dimensional space for visual presentation. Secondly, the dimensionality of the data strongly influences on the computational complexity of processing operations. Thirdly, it is necessary to provide an acceptable level of clustering quality of the explored data sets, and it is often impossible when amount of coordinates is too large.

At the moment, there are methods that solve the problem of processing of multidimensional data [1], [2], [3]. One of them is a linear discriminant analysis (LDA). This is one of the fastest approaches used by researchers to reduce the dimensionality of the processed data in BCI systems [4]. However, its use requires a fairly productive computer, because it is associated with complex calculations.

Therefore, there is a need for new methods of data dimensionality reduction in BCI systems. One of these approaches is described in the article.

2 METHOD DESCRIPTION

The initial data are the averaged amplitude spectrums of the EEG signals measured during the

preparation of the learning sample. These spectrums are considered as vectors. They form clusters that correspond to K different operator commands. Let's denote the number of vectors in each cluster by L. N is the number of components in each vector. It is proposed to represent an arbitrary cluster $\mathbf{K}_i$ in the form of a matrix (1):

$$\mathbf{K}_i = \begin{bmatrix} y_{11} & \cdots & y_{1N} \\ \vdots & \ddots & \vdots \\ y_{L1} & \cdots & y_{LN} \end{bmatrix}, i = 1...K. \quad (1)$$

The rows of this matrix are the vectors included in the cluster, and the columns are the coordinates of these vectors.

The functioning of the proposed method for the data dimensionality reduction is based on the assumption that the values of the same coordinates within the current cluster have some similarity. This assumption follows from the fact that each cluster corresponds to one particular operator command. Similarly, when comparing different clusters, the same coordinates of the vectors belonging to them should differ in some way. Thus, the method is based on the idea of exploring of same columns taken from different matrices for the presence of similarities or differences.

As a specified measure of similarity or difference, it is proposed to use the following concepts: the distance between vectors (2) and the cross-correlation coefficient between two signals (3). These concepts are described in detail in [5].

$$d(\mathbf{f},\mathbf{g}) = \|\mathbf{f}-\mathbf{g}\| = \sqrt{\sum_{k=1}^{N}(f_k - g_k)^2}, \quad (2)$$

$$r_{fg}(j) = \frac{R_{fg}(j)}{\frac{1}{N}\sqrt{\sum_{i=1}^{N} f_i^2 \sum_{i=1}^{N} g_i^2}}. \quad (3)$$

In the given equations, **f** and **g** are signals (vectors) consisting of *N* samples, $R_{fg}(j)$ is the cross-correlation function between signals **f** and **g** at shift *j*.

In the process of the method application a number of coordinates is excluded. The possibility of exclusion is based on a special criterion. These coordinates do not have a significant effect on the differences between the clusters. Therefore, the application of the algorithm should not lead to a downgrade of the clustering quality.

A step-by-step description of the proposed algorithm is given below.

The columns of each matrix $\mathbf{K}_i$ are considered as signals whose number of samples is equal to L. The same columns from different matrices are considered in pairs. Let P be the total number of such pairs (for columns with the same numbers). It depends on the number of clusters K and is calculated by the (4), which determines the number of edges for a complete graph with K vertices:

$$P = \frac{K(K-1)}{2}. \quad (4)$$

For each pair it is necessary to calculate the cross-correlation coefficient (at zero shift) and the distance (in accordance with the equations given earlier). The result is a vector consisting of two components and shown in the (5):

$$\mathbf{f_{ip}} = \left(r_{x_i y_i}(0); d(\mathbf{x_i}, \mathbf{y_i})\right), \quad i = 0...N-1,\ p = 0...P-1 \quad (5)$$

In this equation, $\mathbf{x}_i$ and $\mathbf{y}_i$ are signals composed of elements of same columns of matrices corresponding to two different clusters. The $\mathbf{f}_{ip}$ vectors calculated for different pairs of same columns and summed up as illustrated by (6):

$$\mathbf{f_i} = \sum_{p=0}^{P-1} \mathbf{f_{ip}} = (R_i, D_i), \quad i = 0...N-1. \quad (6)$$

The key idea of the proposed method is to begin the dimensionality reduction at the least important coordinate. Therefore, it is required an objective function, the largest value of which corresponds to the most important coordinate, and the smallest – to the least important. It is proposed to adopt the expression given in (7) as this objective function.

$$Y_i = CR_i + D_i. \quad (7)$$

The coefficient C is calculated in accordance with the (8):

$$C = -\frac{\max\{D_i\}}{\max\{R_i\}} + 1, \quad i = 0...N-1. \quad (8)$$

The function Yi is calculated for each vector fi. The calculated values are sorted in ascending order. After this, the coordinate numbers i (i = 0 ... N-1) are written in the order corresponding to the increase of the values of the function Yi. As a result, an array consisting of coordinates numbers sorted in order of increasing importance (determined according to an accepted criterion) will be obtained. This array will start with the number of the least important coordinate, which can be eliminated first.

The coordinates exclusion is performed step by step in accordance with their order in the described array. At each step it is important to check how it affects on the quality of clustering. If the quality of clustering has not downgraded, then the next coordinate can be excluded. Thus, the data dimensionality is reduced step by step. The values of special indicators must be calculated at each step. As such criteria it is proposed to use the compactness

and isolation index CS and the efficiency index PI. Their descriptions are given in [6], an example of practical application is described in [7].

If there is a downgrade of the clustering quality in comparison with the initial values of criteria (before the dimensionality reducing), then the process ends here. Next, the best combination of values of the criteria is selected, and the corresponding step is remembered. After that, all excluded coordinates are restored, and then, the coordinates are eliminated again (in accordance with the sorted values Yi) until the specified step (including it). Thus, the final values of the clustering quality criteria will match with the selected best values.

The algorithm scheme of the proposed method of the data dimensionality reduction is presented in Figure 1.

Continuously calculation of the clustering quality criteria CS and PI can put a heavy load on the computing device, especially if it has low performance.

Therefore, it is proposed to use in practice a slightly modified version of the developed method.

The number of calculations can be significantly reduced if desired data dimensionality, which should remain after the algorithm operation, is specified before starting. In that way the required number of the least important coordinates will be excluded.

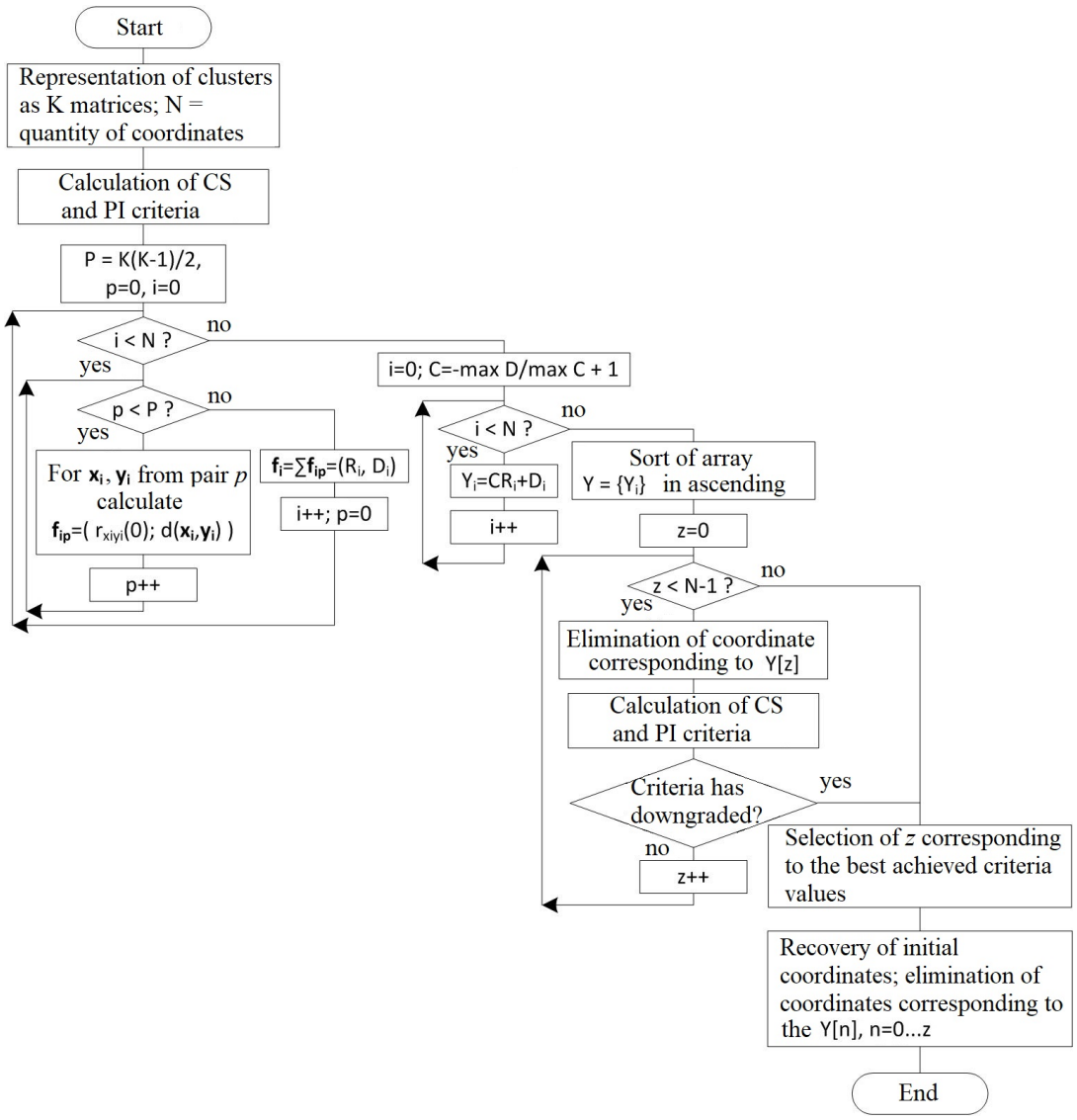

Figure 1: The scheme of the sequential elimination of the least important coordinates.

After that it will be possible to conclude whether it is permissible to use the desired dimensionality of the data or it needs to be increased. This conclusion is based on the values of the CS and PI criteria. If the desired data dimensionality reduction has led to a downgrade of the clustering quality, then it is required to restore the eliminated coordinates one by one and checking the values of the CS and PI criteria in each step of restoring.

The developed approach demonstrates the greatest efficiency in the role of a preprocessor in conjunction with LDA. Schematically, such data processing sequence is illustrated in Figure 2. Experiments have shown that using of this scheme makes it possible to effectively reduce the dimensionality of the processed data faster than in the case when LDA is used without a preprocessor. In addition, the proposed approach allows the use of all practical benefits offered by LDA, which has proven itself in information-measuring systems [8].

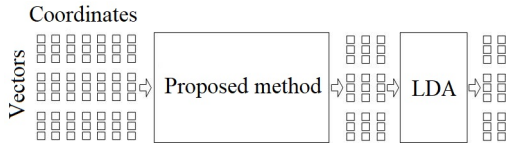

Figure 2: Proposed data processing sequence.

3 EXPERIMENTAL PART

The results of experimental testing of the developed method in conjunction with the LDA are given below.

EEG signals were measured by the NeuroSky MindWave Mobile system (the sampling frequency is 512 Hz).

Software for sampling and digital signal processing was created in the LabVIEW environment [9].

A computer with the following characteristics was used for the experiments:

- processor: Intel Core i7 with a clock speed of 2.2 GHz;
- RAM: 8 GB DDR3.

Analysis of the developed method was carried out in accordance with the following sequence of steps:

- Registration of EEG signals, their primary processing and grouping into clusters corresponding to specific commands of the operator.
- Splitting of each signal into segments consisting of the number of N_{SEG} samples, performing a fast Fourier transform (FFT) for the segments and calculation of the averaged amplitude spectrum for each signal.
- Calculation of compactness and isolation index CS_0 and efficiency index PI_0 for clusters consisting of averaged amplitude spectrums.
- Application of LDA to the clusters. Measurement of LDA processing time t_{LDA}.
- Calculation of the CS_{LDA} and PI_{LDA} criteria for clusters modified after the application of LDA.
- Processing of the original (not affected by LDA) clusters using the developed algorithm (in conjunction with the LDA). Measurement of processing time t_{CORR}.
- Calculation of the CS_{CORR} and PI_{CORR} criteria for clusters modified after the application of the proposed algorithm.
- Analysis of the calculated indicators (t_{LDA} и t_{CORR}, CS_{LDA} и CS_{CORR}, PI_{LDA} и PI_{CORR}).

This sequence of steps was repeated many times with a certain variability (the EEG signals were measured again, different operator commands were used, the sizes of the segments were changed, etc.).

Experiments have shown that the developed approach is many times faster than “pure” LDA, especially for large dimensionality of the original feature space. Graphically, this is illustrated in Figure 3.

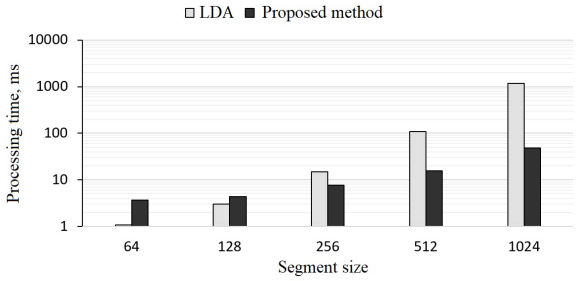

Figure 3: Performance comparison of the developed method and LDA.

Figures 4 and 5 show graphs illustrating the changes in the CS and PI criteria. These criteria are given in relative form.

Both methods improve the values of the CS and PI criteria. It should be noted that smaller values of the CS criterion are preferable because they correspond to more compact, isolated clusters. In the case of the PI criterion, larger values are preferable.

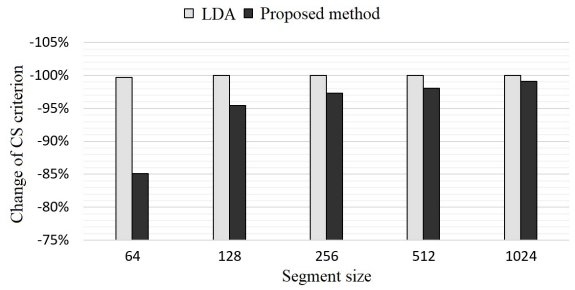

Figure 4: Comparison of the developed method and LDA by criterion CS.

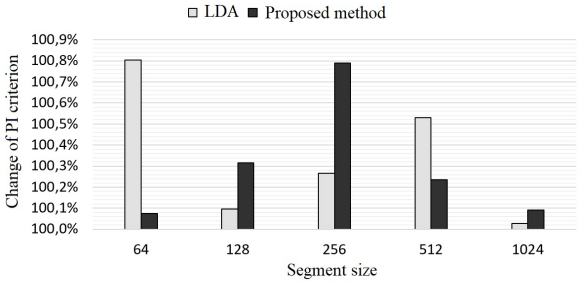

Figure 5: Comparison of the developed method and LDA by criterion PI.

Thus, the developed method of data dimensionality reduction shows a higher performance than the LDA without a preprocessor. At the same time, there are no major differences in the influence of the methods on the clustering quality criteria.

4 CONCLUSIONS

The described algorithm has several advantages. Its functioning does not depend on the statistical characteristics of the processed EEG signals. In addition, it is invariant to such problematic factors as the individual characteristics of the user, the design features of the BCI, the details of the control and feedback organization.

Implementation of the proposed approach does not require large computing resources. The application of the described method in conjunction with the algorithms of the learning sample reduction [10] will allow to design BCI systems based on microprocessor devices with low performance, small size of memory, low power consumption and long battery life.

The proposed method can be used not only in BCI systems. The algorithm can be implemented in any information-measuring and control systems which functioning is associated with the processing of multidimensional data.

REFERENCES

[1] Dzh. Tu, R. Gonsales, “Principles of pattern recognition,” Mir, 1978.

[2] I. Gajdyshev, “Data analysis and processing: special reference book,” Piter, 2001, 752 p.

[3] S. A. Ajvazjan, V. M. Buhshtaber, I. S. Enjukov, L. D. Meshalkin, “Applied Statistics: Classification and Dimension Reduction,” Finansy i statistika, 1989, 607 p.

[4] T. Lan, L. Black, J. Van Santen, D. Erdogmus, “A comparison of different dimensionality reduction and feature selection methods for single trial ERP detection,” Annual international conference of the IEEE engineering in medicine and biology society (EMBC'10), 2010, pp. 6329-6332.

[5] E. C. Ifeachor, B. W. Jervis, “Digital Signal,” Processing: A Practical Approach, Second Edition, Izdatel'skij dom «Vil'jams», 2004, 992 p.

[6] A. A. Barsegjan, M. S. Kuprijanov, I. I. Holod, M. D. Tess, S. I. Elizarov, “Data and process analysis,” Tutorial, Third edition, revised and enlarged, BHV-Peterburg, 2009, 512 p.

[7] R. A. Fajzrahmanov, R. R. Bakunov, O. A. Kashin, “Application of criteria of the clustering quality in information-measuring systems,” Nauchnoe obozrenie, no. 8, pp. 231-238, 2015.

[8] R. A. Fajzrahmanov, R. R. Bakunov, “About improving of the quality of signals clustering in technical systems using linear discriminant analysis,” Elektrotehnika, no. 11, pp. 55-59, 2016.

[9] Dzh. Trjevis, Dzh. Kring, “LabVIEW for everyone,” Fourth edition, revised and enlarged, DMK Press, 2011, 904 p.

[10] R. A. Fayzrakhmanov, R. R. Bakunov, “The reduction of learning sample in information-measuring and control systems based on brain-computer interface technology,” 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM) : Proc., Chelyabinsk, Russia, May 19-20, 2016. [Online] Available: http://ieeexplore.ieee.org/document/7911544/

Beward Android Mobile Awareness Platform about Natural Disasters

Goran Jakimovski[1], Danco Davcev[2] and Marija Kalendar[1]
[1]*Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, PO Box 574, 1000 Skopje, Macedonia*
[2]*Faculty of Electrical Engineering and Information Technology, Ss. Cyril and Methodius University, Karpos II bb, PO Box 574, 1000 Skopje, Macedonia*
{goranj, marijaka}@feit.ukim.edu.mk, dancodavcev@gmail.com

Keywords: Natural Disaster, Earthquake, Flood, Fire, Crowdsourcing, Social Networks, Facebook.

Abstract: Latest developments in crowd-sourcing and inter-user information sharing has led to the idea of sharing crucial information about a near disaster. Having this information can decrease the number of casualties. If people are aware of a near disaster, they can more easily avoid it, which in term, can minimize the damage it makes. The fastest way to get informed, in these situations, is by having peer information and on-time alert. Users of the social networks, have shown, over the years, that they can share information fast in time of a crisis. This type of information sharing is commonly known as crowd-sourcing. In this paper, we propose an android awareness platform called Bewared, which, in term, allows users of the social networks to collaborate, share information and pinpoint a natural disaster with its location. The types of natural disasters, used in this research, are: earthquakes, fires, floods, fire and their level of validity. The application uses some of the existing platforms and social media networks. In our case, Bewared is supported by Facebook due to its option (application interface) to geographically locate each user of the network, by using their mobile device its real time interaction possibilities.

1 INTRODUCTION

Each person (internet user) is equally responsible for the environment, interaction with other people and the way of living. According to latest research finding in [1], the chances of facing a catastrophe or a disaster are increasing. Since we are witnesses of several natural disasters [1], that happened in the near environment without previous awareness, it was the motivation to explore and develop an application to help users stay informed of disasters by other users who can share this information. Since people nowadays are constantly linked on social networks and share information about anything, the idea of this research is to use this massive information to alert users of danger and thus avoid it. Crowd-sourcing can be used as a method to obtain data from different users and reconstruct the object. In our case, we can use crowd-sourced information from different users to increase the liability of the information. This means that if more users report the similar information (at a near geo-location), the information is considered more reliable. With the number of users continually increasing over time, the crowd-sourcing is being more specific and the data that is being generated is more detailed and quantitative [2].

As explained in [3], generating big data of knowledge in a group can increase the validity of the information passed and confirm it as a fact. This is why we need more users to confirm the information and help others. In terms of popularity and number of users and social media ranking, Facebook is the most popular social media network worldwide. A wide selection of social networks also heavily relies on user-generated content: image-heavy Tumblr, Instagram and Pinterest. However, these networks are home-based and content-oriented. Data can be automatically extracted from social media sites via Application Programming. API's are generally used for recording an event sent by an individual or bundle of users, accumulating within collections for later analysis. The data that will be kept in the database using API's can be manually selected and saved in variables as properties.

After the information is obtained from the network, it can be analysed and alert other users of the network. There can be levels of different alerts

or alert types to indicate the level of danger the user is in. Furthermore, the data collected can help the system produce more accurate and more reliable conclusions about the disasters. Since our data acquisition uses the Facebook platform, the data also will have the location of the users who agreed to the terms of use of Facebook. Also, for Facebook users to be able to use the Bewared application, they would have to agree that their data will be used in additional calculations and data generation for the purposes of the Bewared. The main research in this paper is the Bewared application, which uses crowd-sourced data from Facebook users to alert about near disasters. This information is verified by crisis management agencies that marks information as verified of not. The reliability of the crisis information is increased with the number of users that are reporting it. The paper is organized in sections, where Section 2 presents state of the art solutions and existing services that provide access to crisis data. Section 3 of this paper presents the data gathering from Facebook and crowdsourcing of the Bewared application, whereas, Section 4 presents the architecture of the Bewared verification system. Section 5 presents a case study of the system with 15 users and Section 6 concludes the paper.

2 RELATED WORK

There are a numerous of different platforms that help, support and improve the process of Prevention, Preparedness, Response and Recovery (PPRR). Some platforms isolate and focus on particular disasters, whereas, others are more general and tend to alert all sorts of anomalies and possible harmful situations. Some small number of these are open source platforms, whereas, most of them are commercial applications.

As an example, the Gov.uk platform, presented in [4], is a specific website for England and Wales, where one can specify any postcode of any place in one of these countries and obtain an information about a flood or get a flood warning.

Another example is the Palantir platform, presented in [5], which is a global website that tries to prepare the individuals for unexpected harmful circumstances is mainly used for risk management. This platform is developed and used mainly to support and integrate massive volumes of data for crisis response operations on a moment's notice. This data includes publicly available data, damage assessments, satellite imagery, weather reports, geospatial information on key infrastructure and relief resources, as well as, reports from news agencies and governments. Furthermore, this data is available for researchers and users to analyze and use via API (Application Programming Interface) in JSON format, [5]. An example of such information is shown in Figure 1.

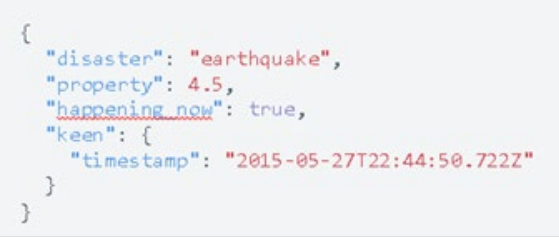

Figure 1: Data collected using API.

The government of Virginia developed and still maintains the VIPER platform (The Virginia Interoperability Picture for Emergency Response) [6]. This platform contains weather reports, flood gages, traffic incidents, wildfires and many other hazards, which is accompanied with the location it covers.

The following sites represent crowd-sourced incident maps, created and maintained mainly by peer users:

Hotosm in [7] creates collaborative maps for humanitarian help, that unlike the previous platforms, in case of a major disaster, this platform helps to gather volunteers. The application creates a map of incidents and maps paths for the volunteers to reach and help victims. It organizes the paths by reaching the victims that require more immediate help. The platform follows the Reach and Rescue protocol. It receives information from victims in a JSON format and creates the routes and maps to reach them.

Climatecolab.org is an organization that works on a #CrowdCriMa platform, where all the unheard voices can be heard.

Case scenario for #CrowdCriMa:

Step 1: A victim or an activist sends SMS to a particular mobile number connected with #CrowdCriMa platform.

Step 2: #CrowdCriMa collects the SMS and store it / directly publish it and forward it to the authority. Or

Step 3: The admin of #CrowdCriMa platform checks SMS and publish it via different social media channels.

Step 4: Once published, SMS will be auto forwarded to particular pre-added email id (e.g. for fire service or for flood and et cetera).

All of the presented thus far have the same general concept, which is, to prevent casualties and save lives by communicating and sharing information. All of the platforms use peer information sharing, which some of the platforms use official reports from agencies.

3 FACEBOOK AND CROWD-SOURCING

It is commonly known fact that people commute on a regular basis every day, and this includes going to work, school, university etc. Each person is connected to friends and acquaintances and communicates using the social networks. Most of the world's population uses Android as a platform for their smartphones as a cheaper version, compared to the iOS. Since the rapid growth of users and bringing the social networks closer in their lives, users tend to share all sorts of information. On one hand, the Twitter social platform is created for short bursts of mostly textual information that has made this platform mainly a deck of personal opinions that the users share with the followers. Facebook, on the other hand, is a social network platform that contains all sorts of information, such as pictures, opinions, activities and so on. However, the main reason why Facebook is chosen as an information sharing platform, for the research, due to the level of sharing and the real time information delivery. Also, Facebook has dedicated an additional section in their platform for disaster information sharing with an API to connect.

Social networks are useful for big data generation, since users are zealous to participate and share information to other users. This behavior further depicts the reason for using Facebook and social networks all-together. Users of the online social networking also share common interests, pages, and by using hashtags, indicate specific subjects of interest. Since most of the social media is based on crowd-sourcing, they present a valuable source of information. On Figure 2 is shown a screen-shot image that displays the interface of the users of the Facebook application, with which, each user has to agree with its Terms of service and Privacy policy.

When developing an application using this Facebook interface, the API requires a token of request per application, which the request has to include the list of parameters that the application will take from the API. In our case, we request user_location and user_action_news. With the user_location is used to pinpoint the news on the map, where the other parameter is the news (disaster) that is announced by the user. The detection of users that are near is done using the Facebook Graph API Explorer, [9].

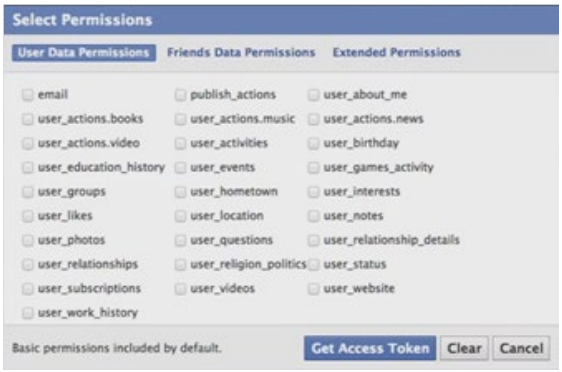

Figure 2: Facebook's API to allow access to user's data.

The response from the Facebook API request includes Facebook login, sharing and sending dialogs, triggering application events and Graph API. When someone connects with an app using Facebook Login, the app will be able to obtain an access token which provides temporary, secure access to Facebook APIs and the application has the right to own the user's credentials or other information on Facebook behalf. Sharing and sending dialogs is crucial part of the process since it is the main target of this application, which is, to share the required information at the right time.

3.1 Google Maps(Marker)

Google maps is a desktop web mapping service developed by Google. It offers satellite imagery, street maps, 360° panoramic views of streets (Street View), real-time traffic conditions (Google Traffic) and route planning for traveling by foot, car or public transportation. In our case, we use Map Android API on which a custom map is created and used by all users of the application. This will allow users to mark all the dangerous areas or places, which in term can serve as a warning information to other users. In our case, the warning is done using three symbols for flood, fire or earthquake, placed on the map.

3.2 Geo Crowd-Sourcing

In its basic form, crowd-sourcing is a method that uses mobile devices from users to gather massive

amount of information, typically named CrowdInfo. This information can be geo-located and can help to determine where to build the next shopping center or obtain an information about the events that are taking place near the user. On the other hand, it can be used by companies to place their products according to user's needs. These types of companies collect massive amount of data from active users, who report the events taking place and reports to the crowd-source database.

Recently, crowd-sourcing is mostly associated with Social Networks, such as, Facebook and Twitter, [10]. These networks generate images, text, Semantic text using hashtags, geo-location and mood markers. All of this information can be used to create the crowd-sourcing picture, where by combining the data, applications can generate new data (e.g. disasters, global events, traffic congestions and health hazards), shown in Figure 3.

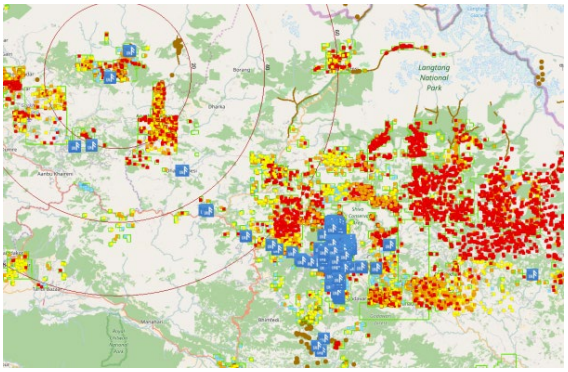

Figure 3: Report of earthquake using CrowdSource.

On Figure 3, it is shown how an earthquake is reported using crowdsourcing. Since many people are using mobile devices, the red markers are where reports are done by many users close together.

4 ARCHITECTURE OF THE OUR AWARENESS PLATFORM

Based on the architecture given in [7], we are proposing our own awareness platform. The architecture of the system is shown in Figure 4, where our system uses two external sources to gather information. These two sources are interfaced by a third part of the application, which is the interfacing API. The first part of the architecture is the crowd-source, where this module connects to Facebook via API and exchanges information about crisis detected from users. This part uses Data Acquisition Package to retrieve information from the crowd. The information that the crowd shares is maintained and acquired by users' permission (shown on the right in Figure 4). The part of the crowd that is willing to participate in this crisis system, voluntarily allows Facebook and Twitter to share their information.

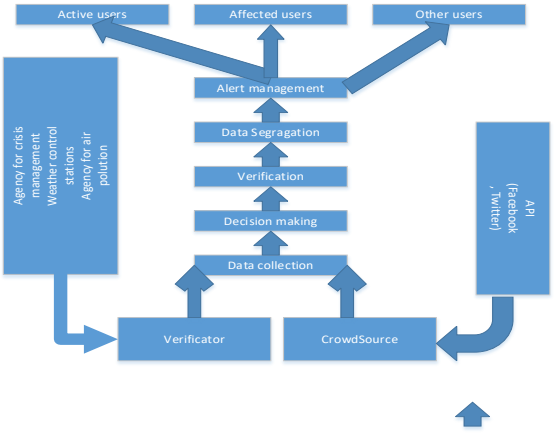

Figure 4: Architecture of our awareness platform.

The second part of the system is the data verification (shown on the left on Figure 4), where data is collected from certified agencies such as Agency for crisis management, weather control stations and air pollution. This data is certified and is used to verify data from the crowd-source. The data from these two parts is collected into a Data Collection Unit, used to store and backup data. This unit sends the data to the Decision Making Unit that uses algorithms that evaluates and classifies data from Crowd-source and Verificatior. This algorithm extracts data from Facebook and Twitter statuses by using key words, location and sentence structure. It determines a crisis based on at least three sources. These three sources have to use similar crisis key words (or word structure), and it has to be done in a certain amount of time and to be close to each other (geo-location). This determines the type of the crisis and the location. Based on the number of sources and how close (by geo-location) they are, it determines the percentage of accuracy of the information of a crisis. Accompanied with the data for that location and type of crisis from the Verificatior, the data is sent to the Verification Unit. This unit simply uses basic algorithms to verify if the information for a crisis is reported by the Verificator. The simple algorithm just checks if some of the stations for crisis actually have reported the determined crisis in the area.

The next part is the Data Segregation, where the crisis detected and validated in the previous step is sent, along with the amount of certainty of detection. The segregator classifies and divides the crisis into geo-location groups and sends the data to the Alert Management part. This part of the system is what connects the second part of the architecture with the third part and delivers alerts to end users. The end users are in the third part and they are divided into three groups: Active Users, Affected Users and Other. Active Users and Affected Users are users that have been registered with our system and these users have registered to receive information about near crisis. The Other Users are the non-registered Crowd-source users, that our system sends global information about crisis in the area. However, these users receive the information only if they explicitly check for crisis (they do not receive push notifications since they are not in immediate danger).

The Active and Affected users, as mentioned above, are registered with the system and they receive immediate information about crisis in their area. Affected users are users that are closest to or in the crisis (according to the decisions made by our system). They receive immediate information about possible crisis, along with the amount of certainty. Our system gets information about registered users via GPS and their location, so it is possible for users to not have their GPS coordinates updated, thus they might get false alarm. That is why our system lets the users choose the frequency of GPS acquisition and the size of the radius that they consider as radius of immediate danger.

5 CASE STUDY

Since Skopje, the capital of Macedonia, is not prone to disasters, our case study uses simulated messages passed between the three parts of the system. The messages sent are shown in Figure 5, where we can see what our system receives from the Crowdsource (Social networks) and from the Verification agencies. In this section, messages shown on Figure 5 are addressed from left to right, so the first message is in the upper left corner and so on. Only the crucial messages are shown in Figure 5. We receive the type of disaster (Flood, Fire or Earthquake), the notification type (1 for CrowdSource and 2 for Verificator), when it happened (timestamp) and where it happened (Latitude or Longitude). As we can see, if the Verification Agency sends us Earthquake, then it additionally sends the size of the Earthquake.

If our system receives several notifications from the same type happening in a 1 km radius, then it creates a single pin on the map by taking the middle point of all those in the radius. This is shown in Figure 6 on the map where messages 2,3 and 5 from Figure 5 are fused in a one alert shown in the middle of the map with yellow triangle for fire.

There are 4 different colors representing different types of alert notifications. The green color means that only one person reported the disaster and that it is not verified. The yellow means that three or more Social Network users reported the disaster but it is still not verified. Orange means that it has been reported by the verification agency but no one from the Crowdsource (Social Networks) reported it. And last, if it is used red as an alert color, then it means that both parts reported the disaster.

Our application was tested by 15 users, who were positioned around the city with their smartphones on different locations throughout Skopje. The users were positioned selectively so that 5 users would be affected, 8 users would be active and 2 users were registered as other type of users. Active and affected users registered to get immediate alert, so when active users entered the radius (1 km), the Alert Management system sent notification about the type of disaster and level of certainty.

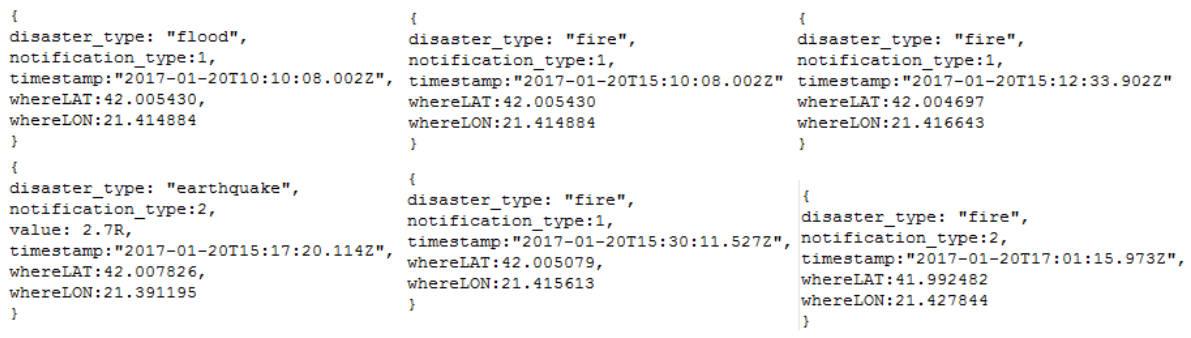

Figure 5: Messages sent to our system from Verificator and Crowdsource.

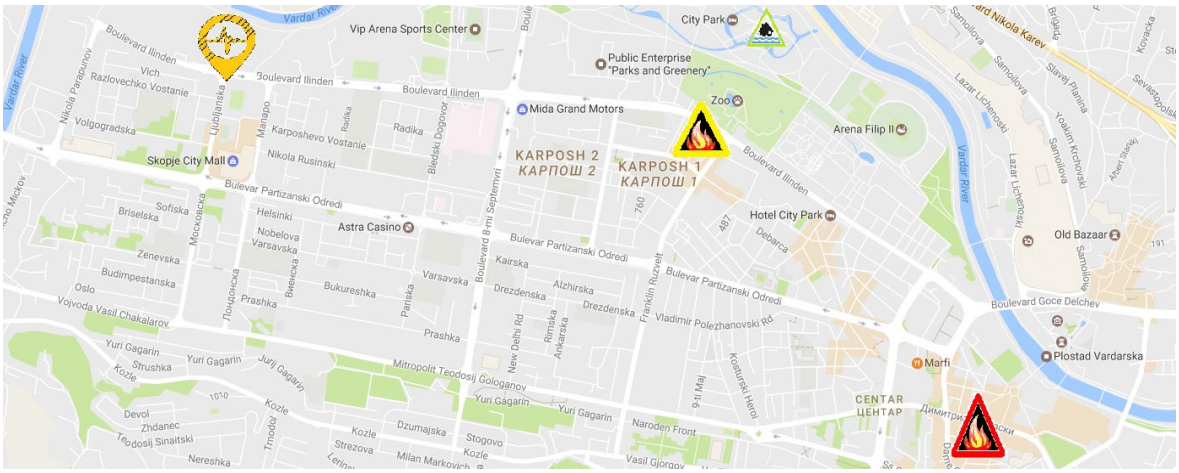

Figure 6: The results that our system sends back to users.

6 CONCLUSION AND FUTURE WORK

This application, can be beneficial to users to get real time alerts about near disasters. Also, this application can be used as a disaster information kit, where users will get information about disasters in the near area. This can be beneficial, in case of fire or flood, for users that are in traffic, to know where the disaster is and to avoid those streets. However, this application mainly relies on the information shared by crowd-sourced users and that information is verified by crisis agencies. For future work, we plan on upgrading the application with information about traffic jam and car accidents to alert drivers which streets are closed (by the authorities) and which streets to avoid (due to accidents).

REFERENCES

[1] A global display of terrorism and other suspicious events. Accessed on: Sep. 18, 2016, [Online]. Available: http://www.globalincidentmap.com/

[2] Th. Buecheler , J. Henrik Sieg , R. M. Füchslin1 and R. Pfeifer, "Crowdsourcing, Open Innovation and Collective Intelligence in the Scientific Method," A Research Agenda and Operational Framework, Proc. of the Alife XII Conference, Odense, Denmark, 2010.

[3] Ch. Eaton, D. Deroos, T. Deutsch, G. Lapis, P. Zikopolous, "Understanding Big Data," USA, The McGraw-Hill companies, 2012, [E-Book].

[4] Flood information service. Accessed on: Sep. 19, 2016, [Online]. Available: https://flood-warning-information.service.gov.uk/

[5] Disaster Preparedness and crisis response. Accessed on: Sep. 19, 2016, [Online]. Available: https://www.palantir.com/disaster-preparedness/

[6] Virginia Deploys Web-Based Emergency Management System. Accessed on: Sep. 19, 2016, [Online]. Available: https://cop.vdem.virginia.gov/dev/viper30/

[7] Humanitarian Open Street Map Team (HOT). Accessed on: Sep. 19, 2016, [Online]. Available: https://hotosm.org/about

[8] A. Broughton, T. Higgins, B. Hicks and A. Cox, "Workplaces and Social Networking," The Implications for Employment Relations, The Institute for Employment Studies, 2010.

[9] The Graph API. Accessed on: Sep. 19, 2016, [Online]. Available: https://developers.facebook.com/docs/graph-api/using-graph-api

[10] W. Tung, G. Jordann, "Crowdsourcing Service Design for Social Enterprise Insight Innovation," IEEE International Congress on Big Data, 2015, pp 367 - 373.

Management and Information Support Issues in the Implementation of Innovation Projects in Production Systems

Leonid Mylnikov
Perm National Research Polytechnic University, Komsomolsky Ave. 29, Perm, Russia
Leonid.Mylnikov@pstu.ru

Keywords: Innovation Project, Production System, Management, Planning, The Review of Management Methods, Information Support of Management Process.

Abstract: The article investigates tasks and approaches related to the management of innovation projects, the current state of production and project management in production environment. The article gives a review of main approaches for the management of project implementation in production systems. It suggests the task of management and selection of project development paths on all the stages of project implementation. The article surveys possibilities and drawbacks of existing theories and approaches regarding planning and management tasks in project implementation. It examines general characteristics of building project management models in production-and-economic systems and their practical application. The main goal of the article is to determine gaps in the development of methods and approaches related to the management of project implementation paths as a study object based on the information about projects and their implementation environment, i.e. systems. The practical significance consists in the possibility to increase the amount of successfully implemented projects, to reduce the time period of project development stages, and to cut expenses for the implementation of project stages.

1 INTRODUCTION

Currently, production systems operate in open market where the markets of innovation products have greatest potential. This situation demands fast decision-making and high quality of decisions as there is a necessity to consider a growing number of factors, multicoupled parameters and criteria of production system and implemented projects. Production systems need to be flexible since innovation products have a short life cycle, a great number of modifications (we can even observe a trend towards short-run and single piece production specifically for customer's needs), have high technology and are highly engineered. At the same time, production systems are accelerative and cannot readjust the ongoing processes in a single step. That is why, in order for the desired outcome to be achieved, it is necessary to make such changes greatly in advance, i.e. make planning and embed changes in production systems beforehand. So, there are attempts to develop methods for the management of projects and their production system environments by taking into consideration the whole project life cycle: idea, the transformation of idea into an innovation project, developing the project for the implementation in production system, its production, and sales.

In practice, it turns out to be complicated to manage this consecutive process as the process is not sufficiently formalized at all the stages. Hence, the application of formalized, theoretically sound approaches is limited in such conditions. Besides at seed stages, project implementation risks are very high and are beyond quantitative estimation. Such risks get much higher if we consider management tasks for long-term perspectives. Today, even sophisticated methods and approaches for production system management are considerably limited in application as they do not allow to make time planning of how innovation projects will be implemented in production systems. For instance, the theory of production functions [1] allows to consider only the implementation principles of projects and production systems on the base of well-known principles of their interaction; the theory of multi-agent systems [2] considers projects as independent elements that compete for resources and production system is taken as their environment, rather than an interaction element, that greatly

affects the principles of project competitiveness; the theory of active systems [3] is focused on studying the principles of production system operation, operation risks, interaction with external environment without distinguishing multiple markets and projects as independent systems; approaches which are based on project portfolio risk management [4] do not consider production systems as an object of management.

When we consider the external environment as multiple markets and projects, it is reasonable to take into account a high level of variability and dynamics of ongoing processes together with the accelerative nature of production systems. This imposes limitations on the application of methods which are used as a basis for the theories and conceptual foundations of production system management. For instance, the application of such approaches as actual data-based management, reflexive management, target management (that does not take into consideration the dynamics of production system environment changing) leads to delayed decisions and actions and shows up in management failures. Such failures trigger incoherent actions of subsystems and disorder production cycles in time.

Decision makers can use different behaviour strategies for their management principles. By the interaction of production systems with market and innovation projects it is possible to choose most optimal strategies based on the existing or unfolding (according to forecast values) situation. Otherwise, there is a possibility to work out effective measures for the external environment that will provide the desired performance of a production system.

Hence, there is a task to determine the application areas of existing theories, approaches and methods by taking into account production system requirements and their operation conditions.

2 PROJECT IMPLEMENTATION TASK SETTING AND THE PLACE OF MODERN APPROACHES AND THEORIES IN THE SOLUTION OF THIS TASK

In project implementation management, it is necessary to consider management processes of project development on different stages and take into account the change from one stage to another one (see the Figure 1) which can be formulated as a set of changes $G_i^{(D)} \cup J_i^{(C)} \exists i$:

$$\begin{gathered} S_{i-3}^{(A)} \cup S_{i-3}^{(B)} \xrightarrow{I_{i-3}^{(B)} \cup I_{i-3}^{(AB)}, P_i^{(B)}, R_{i-3}^{(B)}} \\ S_{i-2}^{(B)} \cup S_{i-2}^{(C)} \xrightarrow{I_i^{(C)} \cup I_{i-2}^{(BC)}, P_{i-2}^{(C)}, R_{i-2}^{(C)}} \\ S_{i-1}^{(C)} \cup S_{i-1}^{(D)} \xrightarrow{I_{i-1}^{(D)} \cup I_{i-1}^{(CD)}, P_{i-1}^{(D)}, R_{i-1}^{(D)}}, \\ S_i^{(D)} \cup S_i^{(E)} \xrightarrow{I_i^{(E)} \cup I_i^{(DE)}, P_i^{(E)}, R_i^{(E)}} \\ S_{i+1}^{(E)} \cup S_{i+1}^{(F)} \xrightarrow{I_{i+1}^{(F)} \cup I_{i+1}^{(EF)}, P_{i+1}^{(F)}, R_{i+1}^{(F)}} S_{i+2}^{(F)} \end{gathered} \quad (1)$$

where i – the number of planning step ($n - 1 > i > 2$), – the planning horizon, I – the set of resources (investments) necessary to make a change, R – the risk evaluation of making a change, P – the potential profit (benefits) expected from making a change, S – the set of possible states.

In the suggested setting, task solution requires an active element, i.e. management subject. More than that, different project stages have different formalization levels. That is why, by tackling the task we cannot apply only one single method or approach, yet we need to think about applying a group of methods or approaches within one theory or strategy.

Currently, scientists consider changes within one stage generally. The most developed stage is project implementation stage in the existing production system environment ($S_{i-1}^{(D)} \xrightarrow{I_{i-1}^{(D)}, P_{i-1}^{(D)}, R_{i-1}^{(D)}} S_i^{(D)}$).

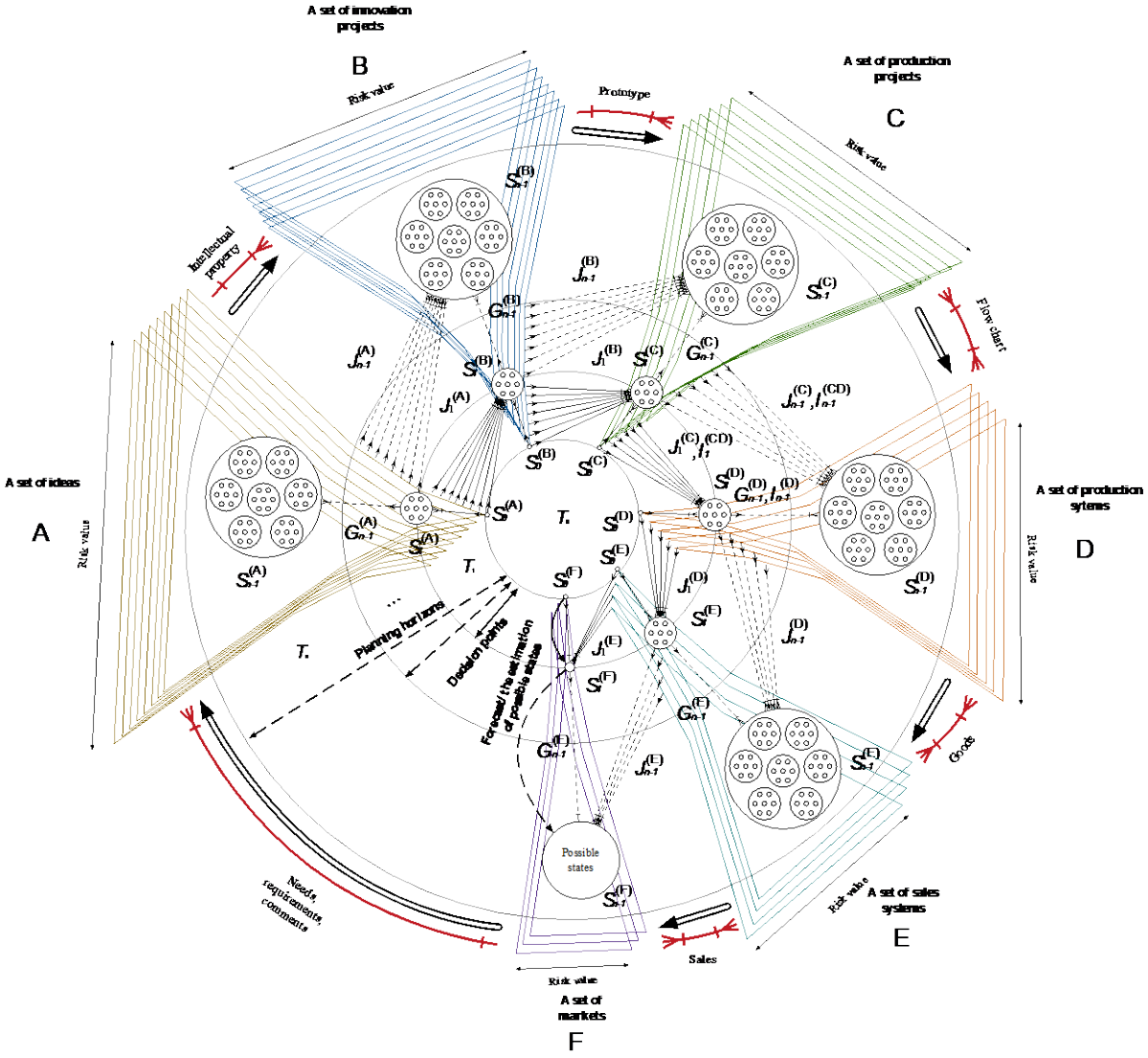

Figure 1: The structural interaction breakdown between innovation project implementation stages by solving the task of their implementation and planning management.

Popular theories which are widely used in the management of production systems and their project activities are given in the Table 1) with their characteristics:

T1) The theory of active systems which is focused on the term «active element» introduced by V.N. Burkov and open management principle [3], as well as the theory of organizational systems (see the works of D.A. Novikov) that developed the idea of cybernetic solution path application for the management of social and economic systems; according to this theory, the ongoing processes in social-and-economic and production environments are also considered in their interaction, including uncertainty conditions of external and internal environments (the set of states $S_i^{(D)}$ and changes $S_{i-1}^{(D)} \xrightarrow{I_{i-1}^{(D)}, P_{i-1}^{(D)}, R_{i-1}^{(D)}} S_i^{(D)}$ by a limited set of production systems D), and multiple management aspects are considered (i.e. financial management, organizational project management, institutional management, information management, etc. (see [5], [6] and [7]). A group of models was implemented under the specified theories: the financial model of innovation projects (see [8] and [9]); the decision making model that is based on rational behaviour and determinism hypotheses (by probabilistic indeterminacy) [10]; the basic model of

organizational (active) system (OS) and its extension (dynamic OS model, multidimensional OS model, multiple-level OS model, OS model with distributed control, OS model with uncertainties, OS model with limited joint activities, OS model with information support) [10] [11]; reflexive model [9] [12]; the basic models of single- and multidimensional active systems (AS) (which also include distributed control) on the basis of the following incentive systems: compensatory, uneven, proportional, unified proportional, and in multidimensional AS by taking into account uncertainty [11] [10]; the models of rational behaviour and bounded rationality [9]; the model of fuzzy control in social-and-economic systems [13], the model that takes into consideration the preferences of decision makers [14], etc.

T2) Related to project management (C), the theory of multi-agent systems became popular after L. Peppal had proposed to use the theory of games for describing backup and improving innovations in 1997 [15] where projects are considered as information agents within the theory of multi-agent systems (see [16] and [17]) that compete for resources ($I_i^{(D)}$ and $I_i^{(CD)}$). In the specified theory, there is a traditional classification of different types of models: deliberative models (as an example see [18] and [19]), reactive models [20], hybrid models [21].

T3) The theory of production functions deals with the investigation and the functional interaction description of production systems (D) and projects (E) that are being implemented in production environment (the set of changes $G_{\mathrm{i}}^{(\mathrm{D})}$ in the Figure 1) by taking into account different factors and, as a rule, in one or a limited set of production systems. In this theory, mathematical model is used as a formula of production output dependence (revenue) from the vector of spent or utilized resources in production (purchased resources) [1]. Here is the list of most popular functions that were developed according to this theory: the function with fixed factor proportions (the Leontief production function), the Cobb-Douglas production function, linear production function, the Allen production function, the CES production function, the production function with a linear factor change elasticity, the Solow and Hilhorst production function, bounded function, multimode function, the production function in linear programming [22], [23], [1].

T4) The results of theoretical and practical efforts in the previous years introduced a vast number of approaches which are based on structuring management processes in production systems, and namely [23]: the methodology of structured analysis and design (SADT (D. Ross), DFD (E. Yourdon), DFD (K. Gane — T. Sarson, DeMarca), object-oriented methods (OOD (Booch/Jacobson/Rumbaugh) [24], OOAD (P. Coad — E. Yourdon) [25] and [26], OODLE (Shlaer — Mellor), Demeter, Henderson-Sellers); information engineering methods (Martin-Finkelstein, Porter, Goldkuhl); project management standards (ISO/IEC 15288; DIN 69901; GOST P54869-2011, etc.).

T5) Machine learning methods related to project management in production systems. There is a steady trend of applying machine learning methods by handling management tasks in production-and-economic systems (see, for instance, [27]). At the same time, the significant role of machine learning methods in production management tasks will be only increasing [28]. Today, machine learning methods are used for solving a group of management and planning tasks (for instance, forecasting machinery breakdown, building empirical models by taking into consideration the changes of machinery characteristics in time, predictive management of accelerative systems (such as head supply systems and processing units), the development of market pricing and planning principles in production [29]).

By counterclockwise movement from the stage C to the stages B and A (see the Figure 1) the formalization level is decreasing. Currently, expert communities examine projects and determine goals for the projects on the stages B and A as a part of competitions. However, different information is collected about projects (analogues, market demand, investment, project team, the presence of prototype, project characteristics compared with analogous versions, etc.), attempts are made to analyze the collected statistical information (see, for instance, [30]). Available statistical data and expert community create a good background for analyzing innovation projects with help of machine learning methods (also on the basis of a new approach, i.e. reinforcement learning techniques (semi-learning methods) [31]).

Table 1: Change management and decision-making support by project implementation in production environment.

	T1	T2	T3	T4	T5
Goal orientation	Searching the ways how to affect the system for achieving its desired behaviour	There is no clear goal definition or the goal is determined only with help of logical means.	Management goal is not formulated in the methodology of production function.	Goal is determined by a decision maker.	Goal is determined on the stage of model design by management subject.
The system of relationships between production system and projects	Performed on the basis of rules, laws and procedures that regulate the interaction of participants.	Each participant operates independently in accordance with own regulations.	Established as interaction between production function parameters.	Established by the regulated structure and principles of interaction.	Laid down during model learning process based on specified goals.
Risk management	Considers different uncertainty types (internal, external, mixed) and uses interval, fuzzy, probabilistic approaches.	The probability of this or that behaviour is determined on the base of simulation modeling by the Monte Carlo method or the Bayes' theorem.	Not performed.	Not performed.	Risk evaluation and the use of any techniques for work with uncertainties.
Time orientation	Models in both statistical and dynamic setting.	Time is discrete and is defined by the emergence of events.	In the classical theory of production function, the time factor is not considered.	Considered as a continuous process.	All models are adjusted in time and are dynamic.
Interaction with management subject	For management subjects, models can be presented as decision-making support systems.	Modeling results are considered as information that is taken into consideration by management subject in decision-making.	Management subject uses the methodology of production function for decision-making based on production function studies with help of only mathematical methods.	Information support of management subject.	Ready decisions are produced which can be used by decision makers.
External environment orientation	Market is considered as a general term that can include other production systems.	The connection of agent and environmental area is not precisely defined. Historical data are not considered. It is not clear what agents the goal will and will not be dependent on.	Bounded by production function factors.	In accordance with the specified principles and rules.	Within a restricted set of observed parameters.
Change management	Recommendations for making changes in system performance.	Not considered.	The study results are used for defining the amount of required resources and production capacity on the basis of production elasticity and maximum capacity determination.	Performed by decision makers by structuring production system activities.	All the decisions are suggestions for the selection of parameters or performance algorithms.
Basic idea	The use of cybernetic approach for managing systems with uncertainty.	System element is considered as an independent active element that operates due to its own internal rules.	Conformity search among the parameters of production system and released products with help of heuristic methods.	There is a possibility to describe system activity with help of a limited set of elements and their interaction rules.	The automated process of building and adjusting empirical models on the basis of empirical and actual data.

3 ADDITIONAL REQUIREMENTS TO THE IMPLEMENTATION OF PROJECT MANAGEMENT MODELS IN PRODUCTION SYSTEMS

The description and analysis of paths (1) sets up a grading problem and, hence, one of the existing management approaches can be used for defining their estimate criteria: project management, management and planning by objective, resource management, information and reflexive management, predictive management, adaptive management. The management mechanism is chosen on the basis of management goals that are set by management subject. In production-and-economic systems, economic efficiency and feasibility of project implementation is considered as a general criterion. In this case, each state S_i and entering this state from the state S_{i-1} can be estimated as follows:

$$(1-R)(P-I) \rightarrow \max, \quad (2)$$

where R – the risk evaluation on making a change, P – the potential profit, I – the required investment (resources) for making a change.

In the planning and management of projects and production systems, it is also possible to use other criteria and consider the consistency of goals in production subsystems and their projects, invariant states in decision points, the complementarity of projects in production environment, irreversibility of managerial decisions. Besides, the information has to be reliable.

The step size $\Delta t_i(t)$ (it depends on decision points and is considered as a discrete variate with a variable step) and the planning horizons T_n. determine the set of possible states (S_i). Such time position narrows a group of techniques that can be used before applying special state principle and case management. However, we encounter the problem of choosing planning horizons. This action is based on the expected project portfolio. The probability of portfolio criteria efficiency is described by a binomial distribution (the planned state S_i), and the Bayes' theorem defines the probability of a successful change into a new state that is dependent on the previous state (the state that we are in) [32]. Invariance in project development path selection will show up as a set of equally obtainable optimal and Pareto states. In this case, the solution of task will be a set of development paths and states that can be conditionally presented as a tree.

Hence, the key problem is the generation of a set of possible states. For this purpose, we need to build a model that includes estimate criteria parameters (2). For the examined stage, the surrounding elements of the stage (see the Figure 1, the set of sales systems E is not marked out in some settings and is considered as a production system element) form the external environment. The management object interacts with the external environment through its variables and the way these variables are used in management model. In order for planning tasks to be solved and accelerative processes to be considered in production environment, the values of these variables [33] and project parameters [34] have to be predicted. The predictions have a certain degree of reliability which is determined on the basis of adequacy evaluation and the range of possible deviations. The latter ones can be calculated into risk estimates for the expected forecast-based states to be obtained [35]. The model parameters can be presented by different types of data (time series, single characteristic values that all together characterize the project, some of them are described by known principles and can be built upon several values [36]). Therefore, another important task is to determine the significance of parameters, certain characteristic values and their combinations for project implementation [37]. Taking into account the differences in implementation stages and data-dependent model characteristics, empirical techniques should be applied in order to build the model.

The model of each project implementation stage has to be designed individually as stages have different implementation environment (as shown in the Figure 1). Besides, it is important to take into consideration the specifics of project itself or its environment system. At the same time, this will be a complex model with a required system optimization [38], that triggers a group of problems, i.e. the problems of selecting behaviour strategy (for instance, the behaviour for the common benefit or for the purposes of certain elements) not only in the interaction of production systems but also in the interaction of production system elements; the problems of model elements' compromising which handle different tasks within one general management and planning task in production environment (for instance, see the breakdown in the Figure 2) on the base of complex modeling and possible states' search by determining the constraints in the area of possible solutions. More

than that, the model has to consider technical and economic tasks jointly (i.e. heterogeneous model) by taking into account the time factor; besides, it should be a computable model that connects different management parameters in one system (the paradigm C^5) [39].

The model should be built on the principles that allow its changes (teaching, adjusting) along with the changes in its external and internal environment operation conditions and different degree of experience that was obtained in different model operation time period. Hence, in order for a complex model to be implemented, we need to use different approaches and methods for its elements by taking into consideration additional requirements of management subject, developer preferences, and statistical data.

4 INSTRUMENTAL CHARACTERISTICS OF COMPLEX MODEL IMPLEMENTATION IN MANAGEMENT

Implemented as information product, the model tackles several DSS tasks: *a)* tool implementation for «searching solutions», that is based on using models as a series of procedures for data and statement processing in decision making [40]; *b)* the implementation of interactive computer-based systems that help use data and models for tackling unstructured problems; *c)* the implementation of computer information systems for the support of diverse activities by making decisions in the

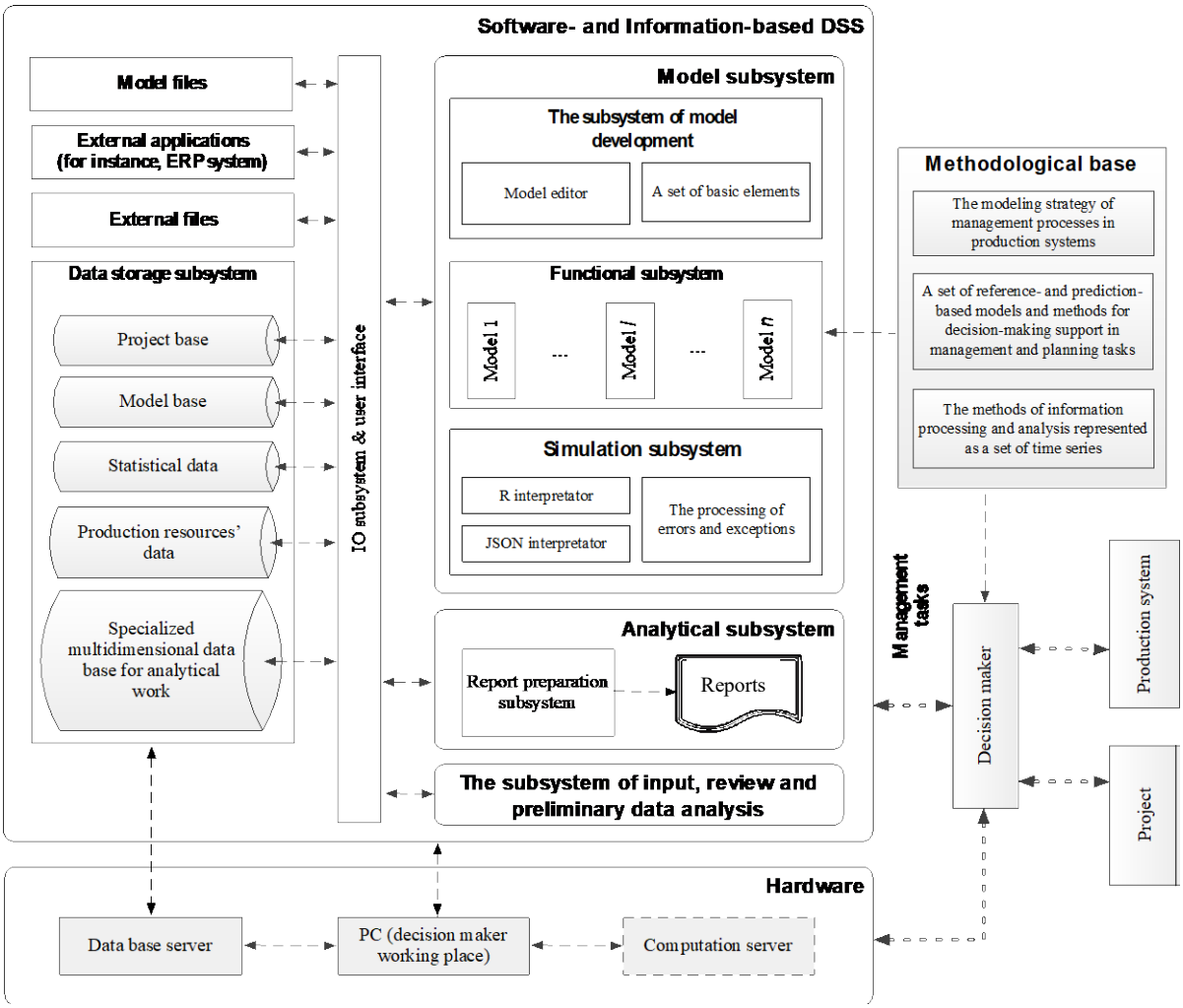

Figure 2: The structural information interaction breakdown for the implementation of models and methods in order to support managerial decision-making process in production systems.

situations where the use of automated systems for the whole decision making process is infeasible or troublesome [41].

The application of one toolset (even the most sophisticated one) is not enough for building heterogeneous systems. That is why, by the development of information system we need to consider the mechanisms of work with data and the ways how model components can be integrated in the consolidated information area [42].

The structure of information area that satisfies the specified requirements is shown in the Figure 2.

5 CONCLUSIONS

Despite a big amount of studies dedicated to different management aspects, the analysis of production and project activities, the issues of studying ongoing processes and the impact of managerial decisions on examined systems, the modern planning and management methods do not allow to consider all factors related to available resources required for production systems as well as technical, economic and financial project criteria and parameters. Today, when production competitiveness is strongly focused on innovations and continual release of new goods, management and planning becomes challenging in production environment, since the level of process automation is increasing due to fast decision-making requirements and human factor deregulation (at the same time, despite a vast number of negative factors, the integration of human workforce into production process allows to provide additional control and handle exceptions).

To summarize, we can say, that today there are interesting theories and approaches that allow to solve management tasks in projects and production systems by taking into consideration certain groups of tasks or specified conditions. However, we encounter a shortage in methodological approaches in marketing management, innovation projects' selection and management formalization.

When we solve individual tasks, it is impossible to solve the task of innovation project management on all the stages of its life cycle even with specific suppositions. Today, there are no approaches which are not bound to singe task characteristics. Moreover, the existing models do not allow to work with several innovation projects simultaneously [43].

REFERENCES

[1] П. М. Симонов, Экономико-математическое моделирование. Пермь: Ред.-изд. отд. Пермского гос. ун-та, 2010.

[2] Б. Я. Советов, В. В. Цехановский, and В. Д. Чертовской, Интеллектуальные системы и технологии. М.: Издательский центр "Академия," 2013.

[3] В. Н. Бурков and Д. А. Новиков, Теория активных систем: состояние и перспективы. М.: Синтег, 1999.

[4] H. Markowitz, Harry Markowitz: selected works. Hackensack, NJ: World Scientific, 2008.

[5] В. Н. Бурков and Д. А. Новиков, Как управлять проектами: Научно-практическое издание. М.: СИНЕРГ ГЕО, 1997.

[6] В. П. Воропаев, Управление проектами в России. М.: Аланс, 1995.

[7] В. В. Цыганов, Адаптивные механизмы в отраслевом управлении. М.: Наука, 1991.

[8] Д. А. Новиков, Управление проектами: организационные механизмы. М.: ПМСОФТ, 2007.

[9] Д. А. Новиков and А. А. Иващенко, Модели и методы организационного управления инновационным развитием фирмы. М.: Ленард, 2006.

[10] В. Н. Бурков, Н. А. Коргин, and Д. А. Новиков, Введение в теорию управления организационными системами. М.: Либроком, 2009.

[11] Д. А. Новиков and А. В. Цветков, Механизмы функционирования организационных систем с распределенным контролем. М.: ИПУ РАН, 2001.

[12] А. Г. Чхартишвили, Теоретико-игровые модели информационного управления. М.: ПМСОФТ, 2004.

[13] М. Б. Гитман, В. Ю. Столбов, and Р. Л. Гилязов, Управление социально-техническими системами с учетом нечетких предпочтений. М.: ЛЕНАНД, 2011.

[14] В. А. Харитонов et al., Интеллектуальные решения обоснования инновационных решений. Пермь: Изд-во Перм. гос. техн. ун-та, 2010.

[15] L. Peppal, "Imitative Competition and Product Innovation in a Duopoly Model," Economica, vol. 64, no. 254, pp. 265-279, May 1997.

[16] M. Wooldridge and N. R. Jennings, "Intelligent agents: theory and practice," The Knowledge Engineering Review, vol. 10, no. 02, p. 115, Jun. 1995.

[17] W. She and D. H. Norrie, "Agent-based systems for intelligent manufacturing: A state-of-the-art survey," Knowledge and Information Systems, vol. 1, no. 2, pp. 129-156, 1999.

[18] M. Wooldridge, N. R. Jennings, and D. Kinny, "The Gaia Methodology for Agent-Oriented

Analysis and Design," Autonomous Agents and Multi-Agent Systems, vol. 3, no. 3, pp. 285-312.

[19] B. Linder, W. Hoek, and J.-J. C. Meyer, "Formalising motivational attitudes of agents," in Intelligent Agents II Agent Theories, Architectures, and Languages, vol. 1037, M. Wooldridge, J. P. Müller, and M. Tambe, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 17-32.

[20] J. Ferber and O. Gutknecht, "A meta-model for the analysis and design of organizations in multi-agent systems," in Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160), Paris, France, 1998, pp. 128-135.

[21] K. Fischer, J. P. Mu¨ller, and M. Pischel, "Cooperative transportation scheduling: An application domain for dai," Applied Artificial Intelligence, vol. 10, no. 1, pp. 1-34, Feb. 1996.

[22] Г. Б. Клейнер, Производственные функции: теория, методы, применение. М: Финансы и статистика, 1986.

[23] И. Л. Туккель, А. В. Сурина, and Н. Б. Культин, Управление инновационными проектами. СПб: БХВ-Петербург, 2011.

[24] G. Booch, Object-oriented analysis and design with applications, 2nd ed. Redwood City, Calif: Benjamin/Cummings Pub. Co, 1994.

[25] P. Coad and E. Yourdon, Object-oriented analysis, 2nd ed. Englewood Cliffs, N.J: Yourdon Press, 1991.

[26] P. Coad and E. Yourdon, Object-oriented design. Englewood Cliffs, N.J: Yourdon Press, 1991.

[27] H. Jalali and I. V. Nieuwenhuyse, "Simulation optimization in inventory replenishment: a classification," IIE Transactions, vol. 47, no. 11, pp. 1217–1235, Nov. 2015.

[28] C. Arnold, D. Kiel, and K.-I. Voigt, "How the industrial internet of things changes business model in different manufactoring industries," International Journal of Innovation Management, vol. 20, no. 08, pp. 1640015-1-1640015-25, Dec. 2016.

[29] A. Paprotny and M. Thess, Realtime data mining: self-learning techniques for recommendation engines, 2013.

[30] Р. Р, Рейтинг инновационных регионов для целей мониторинга и управления. М.: АИРР, 2014.

[31] Л. А. Мыльников, Б. Краузе, М. Кютц, К. Баде, and И. А. Шмидт, Интеллектуальный анализ данных в управлении производственными системами (подходы и методы). М.: БИБЛИО-ГЛОБУС, 2017.

[32] S. J. Russell and P. Norvig, Artificial intelligence: a modern approach, 2nd ed. Upper Saddle River, N.J: Prentice Hall/Pearson Education, 2003.

[33] Л. А. Мыльников, "Особенности решения моделей планирования производственной деятельности и управления в производственных системах с учетом фактора времени," Информационно-измерительные и управляющие системы, vol. 15, no. 9, pp. 29-34, 2017.

[34] В. М. Винокур, Л. А. Мыльников, and Н. В. Перминова, "Подход к прогнозированию успешности инновационного проекта," Проблемы управления, no. 4, pp. 56-59, 2007.

[35] L. Mylnikov and M. Kuetz, "The risk assessment method in prognostic models of production systems management with account of the time factor," European Research Studies Journal, vol. 20, no. 3, pp. 291-310, 2017.

[36] Л. А. Мыльников, Поддержка принятия решений при управлении инновационными проектами. Пермь: Перм. гос. техн. ун-т, 2011.

[37] Л. А. Мыльников and С. А. Колчанов, "Методика выявления ключевых параметров иннова-ционных проектов на основе статистических данных," Экономический анализ: теория и практика, no. 5 (260), pp. 22-28, 2012.

[38] Д. А. Новиков, "Комплексные модели системной оптимизации производственно-экономической деятельности предприятия," Управление большими системами: сборник трудов, no. 65, pp. 118-152, 2017.

[39] Д. А. Новиков, Кибернетика: навигатор. История кибернетики, современное состояние, перспективы развития. М.: ЛЕНАНД, 2016.

[40] W. H. Inmon and R. D. Hackathorn, Using the data warehouse. New York: Wiley, 1994.

[41] M. J. Ginzberg and E. A. Stohr, "Decision Support Systems: Issues and Perspectives," in Processes and Tools for Decision Support, Amsterdam: North-Holland Pub.Co, 1983.

[42] Л. А. Мыльников, "Системный взгляд на проблему моделирования и управления производственными инновациями," Научно-техническая информация. Серия 1: Организация и методика информационной работы., no. 5, pp. 11-23, 2012.

[43] Л. А. Мыльников, "Микроэкономические проблемы управления инновационными проектами," Проблемы управления, no. 3, pp. 2-11, 2011.

Uncertainty Analysis of Oil Well Flow Rate on the Basis of Differential Entropy

Ivan Luzyanin, Anton Petrochenkov and Sergey Bochkarev
Department of Microprocessor Units of Automation, Perm National Research Polytechnic University, 29 Komsomolsky ave. , Perm, Russia
{lis, pab}@msa.pstu.ru, bochkarev@msa.pstu.ru

Keywords: Uncertainty, Differential Entropy, Exponential Distribution Probability Density Function, Well Flow Rate, Oil Field, Oil Production.

Abstract: Oil well production efficiency depends on the accuracy of the flow rate prediction. The electrical submersible pumps are selecting and the well production control is carrying out based on the predicted values of flow rate. Inaccurate prediction may cause limitations of well deliverability or inefficient pumping. The prediction accuracy of flow rate changes in time related to initial data uncertainty that causes deviations between calculated flow rate values and measured ones. To minimize operating costs the same pump selection and control methods are used for groups of wells operating under the same conditions. However, sometimes wells demonstrate very different behavior even under the same conditions. In these wells flow rate changes becomes unpredictable by the common methods and additional studies required for correct prediction. The problem of finding wells with unpredictable flow rates at the early operation stages is very important because their inefficiency can significantly increase in time without special operation methods. The article considers the method of finding wells with potentially unpredictable flow rate changes with use of the entropy concept. The main feature of this method is that it is appropriate for data of any distribution types with given probability density function. The article discusses the relation between the value of joint reduction in uncertainty obtained from entropies of calculated flow rates and measured ones for a single well and the deviations between these flow rates. The novelty of the article is that the joint reduction in uncertainty in calculated value of well rate when knowing measured well rate is proposed as the measure of the well flow rate predictability.

1 INTRODUCTION

At present oil remains the main energy source for many fields of activity. However, its production becomes more difficult and energy intensive every year. This is caused both by oil reserves depletion and by increase in number of fields is being operated under complex geological conditions [1].

Today the main oil producing method is pumping with use of electric driven submersible pumps (ESP). ESP can work with high performance and efficiency in deep wells but its use in new oil fields and under complex geological conditions is limited. The limitations appears mainly when selecting the submersible equipment. The ESP energy efficiency is ultimately depends on accurate equipment selection for specific operating conditions. Incorrect selection can cause inefficient well operation for all pump life cycle (since the ESP replacement is carrying out only in case of its failure).

Large number of parameters are used when selecting ESP equipment. Some of them are found by statistical methods. The most significant parameter that determines energy efficiency is the desired well flow rate [2]. It is usually calculated under conditions of uncertainty and data incompleteness (especially in the early stages of field exploitation). Uncertainty is caused by the inability of detailed study of the reservoir and the data incompleteness is caused by experiment limitations. Various statistical models of wells are developed to solve this problem. Since each well has individual conditions, it requires an individual model. In practice, it is impossible to build individual models for each well, so the generalized models are used. The models are usually based on regression equations, their coefficients are found

either with using experiments or by studying large datasets. These approaches are effective for well-studied fields that have been operating for a long time and for fields without special geological conditions. However, they do not consider geological factors that appear in individual wells and cause additional uncertainty of calculated parameters. This additional uncertainty together with data incompleteness can cause significant deviation between desirable flow rate and actual one. This in turn can be the reason of incorrect equipment selection or ineffective well production control. In addition, regression models usually require large amount of data that is unavailable at the early stage of well operation.

The article presents the investigation of the uncertainty that is existing in desirable (calculated with models) and actual (measured in well) flow rates of various oil wells with use of entropy concept. The ability of using information entropy for estimating the flow rates uncertainty for insufficiently known wells or for wells that operates in special regimes is studied. The main hypothesis is that when the entropies of both desirable and actual well flow datasets are known, the mutual information of these datasets will increase with decreasing flow rate predictability. This dependency will help to classify wells by flow rate predictability and select the most unpredictable wells for additional study.

Since the information entropy was originally introduced for discrete random variables [3], in this study the differential entropy of a continuous random variable is used instead. In general it is not an analogue of information entropy for continuous variables. However, when knowing the differential entropy, it is possible to obtain the mutual information for the case of continuous random variable.

When the above hypothesis is proven, the proposed method of predictability classifying will easily be applied in practice as it requires only knowing the data distribution law and allows data to have any distribution with given probability density function (PDF).

The research aim is to check the above hypothesis on the real data. The article considers the example of exponentially distributed data but the general algorithm of applying the method for any other distribution types is presented in the last part of the article.

The article includes four parts. First part presents short overview of commonly used flow rate calculation model and studies the causes of uncertainty. Second part considers data preparation and preliminary classification of statistical data obtained from oil fields. Third part considers determining the appropriate distribution law for classified datasets and presents the results of uncertainty analysis. The last part presents the generalized algorithm of applying the method for data of any distribution with given PDF.

2 CAUSES OF UNCERTAINTY IN WELL FLOW RATES

The ESP efficiency depends on current load of the motor that is represented by load factor (K_l). Its value can be calculated by (1).

$$K_l = \frac{N}{N_n} \tag{1}$$

where N is an electric power that is currently being consumed by ESP (found by (2)), N_n is rated ESP power. ESP efficiency has maximal value when K_l is one.

$$N = \frac{P_{pump} Q}{3600 \cdot 24 \cdot \eta_{pump} \cdot \eta_{motor}} \cdot 10^3 \tag{2}$$

In the above equation P_{pump} is a pump pressure required to lift oil to the surface, η_{pump}, η_{motor} are efficiencies of pump and motor respectively, Q is desirable (or actual) well flow rate. Pump pressure required for lifting oil to the surface and equipment efficiencies depend on well design and current operational regime. These parameters are usually constants for given regime. Therefore ESP load changes (and ESP efficiency) are ruled mostly by flow rate changes. These changes are typical not only for oil wells [4]. Moreover, desirable flow rate is used for ESP equipment selection. When pump has high performance and well has low flow rate the efficiency becomes significantly less than one. Nevertheless, in this case there is an ability to increase efficiency with using another regime. When pump has low performance and well has high flow rate, the efficiency will be low again but in this case efficiency increasing is more complicated than in previous one. Detailed description of these dependencies is given in [2].

Given considerations illustrate the significance of accurate calculation the desired well flow rate before well starts operating.

Standard well flow calculation model is based on Dupuit equation [2]. This equation represents the flow rate to the cylindrical well placed in the center of an “ideal” reservoir. “Ideal” reservoir must have

regular geometry and be fully saturated with oil. Since there are no "ideal" reservoirs in real life, the equation is only useful for some sections in the real reservoir that fit the above requirements. These sections are usually separated from each other and have individual geometry. The production efficiency reaches its maximum if the reservoir can be divided into homogeneous sections of regular geometric shape with one or more wells operating in each section.

To find such sections, experimental data of similar fields are used. These data have uncertainty caused by experiment limitations and lack of information about field being studied. Analysis of the data obtained at the fields showed that the uncertainty of the reservoir structure and properties has a maximum value at the beginning of the field lifecycle and reaches the minimum at the end of its operation. Besides that, external factors such as rock destruction or changes in fluid properties also affect uncertainty [1], [5].

According to the Dupuit equation the deliverability of a given well is determined by productivity index (PI). In addition, flow rate depends on difference between reservoir pressure and bottomhole pressure (ΔP_f). PI and ΔP_f as well as reservoir geometry are either obtained experimentally or calculated with models.

Thus, the uncertainty of flow rate includes three components: the uncertainty of the reservoir geometry, the uncertainty of PI calculation and the uncertainty of ΔP_f calculation.

The actual flow rate value is measured by special sensors. The sensors have a measurement error that can usually be included in the rate value.

In these conditions, the comparative analysis of the uncertainties appearing in desirable and actual flow rates over long time periods can give significant results for understanding the ways of initial data uncertainty resolution.

It should be noted however that wells could have different operational conditions and work in different operational regimes. At that, desirable and actual flow rates must have different uncertainty.

3 PRELIMINARY DATA ANALYSIS

Statistical data for the study were obtained from 27 oil fields that are operating under different geological conditions. Obtained dataset includes 440 values of average annual well flow rates (220 values corresponds to desirable flow rates, others – actual ones).

At the first stage of the research the accuracy of predicting the actual flow rates was studied. For this purpose, the initial dataset was divided into subsets, each of that included the average annual values of the desirable and actual flow rates of a single well for all years of its operation. After that, the pairs of graphs (desirable rates changes in time and actual rates changes in time) were built for all wells. The graphs were classified according to the form of deviations of the desirable and actual flow rate curves. Figure 1 illustrates obtained classes of curves deviation.

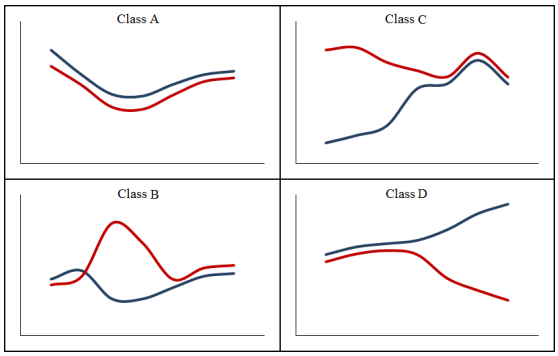

Figure 1: The classes of well flow curves deviations. Blue line corresponds to desirable rates, red line – actual rates.

The following classes were found:

- Class A – the desirable flow rate curve matches the actual flow rate curve;
- Class B – there is a single deviation inside desirable and actual flow rate curves;
- Class C – flow rate curves converges to the same shape;
- Class D – flow rate curves diverges from the same shape.

It should be noted that classification was built only by form of curves but not by value of deviations. For example class B includes both curves where desirable flow rates are below actual ones (as shown on figure) and vice versa. The classification tree (Figure 2) represents the probabilities of getting pairs of graphs into the classes A to D.

The tree includes two layers. The first one defines general form of discrepancy between graphs (classes A and B corresponds to generally coincident graphs; in opposite, graphs B and C correspond to not coincident ones). The second layer determines the belonging of the graphs to the specified class. The classification results can be interpreted as follows: the probability of accurate prediction of

flow rate changes is 40% (class A), the probability of incorrect prediction of flow rate changes after some time period is 37% (classes B and D), the probability of incorrect prediction of flow rate changes in initial time period is 22 %. The overall probability of the prediction error is 59 %. Thus, the probability of accurate prediction is relatively small that possibly indicates the presence of large uncertainty in the initial data.

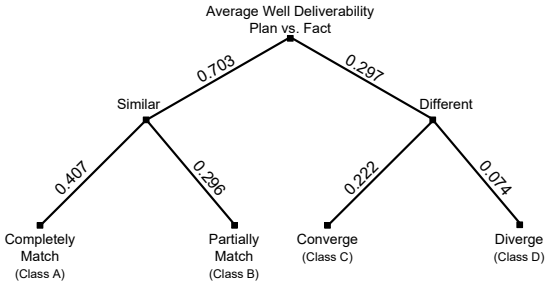

Figure 2: Classification tree for typical graph forms.

For further research, samples for each class and joint samples including samples of more than one class were obtained. Since the size of samples for classes C and D is small, they were combined in one sample. Table 1 includes samples that were used in the study for entropy analysis.

Table 1: Datasets used for uncertainty analysis (P – separate sets of desirable flow rates; F – separate sets of actual flow rates).

Classes included in sample (sample type)
A (P, F)
B (P, F)
C+D (P, F)
B+C+D (P, F)
A+B+C+D (P, F)

4 ENTROPY CALCULATION

The general (3) is commonly used for differential entropy calculation [3].

$$H(x) = -\int_S f(x)\log\left[f(x)\right]dx \tag{3}$$

where S is a support set of the random variable with given continuous distribution, $f(x)$ is a PDF for given X. The logarithm base defines the units of entropy. At the following study the base 2 is used, so the entropy is measuring in bits.

PDFs for statistical data were found by using the probability distribution histogram. The (4) was used for calculating the PDF value in each interval.

$$f_i(x) = \frac{m_i}{hN} \tag{4}$$

where m_i is a number of values from dataset that are included in the i-th interval, h is interval length, N is a number of values in sample. At the next step, the histogram was interpolated with PDF of appropriate theoretical distribution (Figure 3) and the distribution parameters were calculated. In the study it was hypothesized that all distributions have exponential distribution with PDFs given by (5) [6].

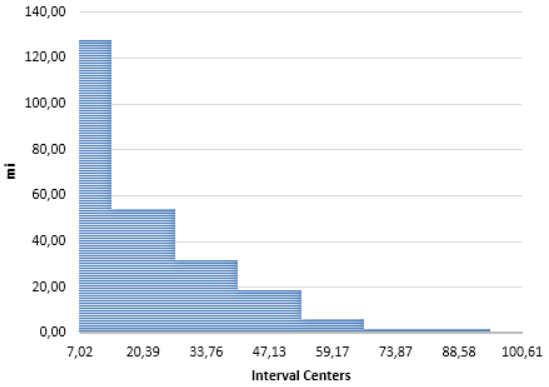

Figure 3: Histogram for probability density function and exponential probability density curve for dataset B+C+D FULL.

$$\begin{cases} f_X(x) = \lambda e^{-\lambda x} : x \geq 0 \\ 0 : x < 0 \end{cases}, \tag{5}$$

where λ is a distribution ratio obtained by the (6) [6]:

$$\lambda = \frac{1}{\bar{x}}, \tag{6}$$

where $\bar{x}$ is the mean value for a given dataset.

The hypotheses of the distribution were proven by F-test.

The distribution ratios for different samples are given in the Table 2.

Table 2: The values of the distribution ratio for different datasets.

Dataset	λ
A P	0.062
A F	0.055
B P	0.070
B F	0.071
C+D P	0.045
C+D F	0.051
B+C+D P	0.060
B+C+D F	0.060
A+B+C+D P	0.058
A+B+C+D F	0.057

Differential entropy for exponential distribution is obtained by the following (7) [7]:

$$H(X) = \log_2\left(\frac{e}{\lambda}\right), \quad (7)$$

The values of differential entropy for the datasets are given in Table 3.

Table 3: The values of the differential entropy for datasets.

Dataset	Differential entropy
A P	5.034
A F	4.945
B P	5.067
B F	5.014
C+D P	5.602
C+D F	5.401
B+C+D P	5.446
B+C+D F	5.351
A+B+C+D P	5.357
A+B+C+D F	5.464

The numeric value of differential entropy for a continuous distribution of random variable is not meaningful in practical tasks. Instead of this, the mutual information (I_{XY}) obtained from one random variable when given another random variable is the most important measure. For better understanding the interrelation between uncertainty and mutual information it is suggested to use term "joint reduction in uncertainty" in case of continuous variables [7]. This term is also more convenient for the following study because it describes the potential ability of reducing the uncertainty when obtaining datasets with different characteristics.

For two continuous random variables X and Y, the I_{XY} can be found by the following (8):

$$I_{XY} = -\int\int f_{XY}(x,y)\log_2\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)dxdy \quad (8)$$

where $f_{XY}(x,y)$ is a joint PDF of X and Y, $f_X(x)$, $f_X(y)$ are marginal distributions of X and Y respectively. If the differential entropies of the X and Y distributions are given, the (9) can be rewritten as follows:

$$I_{XY} = H_X + H_Y - H_{XY} \quad (9)$$

Here H_X and H_Y are the differential entropies of X and Y distributions themselves and H_{XY} is an entropy of joint distribution of X and Y that is obtained by the (10).

$$H_{XY} = -\int_X\int_Y f_{XY}(x,y)\log_2(f_{XY}(x,y))dxdy = \\ = -E(\log_2 f_{XY}(x,y)) \quad (10)$$

The above equation requires knowing the joint PDF of X and Y. For completely independent random variables the joint PDF is simply the product of PDFs for X and Y. If the variables are not independent, their dependency level is estimated by correlation coefficient ρ. In this case calculation of the joint PDF becomes more complicated as the dependency needs to be considered.

Since all samples in the table 1 have exponential distribution with PDFs given by (5), the joint PDFs or all the samples will be the PDFs of two exponentially distributed random variables and will have bivariate exponential distribution. Several studies consider the obtaining the bivariate exponential PDF for different cases [8-11]. In the following study the joint PDF is calculated as follows (11):

$$\begin{cases} f_{XY}(x,y) = \lambda_1\gamma_2 e^{-\lambda_1 x-\lambda_2 y-\lambda_3 y} : (x,y) \mid 0 < x < y \\ f_{XY}(x,y) = \lambda_2\gamma_1 e^{-\lambda_1 x-\lambda_2 y-\lambda_3 x} : (x,y) \mid 0 < y < x \\ f_{XX}(x,x) = \lambda_3 e^{-\lambda_3 x} : (x,y) \mid 0 < x = y \end{cases} \quad (11)$$

The (11) was obtained from general equation for bivariate exponential PDF given in [12] after some mathematical manipulations.

Parameters λ_1 and λ_2 in the equation are ratios of the corresponding PDFs for variables X and Y. λ_3 is the parameter that considers dependence between X and Y. It is obtained by the (12).

$$\rho = \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}, \quad (12)$$

where ρ is the correlation coefficient of X and Y.

Studying changes of I_{XY} for calculated and actual well flow rates in different samples allows finding the sets with the highest joint reduction of uncertainty (JR). These sets as hypothesized include wells with potentially the most unpredictable flow rates. In opposite, the sets with lowest JR values demonstrate the most predictable behavior. Therefore, it is expected that the JR value increases with decreasing of flow rate predictability. The predictability is estimated by deviations between calculated and measured well flow rates.

In the study the values of JR in uncertainty were calculated for datasets from the table 1. The calculation results along with the degree of dependency ρ are presented in Table 4.

The sample of full data including all classes from A to D has the relatively small JR value. This sample has a large part belonging to the classes A and B and a small part belonging to the other classes. In this case the uncertainty of the initial data

is small and the data are high correlated. As a result there is not big uncertainty reduction for the calculated data when knowing the measured ones and this set has relatively good predictability. The JR value for data belonging to the class A itself is only a few smaller than this for class B. The maximal JR value was obtained for dataset of classes C and D. These classes include the potentially unpredictable wells.

Table 4: Joint reduction in uncertainty for pairs of datasets corresponding to calculated and actual (measured) values of flow rate.

Datasets	ρ	H_X+H_Y	H_{XY}	I_{XY}
A P + A F	0.952	9.979	4.781	5.198
B P +B F	0.914	10.081	4.619	5.462
C+D P + C+D F	0.576	11.003	0.0565	10.947
B+C+D P + B+C+D F	0.721	10.797	6.945	3.852
A+B+C+D P + A+B+C+D F	0.788	10.821	9.290	1.531

The described results are able to confirm the hypothesis of existing dependency between JR value and the predictability of well flow rate. It was found that datasets including data of wells that demonstrate deviations between calculated and actual values of flow rate had greater JR values then datasets without deviations. However, the amount of data used in the study is not able to give stable classification. In addition, the study shows that the algorithm is very sensitive to the concrete values of deviations between desirable and actual flow rates in any cases. This sensitivity is a cause of small difference between JR values of classes A and B. In the example presented in the article the concrete values of deviation are not considered.

It also should be noted that the preliminary classification of data was carried out by a single expert. So, the probability of misclassifying between classes A and B is relatively high. In practice it is suggested to use the algorithm presented in the next section for automatic classification based on the JR value.

4 APPLYING ENTROPY CONCEPT FOR ESTIMATION THE WELL FLOW PREDICTABILITY

The dependency between JR value and well flow rate predictability can be used for classification of wells by their flow rate predictability.

The initial dataset for this classification must include the equal amount of desirable (calculated with the models) and actual (measured in real well) flow rate values from any number of wells. The amount of values for each well must also be equal. Since oil production companies measure and recalculate the well rate values every equal time period (e.g. a year, a month etc.) these conditions are easy to match. The required total amount of data depends on how many classes need to be obtained. A number of classes (that determined by an expert before classification procedure) corresponds to a number of datasets are being obtained when classifying.

The classification procedure begins with dividing the initial dataset into parts of desirable and actual flow rate. After that, the distribution laws for each part are found and the parameters of distribution are calculated. Then the joint distribution law parameters are calculated. Finally, the entropies and the JR value for initial dataset are calculated. This value is used as a low constraint for JR value.

In the next iteration the initial dataset is divided into two equal subsets and the previous steps are repeated obtaining two JR values. The next iteration starts with comparing JR values. The dataset that has the lowest value is divided into two equal datasets and the previous steps are repeated again. The classification stops when the number of classes matches the value given by an expert.

As a result, the subsets sorted by the JR value are obtained. The subset with the largest JR value includes wells potentially the most unpredictable flow rates.

It should be noted that the division of the dataset in each iteration is carried out by the way that the values corresponding to one well cannot be divided into several subsets.

5 CONCLUSIONS

The concept of differential entropy presented in the article for estimating the predictability of oil well flow rates proves its usability. The dependency between JR value, calculated based on differential entropies of desirable and actual flow rates, and the flow rate predictability was obtained. The exponential distributed test dataset was used to illustrate work of the proposed method.

The described method of JR value calculation is appropriate not only for exponential distribution but also for the most of distribution types with given PDFs. However, it requires finding the joint PDF of two random variables that is sometimes a complicated task. Solutions of this task for different distributions are considered in [13]-[16].

Obtained dependency can be used for classification of the oil wells by their flow rates predictability. A simple iterative classification algorithm is presented in the article. The algorithm gives a solution of the problem of estimating the predictability of concrete wells at the early stages of their lifecycles. It will help oil production engineers and energetics to find wells that require special operation methods. In general, when using large amount of data describing flow rates of different wells the algorithm will help to correct the flow rate prediction models that are used for pump selection and well operation control.

The further study of the proposed algorithm will be carried out in field of estimation of the algorithm sensitivity and ways of its control. Additional study is also required for the procedures of finding the joint PDF for different distribution types.

The results of the study will be implemented in the software for analysis of the electrical power supply systems of oil fields [17].

The project is aimed at supporting of a new Master's program "Conceptual Design and Engineering to Improve Energy Efficiency" for preparing of engineers, scientific researchers and managers in energetics and related branches [18].

Research is also supported by educational and research grant 573879-EPP-1-2016-1-FR-EPPKA2-CBHE-JP by European program Erasmus+ (Project INSPIRE).

REFERENCES

[1] K. Bjørlykke, Petroleum geoscience: from sedimentary environments to rock physics, Springer-Verlag Berlin Heidelberg, 2010, 508 p.

[2] G. Takacs, Electrical submersible pump manual: design, operations, and maintenance, Gulf Professional Publishing, 2009, 420 p.

[3] T. M. Cover, J. A. Thomas, Elements of information theory, 2nd edition, Wiley, 2006, 748 p.

[4] V. N. Fashchilenko, , S. N. Reshetnyak, Resonant behavior of electric drives of mining machines, Gornyi Zhurnal, vol. 7, 2017, pp. 80-83, doi: 10.17580/gzh.2017.07.15.

[5] N. J. Hyne, Nontechnical guide to petroleum geology, exploration, drilling and production, 2nd edition, Pennwell Books, 2001, 575 p.

[6] N. L. Johnson S. Kotz, N. Balakrishnan, Continuous univariate distributions, 2nd edition, vol. 1, Wiley, 1994, 761 p.

[7] J. V. Mihalowicz et. al., handbook of differential ntropy, CRC Press, 2014, 220 p.

[8] I. Ghosh, A. Alzaatreh, A new class of bivariate and multivariate exponential distributions, Far east journal of theoretical statistics, vol. 50, issue 2, 2015, pp. 77-98, doi: 10.17654/FJTSMar2015_ 077_098.

[9] S. Nadarajah, D. Choi, Arnold and Strauss's bivariate exponential distribution – products and ratios, New Zealand journal of mathematics, vol. 35, 2006, pp. 189-199.

[10] S. K. Iyer, D. Manjunath, R. Manivasakan, Bivariate exponential distributions using linear structures, The Indian journal of statistics, vol. 64, series A, pt. 1, 2002, pp. 156-166

[11] D. Kundu, R. D. Gupta, Bivariate generalized exponential distribution, Journal of multivariate analysis, vol. 100, issue 4, 2009, pp. 581-593, DOI: 10.1016/j.jmva.2008.06.012

[12] B. M. Bemis, Some statistical inferences for the bivariate exponential distribution, dissertation, 1971, 116 p.

[13] A. Seijas–Macías, A. Oliveira, An approach to distribution of the product of two normal variables, Discussions mathematicae probability and statistics, vol. 32, 2012, pp. 87-99, DOI: 10.7151/dmps.1146

[14] Y. A. Tashkandy, M. A. Omair, A. Alzaid, Bivariate and bilateral gamma distributions, International journal of statistics and probability, vol. 7, no. 2, 2018, pp. 66-79.

[15] E. Furman, On a multivariate gamma distribution, Statistics and probability letters, vol. 78, 2008, pp. 2353-2360.

[16] S. Nadarajah, A. K. Gupta, Some bivariate gamma distributions, Applied mathematics letters, vol. 19, 2006, pp. 767-774

[17] I. Luzyanin , A. Petrochenkov, Practical Aspects of Software Developing for the System of Structural and Functional Analysis of Power Supply Systems in Oil Companies, Proceedings Of The 5th International Conference On Applied Innovations In IT, vol. 5, 2017, pp. 65-69, doi: 10.13142/KT10005.31, WOS: 000402660300 009

[18] A. Lyakhomskii et. al., Conceptual design and engineering strategies to increase energy efficiency at enterprises: Research, technologies and personnel, Proceedings of 2015 IV Forum Strategic Partnership of Universities and Enterprises of Hi-Tech Branches (Science. Education. Innovations), 2015, pp. 44–47, doi: 10.1109/IVForum.2015.7388249, wos: 000380529800015.

Question Embeddings Based on Shannon Entropy
Solving intent classification task in goal-oriented dialogue system

Aleksandr Perevalov[1], Daniil Kurushin[1], Rustam Faizrakhmanov[1] and Farida Khabibrakhmanova[2]
[1]*Informational Technologies and Automatic Systems Department, Perm National Research Polytechnic University, 9 Pozdeeva st., Perm, Russia*
[2]*Foreign Languages, Linguistics and Translation Department, Perm National Research Polytechnic University, 29 Komsomolsky prospect st., Perm, Russia*
perevalovproduction@gmail.com, dan973@yandex.ru, fayzrakhmanov@gmail.com, farida@pstu.ru

Keywords: Text Classification, Word Embeddings, Shannon Entropy, Intent Classification, Natural Language Processing, Dialogue Systems, Word2vec, FastText.

Abstract: Question-answering systems and voice assistants are becoming major part of client service departments of many organizations, helping them to reduce the labor costs of staff. In many such systems, there is always natural language understanding module that solves intent classification task. This task is complicated because of its case-dependency – every subject area has its own semantic kernel. The state of art approaches for intent classification are different machine learning and deep learning methods that use text vector representations as input. The basic vector representation models such as Bag of words and TF-IDF generate sparse matrixes, which are becoming very big as the amount of input data grows. Modern methods such as word2vec and FastText use neural networks to evaluate word embeddings with fixed dimension size. As we are developing a question-answering system for students and enrollees of the Perm National Research Polytechnic University, we have faced the problem of user's intent detection. The subject area of our system is very specific, that is why there is a lack of training data. This aspect makes intent classification task more challenging for using state of the art deep learning methods. In this paper, we propose an approach of the questions embeddings representation based on calculation of Shannon entropy. The goal of the approach is to produce low dimensional question vectors as neural approaches do and to outperform related methods, described above in condition of small dataset. We evaluate and compare our model with existing ones using logistic regression and dataset that contains questions asked by students and enrollees. The data is labeled into six classes. Experimental comparison of proposed approach and other models revealed that proposed model performed better in the given task.

1 INTRODUCTION

Developing of domain-specific question-answering system requires solving natural language understanding tasks. One of them is classification of user's intent, which is frequently solved by machine learning methods. Obviously, machine learning models cannot work with a text itself, and it is required to represent the text as a vector. However, there are some questions about how to represent a text as a vector including the fact that the vector has to represent semantic meaning of the text. Classical methods, such as Bag of Words and TF-IDF are always good baseline for text vectorization in classification tasks, but these methods are producing sparse vectors, that are becoming very big as the data grows.

Modern natural language processing science includes a lot of text-to-vector representations. The major part of them is based on distributive hypothesis: Words that occur in the same contexts tend to have similar meanings [1]. Methods of word to vector representations, such as word2vec [2], FastText [3], Vector Space Model [4], etc. are actually formalization of distributed hypothesis, and are called word embeddings. Therefore, sentence to vector representations — sent2vec [5] and document to vector representations — doc2vec [6] are based on methods mentioned before. All the methods above use neural networks to maximize conditional probability between similar words, that is why they are performing well only when there is enough data

for training. However, there are also some cases when researchers or developers come across a lack of data, so modern methods do not work well. Intent Classification on a small dataset is a challenging task for data-hungry state-of-the-art Deep Learning based systems [7].

To summarize previous paragraph, we need to develop a simple, non-data-hungry method of word to dense vector representations which can outperform both classical and modern methods in condition of small and specific dataset. As a solution, we propose the approach of question embeddings based on Shannon entropy calculation, which main idea is to represent word by its' entropy distribution within the questions in given dataset.

The approach will be tested in context of intent classification task within question-answering system for consultation of university students and enrollees. The dataset contains 1300 questions labeled into six classes.

2 RELATED WORK

2.1 TF-IDF

TF-IDF (term frequency-inverse document frequency) is a classical statistic that reflects how important a word is to a document in a given corpus. Given a document collection D, a word w, and an individual document $d \in D$, we calculate (1):

$$w_d = f_{w,d} * \log\left(\frac{|D|}{f_{w,d}}\right), \quad (1)$$

where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of the corpus, and $f_{w,D}$ equals the number of documents in which w appears in D [8]. In this case, document vector represented as a set of TF-IDF statistics for every word in given collection of documents.

2.2 Vector space model

An alternative to TF-IDF is Pointwise Mutual Information (PMI) which is being calculated in vector space model. Let F be a word context (word co-occurrence within window h) matrix. Based on context matrix F we calculate matrix X (2), (3) [4]

$$pmi_{i,j} = \log\left(\frac{p_{i,\mathrm{j}}}{p_{i*}p_{j*}}\right) \quad (2)$$

$$X = [x_{i,\mathrm{j}}], x_{i.\mathrm{j}} = \begin{cases} pmi_{i,\mathrm{j}}, \mathrm{pmi}_{i,j} > 0 \\ 0, \mathrm{pmi}_{i,j} \le 0 \end{cases} \quad (3)$$

In general, X is very sparse that is why truncated singular value decomposition (SVD) is applied (4):

$$\dot{X} = U_k \Sigma_k V_k^T, \mathrm{k} < \mathrm{r}, \quad (4)$$

where U and Σ are orthonormal matrixes and V is diagonal [9], r is rank of X, k is new rank. Given matrix $\dot{X}$ best approximates the original matrix X and minimizes the dimension size. Thus, in matrix $\dot{X}$ i-th row represents a vector of i-th word.

2.3 Word2vec and FastText

Word2vec is technique that can be used for learning high-quality word vectors from huge data sets with billions of words, with low dimensionality of word vectors [2]. It has two architectures: Continuous Bag of Words – predicts the current word based on the context, and the Skip-Gram model predicts surrounding words given the current word [2]. These architectures are shown on Figure 1.

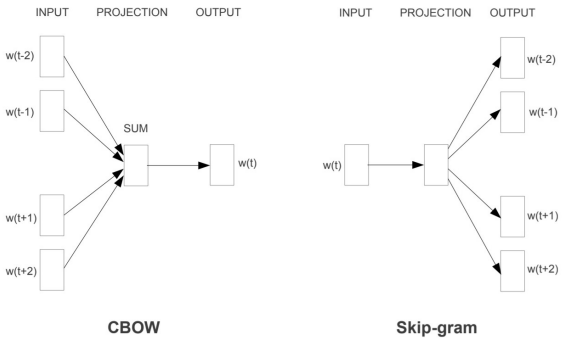

Figure 1: Word2vec architectures.

FastText is an approach proposed by word2vec creators based on Skip-Gram model, where each word is represented as a bag of character n-grams. Let's take word *hello* with n = 3 as an example, it will be represented by the character n-grams:

```
<he, hel, elo, llo, lo>
```

In this case, a word vector is represented as a sum of the vector representations of its n-grams (5):

$$w_t = \sum_{g \in G_w} z_g, \quad (5)$$

where g is an character n-gram, G_w is a set of n-grams appearing in w and z_g is a vector representation of given n-gram.

3 METHODS

3.1 Data collection

Domain-specificness of the intent classification task requires using relevant data – real-life questions that were asked by students and enrolees. Analysis of existing datasets revealed that there is no open source information required for solving our problem. This is due the fact that the solving classification task is highly specific.

Data collection was performed by scraping open data sources, specifically the following websites: pstu.ru, vk.com/politehperm, abiturient.ru. As a result, 7300 questions were collected. Based on this data, question taxonomy has been developed. The taxonomy consists of six classes: *DOC* – questions about documents, *ENTER* – questions about enrolment process, *ORG* – common questions, *PRIV* – questions about privileges during enrolment, *RANG* – questions related to passing score/exam results, *HOST* – questions about student hostels. Taxonomy is shown on the Figure 2.

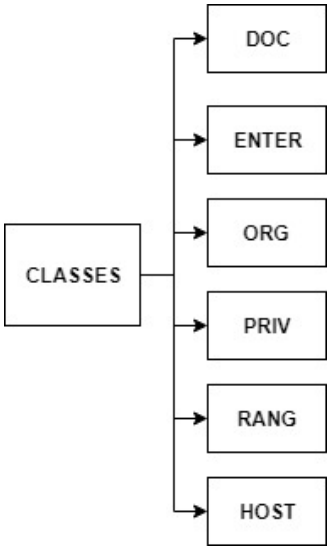

Figure 2: Questions taxonomy.

The example of the training dataset is shown in the table 1. Original data is in Russian, translation is given in parentheses.

Table 1: Training data example.

Question	Class
Можно ли подать документы в субботу? (Is it possible to submit documents on Saturday?)	DOC
Когда день открытых дверей? (When is open doors day?)	ORG
Когда публикуются списки зачисленных? (When lists of enrolled will be posted?)	RANG
Какие документы нужны для заселения в общежитие? (What documents are needed for checking in to students hostel?)	HOST

Data labelling was made using ipyannotate tool (https://github.com/natasha/ipyannotate) manually. As the result, 1300 questions were labelled. Every question was referred to one class from the taxonomy. Question-class distribution is shown on Figure 3.

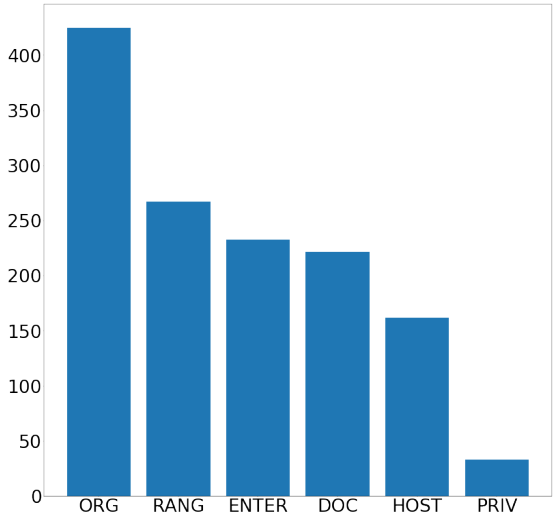

Figure 3: Question-class distribution.

The figure is illustrating imbalance of the dataset: the *ORG* class has much more samples than the average, whereas the *PRIV* class has much less than the average. This aspect has been taken into consideration during classifier evaluation.

3.2 Data Preprocessing

Data preprocessing starts with named entity extraction and its transformation to the normal form. It is needed to remove homonyms – words that have the same meaning but different spelling. For named

entity extraction, rule-based yargy library (https://github.com/natasha/yargy) was used. Next step in preprocessing is tokenization. After that, every token is being checked for multiple rules, for example if the token is one of the stop-words, or if it contains any Latin letters (system works with Russian language). If one of the rules returns "true" value – then the token is removed, otherwise it remains in the question. All tokens in preprocessed questions are separated by the space character. The code of preprocessing functions is shown in Listing 1.

Listing 1: Preprocessing functions implemented in Python.

```
def preprocess_word(word):
      return stemmer.stem(morph.parse(word)
[0].normal_form.lower())

def preprocess_list(list_):
   new_list = []
   for l in list_:

     for rule in list(ner.rules.keys()):
       parser=ner.Parser(ner.rules[rule])
         for match in parser.findall(l):
           for _ in match.tokens:
             l=l.replace(_.value,rule)

      words = tokenizer.tokenize(l)
      new_words = [preprocess_word(word) for
word in words
      if morph.parse(word)[0].normal_form
not in stopwords and not any(char.isdigit()
for char in word) and not
bool(re.search(r'[a-zA-Z]', word)) and
morph.parse(word)[0].normal_form.lower() not
in custom_stopwords]
      new_list.append(' '.join(w for w in
new_words))
   return new_list
```

Because of word2vec interpretability, preprocessed questions were represented as word2vec embeddings [2] and visualized for cluster analysis. The visualization is shown on the Figure 4.

Visualization showed that some of the classes are clearly separated one from another: *RANG* (marked red on the figure), *DOC* (marked yellow on the figure). Such classes as *ORG* (marked violet on the figure) has many intersections with other classes. This fact undoubtedly has negative impact on effectiveness of the classifier and will be resolved in the future work by redesigning questions taxonomy.

Figure 4: 2D visualization of word2vec embeddings.

3.3 Question Embeddings

We propose the approach to the document embeddings or a questions embeddings based on Shannon entropy calculation [10] for every word in the question. One of the Shannon entropy interpretations is a measure of information rate. In this way, the measure of information amount in the word in question is calculated.

First of all, the list of words that appear in the document, is made up. After that, Shannon entropy for every word in the list within every question is calculated (6).

$$e_{ij} = \begin{cases} -p_{ij}\log_2(p_{ij}), p_{ij} = w_{ij} / n_{ij}, w_{ij} > 0 \\ -0.0001, w_{ij} = 0 \end{cases}, \quad (6)$$

where w_{ij} _ number of occurrences of *j-th* word in *i-ih* question, n_{ij} _ number of words in *i-th* question.

Speaking in terms of machine learning, we calculate matrix where rows (or samples) represent questions, and columns (or features) represent words, so the obtained matrix has 1300 rows (questions) and 1212 columns (features or unique words in training set). Thus, the question is represented by a words entropy vectors. In this case, we need to transpose the matrix, so the rows will represent words, and the columns will represent questions or features (7). In this way, a word, or more precisely a word meaning, is represented by its distribution within the questions in given dataset.

$$M^T = \begin{vmatrix} e_{11} & e_{21} & \dots & e_{m1} \\ e_{11} & e_{11} & \dots & e_{m1} \\ \dots & \dots & \dots & \dots \\ e_{1n} & e_{1n} & \dots & e_{m,n} \end{vmatrix} \quad (7)$$

The obtained transposed matrix is sparse (1300 features), thus a dimension reduction using truncated singular value decomposition (4) will be done. In our case, the dimension of the vector will be 200. It is possible by taking first 200 components of decomposed matrixes.

In order to represent a question as a vector, the vectors of words that appear in the question, are to be chosen, and the average of these word vectors, are to be taken (8). The obtained matrix of question vectors and question class will be used as the training set for the classifier.

$$Q_i = \frac{\sum w_j}{count(W)}, w_j \in W \quad (8)$$

where Q_i – vector of *i-th* question, w_j – word vector, W_i – set of the words, that appear in *i-th* question.

4 EXPERIMENTS

For experimental testing of the approach proposed, the linear classifier e.g. logistic regression is used. As the dataset has multiple classes, the one vs rest classification method is used. The final model inspired by [11] is shown on the Figure 5.

The proposed Shannon entropy embeddings have been compared with TF-IDF, word2vec and FastText models. Word2Vec and Fast Text word embeddings were transformed to question embeddings by taking the average vector of words' vectors contained in question. During the experiments, dataset was shuffled and split into 5 folds for cross-validation. The evaluation metric for classifier is F1-score (9).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

The classification algorithm is logistic regression and classification scheme is "One vs Rest". The results of the experiments are presented in Table 2.

As it can be seen, the proposed method has performed better than existing ones on students and enrollees questions dataset. Its F1-score is 2% higher than the best of the others — TF-IDF. FastText showed the worst result —63% (11% lower than the proposed method). It can be explained by the lack of data in the training set.

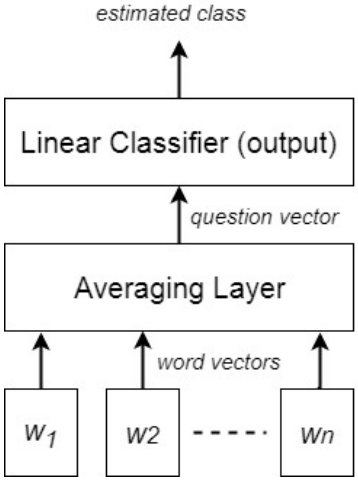

Figure 5: Classification model for the experiments.

Table 2: Classification report – questions dataset.

	F1-score			
	TF-IDF	Word2vec	FastText	**Shannon Entropy**
DOC	0.77	0.76	0.66	**0.76**
ENTER	0.62	0.59	0.58	**0.65**
ORG	0.73	0.66	0.65	**0.74**
PRIV	0.24	0.25	0.31	**0.32**
RANG	0.74	0.65	0.65	**0.75**
HOST	0.91	0.85	0.75	**0.91**
Average	0.72	0.67	0.63	**0.74**

Also, analysis of the results revealed that PRIV class is hardly recognized. This can be explained by the lack of the objects in this class compared with the other classes. To avoid this problem, in future work this class could be merged with other ones.

To make sure in model performance, the proposed approach has been tested on imdb.com reviews dataset. The dataset contains two classes: positive and negative review. Every class contains 10000 reviews written in English. The results of experiments are shown in Table 3.

Table 3: Classification report – IMDB dataset.

	F1-score			
	TF-IDF	Word2vec	Fast Text	**Shannon Entropy**
POSITIVE	0.9	0.88	0.86	**0.9**
NEGATIVE	0.9	0.88	0.86	**0.9**
Average	0.9	0.88	0.86	**0.9**

It can be seen, that TF-IDF and Shannon entropy showed the same result on F1-score, however, there

is difference between dimension sizes in these models: TF-IDF has dimension size equals to 8623 whilst the proposed Shannon entropy model has only 200 (because of applying truncated singular value decomposition, described in 2.2 chapter).

Considering this fact, it can be said that proposed model can store the same information amount with lower dimension size, which can help in improving speed during the data processing.

5 CONCLUSIONS

In this work, the approach of question vector representation based on Shannon entropy, has been proposed. For experimental testing, the intent classification task has been suggested. The task was set in terms of voice assistant system for students and enrollees of the university.

The dataset containing students' and enrollees' questions was collected. After that, the taxonomy of the data was designed; the dataset was labeled by classes according to the taxonomy. The approach of question vector representation was designed, implemented and tested.

As the result, the proposed method performed better comparing to the TF-IDF (F1-score is 2% higher), Word2vec (F1-score is 7% higher) and FastText (F1-score is 11% higher).

There was also one experiment on imdb.com reviews dataset that have proved proposed model performance: TF-IDF and Shannon Entropy showed the same result on F-score – 90%, however Shannon Entropy has lower dimension size rather than TF-IDF. This fact can help in improving speed during the data processing without any information loss.

In future work, the redesign of the existing taxonomy for imbalance reduction is planned. Also, modernization of the approach using weighted averaging is going to be done.

The obtained classifier model and the dataset will be used in voice assistant system for students and enrollees consultation. All the data, source code and models described above are available online: https://github.com/Perevalov/intent_classifier.

REFERENCES

[1] Harris, Zellig, "Distributional structure," In: *Word*, S. , no. 23, pp.146-162, 1954.

[2] Mikolov, Tomas, Chen, Kai, Corrado, Greg and Dean, Jeffrey. "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[3] Joulin, Armand, Grave, Edouard, Bojanowski, Piotr, Douze, Matthijs, Jégou, Hervé and Mikolov, Tomas, "FastText.zip: Compressing text classification models," arXiv preprint arXiv:1612.03651, 2016.

[4] Turney, Peter D, Pantel, Patrick and others, "From frequency to meaning: Vector space models of semantics,"Journal of artificial intelligence research 37 , no. 1, pp. 141-188, 2010.

[5] Pagliardini, Matteo, Gupta, Prakhar and Jaggi, Martin, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," arXiv preprint arXiv:1703.02507, 2017.

[6] Le, Quoc V and Mikolov, Tomas, "Distributed Representations of Sentences and Documents," Paper presented at the meeting of the ICML, 2014.

[7] K. Shridhar, A. Dash, A. Sahu, G. Grund Pihlgren, P. Alonso, V. Pondenkandath, G. Kovacs, F. Simistira, M. Liwicki, "Subword Semantic Hashing for Intent Classification on Small Datasets," arXiv preprint arXiv:1810.07150, 2018.

[8] G. Salton, C. Buckley, "Term-weighing approache sin automatic text retrieval," In Information Processing & Management, no. 24(5), pp. 513-523, 1988.

[9] G. H. Golub, C. F. Van Loan, "Matrix computations. Third Edition," The John Hopkins University Press, 1996.

[10] Vajapeyam, Sriram, "Understanding Shannon's Entropy metric for Information," arXiv preprint arXiv:1405.2061, 2014.

[11] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, "Bag of Tricks for Efficient Text Classification," arXiv preprint arXiv:1607.01759, 2016.

The Improvement of Machine Translation Quality with Help of Structural Analysis and Formal Methods-Based Text Processing

Anna Mylnikova and Aigul Akhmetgaraeva
Perm National Research Polytechnic University, Komsomolsky Ave. 29, Perm, Russia
novikova@yandex.ru, aigul15.ahmetgaraeva@yandex.ru

Keywords: Machine Translation, Machine-Aided Translation, Language Pair, Classification, BLEU Scores, the Frequency of Vocabulary Use, Algorithm, the Evaluation of Translation Quality.

Abstract: This article considers the issues of enhancing the quality of machine translation from one language into another one by structuring linguistic patterns and using identification methods for the situations that cannot be processed by the suggested approach and are subject to individual processing. According to the BLEU score metrics, the described approach allows to increase the quality of machine translation on average by 0.1 and reduce postprocessing time due to the identification of idioms and words with context-dependent meanings by translation. The experiment data base of the study was built upon online available pairs of texts that cover the events of FIFA World Cup 2018 and well-known idioms.

1 INTRODUCTION

Beginning with the industrialization age, there is an ongoing growth in labor efficiency. The issues of sharing knowledge, activity outcomes and technologies have become global and, hence, communication in foreign languages has gradually penetrated from elite environment into routine activity. The emergence of new products and projects leads to a big amount of in-line documentation and correspondence issues. This process entails works on preparing various documents in different languages and their translation.

First, we encountered machine and machine-aided translation systems in science fiction books and movies, but in the middle of the XX century the organizations Warren Weaver of the Rockefeller Foundation and RAND announced the possibility of making translations from one foreign language into another one through a mediator – computer [1] and started to implement that idea as a part of projects. The conceptual guidelines of machine translation system operation are laid down in the works of A. Vakher, W. Weaver, H.P. Edmundson, P.G. Hays, G. Artsrouni and P.P. Smirnov-Troyanskii [2]. The research outcomes of this scientific school distinguished machine translation into a science intensive direction, that got exponential development by introducing such methods as 1) structural grammatical methods (GAT, COMIT, METAL, ESPERANTO), 2) syntactic methods (P. Garvin, E. Brown, A. Lukjanow, etc.) 3) semantic approaches (ETAP-1,2,3, DLT, Rosetta, KANT).

The current studies are focused on the issues of enhancing the quality of machine translation and, as a rule, take advantage of hybrid models that combine the methods of corpus linguistics, statistical analysis and cognitive analysis on the basis of the methods that are developed in the theory of intelligence systems [3], [4], [5].

2 THE CURRENT STATE OF MACHINE-AIDED TRANSLATION METHODS AND SYSTEMS

Meaningful translation from one foreign language into another one underlies the identification of the syntactic structure of source-language and the model that actualizes the in-depth and external semantics of this phrase and, hence, the identification of a single value matching on the syntactic and semantic levels of target-language. This task is challenging due to some reasons. First of all, the difference in syntactic structures of natural languages leads to an effect of rigid and “not rigid” localization effects [6], when,

in particular, one syntactic structure of English or German languages can be assigned to up to 4 variants of syntactic structures in Russian language due to its not rigid theme - rheme based order; however, the semantic content in the latter 4 Russian variants maintains generally equal.

It is assumed that the formalization of linguistic structures for source-language and target-language as well as the development of their match pattern base can help achieve meaningful machine translation. In the 98s, K.A. Papenini suggested that this problem can be tackled by using direct maximum entropy translation models [7]. The drawback of such models is a strict limitation of parallel data. The German scientists Franz J. Och and Hermann Ney developed this idea for statistical machine translation by introducing conventional dynamic programming search algorithms. With help of Bayes' decision rule, they included a dependence parameter on the hidden variable of the translation model [8]. However, this model works only by true probability distributions, which is not always the case due to differences in language systems and the nature of thought unfolding in different languages [9], [10].

Another popular approach today is an approach which is based on the methods of machine learning. For instance, in 2010 the corporation Google developed and embedded the method of cross-language near duplicate detection by using parallel document mining for statistical machine translation system learning [11]. In this approach they extend the local distribution distance of a word or phrase to be translated and apply deep learning methods to teach neural networks. Currently, the system of machine translation Google identifies the local distribution distance within 8 words [12], [13] and does not cover the lexical and grammatical context of the whole phrase. As a result, it entails a number of translation mistakes.

3 THE COMPARISON OF MACHINE TRANSLATIONS AND THE ANALYSIS OF MISTAKES

Based on the analysis of text translations of various thematic scope websites, news blocks devoted to the coverage of FIFA World Cup 2018 events (official texts translated in many languages were taken as most accurate translations since they were translated by professional translators which ensures the accuracy of professional terms, well-known expressions and idioms used in translation), performed by the machine translation systems Google, PROMT, SYSTRAN, Babylon, Microsoft translator, Yandex translator we can observe only a low quality of machine translations. See the results of the BLEU score metrics used for the evaluation of machine translation quality [14], [13] in the Table 1.

Table 1: The evaluation of machine translation quality made by the BLEU score metrics.

Machine translation system	BLEU score metrics (Russian-English)	BLEU score metrics (English-Russian)
Google	0.298	0.5
PROMT	0.232	0.413
SYSTRAN	0.155	0.175
Babylon	0.26	0.45
Microsoft translator	0.307	0.51
Yandex translator	0.304	0.58

Taking into account the fact that according to the BLEU score metrics the highest result corresponds to the value «1», we can conclude that nowadays the problem of producing accurate meaningful translation from the source-language to the target language is not solved yet. Therefore, it is important to understand the reasons of such low quality. For this case, let us analyze the most frequently observed mistakes, see the Figure 1.

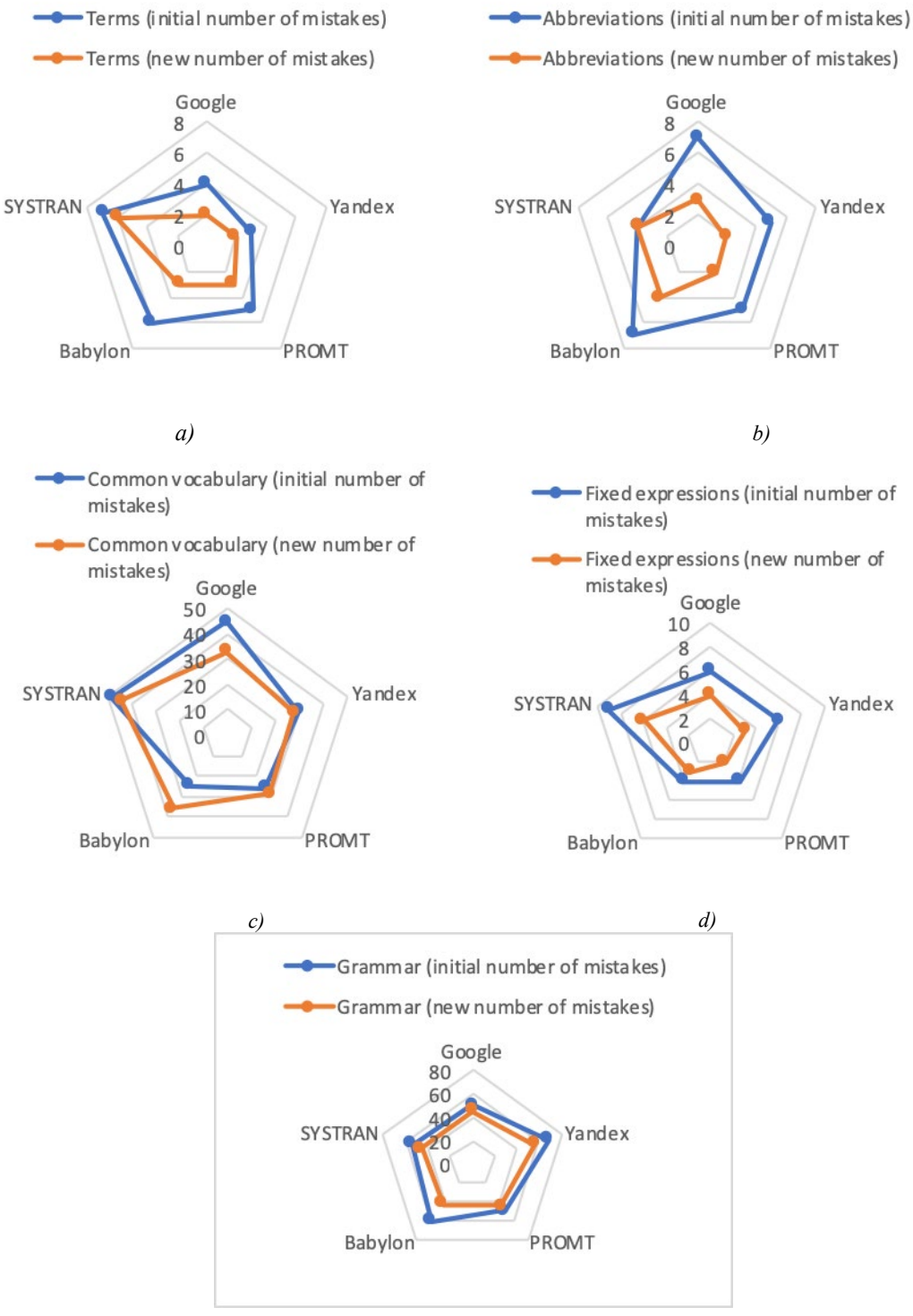

Figure1: Numerical reduction of mistakes by using initial source-language structures and decoded source-language structures: *a)* terms, *b)* abbreviations, *c)* fixed expressions, *d)* common vocabulary, *e)* grammar.

After seeing the given statistics, it is obvious that systems make various errors. In particular, grammar and in some cases semantic errors are dependent on the structure used in the source-language. At the same time, the distributive location of head lexical transducers is significant since it affects the lexical-grammatical phrase realization in English language [9]. We could not agree more with the statement of E. Sumita and H. Iida that «example-retrieval cost is high when the input sentence is syntactically ambiguous» [15].

The decoding of source-language syntactic structures in line with the structures of target-language grammar system allows to reduce not only the number of grammar mistakes, but also semantic mismatches [16], [17].

However, in machine translation this approach does not help avoid all the sense distorting semantic mistakes. The use of idioms, fixed and professional terms leads to a word-by-word translation, that distorts the meaning of the phrase. For instance, the phrase in Russian language «Я не уверен, смогу ли выслать доклад сегодня вечером, он еще совсем сырой» was translated by Google into English language as follows: «I'm not sure if I could send a report tonight, he's still very raw». The sense of the phrase is not given correctly though, as the accurate translation of the Russian phrase into English language corresponds to the following phrase: «Most likely I won't be able to send the report tonight, it's far from done».

It is getting worse when it comes to idioms as whole phrases (see the Table 2).

It is logically to assume that such situations cannot be processed correctly with help of the existing concepts and methods. The identification of such "special" linguistic expressions with further individual processing could be a way to deal with this challenging situation. Hence, we need to determine the attributes which are required for identifying such expressions.

Table 2: Examples of idiom machine translation.

Source-language phrase	**Accurate translation into target-language**	**Google**	**PROMT**	**Yandex**	**Babylon**	**SYSTRAN**
A) English-Russian						
1. Born with a silver spoon in his mouth	**Родившийся под счастливой звездой**	Родился с серебряной ложкой во рту	Терпевший серебряная ложка в его рту	Родился с серебряной ложкой во рту	Родился с серебряной ложкой во рту	Принесенный с серебряной ложкой в его рте
2. An old head on young shoulders	**Мудр не по годам**	Старая голова на молодых плечах	Старая голова на молодых плечах	Старая голова на молодых плечах	Старая голова на молодых плечах	Старая голова на молодых плечах
3. To have one's head in the clouds	**Витать в облаках**	Иметь голову в облаках	Витать в облаках	Чтобы иметь голову в облаках	На голова в облаках	Иметь one голову в облаках
4. To take it on the chin	**Не падать духом**	Взять его на подбородок	Взять его на подбородке	Чтобы взять его на подбородок	Принять его на подбородке	Принять его на подбородке
B) Russian-English						
1. Уйти по-английски	**To take French leave**	Take French leave	To take French leave	Leave in English	Take French leave	To leave in English
2. Подложить свинью	**To play a dirty trick**	Put a pig	Play a dirty trick	A pig in a poke	Send to a pig	To place the pig
3. Ударить в грязь лицом	**To have egg on one's face**	Smash face	To lose face	To strike in a dirt the person	Hit the dirt in the face	To strike into mud by face
4. У чёрта на куличиках	**In the middle of nowhere**	At the damn thing	At the world's end	In the middle of nowhere	The feature on the куличках	In feature on kulichkakh

4 MACHINE TRANSLATION IMPROVEMENT

4.1 The Use of Distributive Localizations to Enhance the Quality of Machine Translation

Distributive localizations are used on the bases of the structuring methods described in [17]. In contrast to the suggested approach, we will use a group of overlapping dependencies that cancel the action of other dependencies by the emergence of certain distributive localizations. This expands the baseline translation system by adding new functional dependencies and, hence, allows to achieve meaningful alignment of the source-language and the target-language without parallel data limitation. For instance, for the case 1A,2A,3A,4B,5A,6A, where 1A – the declarative sentence, 2A – the indicative mood, 3A – the active voice; 4B – the present simple tense, 5A – the affirmation, 6A – the simple predicate (actualized by a notional verb or a copulative verb):

[Parenth][NP3]<NP1>[AdvP1.2][AdvPM][Adv/measure]<VP>[NP2][NP2ext][AdvP3][AdvP2][AdvP1.1]<".">,

where Adv/measure is not applied in one distributional context together with AdvP1.2, AdvP1.3., AdvPM; AdvP2 is not applied in one distributional context together with AdvP1.1, AdvPM, Adv/measure; AdvP1.1 is not applied in one distributional context together with c AdvP2.

Such action will allow to improve machine translation. The Table 3 shows the translation examples from Russian language into English language performed by Google based on the source-language phrase structures and those decoded in accordance with the structure of the target-language. The formalization language suggested in [17] is taken to describe linguistic structures.

4.2 Machine Translation of Idioms and Terms

When working with idiomatic expressions, phrases and terms, which were mentioned in the previous section, the suggested approach helps achieve grammatically correct, but not semantically sound translations. Hence, the outcome does not make sense to the native speaker.

It is assumed that one of the ways to handle this issue is the identification of such phrases, expressions and words and their special processing (post-editing or manual translation).

In this study, idioms are assigned to phrases and expressions, that have similar meanings yet different lexical-grammatical actualization in the source-language and target-language (see examples in the Table 2), terms are assigned to certain words that generate context-dependent meanings (for instance, professional terms).

Table 3: Machine translation examples made by Google without the decoding of the source-language structures and with the decoding of source-language structures.

	Phrase structure	Machine translation	Number of mistakes
Example 1 – «Они в компании всегда быстро проводят обновление программного обеспечения»			
Source-language structure	<NP1>[NP3][AdvP1.2][AdvP2] <VPvf1>[NP2][NP2ext] <".">	They always update the software in the company.	3
Decoded source-language structure	[NP3]<NP1>[AdvP1.2] <VPvf1>[NP2][NP2ext][AdvP2] <".">	In the company, they always carry out software updates quickly.	0
Example 2 – «Раньше ваша компания когда-либо обновляла программное обеспечение для переводчика?»			
Source-language structure	[AdvP1.3/2*]<NP1>[AdvP1.3/4] <VP3> [NP2][NP2ext]<''?''>	Did your company ever update the software for an interpreter?	3
Decoded source-language structure	<NP1>[AdvP1.3/4]<VP3>[NP2] [NP2ext][AdvP1.3/2*]<''?''>	Has your company ever updated the software for an interpreter before?	0

Table 4: Pairwise comparison of the BLEU score metrics for the idiom «To make a push in the development» (English) - «Сделать толчок в развитии» (Russian).

	Google	PROMT	Yandex	Babylon	SYSTRAN
Google	1	0.4	0.14	0.28	0.4
PROMT	0.4	1	0.57	0.57	0.4
Yandex	0.14	0.57	1	0.5	0.16
Babylon	0.28	0.57	0.5	1	0.4
SYSTRAN	1	0.4	0.16	0.4	1

The experiments of translating texts of various stylistic codes with help of the above-mentioned machine translation systems show that when such special situations emerge – i.e. the presence of idioms and terms in the source-language context – we observe differences in translations. If we take formalization language to represent the source-language phrase, it will be possible to identify such special cases.

Case 1. If the difference is observed on the segment <NP1><VP> [NP2], the special situation is assigned to the whole phrase (or sentence).

In order for the differences to be evaluated, let us perform a pairwise comparison of translations' quality with help of the BLEU score metrics (see an example in the Table 4). To specify the situations, we can use the frequency of vocabulary use in the observed text. Idiomatic expressions are generally based on common vocabulary, which means that the frequency of a notional verb use should be above the average value.

Case 2. If the difference is observed on the segment [NP2][NP2ext][AdvP3][AdvP2][AdvP1.1], the special situation is assigned to a word-combination.

Case 3. If the difference is observed on the segments <NP1> or [NP2], the cause of inaccurate translation is a certain word or a stem-compound.

In the latter two cases the frequency of vocabulary use should be not higher than the average value (see the Figure 2).

Therefore, if we introduce 2 classes and check their relations we can identify the situations that require additional processing with help of a translator.

In general, the machine-aided translation algorithm can be demonstrated by the algorithm given in the Figure 3.

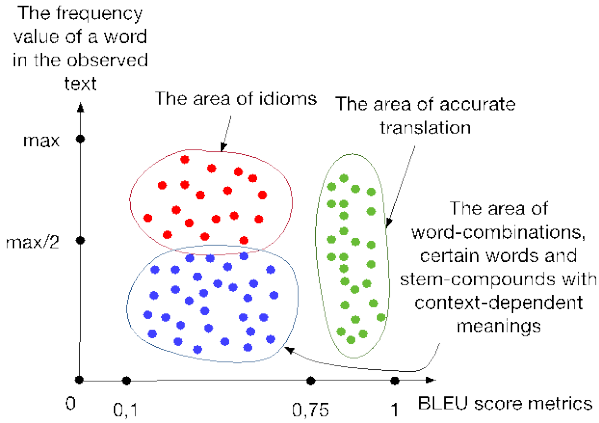

Figure 2: The examples of value distribution built on the analyzed set of values for idioms, word-combinations and certain words and stem-compounds.

Figure 3: The machine-aided translation algorithm.

1. The syntactic parsing of a source-language sentence.
2. The coding of a source-language syntactic structure in the formalization language [17].
3. The decoding of a source-language syntactic structure in accordance with the model of a target-language syntactic structure.
4. The rearrangement of words in a source-language phrase in accordance with the new syntactic structure.
5. The translation of a rearranged phrase with help of the existing machine translation systems.
6. The calculation of the BLEU score metrics under the cases 1-3 specified in the section 4.2. If the evaluation metrics can be assigned to one of the classes in the Figure 2, the corresponding phrase segment is marked with a special label.
7. The selection of a baseline translation (for instance, according to the experiment results of the present study Google delivers best translation outcome).

5 CONCLUSIONS

By manual processing, the suggested algorithm delivers the increase in translation quality by 0,1 of the BLEU score metrics. This evidence is a significant step forward as the existing machine translation systems are competing for basis points. Only in certain cases the difference in the translation quality comes up to tenths among the existing machine translation systems. Besides, the identification of situations that need close attention will considerably save translator's time. Today, the translation algorithm for big texts consists in the use of a machine translation system with further professional proofreading and post-editing.

The highlighted advantages make it clear that the suggested approach will work only with big texts, that will provide a sufficient amount of data for frequency calculation. More than that, the configuration of class memberships will be dependent on the knowledge domain (medicine, law, information technologies, programming, technics, etc.) of an analyzed text and on the language pair. These issues need background investigation. Efficient algorithm operation might need an introduction of a special non-traditional text classification [18], [19]. Besides, not all the phrases can be identified this way (for instance, the idiomatic expression «An old head on young shoulders» from the Table 2 was not identified as a special situation). This requires an additional analysis of the obtained translation result (on how it makes sense to a native speaker) and a possibility of introducing additional classification attributes.

REFERENCES

[1] B. Henisz-Dostert, R. R. Macdonald, and M. Zarechnak, Machine translation. The Hague ; New York: Mouton, 1979.

[2] P. Sojka, Ed., Text, speech and dialogue: 13th international conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010: proceedings. Berlin ; New York: Springer, 2010.

[3] M. R. Costa-Jussà and M. Farrús, "Statistical machine translation enhancements through linguistic levels: A survey," ACM Comput. Surv., vol. 46, no. 3, pp. 1-28, Jan. 2014.

[4] M. R. Costa-jussà, A. Allauzen, L. Barrault, K. Cho, and H. Schwenk, "Introduction to the special issue on deep learning approaches for machine translation," Comput. Speech Lang., vol. 46, pp. 367-373, Nov. 2017.

[5] E. Hasler, A. de Gispert, F. Stahlberg, A. Waite, and B. Byrne, "Source sentence simplification for statistical machine translation," Comput. Speech Lang., vol. 45, pp. 221–235, Sep. 2017.

[6] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," Comput. Linguist., vol. 29, no. 1, pp. 19–51, Mar. 2003.

[7] K. A. Papineni, S. Roukos, and R. T. Ward, "Maximum likelihood and discriminative training of direct translation models," in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181), Seattle, WA, USA, 1998, vol. 1, pp. 189-192.

[8] F. J. Och, C. Tillmann, and H. Ney, "Improved Alignment Models for Statistical Machine Translation," vol. 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[9] H. Alshawi, S. Bangalore, and S. Douglas, "Automatic acquisition of hierarchical transduction models for machine translation," in Proceedings of the 36th annual meeting on Association for Computational Linguistics -, Montreal, Quebec, Canada, 1998, vol. 1, p. 41.

[10] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," Comput. Linguist., vol. 23, no. 3, pp. 377-403, Sep. 1997.

[11] J. Uszkoreit, J. M. Ponte, A. C. Popat, and M. Dubiner, "Large scale parallel document mining for machine translation," COLING 10 Proc. 23rd Int. Conf. Comput. Linguist., pp. 1101–1109, Aug. 2010.

[12] M. Aiken, K. Ghosh, J. Wee, and M. Vanjani, "An Evaluation of the Accuracy of Online Translation Systems," Commun. IIMA, vol. 09, no. 04, 2009.

[13] P. N. Astya et al., Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016): 29-30 April, 2016. 2016.

[14] Association for Computational Linguistics, P. Isabelle, and Association for Computational Linguistics, Eds., Proceedings of the conference, 40th annual meeting of the Association for Computational Linguistics: Philadelphia, [6 - 13] July 2002, University of Pennsylvania, Philadelphia, Pennsylvania. Hauptbd. ... San Francisco: Morgan Kaufmann, 2002.

[15] E. Sumita and H. Iida, "Heterogenous Computing for Example-Based Translation of Spoken Language," Proc. Sixth Int. Conf. Theor. Methodol. Issues Mach. Transl., pp. 273-286, 1995.

[16] A. V. Novikova and L. A. Mylnikov, "Problems of machine translation of business texts from Russian into English," Autom. Doc. Math. Linguist., vol. 51, no. 3, pp. 159-169, Jun. 2017.

[17] A. Novikova, "Direct Machine Translation and Formalization Issues of Language Structures and Their Matches by Automated Machine Translation for the Russian-English Language Pair," Proc. Int. Conf. Appl. Innov. IT, vol. 6, no. 1, p. 85-92., 2018.

[18] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic Text Categorization in Terms of Genre and Author," Comput. Linguist., vol. 26, no. 4, pp. 471-495, Dec. 2000.

[19] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Comput. Linguist., vol. 37, no. 2, pp. 267-307, Jun. 2011.