ISSN 2078-9181 DOI 10.15622/sp.57

РОССИЙСКАЯ АКАДЕМИЯ НАУК
Отделение нанотехнологий и информационных технологий

САНКТ-ПЕТЕРБУРГСКИЙ
ИНСТИТУТ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ РАН

ТРУДЫ СПИИРАН

proceedings.spiiras.nw.ru

ВЫПУСК 2(57)

Санкт-Петербург
2018

18+

SPIIRAS PROCEEDINGS

Issue № 2(57), 2018

Scientific, educational, and interdisciplinary journal primarily specialized
in computer science, automation, and applied mathematics

Trudy SPIIRAN ♦ Founded in 2002 ♦ Труды СПИИРАН

Founder and Publisher

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences

Editor: E. P. Miroshnikova
Technical editor: A. I. Motienko
Translator: P. N. Emeleva

Editorial Board's address

14-th line VO, 39, SPIIRAS, St. Petersburg, 199178, Russia,
e-mail: publ@iias.spb.su, web: http://www.proceedings.spiiras.nw.ru/

The journal is indexed in Scopus

ТРУДЫ СПИИРАН

Выпуск № 2(57), 2018

Научный, научно-образовательный, междисциплинарный журнал с базовой специализацией в области информатики, автоматизации и прикладной математики
Журнал основан в 2002 году

Учредитель и издатель

Федеральное государственное бюджетное учреждение науки
Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)

Адрес редакции

199178, Санкт-Петербург, 14-я линия, д. 39,
e-mail: publ@iias.spb.su, сайт: http://www.proceedings.spiiras.nw.ru/

Журнал индексируется в международной базе данных Scopus

Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук»

CONTENTS

СОДЕРЖАНИЕ

DOI 10.15622/sp.57.1

H.M.A. ABDULLAH, A.V.S. KUMAR

VERTEX SEARCH BASED ENERGY-EFFICIENT OPTIMAL RESOURCE ALLOCATION IN COGNITIVE RADIO AD HOC NETWORKS

Abdullah H.M.A., Kumar A.V.S. **Vertex Search based Energy-efficient Optimal Resource Allocation in Cognitive Radio Ad Hoc Networks.**

Abstract. Cognitive Radio Ad Hoc Networks (CRAHN) is the infrastructure-less network model of Cognitive radios developed in an ad hoc manner. Regulating resource allocation in CRAHN is considered to be an energy constrained problem. Many researches have been carried out for allocating spectrum in an efficient way using various protocols. In this paper, the Spectrum-Map-Empowered Opportunistic Routing (SMOR) model has been utilized as the fundamental model for routing and an energy efficient optimal spectrum allocation solution is provided. In the proposed model, the previously modified SMOR model is enhanced for the main objective of energy efficient and optimal resource allocation using Vertex search algorithm with a gradient-based approximation. Initially, the resource allocation problem is modelled into a non-convex optimization problem. The power allocation, data rate adaptation, channel allocation, and user scheduling policies are optimized for maximization of the energy efficiency during data transmission. The proposed Vortex search algorithm resolves this optimization problem by determining the training interval for the channel estimation and power consumption. The experimental results prove that the proposed Vertex search based modified SMOR (VS-M-SMOR) model provides efficient routing with energy efficient optimal resource allocation.

Keywords: Resource allocation, Spectrum sharing, SMOR, Vertex search, energy efficiency, user scheduling.

1. Introduction. Rapidly rising energy costs and progressively inflexible natural models have prompted a rising pattern of tending to "energy efficiency" part of wireless communication technologies [1]. In an ordinary wireless cell network, the radio access part represents up to more than 70 percent of the aggregate energy utilization [2]. In this way, expanding the energy efficiency of radio networks is imperative to address the difficulties raised by the levels of popularity of activity and energy utilization. *Cognitive radio* technology can assume a critical part in enhancing energy efficiency in radio networks [3]. The subjective capacities have an extensive variety of properties, including spectrum sensing [4], spectrum sharing [5] and versatile transmission [6, 7], which are advantageous to enhance the tradeoff among energy efficiency, spectrum efficiency, transfer speed, and deployment efficiency in wireless networks [8, 9]. CRAHN have been developed to provide better performance than normal cognitive radio networks.

Despite the fact that the Primary Users (PUs) still have needed access to the spectrum, the Secondary Users (SUs) are permitted to have

limited access subject to an obliged corruption on the PUs' execution [10, 11]. In this new worldview of correspondence, the key outline difficulties of a subjective radio network are accordingly to ensure the insurance of the PUs from intemperate obstruction initiated by the SUs and to meet some Quality-of-Service (QoS) prerequisites for the SUs [12, 13]. Then again, spectrum pooling is an opportunistic spectrum access that empowers the community to the officially authorized frequency bands [14, 15]. The fundamental thought is to combine ghastly ranges from various spectrum proprietors into a typical pool, from which the SUs may incidentally lease unearthly assets amid sit still times of the PUs. In actuality, the authorized system should not be changed, though the SUs access unused assets. In subjective radio settings where the PU exercises on the radio spectrum are exceedingly powerful and the genuine open door for the SU access is little, the issue of how we can viably share the briefly accessible frequency bands among the SUs is of special relevance.

In [16], SMOR routing protocol has been introduced for improving the opportunistic routing performance. This model has provided efficient routing with minimal delay; however due to some limitations the model has been modified in [17] with the inclusion of different approximations and Sparsity-based distributed spectrum map as SDS-M-SMOR. There were fewer hybrid models [18, 19] developed in recent past aimed at improving different aspects of SMOR. However in order further improve the user experience, the optimal channel selections, improved security through relay selection and encryption has also been included in the successive improved models of SMOR. Yet the major focus has always been on the energy reduction for spectrum allocation, for which this paper presents a novel energy efficient optimal resource allocation scheme for SMOR. The proposed model utilizes the energy efficiency concept along with the features of SMOR, M-SMOR, SDS-M-SMOR, OCJ-SMOR, and SCJB-M-SMOR. It converts the resource allocation problem as non-convex optimization and employed the vertex search optimization to resolve it. This model is named as VS-M-SMOR and has been explained and evaluated in the following sections. The remainder of this article is organized as: Section 2 describes some of the related research works. Section 3 presents the system model and section 4 explains the proposed methodology. It is evaluated and the results are provided in section 5 while section 6 makes a conclusion about this research model.

2. Related Works. In [20], Qian et al have proposed a power control mechanism to maximize the energy efficiency of the secondary users along with a guarantee of the QoS parameters. The feasibility condition of the power consumption problem is derived and both centralized and distributed solutions are provided. This approach improved the energy conservation considerably for

the spectrum allocation; however, it considers only one cognitive radio which doubts its efficiency of power control for multiple cognitive radios.

Gao et al [21] proposed a framework distributed energy efficient spectrum access for CRAHNs. A multidimensional constrained optimization problem is formulated by minimizing the energy consumption per bit over the entire available subcarrier set for each individual user while satisfying its QoS constraints and power limit. Then a two-step solution is proposed by decoupling it into the unconstrained problem. However, in this model, the lack of consideration for detection errors degrades the performance. Sanchez et al [22] proposed two strategies Rate-Efficient Power Control to maximize the secondary capacity, and Energy-Efficient Power Control to minimize the secondary energy consumption. These strategies adjust the secondary user transmit power with the current transmission probability for improved performance.

Ngo et al [23] proposed two distributed resource allocation mechanisms with the spectrum sharing constraints. This design formulation aims at optimizing the energy efficiency of the power allocation strategy. The devised schemes also take into account the issue of controlling the shares of spectral holes by enforcing lower and upper limits on the number of sub-channels that individual SUs may occupy. This method improves the energy efficient spectrum allocation to the secondary users. Ding et al [24] also proposed a distributed resource allocation scheme with higher energy efficiency using decode-and-forward (DF) and amplify-and-forward (AF), based on convex optimization and arithmetic-geometric mean approximation techniques. This approach utilized a practical medium access control protocol for dynamic spectrum allocation. It maximizes the network throughput through local controls, but the major issue with this technique is the lack of congestion control model.

In [25], the authors proposed a cooperative transmission method for the energy efficient spectrum allocation process. This method uses a heuristic algorithm to solve the resource allocation problem based on the utility-spectrum ratio. The results are provided in a satisfying way however the relay selection problem has not been resolved considerably. In [26], the authors developed a provably convergent distributed algorithm that yields a locally optimal solution for the spectrum allocation problem. An alternative centralized algorithm was also developed for network duality and power control. These approaches provide efficient power control, however, the interference alignment is not precluded. In [27], the authors proposed a cross-layer opportunistic spectrum access and dynamic routing algorithm called ROSA (ROuting and Spectrum Allocation algorithm). ROSA dynamically allocates spectrum resources to maximize the capacity of links without generating harmful interference to other users. However,

the problem with the techniques in literature is that they do not consider the power requirements of the secondary users in a dynamic manner.

3. System Model. Consider a downlink CRAHN comprised of a network of PUs, and a network of SUs with single transmitter Tx and K receivers Rx. The network is presumed that the wireless policies are cognitive and proficient of sensing the environment and adjusting their parameters. A sample of the system model is shown in Figure 1.

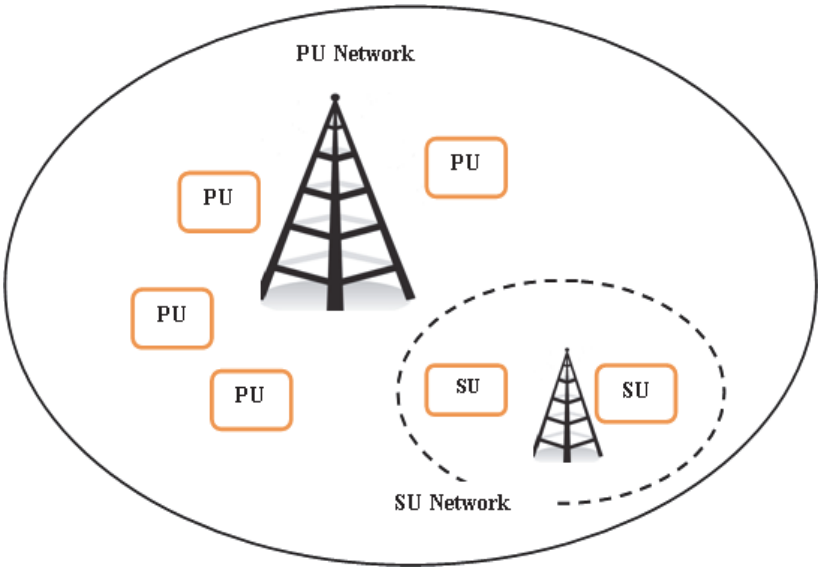


Fig. 1. System Model

In this paper, M PU nodes and N SU nodes with Q orthogonal channels are considered. The transmission energy at i-th PU is denoted by p_i^P and the transmission power at the *j-th* SU is specified by p_j^S while it is presumed that the SU transmitter/receiver pairs are inside the communication range of each other. The communication among the SU or PU pairs can practice intrusion from transmissions stemming from other PUs or SUs that are consuming the same channel and are in the sensing range of the receivers. Given the channel and transmission power as the network resources, the aim is to allot these resources such that the intrusion that SUs origin to the PUs will be decreased and SUs be capable to interconnect to each other. The aim of the PU network is to fulfill its QoS necessities with the minimum energy consumption while

the SU nodes should evade accumulation intrusion to the PU nodes while trying to find a spectrum hole for their own communications.

4. Vertex Search Optimization Based Modified SMOR Routing Model. In the proposed method, the resource allocation problem is modelled into a non-convex optimization problem. It is resolved using vertex search optimization with gradient approximations. The channel between the Tx and Rx is given as h_k and the channel estimation error is given as e_k [28]. At the beginning of each transmission block, Tx receives training signals from Rx for estimating channel states over a training interval T_t. Based on the received signals, the channel with minimum mean square error is estimated and denoted as $\overline{\overline{h}}_k$. With $\overline{\overline{h}}_k$, the achievable sum rate is given as

$$R\left(T_t, s, p_d\right) = \sum_{k=1}^{K} s_k \frac{T - KT_t}{T} \log\left(1 + \frac{p_d \mid \overline{\overline{h}}_k \mid^2}{\sigma_n^2 + p_d \sigma_e^2}\right). \quad (1)$$

Here p_d is the channel training transmission power and p_T is the data transmission power, s is the scheduling status and r is the relay status. The term $\frac{T - KT_t}{T}$ illustrates the loss of time for data transmission caused by the time T_t for channel training σ_n^2 is the additive noise power and $p_d \sigma_e^2$ is the self-interference power.

The net energy dissipation is given as

$$E\left(T_t, s, p_d\right) = P_S + \frac{KT_t}{T} p_T + \frac{T - KT_t}{T} p_d - EH\left(T_t, s, p_d\right). \quad (2)$$

Here P_S is the constant energy consumption in circuits, $\frac{KT_t}{T} p_T$ is the power consumed for estimating the channels, $\frac{T - KT_t}{T} p_d$ is the power consumed for estimating the transmitting data and $EH\left(T_t, s, p_d\right)$ is the harvested energy at the Rx.

The energy efficiency of the entire CRAHN is the achievable rate per unit energy which is represented as

$$E_E = \frac{R(T_t, s, p_d)}{E(T_t, s, p_d)} \tag{3}$$

The optimal resource allocation can be achieved by modelling the E_E into optimization problem as follows:

$$\max_{T_t, s, p_d} \frac{R(T_t, s, p_d)}{E(T_t, s, p_d)} \tag{4}$$

Such that $s_k r_k \geq s_k r_{min}$ $r_k \geq s_k r_{min}$ and $0 \leq p_d \leq p_{max}$. These constraints ensure the scheduled data rate as minimum, amount of energy harvested is larger than energy dissipated and ensures that only one Rx schedules for one ID. Thus effective non-convex optimization problem is formulated.

In order to resolve this problem, vertex search (VS) with gradient approximation is introduced. The VS algorithm is a reasonably new global optimization method initially proposed by Berat Dogan and Tamer O lmez [29]. It is an effective meta-heuristic technique which gives a decent harmony between the exploration and exploitation. But VS is easy to trap into the local optimum and neglects to discover global optimum. Therefore, it can't generally manage the optimization problem effectively. To control the individuals all the more proficiently moving towards to the feasible region, the pre-assessed method can be utilized to recognize the obscure area for conceivable moves. Subsequently, an enhanced VS utilizing gradient-based approximation is proposed. The gradient-based approximation is specifically following up on focuses. In the wake of setting a point as the inside, the gradient course of this point is ascertained and an arbitrary search in the negative gradient bearing of the fact of the matter is finished. In the event that a superior point is found in this procedure, at that point, the inside will be refreshed. VS-G utilizes gradient matrix to decide the heading of search and the gradient-based approximation is utilized as an indicator to discover the route towards the feasible region.

In the two-dimensional optimization problem, the initial solution is computed as

$$u_0 = \frac{upper\,limit + lower\,limit}{2}. \tag{5}$$

Here upper limit and lower limit are dx1 vectors that define the bound constraints of the problem in d-dimension space. Then a number of neighbour solutions are randomly generated using Gaussian distribution as

$$p(z|u,\Sigma) = \frac{1}{\sqrt{(2\pi)^d\ |\Sigma|}} exp\left\{-\frac{1}{2}(z-u)^T \sum (z-u)\right\} \quad (6)$$

Here d represents the dimension, z is the $d \times 1$ vector of a random variable, u denotes $(d \times 1)$ vector of the sample mean and Σ is covariance matrix. Then the solutions that are beyond the boundary limits are neglected and a bounded solution is chosen. If the selected minimum solution is better than the best solution found so far, then this solution is allocated to be the new best solution. Likewise, a number of candidate solutions are randomly generated along the negative gradient direction in the specified d-dimension as:

$$A = C - (D.L). \quad (7)$$

Here A is the $d \times 1$ vector which represents a candidate solution, C is the center solution so far, D is the gradient direction computed by finite difference method [30], L is the step length calculated as $L = rand.(upper\,limit - lower\,limit)$. In this method, the search can be adjusted using

$$a_t = a_0 - \frac{t}{MaxItr}. \quad (8)$$

Here a denotes sampling step values [0, 1]; t is the time and $MaxItr$ is the maximum iterations.

Based on this approach, the non-convex optimization problem is resolved as

$$\max_{s,p_d} f(T_t, s, p_d) = R(T_t, s, p_d) - x[E(T_t, s, p_d) + E_{vs}]. \quad (9)$$

Here x is the maximum energy efficiency and $f(.)$ is the objective function while E_{vs} denotes the energy consumed for performing the VS optimization algorithm. Thus the optimal resource allocation solution is achieved. The overall process of the proposed methodology is given in the following algorithm.

Begin

For $T_t = 1:T$

Initialize x, s, p_d, ε

Repeat

Set x= $\frac{R(T_t, s, p_d)}{E(T_t, s, p_d)}$

End for

Repeat

Update s, p_d

Update K until Convergence of s, p_d

Use VS-G

Initialize solutions u_0

t=0

Repeat

Generate candidate solutions A

Generate C

Select Ci

Adjust search using a_t

Repeat

For Itr=Itr+1

Select center solution u_j

For solution u_j

If $u_i > u_j$

Replace solution u_j

Else

Keep center solution u_j

End for

End for

$\max_{s, p_d} f(T_t, s, p_d)$

End

Listing 1. Algorithm: VS-M-SMOR

The above algorithm clearly explains the overall procedure of VS-M-SMOR. Initially, the parameters are set for VS followed by the limit setting. Then the search operation is performed and the newer results are updated continuously. The search solutions are generated based on the

energy efficient resource allocation is approximated. Finally, the center solution is obtained after many iterations and it is dynamically updated.

5. Performance Evaluation. The performance of the proposed energy efficient optimal resource allocation based routing model of VS-M-SMOR is evaluated using MATLAB. The simulation environment is set as in [16] and the comparisons are made vice versa for large and small scale CRAHNs. As the proposed model is used for both the scalable networks, the model is utilized as VS-M-SMOR-1 and VS-M-SMOR-2 as in [16, 17]. The performance of these models is compared with their corresponding SMOR, M-SMOR, SDS-M-SMOR, OCS-M-SMOR [31] and SCJB-M-SMOR which are the modified models of original SMOR. The evaluations are made in terms of end-to-end delay (EED), Bit Error Rate (BER), Throughput, path loss ratio, Peak Signal-to-Noise Ratio (PSNR) and Mean Square Error (MSE), transmitted power and power consumption.

Figure 2 shows the EED comparison of regular CRAHN while Figure 3 shows EED of large-scale CRAHN comparing SMOR, M-SMOR, SDS-SMOR, OCS-M-SMOR, SCJB-M-SMOR and the proposed VS-M-SMOR. VS-M-SMOR shows a lower delay in all level of the offered load with an average of 16% reduced delay than other models because of the combination of optimal relay selection and optimal resource allocation. The other models show comparatively higher delay due to the limited spectrum availability and inability to allocate resources with better energy efficiency.

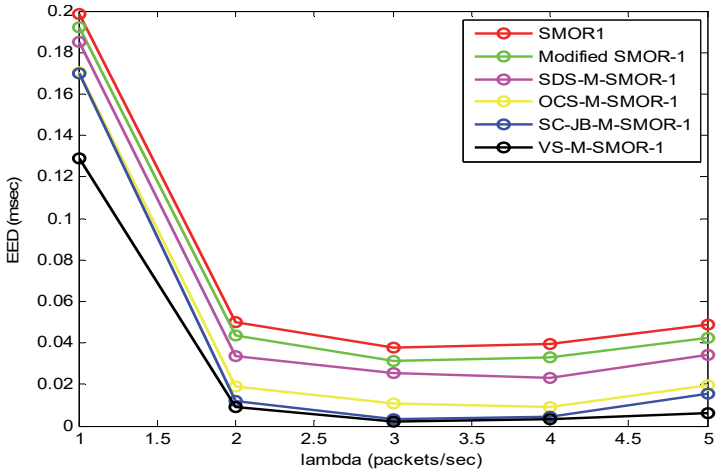


Fig. 2. End to end delay of Regular CRAHN

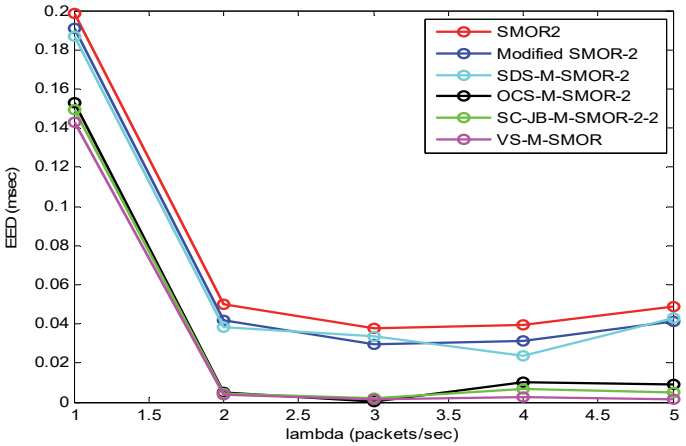


Fig. 3. End to end delay of large-scale CRAHN

Figure 4 shows the BER of regular CRAHN while Figure 5 shows BER of large-scale CRAHN comparing SMOR, M-SMOR, SDS-SMOR, OCS-M-SMOR, SCJB-M-SMOR and the proposed VS-M-SMOR. VS-M-SMOR shows lower error rate with 9 to 11% decrease on average while other models have comparatively higher BER. This can be attributed to the efficient spectrum allocation in the proposed model along with the better selection of the channel and relay for improved opportunistic routing.

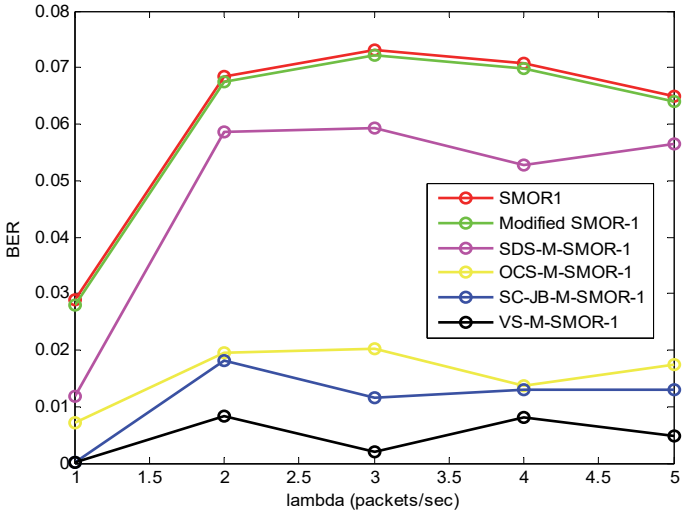


Fig. 4. BER for Regular CRAHN

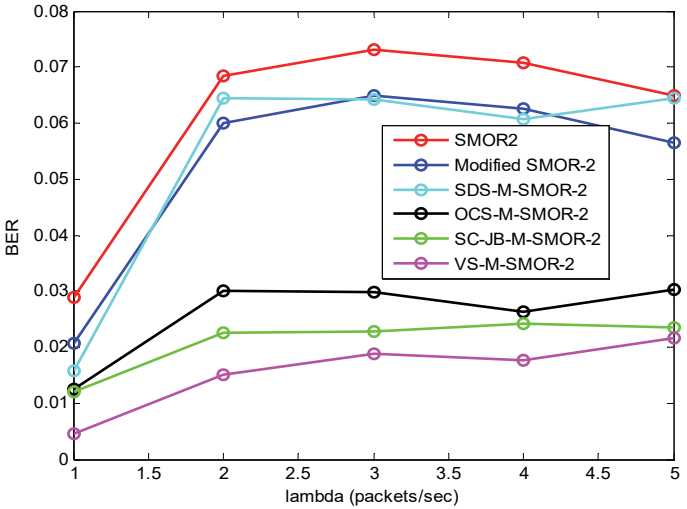


Fig. 5. BER for large-scale CRAHN

Figure 6 shows the throughput of regular CRAHN while Figure 7 shows throughput of large-scale CRAHN comparing SMOR, M-SMOR, SDS-SMOR, OCS-M-SMOR, SCJB-M-SMOR and the proposed VS-M-SMOR.

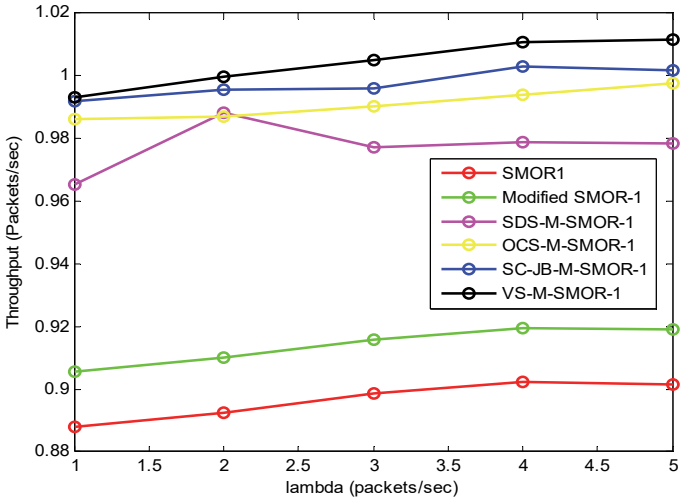


Fig. 6. Throughput for Regular CRAHN

It is seen that the VS-M-SMOR-1 and VS-M-SMOR-2 have higher throughput values of 10% increase on average because of the ability to utilize the whole resources of the multi-channels with energy efficient spectrum

allocation to the users. The selection of optimal resource allocation solution largely influences this positive change in performance.

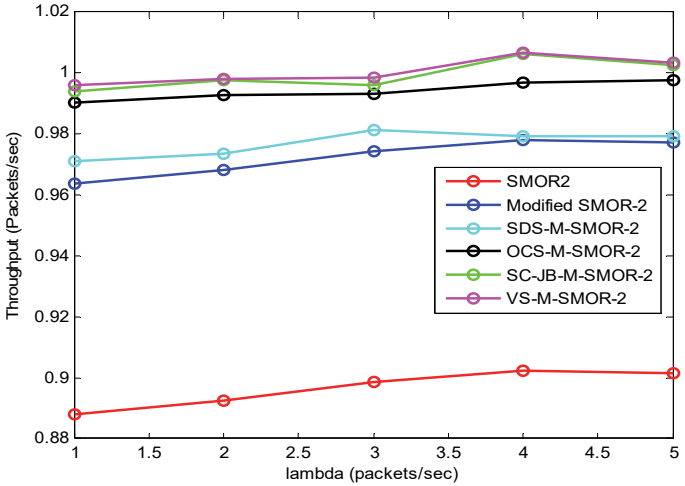


Fig. 7. Throughput for Large-scale CRAHN

Figure 8 shows the path loss ratio of regular CRAHN while Figure 9 shows path loss ratio of large-scale CRAHN comparing SMOR, M-SMOR, SDS-SMOR, OCS-M-SMOR, SCJB-M-SMOR and the proposed VS-M-SMOR. From this evaluation, it is proved that the proposed model of VS-M-SMOR is significantly efficient than the existing models in both the cases.

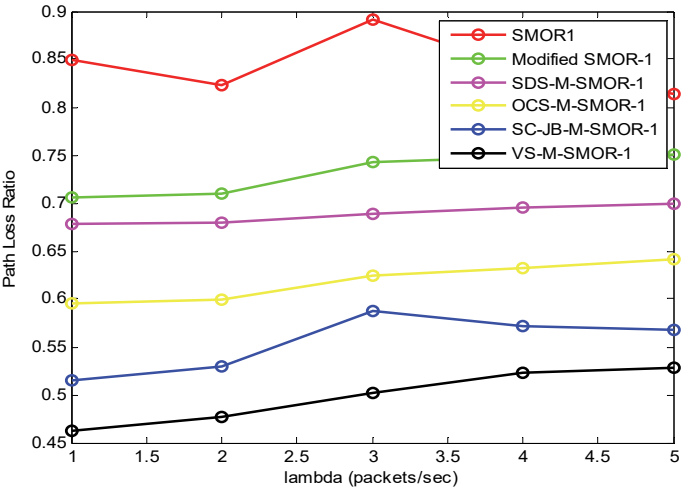


Fig. 8. Path Loss Ratio for Regular CRAHN

The optimal spectrum allocation reduces the loss with an average of 15% decrease in path loss ratio while the relay selection improves the packet delivery. Thus the proposed model satisfies all the QoS requirements for effective routing.

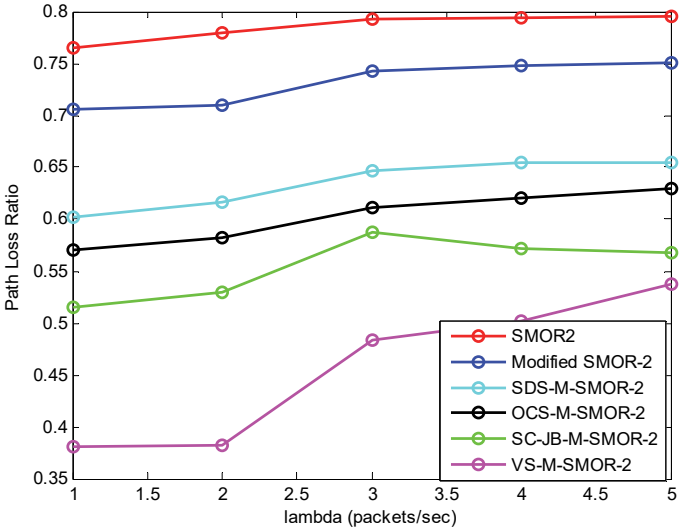


Fig. 9. Path Loss Ratio for Large-scale CRAHN

Figure 10 shows the PSNR of regular CRAHN while Figure 11 shows PSNR of large-scale CRAHN comparing SMOR, M-SMOR, SDS-SMOR, OCS-M-SMOR, SCJB-M-SMOR and the proposed VS-M-SMOR.

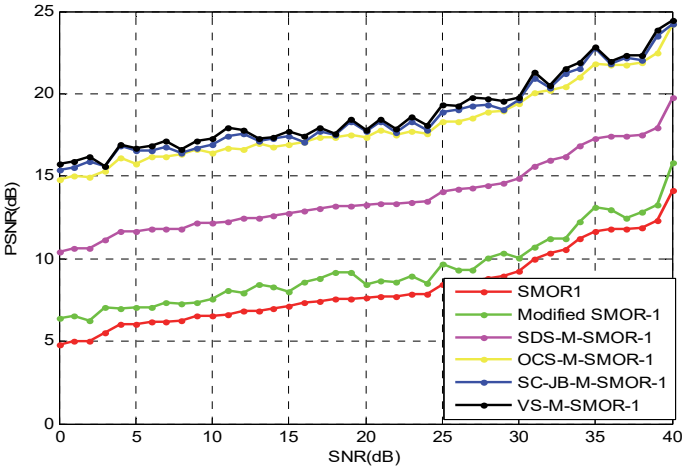


Fig. 10. PSNR for regular CRAHN

VS-M-SMOR has higher PSNR ratio on average 4% increase due to the fact the error rate is reduced considerably by selecting an efficient channel and better spectrum allocation for secondary users in opportunistic routing.

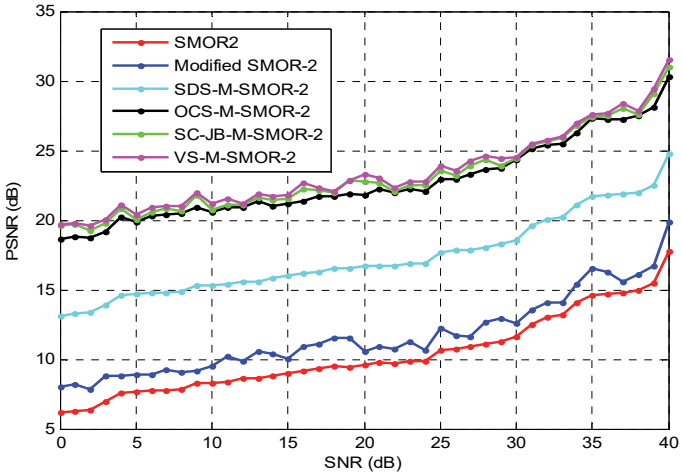


Fig. 11. PSNR for Large-scale CRAHN

The same can be applied to MSE comparison in Figure 12 and 13. VS-M-SMOR has an average of 12% decreased MSE than other models. It notably has better performance and can be stated as the best version of modified SMOR model.

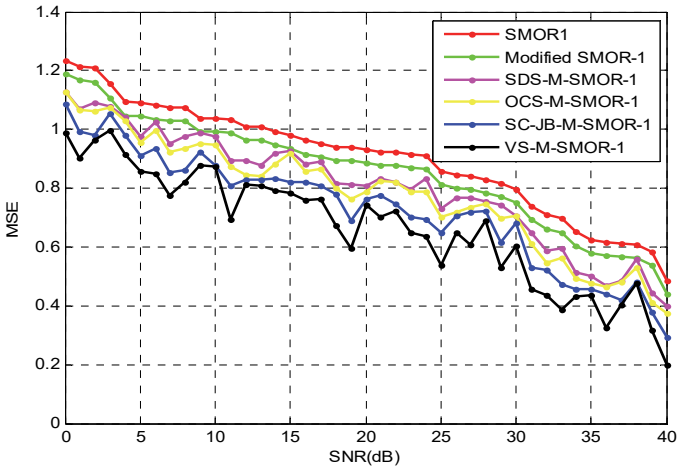


Fig. 12. MSE for regular CRAHN

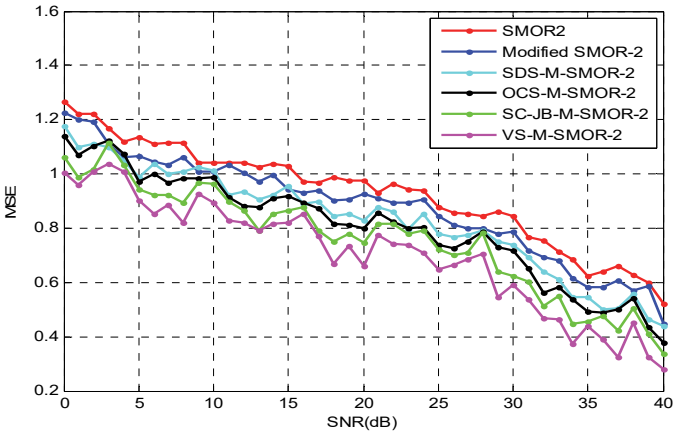


Fig. 13. MSE for Large-scale CRAHN

Figure 14 shows the transmitted power comparison of the proposed VS-M-SMOR and existing SMOR and Modified SMOR models. It can be seen that the proposed model has considerable 6% of lower transmission power on average than the SMOR model. Though at times the Modified SMOR has less power than VS-M-SMOR it can be attributed that the proposed model also improves the security of routing compared to M-SMOR as it includes all the best features of SDS-SMOR, OCS-M-SMOR, and SCJB-M-SMOR.

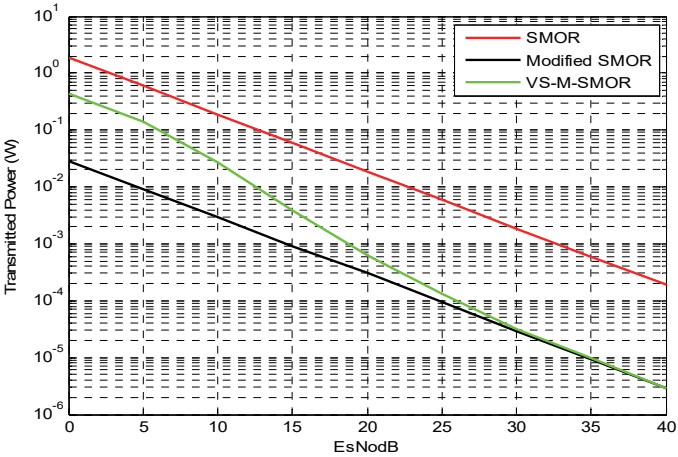


Fig. 14. Transmitted power

Similar reasons can be attributed for the improvement in power conservation of the routing models using the proposed approach. Figure 15

shows the power consumption comparison of the proposed VS-M-SMOR and existing SMOR and Modified SMOR models. The proposed model has 7% of less overall power consumption on average than the other models due to the optimal spectrum allocation.

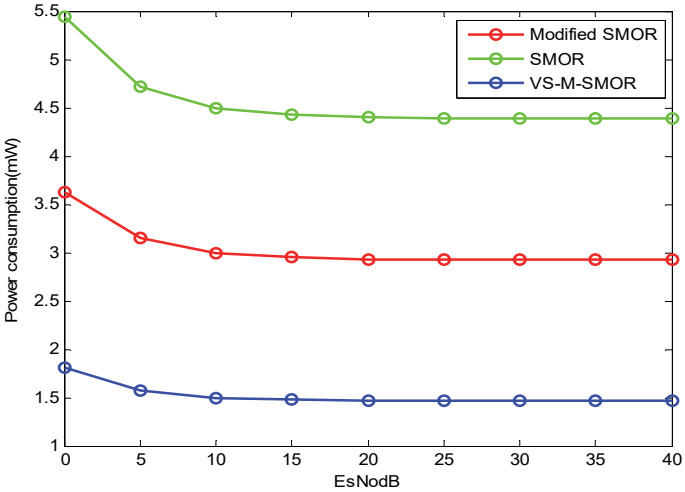


Fig. 15. Overall Power consumption

Hence from the performance comparison results, it can be verified that the proposed VS-M-SMOR model has better performance efficiency compared to the other models of SMOR.

6. Conclusion. Many improved versions of SMOR routing models have been developed in previous researches for the purpose of reducing delay in routing, reducing packet loss in transmission, improving optimal channel selection, improving security and secrecy. This paper focused on developing a modified SMOR model that includes all the above features and also enhances the energy efficiency in the resource allocation optimally. For this purpose, a novel routing strategy is introduced by using Vertex search with gradient approximation for the optimal resource allocation problem. Thus developed VS-M-SMOR model has been explained and has been evaluated for performance comparison. The results indicate that the proposed model outperforms the SMOR and other modified SMOR models with better performance results. VS-M-SMOR has an average of 16% less delay, 9-11% less error rate, 10% higher throughput, 15% lesser path loss, 4% higher PSNR, 12% lesser MSE, and 6-7% lesser power consumption with higher security. In future, more advanced techniques for security and optimal resource allocation can be utilized for improving the performance of opportunistic routing in CRAHNs.

References

1. Yu F.R., Zhang X., Leung V.C. Green communications and networking. CRC Press. 2012. 399 p.
2. Chen Y, Zhang S, Xu S, Li G.Y. Fundamental trade-offs on green wireless networks. *IEEE Communications Magazine*. 2011. vol. 49. no. 6. pp. 30–37.
3. Gür G., Alagöz F. Green wireless communications via cognitive dimension: an overview. *IEEE network*. 2011. vol. 25. no. 2. pp. 50–56.
4. Yu F.R., Huang M., Tang H. Biologically inspired consensus-based spectrum sensing in mobile ad hoc networks with cognitive radios. *IEEE network*. 2010. vol. 24. no. 3. pp. 26–30.
5. Si P., Ji H., Yu F.R., Leung V.C. Optimal cooperative internetwork spectrum sharing for cognitive radio systems with spectrum pooling. *IEEE Transactions on Vehicular Technology*. 2010. vol. 59. no. 4. pp. 1760–1768.
6. Yu F.R., Sun B., Krishnamurthy V., Ali S. Application layer QoS optimization for multimedia transmission over cognitive radio networks. *Wireless Networks*. 2011. vol. 17. no. 2. pp. 371–383.
7. Al-Ariki H.D., Swamy M.S. A survey and analysis of multipath routing protocols in wireless multimedia sensor networks. *Wireless Networks*. 2017. vol. 23. no. 6. pp. 1823–1835.
8. Anikanov G.A., Konovalchik P.M., Morgunov V.M., Ovcharov V.A. [The Multi-Controlled Media Access to Wireless Data Networks]. *Trudy SPIIRAS – SPIIRAS Proceedings*. 2015. vol. 38. pp. 246–268. (In Russ.).
9. Khludova M. Resource Allocation Policies for Smart Energy Efficiency in Data Centers. Proceedings of International Conference on Next Generation Wired/Wireless Networking. 2014. pp. 16–28.
10. Zhao Q., Swami A. A survey of dynamic spectrum access: Signal processing and networking perspectives. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2007. vol. 4. pp. IV-1349-IV-1352.
11. Abdullah H.M.A., Kumar A.V.S. A Survey on Spectrum-Map Based on Normal Opportunistic Routing Methods for Cognitive Radio Ad Hoc Networks. *International Journal of Advanced Networking and Applications*. 2015. vol. 7. no. 3. pp. 2761–2770.
12. Clancy T.C. Formalizing the interference temperature model. *Wireless Communications and Mobile Computing*. 2007. vol. 7. no. 9. pp. 1077–1086.
13. Xing Y. et. Dynamic spectrum access with QoS and interference temperature constraints. *IEEE Transactions on mobile computing*. 2007. vol. 6. no. 4. pp. 423–433.
14. Weiss T.A., Jondral F.K. Spectrum pooling: an innovative strategy for the enhancement of spectrum efficiency. *IEEE communications Magazine*. 2004. vol. 42. no. 3. pp. S8-14.
15. Berthold U., Jondral F.K., Brandes S., Schnell M. OFDM-based overlay systems: A promising approach for enhancing spectral efficiency [Topics in radio communications]. *IEEE Communications Magazine*. 2007. vol. 45. no. 12. pp. 52–58.
16. Lin S.C., Chen K.C. Spectrum-map-empowered opportunistic routing for cognitive radio ad hoc networks. *IEEE Transactions on Vehicular Technology*. 2014. vol. 63. no. 6. pp. 2848–2861.
17. Abdullah H.M.A., Kumar A.V.S. Modified SMOR Using Sparsity Aware Distributed Spectrum Map for Enhanced Opportunistic Routing in Cognitive Radio Adhoc Networks. *Journal of Advanced Research in Dynamical and Control Systems*. 2017. vol. 9. no. 6. pp. 184–196.
18. Abdullah H.M.A., Kumar A.V.S. A Hybrid Artificial Bee Colony Based Spectrum Opportunistic Routing Algorithm for Cognitive Radio Ad Hoc Networks. *International Journal of Scientific and Engineering Research*. 2016. vol. 7. no. 6. pp. 294–303.
19. Abdullah H.M.A., Kumar A.V.S. HB-SOR: Hybrid Bat Spectrum Map Empowered Opportunistic Routing and Energy Reduction for Cognitive Radio Ad Hoc Networks (CRAHNs). *International Journal of Scientific and Research Publications (IJSRP)*. 2017. vol. 7. no. 5. pp. 284–297.

20. Qian L., Li X., Attia J., Gajic Z. Power control for cognitive radio ad hoc networks. Proceedings of 15th IEEE Workshop on Local & Metropolitan Area Networks (LANMAN 2007). 2007. pp. 7–12.
21. Gao S., Qian L., Vaman D.R. Distributed energy efficient spectrum access in cognitive radio wireless ad hoc networks. *IEEE Transactions on Wireless Communications*. 2009. vol. 8. no. 10. pp. 5202–5213.
22. Sanchez S.M., Souza R.D., Fernandez E.M., Reguera V.A. Rate and energy efficient power control in a cognitive radio ad hoc network. *IEEE signal processing letters*. 2013. vol. 20. no. 5. pp. 451–454.
23. Ngo D.T., Le-Ngoc T. Distributed resource allocation for cognitive radio networks with spectrum-sharing constraints. *IEEE Transactions on Vehicular Technology*. 2011. vol. 60. no. 7. pp. 3436–3449.
24. Ding L., Melodia T., Batalama S.N., Matyjas J.D. Distributed resource allocation in cognitive and cooperative ad hoc networks through joint routing, relay selection and spectrum allocation. *Computer Networks*. 2015. vol. 83. pp. 315–331.
25. Jia J., Zhang S. Cooperative transmission in cognitive radio ad hoc networks. *International Journal of Distributed Sensor Networks*. 2012. vol. 8. no. 5. pp. 863634.
26. Kim S.J., Giannakis G.B. Optimal resource allocation for MIMO ad hoc cognitive radio networks. *IEEE Transactions on Information Theory*. 2011. vol. 57. no. 5. pp. 3117–3131.
27. Ding L. et al. Cross-layer routing and dynamic spectrum allocation in cognitive radio ad hoc networks. *IEEE Transactions on Vehicular Technology*. 2010. vol. 59. no. 4. pp. 1969–1979.
28. Lee K., Hong J.P. Energy-efficient resource allocation for simultaneous information and energy transfer with imperfect channel estimation. *IEEE Transactions on Vehicular Technology*. 2016. vol. 65. no. 4. pp. 2775–2780.
29. Doğan B., Ölmez T. A new metaheuristic for numerical function optimization: Vortex Search algorithm. *Information Sciences*. 2015. vol. 293. pp. 125–145.
30. Liszka T., Orkisz J. The finite difference method at arbitrary irregular grids and its application in applied mechanics. *Computers & Structures*. 1980. vol. 11. no. 1-2. pp. 83–95.
31. Abdullah H.M.A., Kumar A.V.S. Proficient opportunistic routing by queuing based optimal channel selection for the primary users in CRAHN. *ARPN Journal of Engineering and Applied Sciences.* 2018. vol. 13. no. 5. pp.1649–1657.

Abdullah Hesham Mohammed Ali — Ph.D. student of Hindusthan College of Arts and science, Bharathiar University. Research interests: Cognitive Radio Ad Hoc Networks, Network Security, Wireless Network. The number of publications — 10. heshammohammedali@gmail.com; Hindusthan Gardens, Behind Nava India, Avanashi Road, Peelamedu, Uppilipalayam Coimbatore, Tamil Nadu, 641028, India; office phone: +919585430114.

Kumar A.V. Senthil — Ph.D., professor, professor of PG & Research Department of Computer Applications of Hindusthan College of Arts and science, Bharathiar University, director of PG & Research Department of Computer Applications of Hindusthan College of Arts and science, Bharathiar University. Research interests: Data Mining, Fuzzy Expert Systems, Networks, Software Engineering, Information Systems (Business Informatics), Artificial Intelligence.. The number of publications — 100. avsenthilkumar@yahoo.com; Hindusthan Gardens, Behind Nava India, Avanashi Road, Peelamedu, Uppilipalayam Coimbatore, Tamil Nadu, 641028, India; office phone: +9109843013009.

УДК 621.39 **DOI** 10.15622/sp.57.1

Х.М.А. Абдулла, А.В.С. Кумар

ЭНЕРГОЭФФЕКТИВНОЕ ОПТИМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ РЕСУРСОВ В КОГНИТИВНЫХ РАДИО- AD-HOC-СЕТЯХ НА ОСНОВЕ ВЕРШИННОГО ПОИСКА

Абдулла Х.М.А, Кумар А.В.С. **Энергоэффективное оптимальное распределение ресурсов в когнитивных радио- ad-hoc-сетях на основе вершинного поиска.**

Аннотация. Когнитивная радио-ad-hoc-сеть (CRAHN) — это безынфраструктурная сетевая модель когнитивного радио, разработанная для ситуативного применения. Регулирование распределения ресурсов в CRAHN может быть рассмотрено как проблема ограничения энергии. Эффективному распределению спектра с использованием различных протоколов посвящено множество исследований. В этой работе модель Spectrum-Map-Empowered Opportunistic Routing (SMOR) была использована в качестве фундаментальной модели маршрутизации данны. Представлено решение по энергоэффективному оптимальному распределению спектра. Улучшена ранее модифицированная модель SMOR для энергоэффективного и оптимального распределения ресурсов с использованием алгоритма вершинного поиска с аппроксимацией на основе градиента. Изначально проблема распределения ресурсов была смоделирована как проблема невыпуклой оптимизации. Распределение мощности, адаптация скорости передачи данных, распределение каналов и политика пользовательского планирования оптимизированы для максимизации энергоэффективности во время передачи данных. Предлагаемый алгоритм вершинного поиска решает проблему оптимизации путем определения интервала обучения для определения канала и распределения энергии. Экспериментальные результаты подтверждают, что предлагаемая модифицированная модель SMOR(VS-M-SMOR), основанная на вершинном поиске, обеспечивает оптимальное распределение ресурсов.

Ключевые слова: распределение ресурсов, распределение спектра, SMOR, вершинный поиск, энергоэффективность, пользовательское планирование.

Абдулла Хишам Мохамед Али — аспирант колледжа искусств и науки Индостана, Университет им. С. Бхарати. Область научных интересов: беспроводная ad-hoc-сеть, сетевая безопасность, беспроводная сеть, IoT, облачные вычисления. Число научных публикаций — 10. heshammohammedali@gmail.com; Индостан Гарденс (за Нава Индиа), Авинаши роуд, Пиламеду, Аппилипалайям, Коимбатур, Тамилнад, 641033, Индия; р.т.: +919585430114.

Кумар А.В. Сентиль — д-р техн. наук, профессор, профессор аспирантуры и научно-исследовательского отдела прикладной информатики колледжа искусств и науки Индостана, Университет им. С. Бхарати, руководитель аспирантуры и научно-исследовательского отдела прикладной информатики колледжа искусств и науки Индостана, Университет им. С. Бхарати. Область научных интересов: интеллектуальный анализ данных, нечеткие экспертные системы, сети, разработка программного обеспечения, информационные системы (бизнес-информатика), искусственный интеллект. Число научных публикаций — 100. avsenthilkumar@yahoo.com; Индостан Гарденс (за Нава Индиа), Авинаши роуд, Пиламеду, Аппилипалайям, Коимбатур, Тамилнад, 641033, Индия; р.т.: +9109843013009.

Литература

1. *Yu F.R., Zhang X., Leung V.C.* Green communications and networking // CRC Press. 2012. 399 p.
2. *Chen Y, Zhang S, Xu S, Li G.Y.* Fundamental trade-offs on green wireless networks // IEEE Communications Magazine. 2011. vol. 49. no. 6. pp. 30–37.
3. *Gür G., Alagöz F.* Green wireless communications via cognitive dimension: an overview // IEEE network. 2011. vol. 25. no. 2. pp. 50–56.
4. *Yu F.R., Huang M., Tang H.* Biologically inspired consensus-based spectrum sensing in mobile ad hoc networks with cognitive radios // *IEEE network*. 2010. vol. 24. no. 3. pp. 26–30.
5. *Si P., Ji H., Yu F.R., Leung V.C.* Optimal cooperative internetwork spectrum sharing for cognitive radio systems with spectrum pooling // IEEE Transactions on Vehicular Technology. 2010. vol. 59. no. 4. pp. 1760–1768.
6. *Yu F.R., Sun B., Krishnamurthy V., Ali S.* Application layer QoS optimization for multimedia transmission over cognitive radio networks // Wireless Networks. 2011. vol. 17. no. 2. pp. 371–383.
7. *Al-Ariki H.D., Swamy M.S.* A survey and analysis of multipath routing protocols in wireless multimedia sensor networks // Wireless Networks. 2017. vol. 23. no. 6. pp. 1823–1835.
8. *Аниканов Г.А., Коновальчик П.М., Моргунов В.М., Овчаров В.А.* Контролируемый многомодельный доступ к среде беспроводных сетей передачи данных // Труды СПИИРАН. 2015. Вып. 1(38). С. 246–268.
9. *Khludova M.* Resource Allocation Policies for Smart Energy Efficiency in Data Centers // Proceedings of International Conference on Next Generation Wired/Wireless Networking. 2014. pp. 16–28.
10. *Zhao Q., Swami A.* A survey of dynamic spectrum access: Signal processing and networking perspectives // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007). vol. 4. pp. IV-1349-IV-1352.
11. *Abdullah H.M.A., Kumar A.V.S.* A Survey on Spectrum-Map Based on Normal Opportunistic Routing Methods for Cognitive Radio Ad Hoc Networks // International Journal of Advanced Networking and Applications. 2015. vol. 7. no. 3. pp. 2761–2770.
12. *Clancy T.C.* Formalizing the interference temperature model // Wireless Communications and Mobile Computing. 2007. vol. 7. no. 9. pp. 1077–1086.
13. *Xing Y. et al.* Dynamic spectrum access with QoS and interference temperature constraints // IEEE Transactions on mobile computing. 2007. vol. 6. no. 4. pp. 423–433.
14. *Weiss T.A., Jondral F.K.* Spectrum pooling: an innovative strategy for the enhancement of spectrum efficiency // IEEE communications Magazine. 2004. vol. 42. no. 3. pp. S8-14.
15. *Berthold U., Jondral F.K., Brandes S., Schnell M.* OFDM-based overlay systems: A promising approach for enhancing spectral efficiency [Topics in radio communications] // IEEE Communications Magazine. 2007. vol. 45. no. 12. pp. 52–58.
16. *Lin S.C., Chen K.C.* Spectrum-map-empowered opportunistic routing for cognitive radio ad hoc networks // IEEE Transactions on Vehicular Technology. 2014. vol. 63. no. 6. pp. 2848–2861.
17. *Abdullah H.M.A., Kumar A.V.S.* Modified SMOR Using Sparsity Aware Distributed Spectrum Map for Enhanced Opportunistic Routing in Cognitive Radio Adhoc Networks // Journal of Advanced Research in Dynamical and Control Systems. 2017. vol. 9. no. 6. pp. 184–196.
18. *Abdullah H.M.A., Kumar A.V.S.* A Hybrid Artificial Bee Colony Based Spectrum Opportunistic Routing Algorithm for Cognitive Radio Ad Hoc Networks // International Journal of Scientific and Engineering Research. 2016. vol. 7. no. 6. pp. 294–303.
19. *Abdullah H.M.A., Kumar A.V.S.* HB-SOR: Hybrid Bat Spectrum Map Empowered Opportunistic Routing and Energy Reduction for Cognitive Radio Ad Hoc Networks

(CRAHNs) // International Journal of Scientific and Research Publications (IJSRP). 2017. vol. 7. no. 5. pp. 284–297.

20. *Qian L., Li X., Attia J., Gajic Z.* Power control for cognitive radio ad hoc networks // Proceedings of 15th IEEE Workshop on Local & Metropolitan Area Networks (LANMAN 2007). 2007. pp. 7–12.
21. *Gao S., Qian L., Vaman D.R.* Distributed energy efficient spectrum access in cognitive radio wireless ad hoc networks // IEEE Transactions on Wireless Communications. 2009. vol. 8. no. 10. pp. 5202–5213.
22. *Sanchez S.M., Souza R.D., Fernandez E.M., Reguera V.A.* Rate and energy efficient power control in a cognitive radio ad hoc network // IEEE signal processing letters. 2013. vol. 20. no. 5. pp. 451–454.
23. *Ngo D.T., Le-Ngoc T.* Distributed resource allocation for cognitive radio networks with spectrum-sharing constraints // IEEE Transactions on Vehicular Technology. 2011. vol. 60. no. 7. pp. 3436–3449.
24. *Ding L., Melodia T., Batalama S.N., Matyjas J.D.* Distributed resource allocation in cognitive and cooperative ad hoc networks through joint routing, relay selection and spectrum allocation // Computer Networks. 2015. vol. 83. pp. 315–331.
25. *Jia J., Zhang S.* Cooperative transmission in cognitive radio ad hoc networks // International Journal of Distributed Sensor Networks. 2012. vol. 8. no. 5. pp. 863634.
26. *Kim S.J., Giannakis G.B.* Optimal resource allocation for MIMO ad hoc cognitive radio networks // IEEE Transactions on Information Theory. 2011. vol. 57. no. 5. pp. 3117–3131.
27. *Ding L. et al.* Cross-layer routing and dynamic spectrum allocation in cognitive radio ad hoc networks // IEEE Transactions on Vehicular Technology. 2010. vol. 59. no. 4. pp. 1969–1979.
28. *Lee K., Hong J.P.* Energy-efficient resource allocation for simultaneous information and energy transfer with imperfect channel estimation // IEEE Transactions on Vehicular Technology. 2016. vol. 65. no. 4. pp. 2775–2780.
29. *Doğan B., Ölmez T.* A new metaheuristic for numerical function optimization: Vortex Search algorithm // Information Sciences. 2015. vol. 293. pp. 125–145.
30. *Liszka T., Orkisz J.* The finite difference method at arbitrary irregular grids and its application in applied mechanics // Computers & Structures. 1980. vol. 11. no. 1-2. pp. 83–95.
31. *Abdullah H.M.A., Kumar A.V.S.* Proficient opportunistic routing by queuing based optimal channel selection for the primary users in CRAHN // ARPN Journal of Engineering and Applied Sciences. 2018. vol. 13. no. 5. pp.1649–1657.

DOI 10.15622/sp.57.2

M.A.S. MOSLEH, G. RADHAMANI

A NOVEL FUZZY QOS BASED IMPROVED HONEY BEE BEHAVIOR ALGORITHM FOR EFFICIENT LOAD BALANCING IN CLOUD

Mosleh M.A.S., Radhamani G. **A Novel Fuzzy QOS Based Improved Honey Bee Behavior Algorithm For Efficient Load Balancing In Cloud.**

Abstract. Nature inspired algorithm based Load balancing of tasks on virtual machines (VMs) has become an area of greater research interest. Honey Bee Behavior Based Load Balancing (HBB-LB) was introduced to balance the load with a maximum throughput. This approach also balances the priorities of the tasks on the VM to minimize the waiting time of the tasks. However, HBB-LB considers only the VM load for balancing the load, which might not be sufficiently effective. This paper proposes an Improved Honey Bee Behavior Based Load Balancing (IHBB-LB), taking into consideration a few more QoS parameters of VM, such as service response time, availability, reliability, cost and throughput to enhance load balancing. Response time is vital in determining the instant activity of a VM while availability determines available resource and state of VM (idle or active) and Reliability determines the level of trust in a VM. Most importantly, Cost for utilizing a VM and Throughput (capability of VM) are also essential in determining the VM efficiency. But, the inclusion of multiple QoS parameters results in multi-objective optimization problem. As a number of QoS parameters are computed, the Fuzzification of the QoS values was performed through the generated fuzzy rules and multi-objective optimization problem was eliminated. The experiments were performed in terms of makespan, response time, degree of imbalance and the number of tasks migrated and results indicate that the IHBB-LB provides a better level of performance.

Keywords: Optimization, QoS parameters, Cloud Computing, Load Balancing, Fuzzification.

1. Introduction. Cloud computing is an internet-based approach processed using the shared resources on a cloud consisting of multiple computers interlinked together. Cloud includes the concepts of distributed and parallel computing to enable shared resources and applications. This approach focuses on the advantages of cost, flexibility and availability of the service users, which in turn increases the demand for the cloud services. The increase in demand increases technical issues such as high availability and scalability, in Service Oriented Architectures (SOA) and Internet of Services (IoS) applications, which can be overcome by the load balancing concept of allocating dynamic workload evenly across all the machines. In [1] developed adaptive cost based task scheduling (ACTS) for efficient task scheduling in cloud However, task scheduling alone cannot improve cloud computing service to the users; exploring several other factors like balanced load, effective VM migration, etc. also contribute to efficient cloud user experience.

Load balancing in cloud computing is essential to tackle the overload and under-load conditions of virtual machines (VM). Efficient load balancing can be achieved using optimization techniques. Whenever an under-load of VM

or overload of VM occurs, the tasks have to be loaded to achieve optimal machine utilization. The Stochastic Hill Climbing approach [2] selects the VM based on a random selection of the uphill move. Stochastic Hill Climbing maintains a VM status table and assigns a random id to each VM. When a task arrives, a VM is chosen randomly and a request is sent to check the status of task allocation, following which the task is assigned. The Simulated Annealing approach [3] obtains all the load balance requests and selectively drops some less important requests from default users in order to maintain the load balance of the SLA premium users. Genetic algorithm [4] allocates tasks by using natural selection strategy of crossover and mutation on the selected VM in order to determine the optimal VM for efficient load balancing. Ant colony optimization [5] follows the natural ant behavior of food detection to allocate the tasks by estimating the computation of the VMs. While honey bee based optimization algorithms are more common in cloud computing. Honey bee mating algorithm [6] can be used for effective and optimal cloud resource selection. The honey bee behavior based load balancing (HBB-LB) [7] selects the overloaded VM, moves the loads to unallocated VMs and then indicates the status of VM to other tasks so that new high priority tasks could be allocated to the unallocated VMs. However, it considers only one parameter (i.e.) the VM load conditions and this is not enough for effective load balancing. The performance of this honey bee behavior based load balancing techniques can be improved by considering the many QoS factors of VMs.

This paper proposes the Improved Honey Bee Behavior Based Load Balancing (IHBB-LB) based on the consideration of multiple QoS parameters, such as service response time, availability, reliability, cost and throughput. These parameters have been included in IHBB-LB for balancing the load when certain VMs are overloaded and the some under-loaded. Multiple QoS parameters result in the multi-objective optimization problem. In general, simple techniques such as crowding distance or weight estimation can be employed to find the pareto-optimal solutions for the multi-objective optimization problem. However, the use of more number of QoS parameters makes it difficult to utilize the general approaches. Hence, the concept of Fuzzification is employed for generating of fuzzy rules in order to resolve the multi-objective problem. Through this, it is also possible to effectively balance the load in the cloud environment.

The rest of the paper is organized as follows: Section II presents the review of related work. Section III explains the HBB-LB and the proposed IHBB-LB methodologies in detail. Section IV presents the performance evaluation of the proposed methodologies while section V concludes the paper.

2. Related Works. Load balancing is one of the most sought after research area in cloud computing. Many researchers have developed their own version of load balancing technique based on the issues they considered

as the prime focus. Analyzing some of the researches related to our research work can make a mile difference in the proposed solution. For example, Dual Directional Files Transfer Protocol (DDFTP) [8] developed for efficient balancing of the downloading servers without duplicating the same files is concept to watch out for. Though this concept seems alien to our concept, deep analysis shows that DDFTP provides efficient load balancing among the multiple heterogeneous data servers with a minimal overhead.

Load balancing using the Bayes theorem (LB-BC) [9] is another example of how well an heuristic clustering concept utilized for task allocation. Bayes theorem was included with clustering algorithm to provide optimal clustering set of physical hosts from which the efficient hosts are selected for balancing the tasks. Hybrid meta-heuristic algorithm for VM scheduling & load balancing [10] and load balancing model based on cloud partitioning [11] are some of the load balancing schemes focusing on reduced computation time. Cache (C^2) [12] for metadata server clusters implements the adaptive cache diffusion to detecting the overloaded nodes and process adaptive replication scheme to split the load in the overloaded nodes. Agent based load balancing [13] for the cloud data centers utilizes the migration heuristics to provide effective load balancing. Another autonomous agent based load balancing algorithm proposed [14] provides comparatively better balancing than other agent based algorithms. However agent based schemes are more commonly used so it becomes mandatory to analyze other options for better load balancing. Game theoretic static load balancing models [15] utilizes equilibrium game concepts either two-player or multi-player to efficiently balance the load with minimal server overload situations. Though these schemes try to resolve load balancing issues, they are static. Dynamic annexed balance method called Cloud load balancing (CLB) [16] considered both server processing power and computer loading to minimize overload and computation complexities. Although these single methods have ruled the load balancing concept, there are some hybrid load balancing concepts. For example DeMS [17] consisting of On-Demand scheduling, Querying and Migrating Task (QMT) and Staged Task Migration (STM) reduced overhead and balances the load effectively than single methods. Energy aware [18] and resource aware [19] schemes were also been largely employed for load balancing in cloud. Though the above described schemes provides better load balancing than their predecessors, in any application the nature inspired schemes have worked better. The same concept can be applied in cloud load balancing.

The foraging behavior of honey bees [20] was considered to develop of an intelligent optimization technique. The intelligent foraging behavior of the bee swarm focuses on the collection of honey for energy. Foraging begins only when suitable conditions such as temperature, weather, etc., are satisfied.

The bees maintains a thumb rule for foraging area which is be fixed around a hive for two miles, as the food obtained within this area will gain weight while the energy spent beyond this area will be greater than the energy obtained. Foraging at the extreme distances wears out the wings of the individual bees and reduces its life expectancy, thus reducing the efficiency of the bee colony. The foraging behavior depends on the communication of distance and direction of food source by round dance, waggle dance and shaking signals of the individual bees. Based on this intelligent foraging behavior, Artificial Bee Colony (ABC) algorithm [21] has been developed. ABC consists of three bees namely: employed bees, onlookers and scouts. Employed bees search for the food source and return to hive and dance on this area. The employed bee whose food source has been abandoned becomes a scout and starts scouting a new food source. Onlookers watch the dances of employed bees and choose food sources depending on the dances. ABC algorithm has the advantage of global search ability, which is achieved by the introduction of the neighborhood source production mechanism. ABC is used in many fields such as signal processing, image processing, scheduling problems and optimization problems. This approach has been employed for cloud load balancing as stated in HBB-LB [7] but as it considers only a single parameter, it cannot be considered efficiently balanced. Hence we intend to include additional load balancing parameters to enhance the load balancing performance of the HBB-LB in this paper.

3. Improved Honey Bee Behavior Based Load Balancing Algorithm. Honey bee behavior inspired load balancing (HBB-LB) algorithm has been employed for efficient balancing of the tasks. HBB-LB exhibits the basic honey bee foraging (wide search) behavior with respect to the tasks loaded in the VMs. However, HBB-LB algorithm considers only the load conditions in the selection of VM, while the other vital QoS factors which are necessary for enhancing the efficiency of computation in cloud are computing are not taken into consideration. In order to balance the available load to the VMs effectively, the VM characteristics are needed to be determined. The major considerations for the VM are makespan, processing time, capacity and load of a VM. While the HBB algorithm considers these parameters in the objective function for sorting the VMs, from the analysis it is found that some important parameters namely response time, reliability, availability, cost and throughput are equally vital in characterizing a VM.

Hence, an enhanced version of HBB-LB called as Improved Honey bee behavior based Load balancing algorithm (IHBB-LB) has been proposed which considers multiple QoS parameters in the selection of the VMs. The VMs are sorted in the ascending order based on the load conditions and QoS parameters such as service response time, availability, reliability, cost, and

throughput. The tasks removed from the overloaded VMs are considered as the honey bees. When the tasks are submitted to the under-loaded VM, and then would update the number of various priority tasks and the load of the respective VM to all other waiting tasks. Based on this input, those tasks with higher priority are allocated to the under-loaded VMs enabled with better QoS parameters based on the list sorted. As the VMs are arranged in an ascending order based on load conditions and QoS parameters, those tasks removed from overloaded VMs will be submitted only to the under-loaded VMs. The main advantage of this approach is the updating of the priorities and load conditions, which is similar to the dance movements of the scout bees. This proposed IHBB-LB approach was found to be very effective in load balancing of tasks in cloud computing environments. Figure 1 shows the load balancing procedure in cloud using the proposed IHBB-LB.

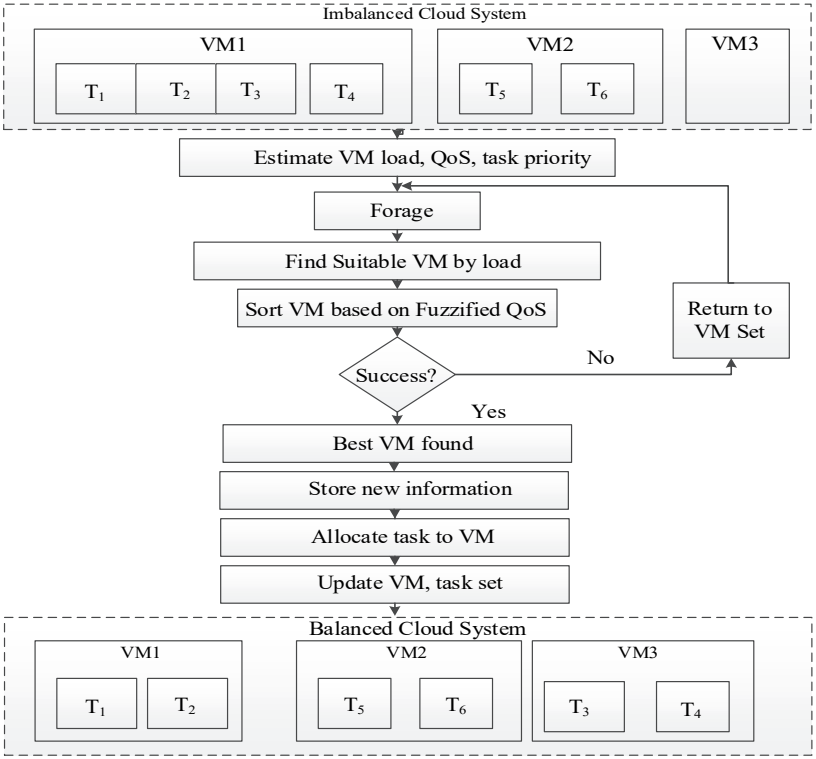


Fig. 1. Procedure of IHBB-LB

Let us analyze the load conditions and the QoS parameters to be considered for VM load balancing in order to clear the air about the importance

of each QoS metric in SLA agreements. Although many QoS aware load balancing concepts are commonly available, no method has achieved the efficiency of HBB algorithm. Hence improving the HBB algorithm, results in maximization of load balancing performance. It is worth mentioning, our proposed concept is suitable for both public and private clouds.

The proposed IHBB-LB is implemented in CloudSim for evaluation. The proposed system follows a data center model (server/client model) to process the cloud user requests. These requests are generally the set-up of different types of tasks. The server is responsible for managing the user requests and allocating the tasks to the VM. The main functions of the cloud server in this aspect are the selection of an efficient physical machine (PM) for VM instance, monitor the task execution in VM through VM monitor and send the computational results to the users. Different tasks have different execution patterns for which they require multiple types of execution resources. A single task can also be split into multiple steps and each allotted to different machines with different I/O bandwidth. Suppose there are n physical machines and a task requires R resources, then the complete set of execution resources are considered in evaluating the performance.

Let us assume the condition that all the PMs are now initiated into VM instances. A set of VMs are initialized as $V = \{V_1, V_2, \ldots . V_m\}$ where m is the total number of VMs. A set of tasks $T = \{T_1, T_2, \ldots . T_n\}$ is initialized where n is the total number of tasks. Each task is assumed to be accepted from the user requests and each task is constructed by multiple sub-tasks. All the sub-tasks are corresponding to the web services where they are data transmission tasks or data read/write through disk. Each task is executed on the respectively allotted VM within the allotted resources (CPU rate and network bandwidth).

The completion time of a task T_i on a virtual machine VM_j is denoted as $CT_{ij}, i = 1, 2, \ldots . n$ and $j = 1, 2, \ldots . m$. Based on CT_{ij}, the makespan is computed. Makespan is an indicator of the general throughput of the system, generally said as the completion time of a VM. Small values of makespan mean that the server is providing good and efficient planning of tasks to resources. It is different from the execution time which means the completion time of tasks in a VM. Makespan is denoted as M and it is given as

$$M = \max \left| CT_{ij} \right|; \qquad i \in T \ . \tag{1}$$

Makespan is generally higher than execution time as it includes all the hold time/ wait time of the VM. However it should be minimized for efficient VM which is done by minimizing the waiting time of a task T_i . The vital condition that should be satisfied in this scenario is that the processing

time must be less than or equal to the maximum time for completion $P_j \le CT_{\max}$. In order to evaluate this condition, P_j is computed as

$$P_j = \sum_{i=1}^{m} P_{ij}, \tag{2}$$

where P_{ij} is the processing time of a task T_i on VM_j.

However, at the time of load balancing, the tasks will be moved from overloaded VMs to other VMs in order to minimize the makespan and reduce the response. The movement of tasks from one VM to another varies the processing time of the tasks based on the VM capacity. $CT_{\max}$ varies with the task migration due to the load balancing.

The tasks removed from the overloaded VMs are considered as honey bees, as in HBB-LB. The main concepts of updation all number of priority tasks and the load allocated to that particular VM. The priority tasks are allocated to the under-loaded VM with better QoS parameters. Therefore, the load conditions and QoS parameters of the VMs are calculated and the VMs are ranked based on these parameters. The priority tasks are allocated to the VMs from the list sorted. In order to balance the workload, the capacity and load are to be calculated. The standard deviation is also calculated to measure the load deviations on VMs.

Capacity and load are considered as the major determinants of the VM characteristics. Capacity determines the total possible resources of a VM that are allotted for task execution. If the minimum number of processors in a VM is denoted as Np with the mips (million instructions per second) for all the processors denoted as Pmips, B denotes the bandwidth requirement of a VM for communication, then the capacity C of a single VM can be calculated as

$$C = Np + Pmips + B. \tag{3}$$

The capacity of a VM denotes the available resources for executing a particular task of length l with sufficient satisfaction of time deadlines. The capacity of a VM server is the summation of all VMs given as $C = \sum_{i=1}^{m} C_i$. As said before, load of a VM determines how much resources are available for newer tasks. The load of a single VM can be estimated based on the total length of the tasks allocated to that VM. If the number of tasks at time t on a service queue of VM is given as T(t) and execution time as S(T,t), then the load $L_{V,t}$ is computed as

$$L_{V,t} = \frac{T(t)}{S(T,t)}. \tag{4}$$

The overall load of all VMs can be obtained by adding the load of all VMs, as $L = \sum_{m} L_{V_m}$. Based on the load and capacity of V_j, the processing time of the VM can be calculated. Processing time of task (given in equation (2)) is different from the processing time of VM. The processing time for V_j is given as

$$PT_j = \frac{L_{V_m}}{C_i}. \quad (5)$$

The same can be computed for as PT for all VMs. Depending upon the processing time, the standard deviation σ of the load of VMs can be computed. σ is the quantity to measure to determine the deviation of load conditions when a new task is accepted in a VM. It is estimated as to determine the effect of a task of length l to the VM j.

$$\sigma = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(PT_j - PT)^2}. \quad (6)$$

Response time is vital in analyzing a VM which determines the readiness of VM to accept the tasks. If the VM is idle or insufficient to process the task, then it does not respond to the cloud user within a desirable time. It is generally the amount of time taken between the submission of a request and the first response which is produced. For a VM j, the response time is denoted as RT_j while the average response time for n VMs is calculated as

$$RT = \sum_{j} \frac{RT_j}{n}. \quad (7)$$

Reliability can be described as the ability of a system or a component essential for performing under steady-state conditions for a specific time period. The reliability of a cloud VM is the measure of success of a task allocated to it. Factors affecting the reliability of VM are Overflow, Timeout, Data resource missing, computing resource missing, Software failure, Database failure, Hardware failure, and Network failure.

For a VM j, if the number of accepted tasks is At while the number of tasks completed is Ct , then the reliability R can be computed as the ratio of Ct to At

$$R = \frac{Ct}{At}. \quad (8)$$

Generally in any cloud service, the resources are said to be accessible or usable if and only if they are free to be utilized at the time needed. Availability is the degree of this usable VM such that when the degree of availability is low, then the VM is either unavailable for users or the VM and resources are in idle state. For a VM which has N tasks to be processed while if At tasks are completed within t time limit, the availability *Avail* is computed as

$$Avail = \frac{At}{N}. \quad (9)$$

Though there are many QoS and other parameters, the foremost one that keeps on haunting the users is the cost. Even if a VM is efficient in all means but not cost, is rejected by the users for affordable VMs. The VM performance is usually compared with the cost parameter to obtain the efficient performance at moderate or even cheaper price. Cost of a VM depends on the cost of one unit of CPU, RAM, network and bandwidth. If a VM is priced at p unit price (p=1), net for network units, data for bandwidth, RAM for memory units, a,b,c,d are weights for each resource attribute and a+b+c+d=1, then the cost of VM is calculated as

$$cost = \frac{p}{cpu^a * net^b * data^c * RAM^d}. \quad (10)$$

Throughput determines whether the estimated capacity of a VM is actually the same and whether it is capable of satisfying the user requests without exceeding the time limits. It is defined as the number of tasks completed by the cloud provider per unit of time. If n tasks are performed on m VMs at completion time CT and T_0 is the time overhead due to various factors such as infrastructure initiation delays and inter task communication delays, then $Throughput$ can be calculated as

$$Throughput = \frac{n}{CT(n,m) + T_0}. \quad (11)$$

Load balancing based on VM characteristics. After computing the load conditions, standard deviation and QoS parameters it is essential to sort the VMs in an ascending order, based on the better values. Initially, the system should decide whether load balancing should be taken up or not. This decision requires the following preliminary: finding whether the system is balanced and also if whether the whole system is overloaded or not. If the whole system is overloaded, then load balancing is not possible. The VMs

then sorted based on the QoS parameters. However, the use of multiple QoS parameters as in our work causes multi-objective optimization problems. Since a number of QoS parameters are involved simple techniques such as crowding distance or weight estimation cannot be employed to find the pareto-optimal solutions for the multi-objective optimization problem. Hence, the Fuzzification process is performed on these QoS values. Using the fuzzy rules generated for the QoS parameters considering the best values for a VM, the VMs are ranked in an order of minimum load and better QoS values with priority are given to load. The fuzzification process generates an affordable solution for calculating the efficiency of each VM in-terms of all QoS parameters without needing to sacrifice performance efficiency. If computes all the parameters to be treated equally in ranking the VM.

When a VM is overloaded and the decision is taken to balance the load, the scheduler triggers the load balancing aspect. The overloaded VMs, load requirement, low-loaded VMs and available load can be determined by analyzing the capacity and the load of VM, along with the supply and demand at time t. Supply S and demand D can be calculated as follows:

$$S = Maximum\,capacity - \frac{L}{C}. \tag{12}$$

$$D = \frac{L}{C} - Maximum\,capacity. \tag{13}$$

The tasks are then removed from the overloaded VMs and the priority of the tasks is found. The task priority can be ranked as follows-: high, medium and low. The tasks are set as scout bees and forager bees and depending on the analysis of scout bee (earlier removed task), the suitable low loaded VMs with better QoS values are found to be effective for the forager bee (current task). The forager bee then becomes the scout bee for the next task and the whole process continues until complete load balancing task is achieved.

Step 1: Initialization

VM initialization $\mathrm{VM} = \{\mathrm{VM}_1, \mathrm{VM}_2, \ldots . \mathrm{VM}_\mathrm{m}\}$

Input Tasks $\mathrm{T} = \{\mathrm{T}_1, \mathrm{T}_2, \ldots . . \mathrm{T}_\mathrm{n}\}$

Step 2: Estimate VM characteristics

For each VM

- Compute M, C, $L_{V,t}$, $\mathrm{PT_j}$& σ
- Set threshold condition set (Ts [0-1]) for VM

//Ts based on M, C, $L_{V,t}$, $\mathrm{PT_j}$& σ

- Check condition for balance VM $\leq$ Ts
- Classify VMs (Overload, Under-load, Normal)
- Check Possibility of balancing

If $L > max\ capacity$
Load balancing aborted
Else
Balancing process begins
End if

End for

Step 3: Estimate QoS parameters for VMs

- Compute *RT, R, Avail, Cost* & $Throughput$
- Fuzzification process
  Generate fuzzy rules
  Fuzzify QoS

Step 4: Load balancing process

For under-load VM

- Compute S
- Rank VM by descending S value
- Adjust VM rank based on Fuzzified QoS values

End for

For overload VM

- Compute D
- Rank VMs by ascending D value
- Adjust VM ranks based on Fuzzified QoS values

End for

- Rank Removed tasks based on priority T_h, T_m, T_l
- Finding suitable VM for each task T

For all tasks$Load_{VM} \leq Capacity_{VM}$; minRT; maxAvail; maxR; minCost; maxThroughput

- Special condition for T_h, T_m, T_l
- $T_h \rightarrow VM \,|\, \min\left(\sum T_h\right) \in VM$
- $T_m \rightarrow VM \,|\, \min\left(\sum T_h + \sum T_m\right) \in VM$
- $T_l \rightarrow VM \,|\, \min\left(\sum T\right) \in VM$

End for

Step 5: Update phase after each iteration

- Update tasks assigned to VM
- Update remaining priority tasks
- Update current load of VM
- Update VM class status (overload, under-load, normal)

Step 6: Termination condition

- If VM class non-empty
- Continue Load balance process
- Else

End the overall process

Listing 1. Algorithm: IHBB-LB

Description: In the above described algorithm, first the VM and tasks are initialized. Each task is constructed by means of several sub-tasks. All the tasks are arithmetic computation programs. For each VM, the makespan, load, processing time and standard deviation of load are estimated. Based on the load conditions, the VMs are classified into overload, under-load and normal load VMs. Then the VMs are checked for the possibility of initiating load balancing. Once the VM load balancing process is possible, the QoS parameters are estimated. As the consideration of different QoS parameters causes pareto-optimal problem, the values are fuzzified. Then the supply for under-load VMs and demand for overload VMs are estimated respectively and then the VMs are ranked. The ranking is re-adjusted based on the fuzzified QoS values. Each task is assigned priority and categorized into high, low and medium priority tasks. Then the tasks are allocated the VMs based on the load and QoS conditions. Finally the remaining tasks, load and VM class structure are updated for the iterative process. Thus the load is balanced efficiently using the proposed IHBB-LB algorithm.

4. Experimental Results. In this section, the performance of the proposed IHBB-LB is evaluated using the CloudSim tool. CloudSim is a library for the simulation of cloud scenarios providing essential classes for describing data centers, computational resources, virtual machines, applications, users, and policies for the management of various parts of the system such as scheduling and provisioning. This performance is compared with that of HBB-LB in terms of makespan, response time, imbalancing degree and number of tasks migrated. The number of VMs is initialized as 10 for the performance evaluation. The results obtained for the proposed IHBB-LB and the existing HBB-LB algorithms are compared in Table 1.

Table. 1. Performance Comparison of HBB-LB & IHBB-LB

No. of tasks	Makespan (in seconds)		Response time (in seconds)		Degree of Imbalance		Number of tasks migrated	
	HBB-LB	IHBB-LB	HBB-LB	IHBB-LB	HBB-LB	IHBB-LB	HBB-LB	IHBB-LB
10	32.576	31.207	12.476	11.107	1.47	1.43	-	-
15	36.378	34.408	16.278	14.308	1.30	1.22	2	1
20	38.576	35.957	18.476	15.857	1.29	1.24	3	1
25	40.424	38.011	20.324	17.911	1.406	1.34	4	2
30	42.253	39.526	22.153	19.426	1.35	1.30	4	3
35	44.115	40.058	24.015	19.958	1.42	1.38	5	2
40	56.915	48.770	36.815	28.670	1.45	1.41	6	3

From the Table 1 it can be seen that for the proposed IHBB-LB has better performance values than HBB-LB in terms of makespan, response time, degree of imbalance and number of tasks migrated. It can be proved that IHBB-

LB balances all the tasks to the VMs with less makespan while also takes less response time for accepting the user requests for the services based on tasks. The degree of imbalance denotes the imbalance in the VM load allocation which is considerably reduced in the IHBB-LB. similarly, the number of tasks migrated from a VM due to lack of resources or inability to handle the task is also minimum in the proposed IHBB-LB algorithm. The graphical comparisons of the evaluation results of HBB-LB and IHBB-LB in terms of the above mentioned performance metrics are given in the following sub-section.

Makespan: Makespan is defined as the overall time taken for task completion. It can be computed as in equation (1). Figure 2 shows the comparison of makespan in HBB-LB and IHBB-LB.

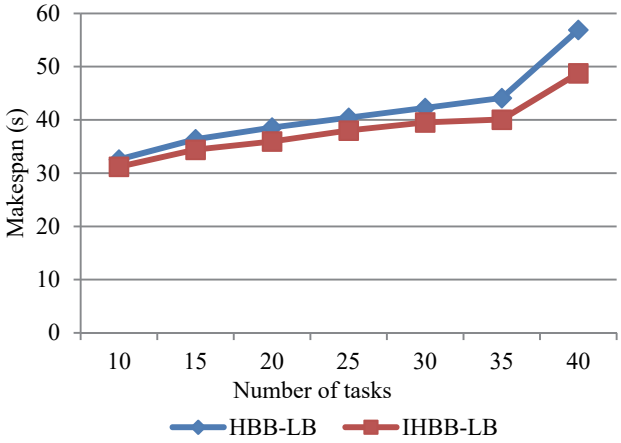


Fig. 2. Makespan

The number of VMs is initialized as 10. The number of tasks is plotted along x-axis while the makespan is plotted along y-axis. Considering the number of tasks as 40, HBB-LB has a makespan of 56.915s and IHBB-LB has a makespan of 48.77s. Thus, from the results it is clear that the proposed IHBB-LB provides better load balancing performance by minimizing the overall makespan of the tasks. The performance of HBB-LB and IHBB-LB in terms of response time is next evaluated and compared as follows:

Response Time: Response time is the amount of time taken between the submission of a request and the first response which is produced. Though makespan evaluates the performance efficiently, it is the time estimated between a task acceptance/allocation to output generation without considering the waiting time for VM response. Response time is computed as in equation (7). Figure 3 shows the comparison of response time in HBB-LB and IHBB-LB.

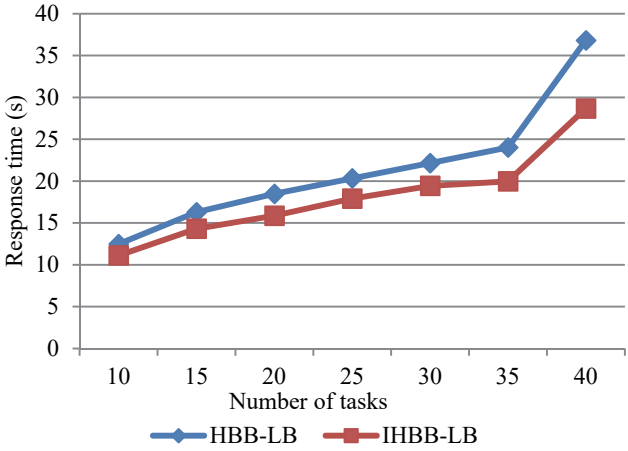


Fig. 3. Response Time

The number of tasks is plotted along x-axis while the response time is plotted along y-axis. Considering the number of tasks as 40, HBB-LB has response time of 36.815s and IHBB-LB has response time of 28.67s. Thus, from the results it is clear that the proposed IHBB-LB provides better load balancing performance. As the evaluation based on makespan and response time is efficient, the proposed scheme balancing rate has to be estimated. Inversely the degree of imbalance is evaluated as in the following sub-section.

Degree of Imbalance: The degree of imbalance is the imbalancing rate of the VM depending on the imbalance in the VM load. It can be calculated as follows

$$DI = \frac{T_{max} - T_{min}}{T_{avg}}, \tag{14}$$

where T_{max} and T_{min} are the maximum and minimum available tasks among all VMs, while T_{avg} is the average of tasks of VMs. Figure 4 shows the comparison of degree of imbalance in HBB-LB and IHBB-LB.

The number of tasks is plotted along x-axis while the DI is plotted along y-axis. Considering the number of tasks as 40, HBB-LB has DI of 1.45 and IHBB-LB has DI of 1.41. Thus, from the results it is clear that the proposed IHBB-LB provides better load balancing performance with less degree of imbalance. Degree of imbalance can also efficient when most tasks are balanced but there should be equal utilization of all VMs. If some VMs are incapable of handling a task, it is due to inefficient load conditions

while the task is migrated. Hence the efficiency of load balancing also depends upon number of migrated tasks.

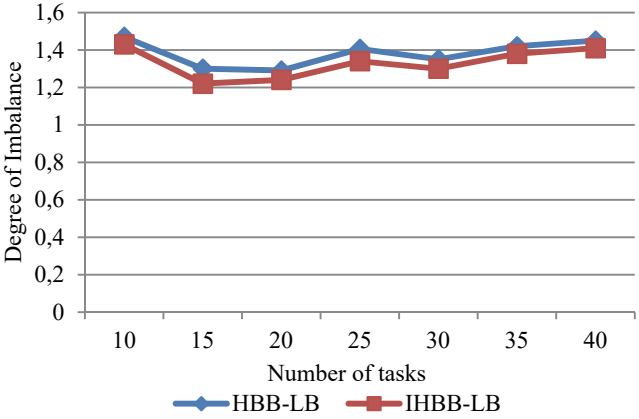


Fig. 4. Degree of Imbalance

Number of tasks migrated: The number of tasks migrated from a VM during load balancing determines the change in completion time. When more number of tasks is migrated, the VMs from which the tasks are migrated have unfavorable load conditions. This scenario must be minimized in order to provide efficient load balancing. Figure 5 shows the comparison of number of tasks migrated in HBB-LB and IHBB-LB.

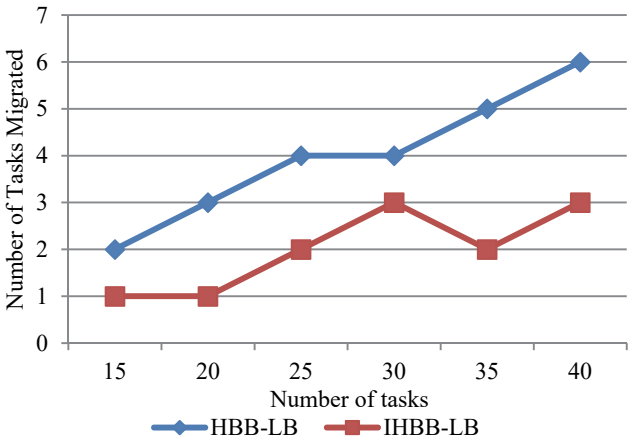


Fig. 5. Number of Tasks Migrated

The number of tasks is plotted along x-axis while the number of tasks migrated is plotted along y-axis. Considering the number of tasks as 40, in HBB-LB 6 tasks are migrated and in IHBB-LB 3 task is migrated. From the comparison results in terms of makespan, response time, degree of imbalance and number of tasks migrated it is clear that the proposed IHBB-LB provides better load balancing performance then HBB-LB. Thus it can be concluded that IHBB-LB can enhance the load balancing performance in cloud computing with high level of user satisfaction.

5. Conclusion. As described in the earlier section Improved Honey Bee Behavior based Load Balancing (IHBB-LB) approach to provide efficient load balancing has been proposed in this paper based on the consideration of multiple QoS parameters. The parameters such as service response time, availability, reliability, cost and throughput are included in IHBB-LB for balancing the load when some VM are overloaded and the remaining VM are under-loaded. This approach was found to perform better than the HBB-LB, which considers only the VM load conditions for the selecting VM for priority tasks. The performance evaluation results show that the proposed IHBB-LB approach more effective in terms of makespan, response time, imbalancing degree and number of migrated tasks which in turn seems to indicate that the IHBB-LB resolves the load balancing problem more effectively.

For more reliable load balancing, it is important for the VMs to successfully execute the tasks, the high priority tasks in particular which require adequate resources than other tasks. Thus, the resources for the VMs needed to be adaptively allocated as per the demands of application, which could be the research direction in future.

References

1. Mosleh M.A., Radhamani G., Hazber M.A., Hasan S.H. Adaptive Cost-Based Task Scheduling in Cloud Environment. *Scientific Programming*. 2016. 9 p.
2. Mondal B., Dasgupta K., Dutta P. Load balancing in cloud computing using stochastic hill climbing-a soft computing approach. *Procedia Technology*. 2012. vol. 4. pp. 783–789.
3. Boone B. et al. SALSA: QoS-aware load balancing for autonomous service brokering. *Journal of Systems and Software*. 2010. vol. 83. no. 3. pp.446–456.
4. Dasgupta K. et al. A genetic algorithm (ga) based load balancing strategy for cloud computing. *Procedia Technology*. 2013. vol. 10. pp. 340–347.
5. Keskinturk T., Yildirim M.B., Barut M. An ant colony optimization algorithm for load balancing in parallel machines with sequence-dependent setup times. *Computers & Operations Research*. 2012. vol. 39. no. 6. pp. 1225–1235.
6. Niknam T. et al. A modified honey bee mating optimization algorithm for multiobjective placement of renewable energy resources. *Applied Energy*. 2011. vol. 88. no. 12. pp. 4817–4830.
7. Dhinesh Babu L.D., Venkata Krishna P. Honey bee behavior inspired load balancing of tasks in cloud computing environments. *Applied Soft Computing*. 2013. vol. 13. no. 5. pp. 2292–2303.

8. Mohamed N., Al-Jaroodi J., Eid A. A dual-direction technique for fast file downloads with dynamic load balancing in the cloud. *Journal of Network and Computer Applications*. 2013. vol. 36. no. 4. pp. 1116–1130.
9. Zhao J. et al. A heuristic clustering-based task deployment approach for load balancing using Bayes theorem in cloud environment. *IEEE Transactions on Parallel and Distributed Systems*. 2016. vol. 27. no. 2. pp. 305–316.
10. Cho K.M., Tsai P.W., Tsai C.W., Yang C.S. A hybrid meta-heuristic algorithm for VM scheduling with load balancing in cloud computing. *Neural Computing and Applications*. 2015. vol. 26. no. 6.pp. 1297–309.
11. Xu G., Pang J., Fu X. A load balancing model based on cloud partitioning for the public cloud. *Tsinghua Science and Technology*. 2013. vol. 18. no. 1. pp. 34–39.
12. Xu Q. et al. Adaptive and scalable load balancing for metadata server cluster in cloud-scale file systems. *Frontiers of Computer Science*. 2015. vol. 9. no. 6. pp. 904–918.
13. Gutierrez-Garcia J.O., Ramirez-Nafarrate A. Agent-based load balancing in cloud data centers. *Cluster Computing*. 2015. vol. 18. no. 3. pp. 1041–1062.
14. Singh A., Juneja D., Malhotra M. Autonomous agent based load balancing algorithm in cloud computing. *Procedia Computer Science*. 2015. vol. 45. pp. 832–841.
15. Siar H., Kiani K., Chronopoulos A.T. An effective game theoretic static load balancing applied to distributed computing. *Cluster Computing*. 2015. vol. 18. no. 4. pp. 1609–1623.
16. Chen S.L., Chen Y.Y., Kuo S.H. CLB: A novel load balancing architecture and algorithm for cloud services. *Computers & Electrical Engineering*. 2017. vol. 58. pp. 154–160.
17. Liu Y., Zhang C., Li B., Niu J. DeMS: A hybrid scheme of task scheduling and load balancing in computing clusters. *Journal of Network and Computer Applications*. 2017. vol. 83. pp. 213–220.
18. Paya A., Marinescu D.C. Energy-aware load balancing and application scaling for the cloud ecosystem. *IEEE Transactions on Cloud Computing*. 2017. vol. 5. no. 1. pp. 15–27.
19. Wang Z. et al. Workload balancing and adaptive resource management for the swift storage system on cloud. *Future Generation Computer Systems*. 2015. vol. 51. pp. 120–131.
20. Randles M., Lamb D., Taleb-Bendiab A. Experiments with Honeybee Foraging Inspired Load Balancing. Proceedings of IEEE Second International Conference on Developments in E-Systems Engineering (DESE). 2009. pp. 240–247.
21. Pan J.S., Wang H., Zhao H., Tang L. Interaction artificial bee colony based load balance method in cloud computing. Proceedings of Genetic and Evolutionary Computing. 2015. pp. 49–57.

Mosleh Mohammed Abdullatef Saeed — Ph.D., research scholar of School of IT & Science, Dr. G.R. Damodaran College of Science. Research interests: Cloud Computing, Data mining, Networking. The number of publications — 4. ma.mosleh2010@gmail.com; Civil Aerodrome Post, Avinashi Road, Peelamedu, Coimbatore, Tamil Nadu 641014, India; office phone: 8122846484, Fax: 8122846484.

Radhamani Govindaraju — Ph.D., director of School of IT & Science, Dr. G.R. Damodaran College of Science, professor of School of IT & Science, Dr. G.R. Damodaran College of Science. Research interests: Internet of things, cloud computing, computer security, databases and mobile computing. The number of publications — 28. radhamani@grd.edu.in; Civil Aerodrome Post, Avinashi Road,Peelamedu, Coimbatore, Tamil Nadu 641014, India; office phone: +91 422-2572719.

УДК 004 **DOI** 10.15622/sp.57.2

М.А.С МОСЛЕХ, Г. РАДХАМАНИ

БАЛАНСИРОВКА ЗАГРУЖЕННОСТИ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ НА ОСНОВЕ УЛУЧШЕННОГО АЛГОРИТМА ПОВЕДЕНИЯ ПЧЕЛИНОЙ КОЛОНИИ

Мослех М.А.С, Радхамани Г. **Балансировка загруженности облачных вычислений на основе улучшенного алгоритма поведения пчелиной колонии.**

Аннотация. На данный момент применение алгоритма балансировки нагрузки задач на виртуальных машинах представляет большой исследовательский интерес. Для балансировки нагрузки с максимальной пропускной способностью была введена балансировка нагрузки на основе поведения медоносных пчел в колонии — Honey Bee Behavior Based Load Balancing (HBB-LB). Этот подход также устанавливает приоритеты выполнения задач на виртуальной машине с целью минимизации времени ожидания задач. Однако он рассматривает только один параметр — нагрузку виртуальных машин, что может оказаться недостаточно эффективным для балансировки. В работе предлагается улучшенный подход к балансировке нагрузки на основе пчелиного поведения, в котором дополнительно учитываются такие параметры качества обслуживания (QoS) виртуальных машин, как время отклика службы, доступность, надежность, стоимость и пропускная способность для улучшения балансировки нагрузки. Время отклика является критически важным для определения мгновенной активности виртуальной машины, доступность определяет доступный ресурс и состояние виртуальной машины (пассивное или активное), а надежность определяет уровень доверия к виртуальной машине. Затраты на использование виртуальной машины и пропускная способность виртуальных машин также необходимы для определения их эффективности. Однако включение нескольких параметров качества обслуживания приводит к многоцелевой оптимизации. По мере вычисления нескольких параметров фаззификация значений качества обслуживания выполнялась с помощью генерируемых нечетких правил, и была устранена проблема многоцелевой оптимизации. Эксперименты проводились с точки зрения времени разрешения задач, времени отклика, степени дисбаланса и количества перенесенных задач, а результаты показывают, что балансировка нагрузки на основе пчелиного поведения обеспечивает лучший уровень производительности.

Ключевые слова: оптимизация, параметры качества обслуживания, облачные вычисления, балансировка нагрузки, фаззификация.

Мослех Мохамед Абдель Латиф Саид — Ph.D., научный сотрудник школы информационных-технологий и науки, Научный колледж им. доктора Г.Р. Дамодарана. Область научных интересов: облачные вычисления, интеллектуальный анализ данных, networking. Число научных публикаций — 4. ma.mosleh2010@gmail.com; Пост гражданского аэродрома, Авинаши роуд, Пиламеду, Коимбатур, Тамилнад, 641014, Индия; р.т.: 8122846484, Факс: 8122846484.

Радхамани Говиндараджу — Ph.D., директор школы информационных-технологий и науки, Научный колледж им. доктора Г.Р. Дамодарана, профессор школы информационных-технологий и науки, Научный колледж им. доктора Г.Р. Дамодарана. Область научных интересов: интернет вещей, облачные вычисления, компьютерная безопасность, базы данных и мобильные вычисления. Число научных публикаций — 28. radhamani@grd.edu.in; Пост гражданского аэродрома, Авинаши роуд, Пиламеду, Коимбатур, Тамилнад, 641014, Индия; р.т.: +91 422-2572719.

Литература

1. *Mosleh M.A., Radhamani G., Hazber M.A., Hasan S.H.* Adaptive Cost-Based Task Scheduling in Cloud Environment // Scientific Programming. 2016. 9 p.
2. *Mondal B., Dasgupta K., Dutta P.* Load balancing in cloud computing using stochastic hill climbing-a soft computing approach // Procedia Technology. 2012. vol. 4. pp. 783–789.
3. *Boone B. et al.* SALSA: QoS-aware load balancing for autonomous service brokering // Journal of Systems and Software. 2010. vol. 83. no. 3. pp.446–456.
4. *Dasgupta K. et al.* A genetic algorithm (ga) based load balancing strategy for cloud computing // Procedia Technology. 2013. vol. 10. pp. 340–347.
5. *Keskinturk T., Yildirim M.B., Barut M.* An ant colony optimization algorithm for load balancing in parallel machines with sequence-dependent setup times // Computers & Operations Research. 2012. vol. 39. no. 6. pp. 1225–1235.
6. *Niknam T. et al.* A modified honey bee mating optimization algorithm for multiobjective placement of renewable energy resources // Applied Energy. 2011. vol. 88. no. 12. pp. 4817–4830.
7. *Dhinesh Babu L.D., Venkata Krishna P.* Honey bee behavior inspired load balancing of tasks in cloud computing environments // Applied Soft Computing. 2013. vol. 13. no. 5. pp. 2292–2303.
8. *Mohamed N., Al-Jaroodi J., Eid A.* A dual-direction technique for fast file downloads with dynamic load balancing in the cloud // Journal of Network and Computer Applications. 2013. vol. 36. no. 4. pp. 1116–1130.
9. *Zhao J. et al.* A heuristic clustering-based task deployment approach for load balancing using Bayes theorem in cloud environment // IEEE Transactions on Parallel and Distributed Systems. 2016. vol. 27. no. 2. pp. 305–316.
10. *Cho K.M., Tsai P.W., Tsai C.W., Yang C.S.* A hybrid meta-heuristic algorithm for VM scheduling with load balancing in cloud computing // Neural Computing and Applications. 2015. vol. 26. no. 6.pp. 1297–309.
11. *Xu G., Pang J., Fu X.* A load balancing model based on cloud partitioning for the public cloud // Tsinghua Science and Technology. 2013. vol. 18. no. 1. pp. 34–39.
12. *Xu Q. et al.* Adaptive and scalable load balancing for metadata server cluster in cloud-scale file systems // Frontiers of Computer Science. 2015. vol. 9. no. 6. pp. 904–918.
13. *Gutierrez-Garcia J.O., Ramirez-Nafarrate A.* Agent-based load balancing in cloud data centers // Cluster Computing. 2015. vol. 18. no. 3. pp. 1041–1062.
14. *Singh A., Juneja D., Malhotra M.* Autonomous agent based load balancing algorithm in cloud computing // Procedia Computer Science. 2015. vol. 45. pp. 832–841.
15. *Siar H., Kiani K., Chronopoulos A.T.* An effective game theoretic static load balancing applied to distributed computing // Cluster Computing. 2015. vol. 18. no. 4. pp. 1609–1623.
16. *Chen S.L., Chen Y.Y., Kuo S.H.* CLB: A novel load balancing architecture and algorithm for cloud services // Computers & Electrical Engineering. 2017. vol. 58. pp. 154–160.
17. *Liu Y., Zhang C., Li B., Niu J.* DeMS: A hybrid scheme of task scheduling and load balancing in computing clusters // Journal of Network and Computer Applications. 2017. vol. 83. pp. 213–220.
18. *Paya A., Marinescu D.C.* Energy-aware load balancing and application scaling for the cloud ecosystem // IEEE Transactions on Cloud Computing. 2017. vol. 5. no. 1. pp. 15–27.
19. *Wang Z. et al.* Workload balancing and adaptive resource management for the swift storage system on cloud // Future Generation Computer Systems. 2015. vol. 51. pp. 120–131.
20. *Randles M., Lamb D., Taleb-Bendiab A.* Experiments with Honeybee Foraging Inspired Load Balancing // Proceedings of IEEE Second International Conference on Developments in E-Systems Engineering (DESE). 2009. pp. 240–247.
21. *Pan J.S., Wang H., Zhao H., Tang L.* Interaction artificial bee colony based load balance method in cloud computing // Proceedings of Genetic and Evolutionary Computing. 2015. pp. 49–57.

УДК 004.93 **DOI** 10.15622/sp.57.3

А.Л. Олейник, Г.А. Кухарев

АЛГОРИТМЫ ВЗАИМНОЙ РЕКОНСТРУКЦИИ ИЗОБРАЖЕНИЙ ЛИЦ НА ОСНОВЕ МЕТОДОВ ПРОЕКЦИИ В СОБСТВЕННЫЕ ПОДПРОСТРАНСТВА

Олейник А.Л., Кухарев Г.А. **Алгоритмы взаимной реконструкции изображений лиц на основе методов проекции в собственные подпространства.**

Аннотация. Обсуждается проблема взаимной реконструкции изображений лиц в соответствующих друг другу парах. Эта проблема была поставлена в предыдущей статье авторов, а предложенные в ней решения обсуждались с приложением к задачам гетерогенного распознавания изображений лиц (Heterogeneous Face Recognition) и кросс-модального мультимедийного поиска (Cross-Modal Multimedia Retrieval). Эти решения основаны на одномерных и двумерных методах анализа главных компонент для двух исходных наборов изображений лиц и проекции их в независимые собственные подпространства, вычислении матриц взаимной трансформации между этими подпространствами и взаимной реконструкции изображений лиц на основе одномерного и двумерного преобразований Карунена — Лоэва.

В настоящей статье предлагаются новые подходы и решения, основанные исключительно на двумерных методах проекции в собственные подпространства и двух вариантах моделей регрессии — множественной линейной регрессии и регрессии частичных наименьших квадратов.

Приведены результаты экспериментов по взаимной реконструкции изображений лиц в парах скетч/фотографии, в парах лиц с возрастными изменениями, а также в парах изображений лиц в формах 2D/3D. Для проведения экспериментов выбраны два варианта реализации предложенного подхода. Первый из них основан на двумерном анализе главных компонент и регрессии частичных наименьших квадратов, второй — на двумерном методе частичных наименьших квадратов и множественной линейной регрессии. Оба варианта показали приемлемые для практики результаты при решении задачи взаимной реконструкции изображений лиц. Кроме того, рассмотрен способ повышения качества реконструируемых изображений лиц при работе со смешанными базами. Он заключается в классификации на основе двумерного линейного дискриминантного анализа и построении регрессионной зависимости в рамках каждого класса.

Показано также, что в общем случае взаимная реконструкция изображений лиц достижима и в условиях, когда исходные изображения не входили в состав обучающих наборов изображений лиц.

Ключевые слова: изображение лица, скетч, фоторобот, взаимная реконструкция изображений лиц мультисенсорной природы, кросс-модальный мультимедийный поиск, метод главных компонент, метод частичных наименьших квадратов, двумерные проекционные методы, регрессия.

1. Введение. Одной из проблем поиска и распознавания изображений лиц (ИЛ) является то, что на входе соответствующих систем могут использоваться различные датчики ИЛ, различные методы предобработки ИЛ и различные формы их представления. В современной практике ИЛ могут быть представлены различными способами — в форме 2D изображений в видимом (VIS), тепловом (NIR) и инфракрасном свете (IR), в виде композиционных фотороботов, рисованных

скетчей и популяций из них; в форме карт глубины («range image», форма 2.5D), в форме контурных моделей области лица (Active Shape Model — ASM), моделей внешнего вида (Active Apperance Model — AAM), определяющих текстуру ИЛ, и, наконец, в форме моделей 3D.

При этом исходные данные на входе системы могут быть одновременно представлены несколькими наборами ИЛ, связанными «в пары», «тройки» или даже «группы из более 3 ИЛ». Примером последнего варианта является пять наборов ИЛ, представляющих лицо одного и того же человека, из которых первый набор содержит ИЛ(VIS), второй — ИЛ(NIR), а третий — ИЛ(IR), четвертый — ИЛ в форме скетчей и, наконец, пятый набор содержит ИЛ в форме 3D. В отечественной биометрии такие наборы относят к данным «мультисенсорной природы». В зарубежной литературе их относят к группе «гетерогенных данных».

С одной стороны, наличие разнообразных способов представления ИЛ существенно расширяет возможности и области применения систем поиска и распознавания людей. С другой стороны, такое разнообразие значительно усложняет структуру соответствующих систем распознавания ИЛ, алгоритмы их функционирования, интерпретацию результатов распознавания и саму реализацию таких систем.

Отметим, что в отечественной практике подобные системы поиска и распознавания ИЛ еще недостаточно исследованы и не представлены в научной литературе.

Исходя из этого, актуальной становится задача исследования подходов для построения методов поиска и распознавания ИЛ, предназначенных для одновременной обработки нескольких наборов исходных данных «мультисенсорной» природы. Примерами этих подходов могут быть широко обсуждаемые в последнее время в зарубежной технической литературе задачи «Heterogeneous Face Recognition and Matching», «Cross-Modal Face Matching», «Face Image Indexing and Retrieval», а также и более общие подходы для поиска информации, например «Cross-Modal Multimedia Retrieval».

Настоящая статья предлагает способы, которые призваны приблизить решение этих задач.

2. Обзор литературы. Класс задач, связанных с обработкой ИЛ различной физической природы, невероятно широк. По этой причине в научной литературе практически отсутствуют обзорные статьи, в полной мере раскрывающие эту тему. В некоторой степени она освещена в рамках класса задач, объединенных общим названием «Heterogeneous Face Recognition» (HFR). В него входит распознавание лиц по скетчам, ИЛ в форме 3D моделей, ИЛ в инфракрасном спектре и ИЛ в низком разрешении [1]. Основная особенность этих задач — наличие исходных

данных, состоящих из связанных пар, представленных двумя (или более) наборами ИЛ или другими числовыми матрицами.

С учетом того, что структура исходных данных — это два связанных между собой набора ИЛ, мы рассмотрим лишь решения, основанные на методах проекции, базовыми из которых являются *анализ главных компонент* (Principal Component Analysis, PCA), *линейный дискриминантный анализ* (Linear Discriminant Analysis, LDA), *канонический корреляционный анализ* (Canonical Correlation Analysis, CCA) и, наконец, *метод частичных наименьших квадратов* (Partial Least Squares, PLS). Их преимуществами являются универсальность, развитый математический аппарат, наличие реализующих эти методы численных алгоритмов. Это позволяет применять проекционные методы для решения широкого спектра задач обработки данных различной природы (которые иногда называют *мультисенсорными*).

Кроме того, в последние годы были предложены обобщения этих методов на двумерный случай [2, 3]: 2DPCA/2DKLT (Two-Dimensional Principal Component Analysis / Two-Dimensional Karhunen–Loève Transform), 2DLDA/2DKLT, 2DCCA/2DKLT, 2DPLS/2DKLT. При решении задач обработки изображений они обладают рядом преимуществ по сравнению с одномерными методами, основным из которых являются отсутствие *проблемы малой выборки* (Small Sample Size, SSS), существенное *сокращение вычислительных затрат* и возможность устойчивого решения задач на собственные значения в этих методах [3].

Ниже представим краткий обзор научной литературы по применению одномерных и двумерных методов проекции в собственные подпространства для решения задач обработки ИЛ различной физической природы.

Говоря о *сценариях и алгоритмах* обработки ИЛ различной физической природы, можно выделить два направления. Первое представляет конкретные решаемые задачи и соответствующие им сценарии (например, кросс-модальное распознавание и сверхразрешение). Второе направление охватывает конкретные способы представления исходных ИЛ — например, фотоизображения лиц и скетчи.

Проанализируем далее эти направления, выделив следующие задачи:

– *Индексирование* и *распознавание* (*matching*). Индексирование предполагает взаимное сопоставление ИЛ, представленных разными способами, в рамках некоторого фиксированного набора, что позволяет эффективно решать задачи кросс-модального поиска. Распознавание включает сравнение и поиск произвольных ИЛ, представленных разными способами, что необходимо для идентификации

и верификации. Так, в работе [4] предлагается использовать метод CCA для объединения высокоуровневых (пол, возраст и др.) и низкоуровневых признаков (гистограммы ориентированных градиентов) в единое признаковое представление, пригодное для распознавания. В [5] авторы применяют метод PLS для проекции исходных ИЛ (фотографии в высоком и низком разрешении, скетчи) на промежуточные подпространства. Полученные проекции используются для решения задач индексирования и распознавания.

– *Трансформация* (*взаимная реконструкция*) [6, 7] предполагает переход между ИЛ, представленными различными способами. Так, например, из ИЛ(NIR) реконструируется ИЛ(VIS), которое может быть использовано для автоматического поиска в базе фотоизображений лиц. Такой подход позволяет использовать готовые (в том числе коммерческие) системы распознавания ИЛ, а реконструкцию ИЛ выделить в отдельную задачу. При этом результат реконструкции также может быть предъявлен человеку (например, свидетелю).

– *Сверхразрешение* (*Super Resolution*, SR) и *Face Hallucination* (FH) представляют собой два класса методов реконструкции изображений в высоком разрешении из изображений в низком разрешении. В рамках SR ставится задача получения одного изображения (не обязательно ИЛ) в высоком разрешении из набора изображений в низком разрешении. Методы SR чаще всего применяются в системах анализа видеопотока с камер низкого разрешения. С другой стороны, во многих случаях доступно только одно изображение в низком разрешении, а не их набор. В случае ИЛ для получения изображения в высоком разрешении может быть использована априорная информация о структуре лица, как это делается в рамках методов FH [8]. В работе [9] предложен подход на основе 2DCCA, выполняемом между ИЛ в низком и высоком разрешениях. Реализуемая далее реконструкция ИЛ производится посредством метода K ближайших соседей.

Перечисленные выше задачи могут решаться применительно к ИЛ, представленным разными способами. В большинстве публикаций это *фотоизображение лица в видимом свете* — ИЛ(VIS) — и один (или несколько) из следующих способов представления:

– *Скетч* («набросок», «эскиз»). Скетчи могут быть нарисованы в присутствии человека или по его фотографии (Viewed Sketch), выполнены на компьютере и доработаны художником (Artist Sketch), составлены из готовых фрагментов (рты, носы, глаза и др.) художником (Composite Sketch) или криминалистом (Composite Forensic Sketch) [10]. Один из первых алгоритмов взаимной реконструкции ИЛ(VIS) и скетчей основан на PCA [6]. Данный подход является расширением

метода собственных лиц (*Eigenfaces*) на случай двух наборов изображений: фотографий и скетчей. Недавно этот алгоритм был улучшен и обобщен в работах [7, 11], где предложено использовать его не только для трансформации ИЛ, но и для генерации их популяций. Кроме того, в [7] впервые представлен метод трансформации на основе двумерного анализа главных компонент.

– *ИЛ в ближнем инфракрасном (ИК) спектре* (NIR) и *в тепловом ИК спектре* (IR). В работе [12] предложен подход к распознаванию лиц по ИЛ(IR) на основе метода PLS, а в [13] CCA применен для реконструкции ИЛ(NIR) из ИЛ(VIS).

– ИЛ в форме *карты глубины* («depth map», «range image», 2.5D) и *трёхмерной модели* (3D) включают *информацию о глубине* (т.е. z-координату). В общем случае задача восстановления этой информации по 2D изображению достаточно сложна. Подходы к ее решению включают, например, методы «shape from shading» и «structure from motion». В случае ИЛ обычно используют априорную информацию о структуре лица, представленную в форме среднего «range image» или некоторой статистической модели. Так, в работах [13, 14] карты глубины реконструируются из RGB-изображения с помощью CCA и линейной регрессии, выполненных между парами RGB-изображений и «range image» из обучающей выборки.

Кроме отмеченного выше, список способов представления ИЛ можно дополнить *ИЛ в низком разрешении* (что соответствует рассмотренным выше задачам SR и FH), а также *ИЛ с возрастными изменениями* [15]. Как правило, моделирование и анализ возрастных изменений сводятся к изучению зависимости тех или иных антропометрических параметров от возраста. С другой стороны, эта задача может быть представлена как анализ связанных наборов ИЛ (в качестве которых, например, могут выступать пары ИЛ с разницей в заданное количество лет). В настоящей статье обсуждается подход к решению этой задачи на основе проекционных методов.

Подводя итог краткому обзору и суммируя представленные результаты, отметим, что вариантов представления ИЛ существует достаточно много. Однако, в большинстве статей рассматривается какая-то одна пара, например, ИЛ(VIS) и ИЛ(IR). Отметим также, что в случаях использования *смешанных баз ИЛ* возникают дополнительные трудности, связанные с их неоднородностью. Возможным решением может быть разбиение таких баз на группы, в рамках которых создаются «локальные» модели взаимосвязи ИЛ, сгруппированных, например, по признакам сенсорной принадлежности разными способами [16].

При этом следует отметить, что именно методы проекции в подпространства применяются наиболее часто для решения задач совместной обработки ИЛ, представленных совершенно различными способами и при условии полного неподобия в парах. Кроме того, эти методы используются и для решения других задач, таких как Cross-Modal Multimedia Retrieval [11].

Кроме того, в последнее время появился ряд других подходов к решению рассматриваемого класса задач обработки ИЛ. К ним относятся, например, глубокие нейронные сети (Deep Neural Network, DNN), Марковские случайные поля (Markov Random Field, MRF) и скрытые Марковские модели (Embedded Hidden Markov Model, E-HMM). В некоторых задачах они могут превосходить проекционные методы по ряду характеристик, например, по качеству реконструкции. Однако такие решения зачастую узкоспециализированны, в то время как проекционные методы отличаются универсальностью. Это позволяет использовать их в совершенно различных сценариях при неизменном математическом и алгоритмическом описании. В настоящее время уже существуют попытки применения методов проекции в собственные подпространства в составе слоев глубоких нейронных сетей [17]. Этим, например, достигается и уменьшение исходного пространства признаков, и использование в слоях сети только наиболее важной информации, представляющей исходные и промежуточные данные. Именно поэтому исследование различных способов применения методов проекции в собственные подпространства представляет особый интерес. И каждое решение, полученное здесь, обладает особой практической ценностью и актуальностью.

Уместно отметить, что на сегодняшний день известны лишь единичные публикации по данной теме [9, 11, 18], в то время как полномасштабные исследования в научной литературе на сегодняшний день не представлены. Данная статья нацелена на устранение этого пробела и представляет подход к взаимной реконструкции ИЛ на основе двумерных проекционных методов.

3. Предлагаемый подход. На рисунке 1 представлена общая схема, описывающая предлагаемый подход к взаимной реконструкции ИЛ, представленных двумя наборами данных и обозначенных как «ИЛ 1» и «ИЛ 2». В роли «ИЛ 1» и «ИЛ 2» могут выступать ИЛ (VIS) + скетчи, ИЛ (IR) + ИЛ (VIS) или другие комбинации различных способов представления ИЛ.

Предлагаемый подход реализуется в два этапа. На первом этапе (верхняя часть рисунка 1) выполняется анализ исходных наборов ИЛ, включающий вычисление матриц проекции, трансформацию исходных данных в собственное подпространство с использованием двумерного преобразования Карунена-Лоэва (Two-Dimensional Karhunen–Loève transform, 2DKLT) и вычисление параметров регрессии для взаимной реконструкции ИЛ в собственном подпространстве. На втором этапе (нижняя часть рисунка 1) выполняется взаимная реконструкция (на рисунке 1: $ИЛ1 \to \widetilde{ИЛ2}$ и $ИЛ2 \to \widetilde{ИЛ1}$), реализуемая с помощью прямого 2DKLT, реконструкции в подпространстве и перехода в исходное пространство признаков (т.е. пространство ИЛ) с помощью обратного 2DKLT.

Результатом анализа исходных наборов ИЛ являются *матрицы проекции* для каждого из наборов и *матрицы регрессии*. Вычисление матриц проекции выполняется с помощью двумерных проекционных методов: 2DPCA, 2DCCA и 2DPLS. В результате для каждого исходного набора ИЛ вычисляются по две матрицы проекции: для строк и для столбцов. Процедура 2DKLT сводится к умножению каждого из исходных изображений на матрицы проекции (слева и справа). На основе полученных таким образом двумерных проекций вычисляются две матрицы регрессии (по одной для реконструкции в каждом направлении). Это может быть сделано, например, с помощью множественной линейной регрессии (Multiple Linear Regression, MLR) или регрессии PLS [19].

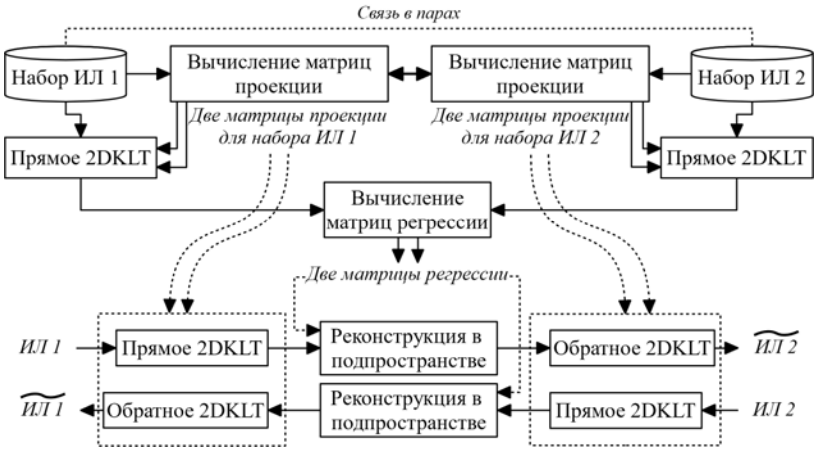


Рис. 1. Общая схема предлагаемого подхода к взаимной реконструкции ИЛ

Представленное выше решение является обобщением метода регрессии на главные компоненты (Principal Component Regression, PCR) [19], представляющего собой комбинацию PCA и MLR. Альтернативой PCR является регрессия PLS, сочетающая PLS с MLR. Использование двумерных проекционных методов позволяет эффективно решать задачи обработки, распознавания и преобразования изображений, так как в этом случае не требуется векторизация (т.е. конкатенация строк или столбцов изображений).

Отображение исходных наборов ИЛ в собственное подпространство с помощью двумерных проекционных методов преследует сразу несколько целей. Использование метода 2DCCA позволяет достичь высокой *взаимной* корреляции между проекциями исходных наборов ИЛ (в случае 2DPLS максимизируется ковариация). Более того, полученные в результате проекции признаки являются декоррелированными (*в рамках каждого из наборов ИЛ)*, что обеспечивает устойчивость вычислений на этапе расчета матрицы регрессии. Кроме того, дополнительно может быть выполнена редукция размерности пространства признаков, что позволяет удалить из исходных данных шумовую составляющую и снизить вычислительные затраты.

Если ИЛ представлены тремя и более способами, предлагается группировать их в связанные пары и анализировать эти пары отдельно. В таком случае двумерный анализ главных компонент может оказаться предпочтительным, так как с его помощью достаточно вычислить матрицы проекции для каждого набора ИЛ, а не для каждой связанной пары наборов.

Двумерные проекционные методы. Обозначим два исходных набора ИЛ как $\{X_k\}$ и $\{Y_k\}$, $k = 1 \ldots K$. Здесь X_k и Y_k — изображения размера $M \times N$. Заметим, что исходные данные, как правило, центрируют относительно средних изображений. Поэтому далее предполагается, что ИЛ отцентрированы.

Двумерные проекционные методы, подобно одномерным аналогам, позволяют выделять из исходных данных составляющие с наибольшей дисперсией, ковариационной или корреляционной связью. Отличительной особенностью двумерных методов является то, что при решении задач обработки изображений они обладают значительно большим быстродействием и вычислительной устойчивостью по сравнению с одномерными методами. Это достигается за счет замены одной задачи на собственные значения большой размерности MN двумя задачами меньшей размерности (соответственно, M и N). В таблице 1 представлено формальное описание этих методов.

Таблица 1. Двумерные проекционные методы

Операция	2DPCA/2DKLT	2DPLS/2DKLT	2DCCA/2DKLT
Вычисление матриц рассеяния по строкам	$C_{XX}^{(r)} = \sum_{k=1}^{K} X^{(k)} \left(X^{(k)} \right)^T$, $C_{YY}^{(r)} = \sum_{k=1}^{K} Y^{(k)} \left(Y^{(k)} \right)^T$		
Вычисление матриц рассеяния по столбцам	$C_{XX}^{(c)} = \sum_{k=1}^{K} \left(X^{(k)} \right)^T X^{(k)}$, $C_{YY}^{(c)} = \sum_{k=1}^{K} \left(Y^{(k)} \right)^T Y^{(k)}$		
Вычисление взаимных матриц рассеяния по строкам	—	$C_{XY}^{(r)} = \sum_{k=1}^{K} X^{(k)} \left(Y^{(k)} \right)^T$ $C_{YX}^{(r)} = \left(C_{XY}^{(r)} \right)^T$	
Вычисление взаимных матриц рассеяния по столбцам	—	$C_{XY}^{(c)} = \sum_{k=1}^{K} \left(X^{(k)} \right)^T Y^{(k)}$ $C_{YX}^{(c)} = \left(C_{XY}^{(c)} \right)^T$	
Вычисление полных матриц рассеяния	—	$S^{(tot1,r)} = C_{XY}^{(r)} C_{YX}^{(r)}$ $S^{(tot2,r)} = C_{YX}^{(r)} C_{XY}^{(r)}$ $S^{(tot1,c)} = C_{XY}^{(c)} C_{YX}^{(c)}$ $S^{(tot2,c)} = C_{YX}^{(c)} C_{XY}^{(c)}$	$S^{(tot1,r)} = \left(C_{XX}^{(r)} \right)^{-1} C_{XY}^{(r)} \left(C_{YY}^{(r)} \right)^{-1} C_{YX}^{(r)}$ $S^{(tot2,r)} = \left(C_{YY}^{(r)} \right)^{-1} C_{YX}^{(r)} \left(C_{XX}^{(r)} \right)^{-1} C_{XY}^{(r)}$ $S^{(tot1,c)} = \left(C_{XX}^{(c)} \right)^{-1} C_{XY}^{(c)} \left(C_{YY}^{(c)} \right)^{-1} C_{YX}^{(c)}$ $S^{(tot2,c)} = \left(C_{YY}^{(c)} \right)^{-1} C_{YX}^{(c)} \left(C_{XX}^{(c)} \right)^{-1} C_{XY}^{(c)}$

Продолжение таблицы 1. Двумерные проекционные методы

	$C_{XX}^{(r)}W_X^{(r)} = \Lambda_X^{(r)}W_X^{(r)}$ $C_{YY}^{(r)}W_Y^{(r)} = \Lambda_Y^{(r)}W_Y^{(r)}$ $C_{XX}^{(c)}W_X^{(c)} = \Lambda_X^{(c)}W_X^{(c)}$ $C_{YY}^{(c)}W_Y^{(c)} = \Lambda_Y^{(c)}W_Y^{(c)}$	$S^{(tot1,r)}W_X^{(r)} = \Lambda_X^{(r)}W_X^{(r)}$ $S^{(tot2,r)}W_Y^{(r)} = \Lambda_Y^{(r)}W_Y^{(r)}$ $S^{(tot1,c)}W_X^{(c)} = \Lambda_X^{(c)}W_X^{(c)}$ $S^{(tot2,c)}W_Y^{(c)} = \Lambda_Y^{(c)}W_Y^{(c)}$
Решение задач на собственные значения	*Матрицы собственных векторов для строк* $W_X^{(r)}$ и $W_Y^{(r)}$, а *также матрицы собственных чисел для строк* $\Lambda_X^{(r)}$ и $\Lambda_Y^{(r)}$ имеют размеры $M \times M$. *Матрицы собственных векторов для столбцов* $W_X^{(c)}$ и $W_Y^{(c)}$, а также *матрицы собственных чисел для столбцов* $\Lambda_X^{(c)}$ и $\Lambda_Y^{(c)}$ имеют размеры $N \times N$. Матрицы $W_X^{(r)}$, $W_Y^{(r)}$, $W_X^{(c)}$ и $W_Y^{(c)}$ в качестве столбцов содержат собственные векторы соответствующих матриц рассеяния. Матрицы $\Lambda_X^{(r)}$, $\Lambda_Y^{(r)}$, $\Lambda_X^{(c)}$ и $\Lambda_Y^{(c)}$ являются диагональными и содержат собственные числа, соответствующие столбцам матриц собственных векторов (в том же порядке).	
Выбор главных компонент (редукция размерности пространства признаков)	Выбор количества компонент по строкам и столбцам: $0 < d_r \le M$, $0 < d_c \le N$. *Матрицы проекции* $F_X^{(r)}$ и $F_Y^{(r)}$ (размера $M \times d_r$) составляются из первых d_r столбцов матриц $W_X^{(r)}$ и $W_Y^{(r)}$ (то есть из собственных векторов), упорядоченных по убыванию соответствующих им собственных чисел. *Матрицы проекции* $F_X^{(c)}$ и $F_Y^{(c)}$ (размера $N \times d_c$) составляются из первых d_c столбцов матриц $W_X^{(c)}$ и $W_Y^{(c)}$ (то есть из собственных векторов), упорядоченных по убыванию соответствующих им собственных чисел.	
2DKLT	$U^{(k)} = \left(F_X^{(r)}\right)^T X^{(k)} F_X^{(c)}$ $V^{(k)} = \left(F_Y^{(r)}\right)^T Y^{(k)} F_Y^{(c)}$ *Двумерные проекции исходных данных* $U^{(k)}$ и $V^{(k)}$ представляют собой матрицы размера $d_r \times d_c$.	

Полное описание двумерных проекционных методов, способов их реализации и анализ характеристик можно найти в монографии [3].

Отметим, что проекции исходных данных U_k и V_k, полученные методом 2DPCA/2DKLT, зависят только от «своих» наборов данных (соответственно, X_k и Y_k) Иными словами, 2DPCA выполняется два раза: для X_k и Y_k. В случае же 2DPLS/2DKLT и 2DCCA/2DKLT как U_k, так и V_k зависит от обоих наборов данных, то есть 2DPLS и 2DCCA выполняются один раз для двух наборов данных. Таким образом, если предполагается обработка большого количества связанных наборов данных, то метод 2DPCA/2DKLT позволит проводить их анализ независимо, без перебора всевозможных пар этих наборов данных. С другой стороны, методы 2DPLS/2DKLT и 2DCCA/2DKLT учитывают взаимосвязь между исходными данными, что снижает ошибку реконструкции при построении регрессионной зависимости и позволяет использовать меньше компонент, снижая таким образом вычислительные затраты.

Заметим также, что метод 2DCCA/2DKLT предполагает обращение матриц ковариации, сформированных по исходным данным. Поскольку такими данными являются изображения, то во многих случаях матрицы ковариации могут быть сингулярными, что приводит либо к невозможности их обращения, либо к неустойчивости дальнейших вычислений. Для того чтобы обойти эту проблему, в 2DCCA применяют методы регуляризации как матриц ковариации, так и общих матриц рассеяния. Подробное описание данного решения можно найти в [3].

Регрессионные модели между двумерными проекциями изображений. Анализ исходных наборов ИЛ завершается вычислением матриц регрессии для проекций, полученных с помощью 2DKLT. Существуют различные методы построения регрессионной зависимости. В настоящей статье представлены варианты базового подхода, основанные на множественной линейной регрессии (MLR) и регрессии PLS.

Метод MLR достаточно прост и эффективен. Если размерность пространства признаков после редукции $D = d_r \times d_c$ достаточно мала, то метод MLR не требует значительных вычислительных затрат, а полученное решение обладает вычислительной устойчивостью. Как правило, это может быть достигнуто при использовании 2DPLS/2DKLT и 2DCCA/2DKLT на этапе вычисления матриц проекции.

Как было отмечено выше, метод 2DPCA/2DKLT чаще всего предполагает использование большего количества компонент D, чем

2DPLS/2DKLT или 2DCCA/2DKLT. По этой причине метод MLR может оказаться неприменимым. В таком случае возникает необходимость в дальнейшем снижении размерности пространства признаков, что может быть достигнуто с помощью регрессии PLS. Пример совместного использования 2DPCA/2DKLT и регрессии PLS представлен ниже, в разделе экспериментальных исследований.

В настоящей статье использована реализация регрессии PLS, основанная на алгоритме NIPALS (Non-linear Iterative PArtial Least Squares). По сути, это вариант степенного метода, представляющего собой итеративную процедуру решения задачи на собственные значения.

4. Экспериментальные исследования. Экспериментальные исследования проведены на трех базах ИЛ:

1. База фотографий лиц и соответствующих им скетчей, выполненных художниками: CUFS/CUFSF [20, 21];
2. ИЛ для заданной группы людей в различном возрасте [22];
3. Фотографии лиц и соответствующие им карты глубины («range image»): база Texas 3D [23-25].

Реконструкция фотоизображений лиц по скетчам на основе 2DPCA/2DKLT и PLS-регрессии. Первый вариант представленного подхода основан на методах 2DPCA/2DKLT и PLS-регрессии; его схема приведена на рисунке 2.

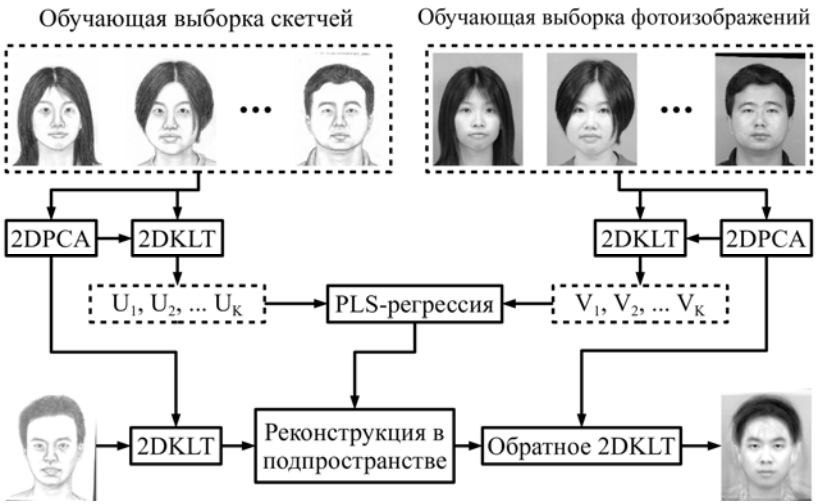


Рис. 2. Реконструкция фотоизображений по скетчам на основе метода 2DPCA/2DKLT и PLS-регрессии

В рамках экспериментов на базах CUFS/CUFSF были использованы следующие параметры:

1. Размеры ИЛ: $M = 250,\ \ N = 200$. змерность исходного пространства признаков $M \times N = 50000$;

2. Количество элементов по строкам и столбцам: $d_r = 60,$ $d_c = 50.$ Размерность промежуточного пространства признаков $D = d_r \times d_c = 3000$;

3. Количество компонент в регрессии PLS: 17.

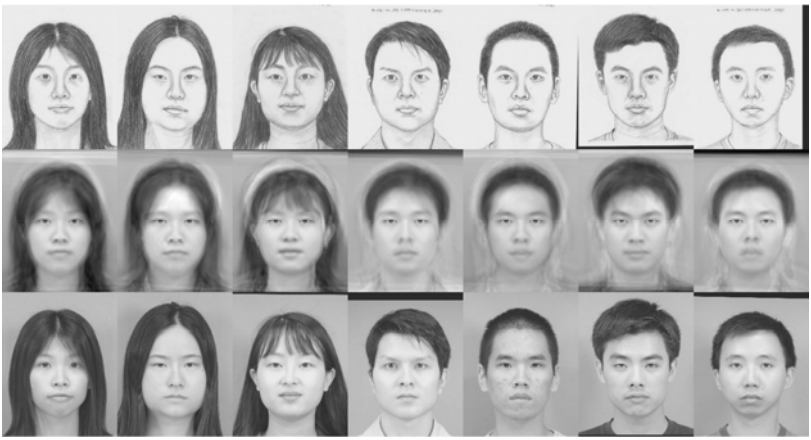
Таким образом, размерность пространства признаков поэтапно снижается с 50000 до 3000 и далее до 17. Попытка применения PLS-регрессии без использования 2DPCA/2DKLT приводит к неоправданно высоким вычислительным затратам и неустойчивости решения регрессионной задачи. Кроме того, при добавлении нового набора ИЛ для тех же людей (например, IR), собственные базисы, построенные с помощью 2DPCA/2DKLT, могут быть использованы повторно.

На рисунке 3 показаны примеры реконструкции фотографий по соответствующим скетчам на обучающей и тестовой выборках. Можно отметить приемлемое качество реконструкции как для обучающей, так и для тестовой выборки. Ухудшение качества реконструкции на тестовой выборке можно объяснить уникальными особенностями каждого скетча, зачастую не подобного исходной фотографии.

На рисунке 4 представлены результаты реконструкции фотоизображений по скетчам в низком разрешении с добавлением шума. Для этого размер скетчей был уменьшен до 9% от исходного (т.е. более чем в десять раз) и к ним был добавлен равномерный шум со значениями из диапазона ±5% от максимальной яркости изображения. Результаты, представленные на рисунке 4, свидетельствуют об устойчивости предложенного подхода к шумам и понижению разрешения исходного изображения. Здесь важно то, что обучение проводилось только на ИЛ в высоком разрешении. Таким образом, можно говорить о применимости предложенного решения к задачам сверхразрешения (SR) и Face Hallucination (FH).

Моделирование возрастных изменений с помощью 2DPCA/2DKLT и PLS-регрессии. Набор ИЛ людей различных возрастов включает ИЛ в возрасте 20, 30, 40, 50, 60 лет. Исходные данные

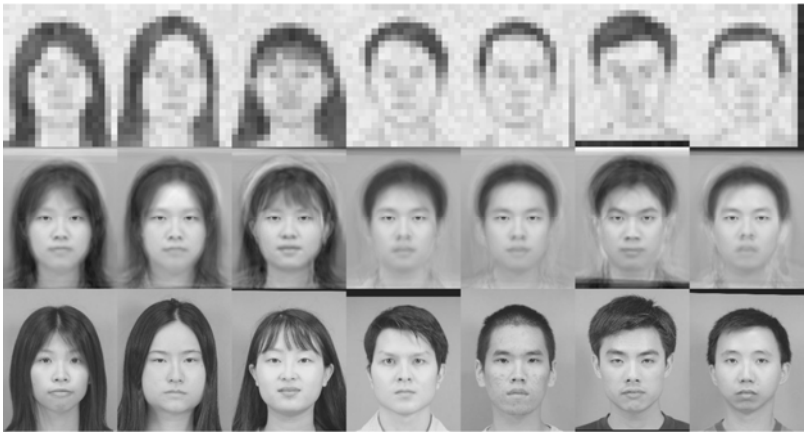
выбраны из статьи [22]. Предлагаемый нами подход применен для «состаривания» ИЛ человека на 10 лет.



а)

б)

Рис. 3. Реконструкция фотографий лиц по скетчам, выполненная на тестовой (б) выборке. В первом ряду приведены исходные скетчи, во втором — реконструированные из них фотографии, в третьем — исходные фотографии



а)

б)

Рис. 4. Реконструкция фотографий лиц по зашумлённым скетчам в низком разрешении, выполненная на обучающей (а) и тестовой (б) выборках. В первом ряду приведены исходные скетчи, во втором — реконструированные из них фотографии, в третьем — исходные фотографии

Для этого для каждого человека соответствующие ИЛ были сгруппированы в пары по (30, 40), (40, 50), (50, 60) лет (как показано на рисунке 5) и использованы в качестве обучающей выборки. ИЛ в возрасте 20 лет играют роль тестовой выборки. Полученные таким образом наборы данных могут быть обработаны в соответствии с описанным выше подходом.

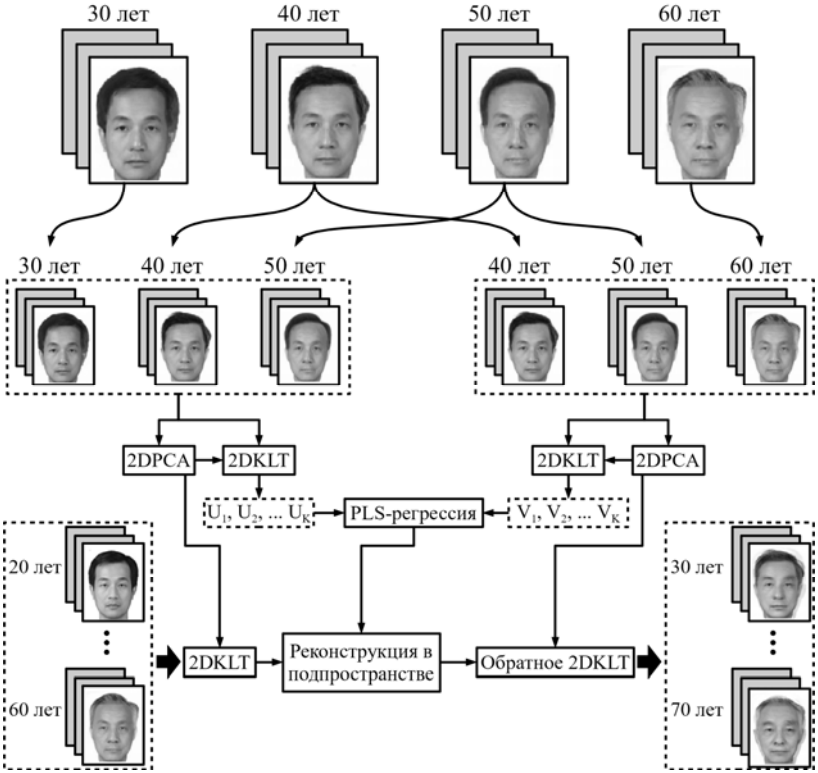


Рис. 5. Сведение задачи моделирования возрастных изменений к задаче реконструкции ИЛ в связанных парах

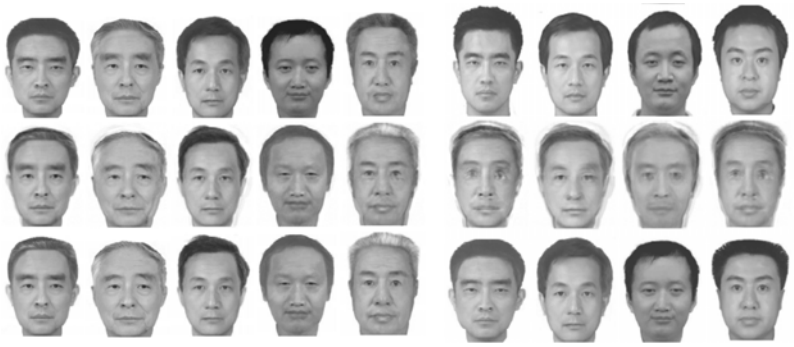
В рамках этих экспериментов были использованы следующие параметры:

1. Размеры ИЛ: $M = 338$, $N = 264$. Размерность исходного пространства признаков $M \times N = 89232$;

2. Количество элементов по строкам и столбцам: $d_r = 60$, $d_c = 50$. Размерность промежуточного пространства признаков $D = d_r \times d_c = 3000$;

3. Количество компонент в регрессии PLS: 8.

На рисунке 6 показаны примеры реконструкции ИЛ старшего возраста из ИЛ младшего возраста. Несмотря на то что качество реконструкции на тестовой выборке несколько ниже, чем на обучающей, можно отметить характерные возрастные изменения, вносимые в реконструированные изображения. Для повышения качества результирующих ИЛ необходима более объемная и репрезентативная выборка ИЛ.



а) *б)*

Рис. 6. Реконструкция ИЛ старшего возраста по ИЛ младшего возраста, выполненная на обучающей (а) и тестовой (б) выборках. В первом ряду приведены ИЛ младшего возраста, во втором — реконструированные из них ИЛ старшего возраста, в третьем — исходные ИЛ старшего возраста

Реконструкция фотоизображений лиц по скетчам на основе 2DPLS/2DKLT и MLR. Здесь представлены эксперименты по взаимной реконструкции фотоизображений лиц и скетчей, а также фотоизображений лиц и карт глубины («range image») с помощью комбинации методов 2DPLS/2DKLT и MLR.

Реконструкция выполнялась в двух режимах:

– «в классе» — регрессия по столбцам двумерных проекций выполняется отдельно для каждой пары ИЛ (рисунок 7);

– «по базе» — регрессия по столбцам двумерных проекций ИЛ выполняется для всей обучающей выборки (рисунок 8).

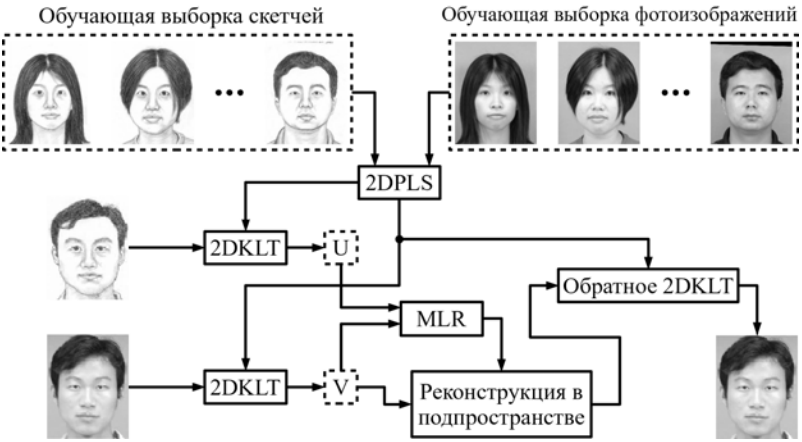


Рис. 7. Реконструкция фотоизображений лиц по скетчам «в классе»

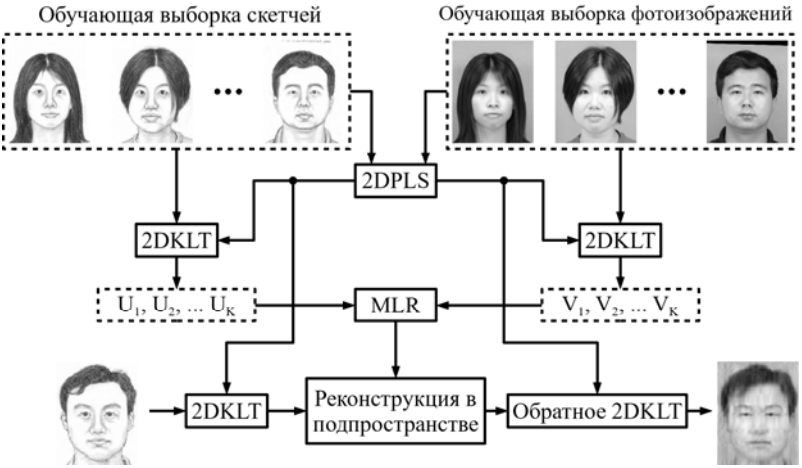


Рис. 8. Реконструкция фотоизображений лиц по скетчам «по базе»

На рисунке 9 представлены примеры изображений исходной базы CUFS, полученные с помощью 2DPLS наборы собственных чисел, а также фазовая корреляция в собственном подпространстве между переменными U и V, полученными с помощью 2DKLT. Достаточно высокое значение, достигаемое в области пика фазовой корреляции, говорит об относительно высоком подобии исходных данных в собственном подпространстве признаков.

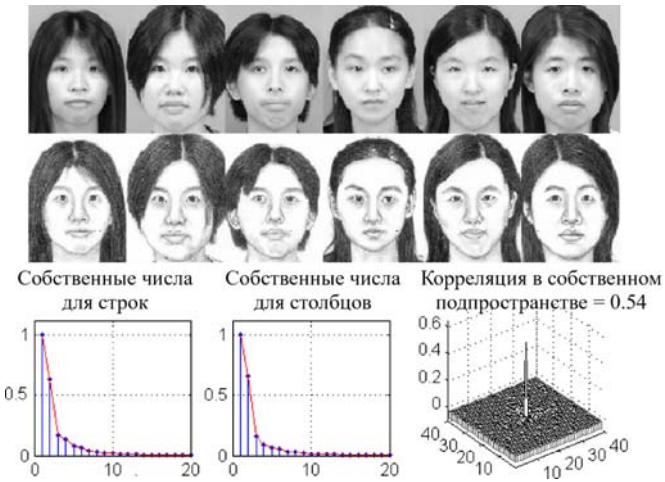


Рис. 9. Примеры обучающей выборки (фото и скетчи) из базы CUFS и результаты их обработки в рамках метода 2DPLS/2DKLT: собственные числа и фазовая корреляция в подпространстве

На рисунке 10 представлены результаты реконструкции тестовых ИЛ «в классе» и «по базе». Отметим, что реконструкция «в классе» всегда выполняется точнее, чем «по базе», и ограничена лишь точностью реконструкции на этапе обратного 2DKLT.

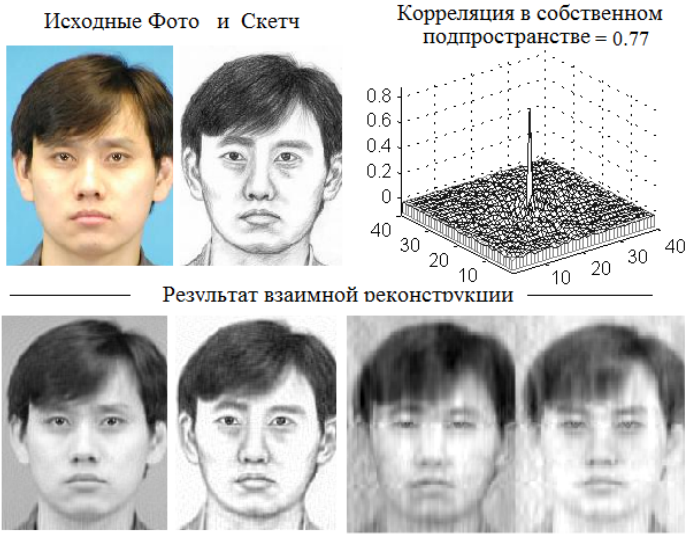
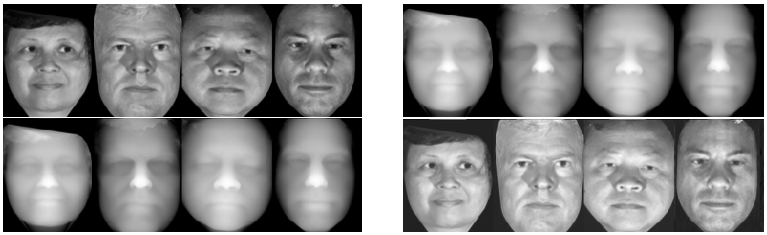


Рис. 10. Примеры реконструкции ИЛ «в классе» (слева) и «по базе» (справа)

На рисунке 11 представлены результаты взаимной реконструкции фотографий и карт глубины («range image»), выполненной на базе Texas 3D.

Эксперимент показывает, что предложенный метод позволяет решать задачу взаимной реконструкции ИЛ различной физической природы, что позволяет сделать вывод о его универсальности.



а) *б)*

Рис. 11. Реконструкция ИЛ «в классе» для базы Texas 3D, содержащей фотографии лиц и соответствующие им карты глубины («range image»). Первый ряд — исходные ИЛ, второй ряд — результат реконструкции

На рисунке 12 показана взаимная реконструкция фотографий и скетчей, входящих в обучающую и тестовую базы.

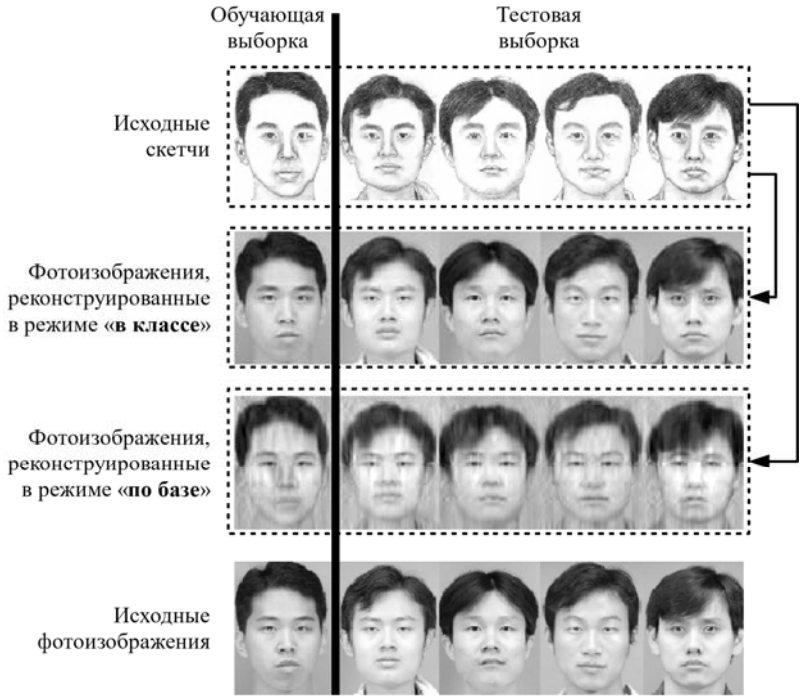


Рис. 12. Реконструкция ИЛ «в классе» и «по базе», выполненная на обучающей и тестовой выборках

Из представленных выше экспериментов видно, что режим реконструкции «в классе» позволяет получить наилучшие результаты. Это достигается за счет того, что в роли «класса» (в рамках которого строится модель и выполняется реконструкция) выступает единственная пара ИЛ.

Для практического использования можно предложить следующий подход. Исходные наборы ИЛ разбиваются на классы по некоторому признаку (фенотипические признаки, пол и др.), после чего для каждого из классов строится отдельная регрессионная модель. На этапе реконструкции сначала выполняется классификация входного ИЛ, а затем выполняется собственно реконструкция с использованием соответствующей матрицы регрессии. Для классификации ИЛ может быть использован метод 2DLDA/2DKLT [3] и некоторый алгоритм классификации (например, алгоритм ближайшего соседа). На рисунке 13 представлен один из вариантов такого подхода.

Заметим, что для реализации такого подхода необходима выборка, снабженная метками классов. Если метки классов отсутствуют, можно применить некоторый алгоритм кластеризации. Такие подходы существуют и представлены в научной литературе [16]. Их применение к решаемой задаче является предметом дальнейших исследований.

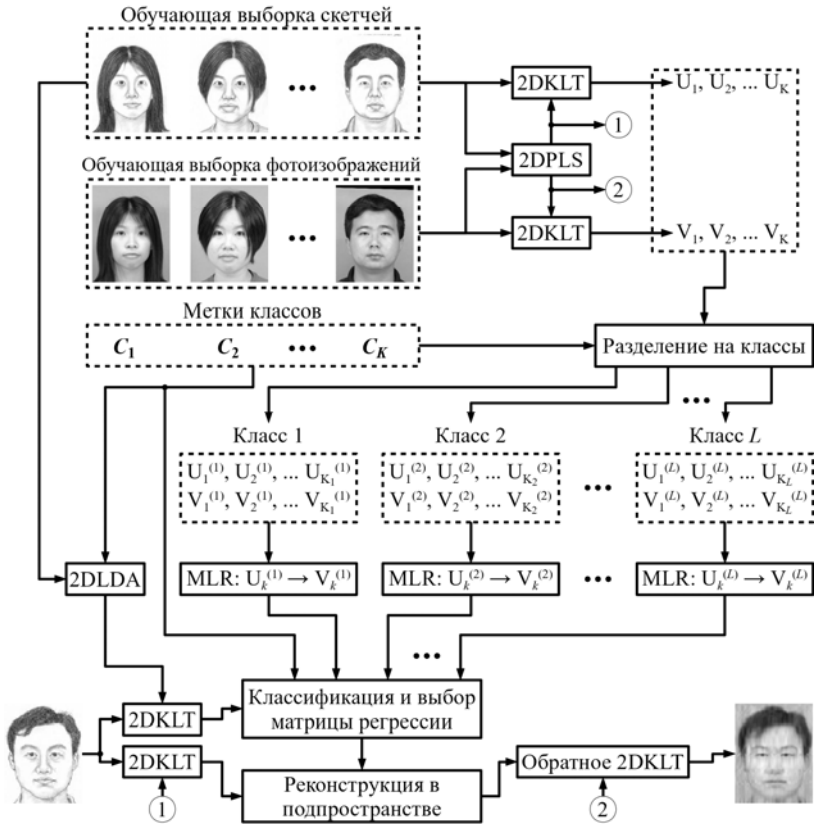


Рис. 13. Реконструкция ИЛ на основе 2DPLS/2DKLT, 2DLDA/2DKLT и MLR

5. Оценка качества реконструкции. Количественная оценка качества реконструкции ИЛ — задача нетривиальная. Поэтому довольно часто в научной литературе вместо прямой оценки качества приводят результаты по точности распознавания ИЛ по реконструированным изображениям. Несмотря на то, что такой способ позволяет получить конкретные количественные результаты, они имеют смысл только для ограниченного набора сценариев. При этом использование других методов распознавания может привести к иным результатам.

Другой подход использует сравнительную оценку результатов реконструкции ИЛ на основе субъективного сравнения собственных результатов с известными из литературных источников по решаемой проблеме. Этот подход применяется в рамках одинаковых сценариев реконструкции ИЛ и относительно близких методов реконструкции.

В настоящей статье рассмотрены новые (оригинальные) сценарии реконструкции ИЛ, а представленные решения основаны на двумерных методах проекции изображений в собственные подпространства. При этом какая-либо связь — корреляция, семантическое подобие или сенсорная природа — между исходными наборами может отсутствовать (несмотря на то, что в собственных пространствах возникает устойчивая корреляционная или ковариационная связь). Аналоги этих сценариев и решений для них, насколько известно авторам, в литературе не описаны, а следовательно, провести сравнительный анализ качества реконструкции ИЛ не представляется возможным.

В настоящей статье качество реконструкции оценивается на основе метрик подобия, не привязанных ни к системам распознавания, ни к известным результатам. Примерами таких метрик являются *фазовая корреляция* и *индекс структурного подобия* SSIM (Structural SIMilarity).

Фазовая корреляция, в отличие от амплитудной, чувствительна к текстуре изображения, а основные различия между изображениями «мультисенсорной» природы заключаются именно в текстуре. Высота пика фазовой корреляции и определяет степень сходства двух ИЛ.

Индекс SSIM первоначально был предложен для оценки качества фото- и видеоизображений путем сравнения различных версий изображения. Однако в научной литературе индекс SSIM используется также и для сравнения различных ИЛ [26]. Таким образом, в рамках задачи взаимной реконструкции ИЛ использование индекса SSIM как меры подобия исходных и реконструированных ИЛ оправдано.

На рисунке 14 представлены фазовая корреляция и значения индекса SSIM для идентичных фотографий и фотографий лиц разных людей. Видно, что для одного и того же изображения фазовая корреляция имеет единственный пик, достигающий единицы; индекс SSIM также равен единице. При сравнении изображений лиц разных людей индекс SSIM составляет 0,53, а главный пик фазовой корреляции практически не выражен на фоне «шумовой составляющей».

На рисунке 15*а* аналогичным способом выполнено сравнение фотографии и скетчей (в оригинальном и в низком разрешении) из базы CUFS, соответствующих одному и тому же человеку. На рисунке 15*б* показано сравнение той же исходной фотографии с фотографиями, реконструированными из соответствующих скетчей рисунка 15*а*. Ре-

конструкция выполнена на основе 2DPCA/2DKLT и PLS-регрессии. Как фазовая корреляция, так и индекс SSIM указывают на то, что реконструкция значительно повышает подобие между ИЛ.

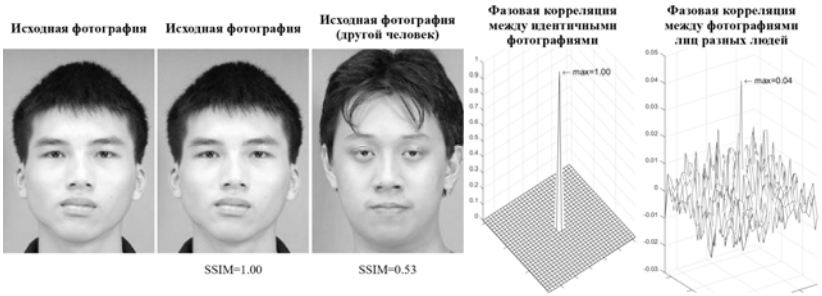


Рис. 14. Значения индекса SSIM и фазовая корреляция для идентичных фотографий и для фотографий лиц разных людей

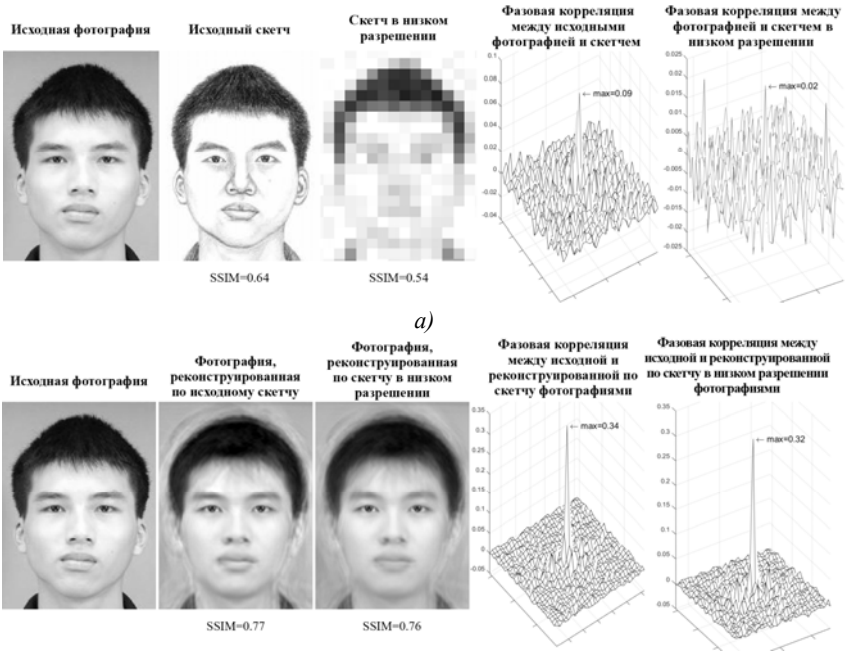


Рис. 15. Значения индекса SSIM и фазовая корреляция для исходной фотографии и соответствующих ей скетчей в оригинальном и в низком разрешении (а) и для исходной и реконструированных фотографий (б)

На рисунке 16 показаны результаты сравнения реконструированных фотографий с фотографией другого человека. Можно видеть, что реконструированные фотографии *не* подобны фотографиям другого человека.

Отметим, что даже при использовании скетча в низком разрешении индекс SSIM и фазовая корреляция ведут себя аналогично тому, как это происходит в случае скетчей в оригинальном разрешении. Здесь важно отметить то, что для построения матриц проекции и регрессии в наших решениях *изображения в низком разрешении на этапе обучения не использовались.*

На основании приведенных рисунков можно сделать вывод о согласованности визуальной оценки подобия изображений и результатов, полученных на основании индекса SSIM и фазовой корреляции.

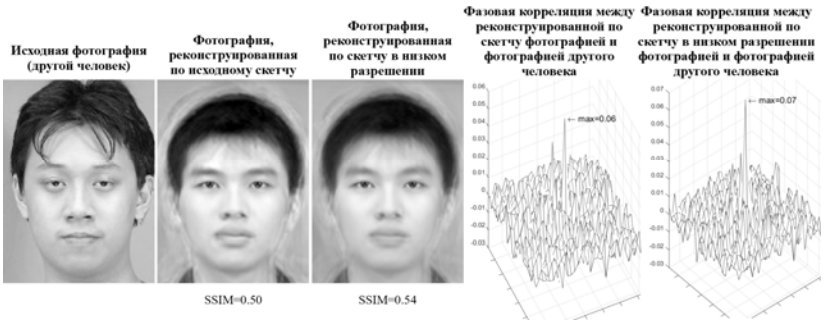


Рис. 16. Значения индекса SSIM и фазовая корреляция для реконструированных фотографий и исходной фотографии другого («постороннего») человека из базы CUFS

В таблице 2 приведены результаты сравнения исходных фотографий со скетчами и реконструированными из них фотографиями для метода на основе 2DPCA/2DKLT и PLS-регрессии, а также метода на основе 2DPLS/2DKLT и MLR. Показаны средние значения индекса SSIM (mean) и его среднеквадратического отклонения (std), полученные по базе CUFS. Кроме того, для результатов, полученных при сравнении исходных фотографий и скетчей, указано максимальное (то есть наилучшее, «наиболее удачное») значение (max). Для значений индекса SSIM, полученных при сравнении исходных и реконструированных из скетчей фотографий, указано минимальное («наименее удачное») значение (min).

Таблица 2. Результаты оценки качества реконструкции на основе индекса SSIM

	Сравнение исходных фотографий и скетчей			Сравнение исходных и реконструированных из скетчей фотографий		
	mean	std	max	mean	std	min
2DPCA/2DKLT + регрессия PLS	0.48	0.07	0.65	0.81	0.04	0.71
2DPCA/2DKLT + регрессия PLS (скетчи в низком разрешении)	0.48	0.05	0.59	0.73	0.03	0.64
2DPLS/2DKLT + MLR	0.48	0.07	0.65	0.77	0.03	0.69

На основании рисунков 14-16 и таблицы 2 можно сделать следующие выводы:

– визуальная (субъективная) оценка подобия изображений и результаты, полученные на основании индекса SSIM и фазовой корреляции (объективная оценка), согласуются друг с другом и подтверждают тот факт, что реконструкция достижима;

– фазовая корреляция и индекс SSIM позволяют отличить ИЛ различных людей;

– качество реконструкции остается приемлемым при реконструкции фотографий из зашумлённых скетчей в низком разрешении, даже в том случае, когда для построения матриц проекции и регрессии изображения в низком разрешении не используются на этапе обучения. Это существенно отличает представленный в статье сценарий от рассматриваемых в научной литературе;

– эксперименты на базе CUFS показали, что подобие реконструированных и исходных фотографий по индексу SSIM существенно выше, чем в случае исходных фотографий и скетчей. При использовании фазовой корреляции этот эффект выражен в большей степени, так как в случае исходных фотографий и скетчей фазовая корреляция между ними полностью отсутствует.

6. Заключение. В настоящей работе предложен подход к взаимной реконструкции ИЛ мультисенсорной природы. Этот подход реализуется на основе ряда представленных выше двумерных проекционных методов и регрессионных моделей. Подробно рассмотрены и описаны двумерные проекционные методы, выделены их основные свойства, даны рекомендации по применению.

Для подтверждения практической значимости предложенного подхода проведены экспериментальные исследования на различных базах ИЛ, включающих фотографии лиц, скетчи, карты глубины («range images»). Для проведения экспериментов выбраны два варианта реализации предложенного подхода. Первый из них основан на 2DPCA/2DKLT и PLS-регрессии, второй — на 2DPLS/2DKLT и MLR. Оба варианта показали хорошие результаты при решении задачи реконструкции.

Проведена визуальная (субъективная) и количественная (объективная) оценки качества реконструкции. Для количественной оценки использованы фазовая корреляция и индекс структурного подобия. Оба показателя указывают на достаточное качество реконструкции. Более того, результаты остаются хорошими даже в том случае, когда входные изображения имеют низкое разрешение, несмотря на то, что для построения регрессионных моделей использовались только изображения в высоком разрешении. Это выгодно отличает предложенное решение от ряда методов сверхразрешения, которые обычно применяют в таких ситуациях.

Кроме того, были получены результаты по моделированию возрастных изменений лица человека с помощью методов 2DPCA/2DKLT и регрессии PLS. То, что предложенный подход позволяет решать данную задачу, указывает на то, что область его применения не ограничивается взаимной реконструкцией различных ИЛ и может включать, например, генерацию популяций ИЛ и деидентификацию. Также вероятно, что предложенное решение может быть использовано не только для обработки и преобразования изображений, но и в других задачах анализа связанных данных, таких как, например, разработка многомодальных человеко-машинных интерфейсов.

Литература

1. *Ouyang S. et al.* A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution // ArXiv Prepr. ArXiv14095114. 2014.
2. *Kukharev G., Kamenskaya E.* Application of two-dimensional canonical correlation analysis for face image processing and recognition // Pattern Recognition and Image Analysis. 2010. vol. 20. no. 2. pp. 210–219.
3. *Кухарев Г.А., Каменская Е.И., Матвеев Ю.Н., Щеголева Н.Л.* Методы обработки и распознавания изображений лиц в задачах биометрии / под ред. Хитрова М.В. // СПб: Политехника. 2013. 388 с.
4. *Ouyang S. et al.* Cross-modal face matching: Beyond viewed sketches // 12th Asian Conference on Computer Vision (ACCV). 2014. pp. 210–225.
5. *Sharma A., Jacobs D.W.* Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011. pp. 593–600.
6. *Tang X., Wang X.* Face sketch recognition // IEEE Trans. Circuits Syst. Video Technol. 2004. vol. 14. no. 1. pp. 50–57.
7. *Kukharev G., Oleinik A.* Face Photo-Sketch Transformation and Population Generation // International Conference on Computer Vision and Graphics (ICCVG 2016). 2016. LNCS 9972. pp. 329–340.
8. *Baker S., Kanade T.* Hallucinating faces // Fourth IEEE International Conference on Automatic Face and Gesture Recognition. 2000. pp. 83–88.
9. *An L., Bhanu B.* Face image super-resolution using 2D CCA // Signal Process. 2014. vol. 103. pp. 184–194.
10. *Кухарев Г.А., Матвеев Ю.Н., Форчманьски П.* Поиск людей по фотороботам: методы, системы и практические решения // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 4. С. 640–653.
11. *Кухарев Г.А., Матвеев Ю.Н., Олейник А.Л.* Алгоритмы взаимной трансформации изображений для систем обработки и поиска визуальной информации // Научно-

технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 1. С. 62–74.
12. *Xie Z.* Partial least squares regression on DCT domain for infrared face recognition // Twelfth International Conference on Photonics and Imaging in Biology and Medicine (PIBM 2014). International Society for Optics and Photonics. 2014. vol. 9230. id 92301I. pp. 6.
13. *Reiter M. et al.* 3D and Infrared Face Reconstruction from RGB data using Canonical Correlation Analysis // 18th International Conference on Pattern Recognition (ICPR 2006). 2006. vol. 1. pp. 425–428.
14. *Reiter M. et al.* Estimation of face depth maps from color textures using canonical correlation analysis. Computer Vision Winter Workshop. 2006. pp. 1–5.
15. *Ramanathan N., Chellappa R., Biswas S.* Computational methods for modeling facial aging: A survey // J. Vis. Lang. Comput. 2009. vol. 20. no. 3. pp. 131–144.
16. *Liang J. et al.* Group-invariant cross-modal subspace learning // Proc. IJCAI. 2016. pp. 1739–1745.
17. *Tian L., Fan C., Ming Y.* Multiple scales combined principle component analysis deep learning network for face recognition // J. Electron. Imaging. 2016. vol. 25. no. 2. pp. 16.
18. *Kim H., Fyfe C., Ko H.* Feature locations in images // International Conference on Intelligent Data Engineering and Automated Learning. 2008. pp. 459–463.
19. *Эсбенсен К.* Анализ многомерных данных // Черноголовка: ИПХФ РАН. 2005. 160 с.
20. CUFS dataset. URL: http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html (дата обращения: 25.04.2015).
21. CUFSF dataset. URL: http://mmlab.ie.cuhk.edu.hk/archive/cufsf/ (дата обращения: 16.04.2016).
22. *Suo J. et al.* A compositional and dynamic model for face aging // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010. vol. 32. no. 3. pp. 385–401.
23. *Gupta S. et al.* Texas 3D Face Recognition Database. URL: http://live.ece.utexas.edu/research/texas3dfr/index.htm (дата обращения: 23.04.2017).
24. *Gupta S. et al.* Texas 3D face recognition database // IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI). 2010. pp. 97–100.
25. *Gupta S., Markey M.K., Bovik A.C.* Anthropometric 3D Face Recognition // Int. J. Comput. Vis. 2010. vol. 90. no. 3. pp. 331–349.
26. *Sun Y., Tistarelli M., Maltoni D.* Structural similarity based image quality map for face recognition across plastic surgery // 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8.

Олейник Андрей Леонидович — аспирант кафедры речевых информационных систем, ФГАОУ ВО "Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики" (Университет ИТМО), инженер кафедры речевых информационных систем, ФГАОУ ВО "Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики" (Университет ИТМО). Область научных интересов: машинное обучение, цифровая обработка изображений, лицевая биометрия, распознавание образов. Число научных публикаций — 15. aoleinik@corp.ifmo.ru; ул. Красуцкого, 4, Санкт-Петербург, 196084; р.т.: +7(812)325-88-48.

Кухарев Георгий Александрович — д-р техн. наук, профессор, профессор кафедры мультимедийных систем, Западнопоморский технологический университет, профессор кафедры математического обеспечения и применения ЭВМ, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина) (СПбГЭТУ "ЛЭТИ"). Область научных интересов: цифровая обработка изображений, распознавание образов, лицевая биометрия. Число научных публикаций — 255. kuga41@gmail.com; Зольнерская ул., 49, Щецин, 71-210, Польша; р.т.: +48 91-449-56-60.

Поддержка исследований. Исследование проводится при частичной финансовой поддержке Правительства Российской Федерации (грант № 074-U01).

DOI 10.15622/sp.57.3

A.L. OLEINIK, G.A. KUKHAREV

ALGORITHMS FOR FACE IMAGE MUTUAL RECONSTRUCTION BY MEANS OF TWO-DIMENSIONAL PROJECTION METHODS

Oleinik A.L., Kukharev G.A. **Algorithms for Face Image Mutual Reconstruction by Means of Two-Dimensional Projection Methods.**

Abstract. In this paper, we consider the problem of mutual reconstruction of face image pairs. We addressed this problem in our previous article, where the proposed solutions were discussed in connection with Heterogeneous Face Recognition and Cross-Modal Multimedia Retrieval problems. Those solutions are based on one-dimensional and two-dimensional Principal Component Analysis performed over two original face images followed by their projection on independent eigenspaces, estimation of a transformation matrix and mutual reconstruction of the face image by means of one-dimensional and two-dimensional Karhunen-Loève Transform.

In this article, we propose new approaches and solutions, which are based solely on the two-dimensional eigenspace projection methods, and two regression models — Multiple Linear Regression and Partial Least Squares regression.

We present the experiments on mutual reconstruction of face images in sketch/photo pairs, in pairs of face images with age-related changes, and in pairs of 2D/3D face images. In order to conduct the experiments, we selected two variants of the proposed approach. First one is based on two-dimensional Principal Component Analysis and Partial Least Squares regression, and the second one is based on two-dimensional Partial Least Squares and Multiple Linear Regression. Both variants showed acceptable performance for practical applications involving the mutual reconstruction of face images. Furthermore, we consider the method to improve the quality of reconstructed face images in the case of mixed datasets. This method involves classification of the dataset by means of two-dimensional Linear Discriminant Analysis and fitting of a separate regression model for each class.

In addition, we show that generally, mutual reconstruction of face images is also achievable in conditions when original images are not a part of training sets of face images.

Keywords: face image, sketch, facial composite, mutual reconstruction of multisensory face images, cross-modal multimedia retrieval, principal component analysis, partial least squares, two-dimensional projections, regression.

Oleinik Andrei Leonidovich — Ph.D. student of the department of speech information systems, ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics), engineer of the department of speech information systems, ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics). Research interests: machine learning, digital image processing, face biometrics, pattern recognition. The number of publications — 15. aoleinik@corp.ifmo.ru; 4, Krasutskogo str., St. Petersburg, 196084, Russia; office phone: +7(812)325-88-48.

Kukharev Georgy Aleksandrovich — Dr. Sci., professor, professor of multimedia systems department, West Pomeranian University of Technology, professor of software engineering and computer applications department, Saint Petersburg Electrotechnical University "LETI". Research interests: Digital image processing, pattern recognition, face biometrics. The number of publications — 255. kuga41@gmail.com; 49, Zolnierska St., Szczecin, 71-210, Poland; office phone: +48 91-449-56-60.

Acknowledgements. This research is supported by the Government of the Russian Federation (grant 074-U01).

References

1. Ouyang S. et al. A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution. ArXiv Prepr. ArXiv14095114. 2014.
2. Kukharev G., Kamenskaya E. Application of two-dimensional canonical correlation analysis for face image processing and recognition. *Pattern Recognition and Image Analysis*. 2010. vol. 20. no. 2. pp. 210–219.
3. Kukharev G.A., Kamenskaya E.I., Matveev Yu.N., Shchegoleva N.L. *Metody obrabotki i raspoznavanija izobrazhenij lic v zadachah biometrii. Pod red. Hitrova M.V.* [Methods of facial images processing and recognition in biometrics. Edited by Khitrov M.V.]. SPb.: Politechnika. 2013. 388 p. (In Russ.).
4. Ouyang S. et al. Cross-modal face matching: Beyond viewed sketches. 12th Asian Conference on Computer Vision (ACCV). 2014. pp. 210–225.
5. Sharma A., Jacobs D.W. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011. pp. 593–600.
6. Tang X., Wang X. Face sketch recognition. *IEEE Trans. Circuits Syst. Video Technol*. 2004. vol. 14. no. 1. pp. 50–57.
7. Kukharev G., Oleinik A. Face Photo-Sketch Transformation and Population Generation. International Conference on Computer Vision and Graphics (ICCVG 2016). 2016. LNCS 9972. pp. 329–340.
8. Baker S., Kanade T. Hallucinating faces. Fourth IEEE International Conference on Automatic Face and Gesture Recognition. 2000. pp. 83–88.
9. An L., Bhanu B. Face image super-resolution using 2D CCA. *Signal Process*. 2014. vol. 103. pp. 184–194.
10. Kukharev G.A., Matveev Yu.N., Forczmański P. [People retrieval by means of composite pictures – methods, systems and practical decisions]. *Nauchno-Tehnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 2015. vol. 15. no. 4. pp. 640–653. (In Russ.).
11. Kukharev G.A., Matveev Yu.N., Oleinik A.L. [Mutual image transformation algorithms for visual information processing and retrieval]. *Nauchno-Tehnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 2017. vol. 17. no. 1. pp. 62–74. (In Russ.).
12. Xie Z. Partial least squares regression on DCT domain for infrared face recognition. Twelfth International Conference on Photonics and Imaging in Biology and Medicine (PIBM 2014). International Society for Optics and Photonics. 2014. vol. 9230. id 92301I. pp. 6.
13. Reiter M. et al. 3D and Infrared Face Reconstruction from RGB data using Canonical Correlation Analysis. 18th International Conference on Pattern Recognition (ICPR 2006). 2006. vol. 1. pp. 425–428.
14. Reiter M. et al. Estimation of face depth maps from color textures using canonical correlation analysis. Computer Vision Winter Workshop. 2006. pp. 1–5.
15. Ramanathan N., Chellappa R., Biswas S. Computational methods for modeling facial aging: A survey. *J. Vis. Lang. Comput*. 2009. vol. 20. no. 3. pp. 131–144.
16. Liang J. et al. Group-invariant cross-modal subspace learning. Proc. IJCAI. 2016. pp. 1739–1745.
17. Tian L., Fan C., Ming Y. Multiple scales combined principle component analysis deep learning network for face recognition. *J. Electron. Imaging*. 2016. vol. 25. no. 2. pp. 16.
18. Kim H., Fyfe C., Ko H. Feature locations in images. International Conference on Intelligent Data Engineering and Automated Learning. Springer. 2008. pp. 459–463.
19. Esbensen K. *Analiz mnogomernyh dannyh* [Multivariate Data Analysis]. Chernogolovka: IPCP RAS. 2005. 160 p. (In Russ.).

20. CUFS dataset. Available at: http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html (accessed: 25.04.2015).
21. CUFSF dataset. Available at: http://mmlab.ie.cuhk.edu.hk/archive/cufsf/ (accessed: 16.04.2016).
22. Suo J. et al. A compositional and dynamic model for face aging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010. vol. 32. no. 3. pp. 385–401.
23. Gupta S. et al. Texas 3D Face Recognition Database. Available at: http://live.ece.utexas.edu/research/texas3dfr/index.htm (accessed: 23.04.2017).
24. Gupta S. et al. Texas 3D face recognition database. IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI). 2010. pp. 97–100.
25. Gupta S., Markey M.K., Bovik A.C. Anthropometric 3D Face Recognition. *Int. J. Comput. Vis*. 2010. vol. 90. no. 3. pp. 331–349.
26. Sun Y., Tistarelli M., Maltoni D. Structural similarity based image quality map for face recognition across plastic surgery. 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8.

УДК 621.391 **DOI** 10.15622/sp.57.4

Ф.А. Таубин, А.Н. Трофимов

КАСКАДНОЕ КОДИРОВАНИЕ НА ОСНОВЕ МНОГОМЕРНЫХ РЕШЕТОК И КОДОВ РИДА — СОЛОМОНА ДЛЯ МНОГОУРОВНЕВОЙ ФЛЭШ-ПАМЯТИ

Таубин Ф.А.,Трофимов А.Н. **Каскадное кодирование на основе многомерных решеток и кодов Рида — Соломона для многоуровневой флэш-памяти.**

Аннотация. В работе рассмотрена каскадная схема кодирования для многоуровневой флэш-памяти, внутренняя ступень которой представляет собой конечное подмножество многомерной целочисленной решетки (lattice code), а в качестве внешней ступени используется код Рида — Соломона.

Анализ помехоустойчивости предложенной каскадной схемы выполнен применительно к модели, отражающей основные физические особенности ячейки флэш-памяти с неравномерно расположенными целевыми уровнями напряжения в ячейке и дисперсией шума, зависящей от записанного значения (input-dependent additive Gaussian noise, ID-AGN). Для этой модели в работе развит новый подход к вычислению вероятности ошибки декодирования внутреннего кода на основе одномерного численного интегрирования произведений характеристических функций случайных величин, используемых декодером при вынесении решения. Показано, как при увеличении времени хранения и/или числа циклов перезаписи адаптировать параметры предложенной каскадной конструкции с тем, чтобы сохранить требуемый уровень вероятности ошибки.

Ключевые слова: каскадный код, многомерные решетки, код Рида — Соломона, многоуровневая флэш-память, вероятность ошибки декодирования.

1. Введение. Многоуровневая флэш-память благодаря более высокой плотности записи занимает доминирующее место на рынке энергонезависимых устройств хранения данных [1]. В настоящее время стандартом де-факто является флэш-память с четырьмя состояниями (уровнями записи), что позволяет хранить два бита в ячейке памяти; при этом в литературе сообщается о дальнейшем прогрессе технологии производства флэш-памяти, который позволит использовать 8 и даже 16 уровней [2-4]. Оборотной стороной повышения плотности записи, достигаемого за счет возрастания количества используемых в ячейке состояний, является снижение надежности хранения данных. Основными факторами, определяющими снижение надежности при увеличении числа уровней, являются повторяющиеся в процессе эксплуатации циклы записи/стирания и утечка заряда плавающего затвора с течением времени. Воздействие указанных факторов проявляется в виде появления шума записи/считывания, что приводит к случайному

отклонению реального порогового уровня напряжения от целевого уровня (target value) напряжения в ячейке. В результате в процессе считывания содержимого флэш-памяти возникают ошибки. Экспериментальные исследования показывают [5], что в многоуровневой памяти (в считываемой области) доминирующим является близкое к равномерному распределение ошибок без заметной тенденции к пакетированию. С учетом того, что вероятность ошибки (raw bit error rate) в многоуровневой флэш-памяти оказывается неприемлемо высокой, порядка $10^{-4}...10^{-3}$ и более, тогда как требуемая вероятность ошибки лежит в диапазоне $10^{-12}...10^{-16}$, введение помехоустойчивого кодирования оказывается неизбежным.

Один из возможных эффективных подходов к организации введения помехоустойчивого кодирования связан с использованием каскадных конструкций. Среди возможных вариантов внутреннего кода следует выделить многомерные сигнальные множества, обладающие гибкой структурой с широким диапазоном варьирования параметров и допускающие, как правило, сравнительно простую организацию мягкого декодирования, что может существенно повысить эффективность внешнего кодирования.

Несмотря на указанные достоинства многомерных сигнальных множеств, возможность их использования для повышения надежности хранения данных во флэш-памяти рассматривалась лишь в нескольких публикациях. Одной из первых, по-видимому, является работа [6]. В работе [7] было показано, что использование многомерных сигнальных множеств позволяет в ряде случаев заметно снизить вероятность ошибки. Каскадное кодирование с многомерным сигнальным множеством в качестве внутреннего кода рассматривалось в работах [8-10]. В этих публикациях в качестве внешних кодов рассматривались, соответственно, БЧХ коды, коды Рида — Соломона и низкоплотностные коды. Во всех работах была продемонстрирована существенная эффективность такого рода каскадных конструкций для повышения надежности хранения данных во флэш-памяти.

Вместе с тем в указанных работах анализ каскадных конструкций проводился применительно к упрощенной модели ячейки флэш-памяти — с равномерно расположенными целевыми уровнями напряжения в ячейке и аддитивным гауссовским шумом с фиксированной дисперсией. В настоящей работе рассматривается

более реалистичная модель ячейки флэш-памяти [11, 12] — с неравномерно расположенными целевыми уровнями напряжения в ячейке и дисперсией шума, зависящей от записанного значения (input-dependent additive Gaussian noise, ID-AGN). Применительно к этой модели развит новый подход к вычислению вероятности ошибки декодирования слов внутреннего кода. В результате оценка вероятности ошибки декодирования слова внутреннего кода по максимуму правдоподобия вычисляется путем одномерного численного интегрирования некоторого выражения, включающего характеристические функции случайных величин, на основе которых формируется решение.

2. Модель ячейки многоуровневой флэш-памяти (гауссовская аппроксимация). В рассматриваемой модели блок флэш-памяти рассматривается как множество независимых ячеек. Далее полагается, что физический носитель может рассматриваться как стационарный канал без заметной тенденции к пакетированию ошибок, поэтому математическая модель представляет собой модель канала без памяти, которая полностью определяется моделью одиночной ячейки флэш-памяти. При описании упрощенной математической модели одиночной ячейки флэш-памяти мы опираемся на публикации [11, 12]. Входные уровни каждой ячейки принимают некоторые фиксированные значения $x_0, x_1, ..., x_{q-1}$, а выходные значения представляют собой случайные величины. Распределения этих случайных величин описываются условными функциями плотности вероятности (ф.п.в.) $p_{y|x}(y \mid x)$, $-\infty < y < \infty$, $x = x_0, x_1, ..., x_{q-1}$.

В публикациях [11-13] перечисляются следующие факторы, влияющие на распределение значений выходных уровней одиночных ячеек: а) начальное распределение пороговых значений; б) влияние циклов записи/стирания (program/erasure cycling, P/E cycling); в) взаимная интерференция ячеек (cell-to-cell interference); г) влияние времени хранения (retention). Оставляя в стороне описание физических процессов, влияющих на распределение выходных уровней одиночной ячейки флэш-памяти, будем считать, что модель ячейки определяется ф.п.в $p_{y|x}(y \mid x_i)$, $i = 0,1,...,q-1$, которые могут аппроксимированы гауссовскими плотностями, то есть будем считать, что:

$$p_{y|x}(y \mid x_i) = \frac{1}{\sqrt{2\pi}\sigma(x_i)} \exp\left(-\frac{(y - x_i)^2}{2\sigma^2(x_i)} \right). \quad (1)$$

Такое описание называется также моделью с гауссовским шумом, стандартное отклонение которого $\sigma(x_i)$ зависит от входного значения x_i [14], или ID-AGN моделью (input-dependent additive Gaussian noise). Мы будем использовать также обозначение $\sigma_i = \sigma(x_i)$. Распространенным примером, используемым далее в качестве основного, служит значение $q = 4$ [11-13]. Заметим, что эта модель допускает обобщение на большее число входных уровней, в частности на шесть, восемь и двенадцать уровней [14]. Важной особенностью модели ячейки многоуровневой флэш-памяти является то, что с ростом числа циклов перезаписи N и времени хранения T значения x_i (кроме x_0) уменьшаются, а значения σ_i (кроме σ_0) увеличиваются, что соответствует ухудшению канала с ростом величин T и N. Зависимость параметров x_i и σ_i, $i = 0,1,...,q-1$ от значений T и N выражается сложным образом и здесь не приводится. Детали описания этой зависимости могут быть найдены в [11, 12] и в [15], где также приводится обоснование методики гауссовской аппроксимации. Здесь мы будем пользоваться зависимостями значений x_i и σ_i от T и N, которые были численно найдены, исходя из описания физической модели [11, 12]. На рисунке 1 в графической форме представлены значения параметров x_i и σ_i, $i = 0,1,2,3$ для ряда значений числа циклов перезаписи N и времени хранения T, а на рисунке 2 — примеры графиков ф.п.в. $p_{y|x}(y \mid x_i)$.

Нетрудно оценить, что вероятность ошибочного решения о записанном уровне составляет величину порядка $10^{-4}...10^{-3}$ и более в зависимости от числа циклов перезаписи N и времени хранения T. Отсюда следует, что вводимое помехоустойчивое кодирование должно обеспечить снижение вероятности ошибки по меньше мере на 8…11 порядков, поскольку требуемое значение вероятности ошибки представляет собой величину порядка 10^{-12}.

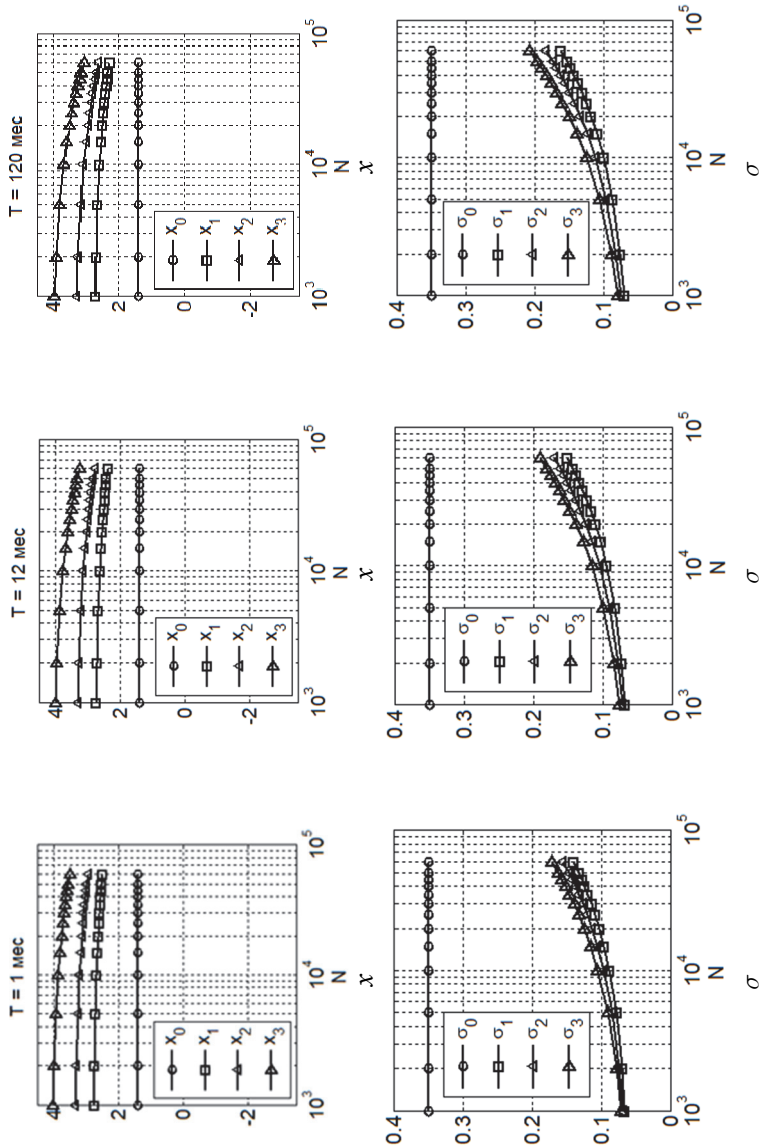


Рис. 1. Значения x_i и σ_i в зависимости от N и T

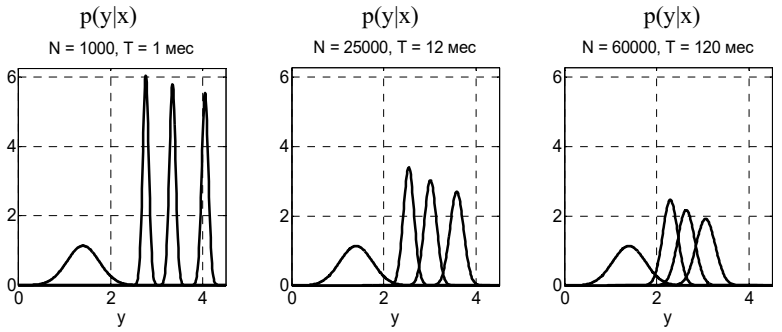


Рис. 2. Функции плотности вероятности $p_{y|x}(y \mid x_i)$ считываемых значений

3. Внутренний код каскадной кодовой конструкции на основе многомерной решетки. Обозначим через элементарное сигнальное множество, состоящее из q целых чисел, $A = \{0,1,\ldots,q-1\}$, и будем полагать, что $m = \log_2 q$ целое. Множество A связано с реальным сигнальным множеством (множеством уровней записи) $X = \{x_0, x_1, ..., x_{q-1}\}$ посредством взаимно однозначного отображения I множества A на множество X вида $I(i) = x_i$, $0 \le i \le q-1$. Иными словами, множество A представляет собой множество индексов уровней записи. Пусть B есть прямое произведение n экземпляров элементарного сигнального множества A, $B = A^n$. Множество B будем называть сигнальным множеством в рассматриваемой конструкции. Общее число точек в сигнальном множестве B равно $q^n = 2^{mn}$, так что множество B соответствует n ячейкам памяти, хранящим $m \times n$ бит.

3.1 Разбиение сигнального множества. При построении и описании разбиения сигнального множества на вложенные подмножества удобно рассматривать исходное сигнальное множество как *конечное подмножество* многомерной целочисленной решетки (lattice).

Пусть $\mathbf{Z}^n$ — прямое произведение n экземпляров одномерной целочисленной решетки $\mathbf{Z}$ и пусть Λ_0 — подрешетка решетки $\mathbf{Z}^n$, $\Lambda_0 \subseteq \mathbf{Z}^n$. Подрешетка Λ_0 определяет разбиение $\mathbf{Z}^n / \Lambda_0$ решетки $\mathbf{Z}^n$ на подрешетку Λ_0 и ее смежные классы; пусть порядок этого разбиения $\left|\mathbf{Z}^n / \Lambda_0\right| = 2^{r_0}$, $r_0 \ge 0$. Значение $r_0 = 0$ означает, что $\Lambda_0 = \mathbf{Z}^n$, что соответствует отсутствию разбиения. Обозначим через

B_0 пересечение подрешетки Λ_0 и конечного сигнального множества B, $B_0 = \Lambda_0 \cap B$. Очевидно, что $|B_0| = 2^{mn-r_0}$, и в этом случае $0 \le r_0 \le mn$. Множество B_0 в общем случае разбивается далее на вложенные подмножества с использованием разбиения решетки Λ_0 на подходящие вложенные подрешетки.

Довольно жесткие требования к быстродействию флэш-памяти и к сложности реализации декодера диктуют выбор в качестве допустимых сравнительно простых вариантов разбиения решетки Λ_0. Поэтому ограничимся далее рассмотрением разбиений вида:

$$\Lambda_0 / 2\mathbf{Z}^n / 4\mathbf{Z}^n / \cdots / 2^L \mathbf{Z}^n , \quad L = m-1 . \tag{2}$$

Будем полагать, что Λ_0 в (2) есть mod 2 решетка, то есть $\Lambda_0 \subseteq 2\mathbf{Z}^n$. Рассмотрим вначале разбиение $\Lambda_0 / 2\mathbf{Z}^n$. Порядок разбиения $\left| \Lambda_0 / 2\mathbf{Z}^n \right| = 2^{n-r_0}$. Можно показать [16], что l-й смежный класс в разбиении $\Lambda_0 / 2\mathbf{Z}^n$ может быть представлен в виде $2\mathbf{Z}^n + \mathbf{a}_l \mathbf{G}$, где $\mathbf{a}_l$ — двоичный (над алфавитом $\{0, 1\}$) вектор размера $n - r_0$, и

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_{11} \\ \vdots \\ \mathbf{g}_{1,\, n-r_0} \end{pmatrix}$$

— двоичная матрица, строками которой являются генераторы совокупности смежных классов в разбиении $\Lambda_0 / 2\mathbf{Z}^n$. Отсюда следует, что решетка Λ_0 может быть представлена как:

$$\Lambda_0 = 2\mathbf{Z}^n + \{\mathbf{a}_l \mathbf{G} \;\; | 1 \le l \le 2^{n-r_0}\}. \tag{3}$$

Матрицу $\mathbf{G}$ в (3) можно интерпретировать как порождающую матрицу линейного двоичного $(n, n-r_0)$ кода C, поэтому решетка Λ_0 представляется в виде:

$$\Lambda_0 = 2\mathbf{Z}^n + C. \tag{4}$$

Для представления (4) нетрудно заметить, что минимальное евклидово расстояние Δ_0 между точками решетки Λ_0 может быть

вычислено как $\Delta_0 = \min(2, \sqrt{d_H})$, где d_H — минимальное хэммингово расстояние кода C.

Приведем три важных примера использования представления (4). Для безызбыточного (полного) кода C $\Lambda_0 = \mathbf{Z}^n$ и $\Delta_0 = 1$. Если код C в (4) есть $(n, n-1)$ код с проверкой на четность, то Λ_0 будет решеткой Schlafli D_n, и $\Delta_0 = \sqrt{2}$. Третий пример: если код C в (4) есть код Рида — Маллера (8, 4), $n = 8$, то решетка Λ_0 будет представлять собой решетку Gosset E_8, и $\Delta_0 = 2$.

Теория кодов на основе целочисленных решеток имеет своей целью построение сигнально-кодовых конструкций с максимально возможным евклидовым расстоянием. Однако евклидова метрика не является в полной мере согласованной с моделью канала записи/считывания с ID-AGN. Построить же строго согласованную метрику для канала с неравномерно расположенными уровнями сигналов и шумом, дисперсия которого зависит от переданного (записанного) значения, не представляется возможным. Поэтому в дальнейшем используется компромиссный подход – использование для рассматриваемой модели флэш-памяти кодов, построенных на основе целочисленных решеток и ориентированных на использование в канале с аддитивным гауссовским шумом с постоянной дисперсией.

Разбиения $2^k\mathbf{Z}^n / 2^{k+1}\mathbf{Z}^n$, $1 \le k \le L-1$, в цепочке (2) имеют весьма простую структуру. Нетрудно видеть, что порядок разбиения $\left|2^k\mathbf{Z}^n / 2^{k+1}\mathbf{Z}^n\right| = 2^n$, а l-й смежный класс в разбиении $2^k\mathbf{Z}^n / 2^{k+1}\mathbf{Z}^n$, $1 \le k \le L-1$, может быть представлен в виде $2^{k+1}\mathbf{Z}^n + 2^k\mathbf{a}_l$, где $\mathbf{a}_l$ — двоичный (над алфавитом $\{0,1\}$) вектор размера n. Это означает, что решетка $2^k\mathbf{Z}^n$ может быть представлена в виде:

$$2^k\mathbf{Z}^n = 2^{k+1}\mathbf{Z}^n + \{2^k\mathbf{a}_l \,|\, 1 \le l \le 2^n\}. \qquad (5)$$

Минимальное евклидово расстояние Δ_k между точками подрешетки $2^k\mathbf{Z}^n$, $1 \le k \le L-1$ равно, очевидно, 2^k.

Представления (3) и (5) разбиений решеток в цепочке (2) позволяют полностью описать структуру разбиений сигнального множества B_0 на вложенные подмножества. Пусть $B_1 = B_0 \cap 2\mathbf{Z}^n$ и пусть B_{1l} — пересечение множества B_0 и l-го смежного класса в

разбиении $\Lambda_0 / 2\mathbf{Z}^n$. Тогда из представления (3) следует, что подмножество B_{1l} (в некотором смысле *l*-й «смежный класс» в разбиении B_0 / B_1) может быть представлено в виде $B_{1l} = B_1 + \mathbf{a}_l \mathbf{G}$, где $\mathbf{a}_l$ — двоичный (над алфавитом $\{0,1\}$) вектор размера $n - r_0$, $\mathbf{G}$ — двоичная матрица, строками которой являются генераторы совокупности смежных классов в разбиении $\Lambda_0 / 2\mathbf{Z}^n$, то есть:

$$B_0 = B_1 + \{\mathbf{a}_l \mathbf{G} \,|\, 1 \le l \le 2^{n-r_0}\}. \qquad (6)$$

Пусть $B_k = B_0 \cap 2^k \mathbf{Z}^n$, $2 \le k \le L$, и пусть B_{kl} — пересечение множества B_0 и *l*-го смежного класса в разбиении $2^{k-1}\mathbf{Z}^n / 2^k \mathbf{Z}^n$. Тогда из представления (5) следует, что подмножество B_{kl} (в некотором смысле *l*-й «смежный класс» в разбиении B_{k-1} / B_k) может быть представлено в виде $B_{kl} = B_k + 2^{k-1}\mathbf{a}_l$, где $\mathbf{a}_l$ — двоичный (над алфавитом $\{0,1\}$) вектор размерности *n*. Отсюда следует, что:

$$B_{k-1} = B_k + \{2^{k-1}\mathbf{a}_l \,|\, 1 \le l \le 2^n\}. \qquad (7)$$

3.2. Кодирование и декодирование внутреннего кода. Представления (6) и (7), в явном виде характеризующие структуру рассматриваемых разбиений сигнального множества, позволяют указать весьма простую процедуру кодирования для внутреннего кода B_0. Двоичный блок $\mathbf{u}$, кодируемый кодом B_0, состоит, очевидно, из $mn - r_0$ символов. Представим блок $\mathbf{u}$ в виде набора *m* подблоков: $\mathbf{u} = (\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{m-1})$, где подблок $\mathbf{u}_0$ имеет длину $n - r_0$, а остальные подблоки $\mathbf{u}_1, \ldots, \mathbf{u}_{m-1}$ имеют длину *n*. Подблок $\mathbf{u}_0$ определяет *l*-й смежный класс, $1 \le l \le 2^{n-r_0}$, в разбиении $\Lambda_0 / 2\mathbf{Z}^n$, а значит, и подмножество B_{1l} в разбиении B_0 / B_1; подблок $\mathbf{u}_1$ определяет одно из 2^n подмножеств в разбиении подмножества B_{1l} и так далее. Положим:

$$\mathbf{b} = \mathbf{u}_0 \mathbf{G} + \sum_{k=1}^{m-1} 2^k \mathbf{u}_k, \qquad (8)$$

где $\mathbf{G}$ — порождающая матрица линейного двоичного $(n, n - r_0)$ кода C в (4). Нетрудно видеть, что: 1) совокупность *n*-мерных векторов над

алфавитом A, порождаемых согласно (8), совпадает с множеством B_0, и 2) подблоки с разными номерами кодируемого блока $\mathbf{u} = (\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{m-1})$ получают, в результате кодирования, неравную защиту (в смысле минимального евклидова расстояния).

Для рассматриваемого варианта разбиения сигнального множества (2), внутренний код, как видно из выражений (4) и (6), можно рассматривать как прямую сумму (кодовых слов) линейного двоичного кода C (с порождающей матрицей $\mathbf{G}$) и подмножества B_1, представляющего собой прямое произведение n экземпляров прореженного элементарного сигнального множества A, то есть $B_1 = \{0, 2, \ldots, q-2\}^n$. Как известно, кодер линейного двоичного кода C (с порождающей матрицей $\mathbf{G}$) может быть представлен посредством решетчатой диаграммы (code trellis), состоящей из n ярусов. Принимая во внимание, что внутренний код есть прямая сумма кода C и подмножества B_1, получаем, что кодер внутреннего кода также может быть представлен в виде решетчатой диаграммы, состоящей из n ярусов. Эта диаграмма получается из решетчатой диаграммы кода C путем введения m параллельных ребер для каждой пары смежных состояний. Ребра диаграммы маркируются символами из элементарного сигнального множества A, при этом, как следует из (8), параллельные ребра, соединяющие пару смежных состояний, маркируются символами либо из множества $\{0, 2, \ldots, q-2\}$ либо из множества $\{1, 3, \ldots, q-1\}$.

Представление кодера внутреннего кода в виде решетчатой диаграммы, состоящей из n ярусов, оказывается полезным при реализации процедуры декодирования по максимуму правдоподобия (МП). Пусть $\mathbf{b} = (b^{(1)}, b^{(2)}, \ldots, b^{(n)})$ — кодовое слово внутреннего кода. В процессе записи в n ячеек флэш-памяти это слово преобразуется в вектор $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$, где $x^{(i)} = I(b^{(i)}), \quad 1 \le i \le n.$ Пусть $\mathbf{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(n)})$ — вектор считываемых значений из n ячеек памяти. В соответствии с рассматриваемой моделью многоуровневой флэш-памяти, плотность распределения вероятности вектора $\mathbf{y}$ при условии записи вектора $\mathbf{x}$ представляет собой произведение n одномерных гауссовских ф.п.в. вида (1). Исходя из (1) и независимости компонентов вектора $\mathbf{y}$, декодер МП, получив вектор $\mathbf{y}$, присваивает каждому ребру решетчатой диаграммы внутреннего кода свой вес, а именно: ребру i-го яруса, $1 \le i \le n$, маркированному символом z, присваивается вес $(y^{(i)} - I(z))^2 + 2\sigma(z)^2 \ln(\sigma(z))$. Затем в размеченной таким образом

решетчатой диаграмме отыскивается путь с минимальным суммарным весом, который и определяет наиболее правдоподобное слово внутреннего кода.

4. Введение внешнего кодирования. Будем полагать, что внутренний код B_0 с минимальным евклидовым расстоянием Δ_0 используется для кодирования одного символа кодового слова внешнего кода Рида — Соломона. В этом случае, очевидно, блок **u** , кодируемый кодом B_0, есть двоичное представление одного символа кодового слова внешнего кода Рида — Соломона. Учитывая, что блок **u** состоит из $mn - r_0$ символов, максимально возможная длина внешнего кода Рида — Соломона составляет 2^{mn-r_0} . Отображение кодового символа внешнего кода в кодовое слово внутреннего кода **b** производится в соответствии с правилом (8). Плотность записи при таком варианте внешнего кодирования составляет $(m - r_0 / n)R$, бит/ячейка, где R — скорость кода Рида — Соломона. В таблице 1 приведены параметры ряда каскадных конструкций с внешним кодом Рида — Соломона.

Таблица 1. Параметры каскадных конструкций

Число уровней записи в ячейке q	Размерность сигнального множества n	Исходная решетка Λ_0	r_0	Объем внутреннего кода $\lvert B_0 \rvert$	Максимальная длина кода Рида—Соломона	Плотность записи, бит/ячейка R_d
4	5	$\mathbf{Z}^5$	0	1024	1024	$2R$
4	4	$\mathbf{Z}^4$	0	256	256	$2R$
4	5	D_5	1	512	512	$(9/5)R = 1.8R$
4	4	D_4	1	128	128	$(7/4)R = 1.75R$
4	7	E_7	3	2048	2048	$(11/7)R = 1.57R$
4	8	E_8	4	4096	4096	$(3/2)R = 1.5R$

5. Анализ помехоустойчивости. Основная проблема, возникающая при анализе помехоустойчивости предложенных каскадных схем кодирования, связана с необходимостью как можно более точного вычисления вероятности ошибки декодирования внутреннего кода. В силу того, что эта вероятность, как правило, сравнительно велика, использование стандартных верхних границ для модели канала записи/считывания с ID-AGN может привести, с учетом требуемой результирующей вероятности ошибки порядка 10^{-12} … 10^{-16}, либо к существенному занижению реально достигаемой плотности записи, либо к излишне завышенной сложности каскадного

кодирования. В этой связи далее в п. 5.1 приводится новый подход к оцениванию вероятности ошибки декодирования, основанная на *точном* вычислении попарной (кода из двух слов) вероятности ошибки для модели канала записи/считывания с ID-AGN. Развиваемый в этой части подход применим для любого канала без памяти.

5.1 Вероятность ошибки декодирования внутреннего кода. Будем полагать, что решение относительно слов внутреннего кода принимается по МП, то есть, как $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x})$, где $p(\mathbf{y} \mid \mathbf{x})$ — n-мерная условная ф.п.в., задающая распределение считанных значений $\mathbf{y} = (y^{(1)}, ..., y^{(n)})$ ячейки флэш-памяти при условии, что была записана последовательность уровней $\mathbf{x}$. Для рассматриваемой модели можно записать, что:

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \mid \mathbf{x}) = \prod_{l=1}^{n} p_{y|x}(y^{(l)} \mid x^{(l)}), \tag{9}$$

где ф.п.в. $p_{y|x}(y \mid x)$ определена равенством (1). Вероятность ошибки при считывании некоторой записанной последовательности $\mathbf{x}$ может быть ограничена сверху с использованием аддитивного неравенства как:

$$P_e(\mathbf{x}) \le \sum_{\mathbf{x}' \neq \mathbf{x}} P_e(\mathbf{x} \to \mathbf{x}' \mid \mathbf{x}), \tag{10}$$

где

$$P_e(\mathbf{x} \to \mathbf{x}' \mid \mathbf{x}) = \Pr\left[p(\mathbf{y} \mid \mathbf{x}') > p(\mathbf{y} \mid \mathbf{x}) \,\middle|\, \mathbf{x} \right] \tag{11}$$

— вероятность ошибки декодирования для кода из двух слов $\mathbf{x}$ и $\mathbf{x}'$ при условии, что было передано (записано) слово $\mathbf{x}$. Вероятность (11) с использованием равенства (9) может быть записана в эквивалентной форме $P_e(\mathbf{x} \to \mathbf{x}' \mid \mathbf{x}) = \Pr\left[Z(\mathbf{y} \mid \mathbf{x}, \mathbf{x}') > 0 \;\middle|\; \mathbf{x} \right]$, где**:**

$$Z(\mathbf{y} \mid \mathbf{x}, \mathbf{x}') = \sum_{l=1}^{n} z(y^{(l)} \mid x^{(l)}, x'^{(l)}),$$

и

$$z(y \mid x, x') = \ln\left(p_{y|x}(y \mid x') / p_{y|x}(y \mid x) \right). \tag{12}$$

В дальнейшем используется следующее утверждение, доказательство которого дано в Приложении.

Утверждение. Пусть Z — вещественная случайная величина и $w_Z(\cdot)$ ее ф.п.в. Пусть $C_Z(\omega)$ — характеристическая функция случайной величины Z, то есть:

$$C_Z(\omega) = \overline{e^{j\omega Z}} = \int_{-\infty}^{\infty} e^{j\omega x} w_Z(x) dx. \tag{13}$$

Черта сверху здесь и далее обозначает усреднение. Тогда

$$\Pr[Z > 0] = \frac{1}{\pi} \int_0^{\infty} \mathrm{Re} \frac{C_Z(\alpha - j\beta)}{\beta + j\alpha} d\alpha \ , \ 0 < \beta < \beta_0,$$

где β_0 — максимальное значение $\mathrm{Im}\,\omega$, при котором сходится интеграл (13).

Используя это утверждение, а также то обстоятельство, что канал записи описывается моделью без памяти (см. (9)), можно записать, что:

$$P_e(\mathbf{x} \to \mathbf{x}' \mid \mathbf{x}) = \frac{1}{\pi} \int_0^{\infty} \mathrm{Re} \frac{\prod_{l=1}^{n} c_{z(x^{(l)}, x'^{(l)})}(\alpha - j\beta)}{\beta + j\alpha} d\alpha, \tag{14}$$

где $c_{z(x,x')}(\omega) = \overline{\exp(j\omega z(y \mid x, x'))}$ — характеристическая функция случайной величины $z(y \mid x, x')$, определенной равенством (12), $0 < \beta < \beta_0(\mathbf{x}, \mathbf{x}')$, а верхняя граница $\beta_0(\mathbf{x}, \mathbf{x}')$ следует из условия сходимости интеграла (14).

Вероятность ошибки p_e при декодировании символа внутреннего кода оценивается сверху с использованием аддитивного неравенства (10):

$$\begin{aligned} p_e \le \sum_{\mathbf{x}} \sum_{\mathbf{x}' \ne \mathbf{x}} P_e(\mathbf{x} \to \mathbf{x}' \mid \mathbf{x}) P(\mathbf{x}) = \frac{1}{\pi M} \sum_{\mathbf{x}} \sum_{\mathbf{x}' \ne \mathbf{x}} \int_0^{\infty} \mathrm{Re} \frac{\prod_{l=1}^{n} c_{z(x^{(l)}, x'^{(l)})}(\alpha - j\beta)}{\beta + j\alpha} d\alpha = \\ = \frac{1}{\pi M} \int_0^{\infty} \mathrm{Re} \frac{\sum_{\mathbf{x}} \sum_{\mathbf{x}' \ne \mathbf{x}} \prod_{l=1}^{n} c_{z(x^{(l)}, x'^{(l)})}(\alpha - j\beta)}{\beta + j\alpha} d\alpha \ , \end{aligned} \tag{15}$$

где $P(\mathbf{x})$ — вероятность использования последовательности $\mathbf{x}$. При записи равенства в выражении (15) использовано предположение о

равновероятном использовании последовательностей $\mathbf{x}$, то есть, что $P(\mathbf{x}) = 1/M$. Поскольку $\forall\omega\ c_{z(x,x)}(\omega) = 1$, то:

$$\sum_{\mathbf{x}}\sum_{\mathbf{x}'\neq\mathbf{x}}\prod_{l=1}^{n} c_{z(x^{(l)},x'^{(l)})}(\omega) = \sum_{\mathbf{x}}\sum_{\mathbf{x}'}\prod_{l=1}^{n} c_{z(x^{(l)},x'^{(l)})}(\omega) - \sum_{\mathbf{x}}\prod_{l=1}^{n} c_{z(x^{(l)},x^{(l)})}(\omega) =$$

$$= \sum_{\mathbf{x}}\sum_{\mathbf{x}'}\prod_{l=1}^{n} c_{z(x^{(l)},x'^{(l)})}(\omega) - M.$$

Поэтому можно записать, что

$$p_e \le \frac{1}{\pi M}\int_0^{\infty} \mathrm{Re}\frac{D(\alpha - j\beta) - M}{\beta + j\alpha}\,d\alpha, \tag{16}$$

где

$$D(\omega) = \sum_{\mathbf{x}}\sum_{\mathbf{x}'}\prod_{l=1}^{n} c_{z(x^{(l)},x'^{(l)})}(\omega). \tag{17}$$

При условии, что было передано (записано) значение уровня x, выходная случайная величина (считанное значение) y имеет гауссовское распределение с нулевым математическим ожиданием x и дисперсией $\sigma(x)$. В этом случае можно показать, что при $\sigma(x) \neq \sigma(x')$

$$c_{z(x,x')}(\omega) =$$

$$= \frac{\sigma(x)^{j\omega}\sigma(x')^{1-j\omega}}{\sqrt{j\omega\sigma^2(x) + (1-j\omega)\sigma^2(x')}} \exp\left(-\frac{j\omega(1-j\omega)(x-x')^2}{2(j\omega\sigma^2(x) + (1-j\omega)\sigma^2(x'))}\right), \tag{18}$$

при условиях:

$$\mathrm{Im}\,\omega < \frac{\sigma^2(x')}{\sigma^2(x) - \sigma^2(x')}, \text{ если } \quad \sigma(x) > \sigma(x'), \tag{19}$$

и

$$-\mathrm{Im}\,\omega < \frac{\sigma^2(x')}{\sigma^2(x') - \sigma^2(x)}, \text{ если } \sigma(x) < \sigma(x'). \tag{20}$$

В случае, когда $\sigma(x) = \sigma(x')$, характеристическая функция $c_{z(x,x')}(\omega)$ вычисляется согласно выражению:

$$c_{z(x,x')}(\omega) = \exp\left(-\frac{j\omega(1-j\omega)(x-x')^2}{2\sigma^2(x)} \right). \quad (21)$$

Нетрудно показать, что условия (19) и (20) могут быть преобразованы в условия $0 < \beta < \beta_0(\mathbf{x}, \mathbf{x}')$ для параметра β в равенстве (14), где:

$$\beta_0(\mathbf{x}, \mathbf{x}') = \begin{cases} \infty, \text{ если } \sigma(x^{(l)}) \geq \sigma(x'^{(l)}) \quad \text{для всех } l = 1,\ldots,n, \\ \min\limits_{l:\sigma(x^{(l)}) < \sigma(x'^{(l)})} \dfrac{\sigma^2(x'^{(l)})}{\sigma^2(x'^{(l)}) - \sigma^2(x^{(l)})}. \end{cases} \quad (22)$$

Несмотря на кажущуюся сложность, вычисление верхней границы (16) вероятности p_e оказывается сравнительно простой задачей. Оно сводится к вычислению подынтегрального выражения в правой части (16) с использованием формул (18) и (21) и численного интегрирования. Параметр β в (16) при практических вычислениях следует выбирать с учетом ограничения $0 < \beta < \min_{\mathbf{x},\mathbf{x}'} \beta_0(\mathbf{x}, \mathbf{x}')$ (см. (22)) таким образом, чтобы подынтегральное выражение представляло собой функцию, удобную для численного интегрирования. В частности, подходящим выбором параметра β можно добиться малого числа осцилляций интегрируемой функции в области интегрирования и за счет этого повысить скорость и точность численного интегрирования.

Для иллюстрации вычисления функции $D(\omega)$, определенной равенством (17), рассмотрим примеры.

Пример 1. Пусть множество $B = A^n$ представляет собой подмножество $\mathbf{Z}^n$ с элементами 0,1 ,…,$q-1$. Тогда очевидно, что $D(\omega) = \left(\sum_x \sum_{x'} c_{z(x,x')}(\omega) \right)^n$ для любого значения аргумента ω.

Пример 2. Пусть множество $B = A^n$, а $B_0 = B \cap D_n$ то есть представляет собой подмножество решетки D_n с элементами 0,1,…,$q-1$. Тогда множество всех векторов $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ представляет собой множество векторов с компонентами из множества $\{x_0, x_1, \ldots, x_{q-1}\}$, сумма индексов которых четна. Такое

множество может быть представлено в виде графа (кодовой решетки). Каждый путь в этом графе из начального узла (начального состояния) в конечный узел (конечное состояние) соответствует одной из кодовых последовательностей. Пример для $n = 5$, $q = 4$ показан на рисунке 3а.

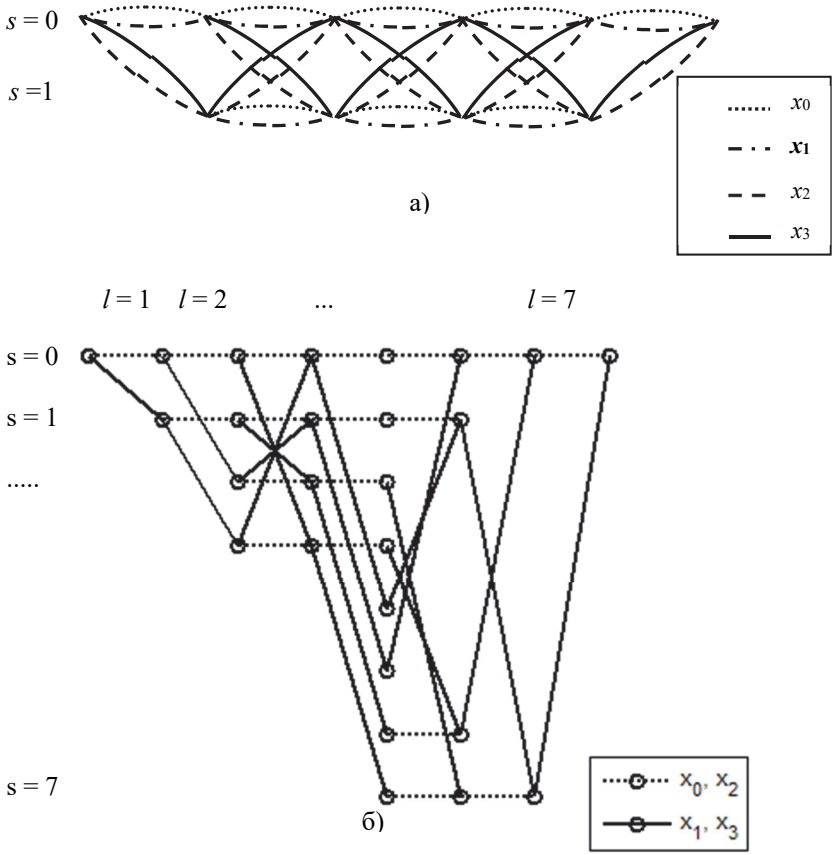


Рис. 3. Кодовые решетки внутреннего кода: а) на основе решетки D_5; б) на основе решетки E_7

Чтобы вычислить функцию $D(\omega)$, определенную равенством (17), надо рассмотреть все пары кодовых последовательностей или все возможные пары путей в решетке,

показанной на рисунке 3а. Очевидно, что две различные последовательности внутреннего кода определяются двумя последовательностями состояний решетки (рисунок 3а) $(0, s^{(1)}, ..., s^{(n-1)}, 0)$ и $(0, s'^{(1)}, ..., s'^{(n-1)}, 0)$, или одной последовательностью индексов $(0, i^{(1)}, ..., i^{(n-1)}, 0)$, где $i^{(l)} = 2s^{(l)} + s'^{(l)}$. Пусть s и s' — пара смежных состояний в решетке, $s, s' = 0,1$, и $\chi(s, s')$ — множество меток, соответствующих переходам из s в s'; например, $\chi(1,0) = \{x_1, x_3\}$. Рассмотрим две пары переходов $s_a \to s'_a$, $s_b \to s_b'$ и определим индексы i, j как $i = 2s_a + s_b$ и $j = 2s_a' + s_b'$, при этом $i, j = 0,1,2,3$. Далее введем в рассмотрение величины:

$$P_{ij}(\omega) = \sum_{x \in \chi(s_a, s_a')} \sum_{x' \in \chi(s_b, s_b')} c_{z(x,x')}(\omega). \tag{23}$$

Тогда нетрудно заметить, что:

$$D(\omega) = \mathbf{a}(\omega)\mathbf{P}(\omega)^{n-2}\mathbf{b}(\omega), \tag{24}$$

где $\mathbf{P}(\omega) = [P_{ij}(\omega)]$, $i, j = 0,1,2,3$, матрица порядка 4, и $\mathbf{a}(\omega) = [P_{00}(\omega)P_{01}(\omega)P_{02}(\omega)P_{03}(\omega)]$, $\mathbf{b}(\omega) = [P_{00}(\omega)P_{10}(\omega)P_{20}(\omega)P_{30}(\omega)]^T$. Заметим, что этот пример очевидным образом обобщается и на более сложные кодовые решетки, например на решетку E_7, рассмотренную в следующем примере.

Пример 3. Пусть множество $B = A^7$, а $B_0 = B \cap E_7$, то есть представляет собой подмножество решетки E_7 с элементами 0,1,…,q – 1. Пусть $\mathbf{i} = (i_1, i_2, ..., i_7)$ — набор индексов, задающих один из векторов $\mathbf{x}(\mathbf{i}) = (x_{i_1}, x_{i_2}, ..., x_{i_7})$, где $i_l \in \{0,1,...,q-1\}$, и следовательно, $x_{i_l} \in \{x_0, ..., x_{q-1}\}$. Тогда множество всех векторов $\mathbf{x}(\mathbf{i})$, построенных на основе решетки E_7, представляет собой множество векторов с компонентами из множества $\{x_0, x_1, \ldots, x_{q-1}\}$, таких, что $\mathbf{i} \bmod 2 = \mathbf{c}$, где $\mathbf{c}$ — слово двоичного кода Хэмминга (7,4). Множество слов кода Хэмминга (7,4) возможно представить в виде графа, который может быть построен как синдромная решетка [17] (см. рисунок 3б). Далее будем считать, что, как и ранее, $q = 4$. Если ребра, показанные пунктиром на рисунке 3б,

заменить парой параллельных ребер, отмеченных символами x_0 и x_2, а ребра, показанные сплошными линиями, — парой параллельных ребер, отмеченных символами x_1 и x_3, то получится решетка, соответствующая словам внутреннего кода. Каждый путь в этой решетке, соединяющий начальный узел с конечным, соответствует одной из кодовых последовательностей. Нетрудно увидеть, что внутренний код содержит 2048 последовательностей.

Пусть s и s' — пара смежных состояний в решетке, показанной на рисунке 3б, $s, s' = 0,1,...7,$ и $\chi_l(s,s')$ — множество меток, соответствующих переходам из s в s' на l-ом уровне, $l = 1,...7$; например, $\chi_2(1,1) = \{x_0, x_2\}$, $\chi_3(3,7) = \varnothing$. По аналогии с (23) определим величины:

$$P_{ij}^{(l)}(\omega) = \sum_{x \in \chi_l(s_a, s_a')} \sum_{x' \in \chi_l(s_b, s_b')} c_{z(x,x')}(\omega),$$

где $i = 8s_a + s_b$ и $j = 8s_a' + s_b'$. Тогда, очевидным образом обобщая равенство (24), можно записать, что $D(\omega) = \mathbf{a}(\omega)\mathbf{P}_2(\omega)...\mathbf{P}_6(\omega)\mathbf{b}(\omega)$, где $\mathbf{P}_l(\omega) = [P_{ij}^{(l)}(\omega)]$, $i, j = 0,1,..,63,$ — матрица порядка 64, $\mathbf{a}(\omega) = [P_{00}^{(1)}(\omega),..., P_{0\,63}^{(1)}(\omega)]$, $\mathbf{b}(\omega) = [P_{00}^{(7)}(\omega),..., P_{63\,0}^{(7)}(\omega)]^T$.

На рисунке 4 показаны примеры зависимостей вероятности ошибки декодирования символа внутреннего кода p_e от числа циклов перезаписи N. Для указанных примеров скорость кодирования (плотность записи) составляет 2 бит/ячейка для безызбыточного кода, построенного на основе решетки $\mathbf{Z}^5$, 1.8 бит/ячейка для внутреннего кода, построенного на основе решетки D_5, и 1.57 бит/ячейка для внутреннего кода, построенного на основе решетки E_7.

Из представленных данных следует, что приемлемая вероятность ошибки не может быть достигнута при использовании только внутреннего кодирования.

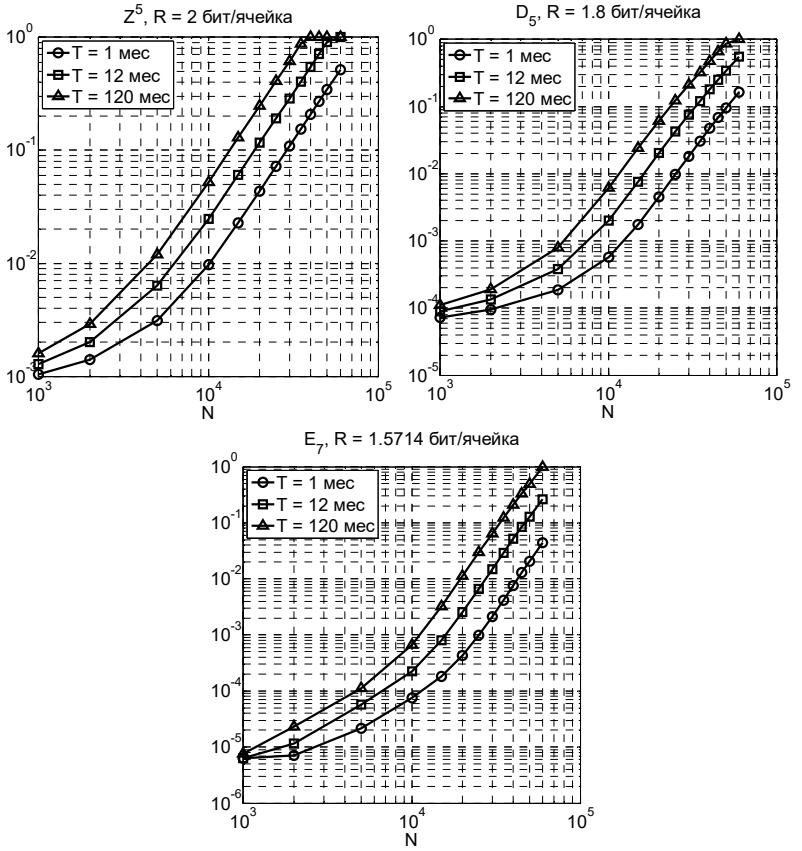


Рис. 4. Вероятность ошибки декодирования p_e символа внутреннего кода, построенного на основе решеток $\mathbf{Z}^5$, D_5 и E_7

5.2. Вероятность ошибки декодирования внешнего кода. Численные результаты. Рассмотрим каскадную конструкцию с одним компонентным кодом, в качестве которого используется *расширенный* код Рида — Соломона. Примеры таких конструкций перечислены в таблице 1. Вероятность ошибки декодирования блокового кода, исправляющего t ошибок при использовании в канале без памяти, оценивается очевидным образом как:

$$P_e \le \sum_{l=t+1}^{N_1} C_{N_1}^l p_e^l (1-p_e)^{N_1-l},$$

где N_1 — длина внешнего кода, p_e — вероятность ошибочного приема символа внешнего кода (вероятность ошибочного решения относительно слова внутреннего кода). При использовании в качестве внутреннего кода последовательностей, построенных на основе некоторой решетки, вероятность p_e можно оценить с помощью неравенства (16). Вероятность ошибки декодирования символа внешнего кода P_s может быть получена из вероятности ошибки декодирования слова внешнего кода P_e как $P_s \approx (d_1 / N_1)P_e$, где d_1 — минимальное расстояние внешнего кода. Это приближенное равенство основано на том наблюдении, что наиболее вероятным будет ошибочное декодирование в пользу ближайшего слова кода, то есть отличающегося от истинного в d_1 символах. Полагая в первом приближении, что при этом примерно половина битов будет декодирована неправильно, получаем, что вероятность ошибки на бит может быть оценена как $P_b \approx (d_1 / 2N_1)P_e$. Далее рассмотрим несколько конкретных примеров каскадной конструкции (они приведены в таблице 1):

1) внутренний код построен на основе решетки $\mathbf{Z}^5$, (безызбыточный код), а внешний — это код Рида — Соломона с параметрами N_1=1024, K_1=1014,1012,1010,..., $d_1 = N_1 - K_1 + 1$=11,13,15,...;

2) внутренний код построен на основе решетки D_5, а внешний — это код Рида — Соломона с параметрами N_1=512, K_1=502,500,498,..., $d_1 = N_1 - K_1 + 1$=11,13,15,...;

3) внутренний код построен на основе решетки E_7, а внешний — это код Рида — Соломона с параметрами N_1=2048, K_1=2038, 2036,..., $d_1 = N_1 - K_1 + 1$=11,13,15...

На рисунке 5 представлены графики вероятности ошибки на бит в зависимости от числа циклов перезаписи N для некоторых примеров кодов Рида — Соломона из списка рассмотренных вариантов. Показаны графики для примеров, которые обеспечивают при сравнительно небольшом числе циклов перезаписи вероятность ошибки на бит $< 10^{-12}$. Внешнее кодирование для этих примеров вызывает снижение плотности записи приблизительно на 3-4%. Рассмотрев полученные данные, можно сделать вывод, что при ухудшении характеристик канала из-за увеличения времени хранения данных T и/или числа циклов перезаписи N возможно изменение параметров схемы каскадного кодирования (адаптация) с тем, чтобы сохранить требуемый уровень надежности хранения данных. Очевидный путь адаптации состоит в увеличении избыточности внешнего кода по мере старения физического носителя. Такое

изменение параметров внешнего кода приводит к уменьшению его скорости и, следовательно, к уменьшению скорости всей каскадной схемы, то есть к снижению плотности записи.

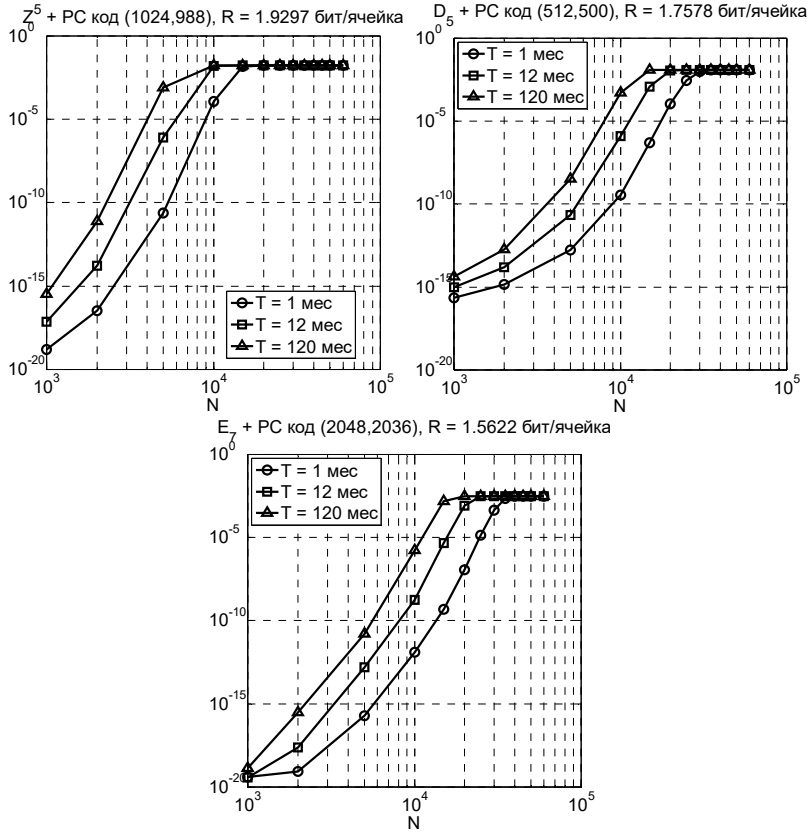


Рис. 5. Вероятность ошибки на бит для каскадных кодов $\mathbf{Z}^5$ +(1024, 988) , D_5 + (512,498) и E_7 + (2048,2036) в зависимости от числа циклов перезаписи

На рисунке 6 показаны зависимости скорости кодирования рассмотренной каскадной схемы в зависимости от числа циклов перезаписи при условии, что вероятность ошибки на бит не превышает 10^{-12}.

Из приведенных на рисунке 6 данных следует, что в области сравнительно малых значений числа циклов перезаписи N предпочтительным оказывается применение безызбыточного

внутреннего кода, построенного с использованием решетки $\mathbf{Z}^5$, и внешнего кода с длиной $N_1 = 1024$. Число информационных символов K_1 в этом случае меняется в пределах от 998 до 950.

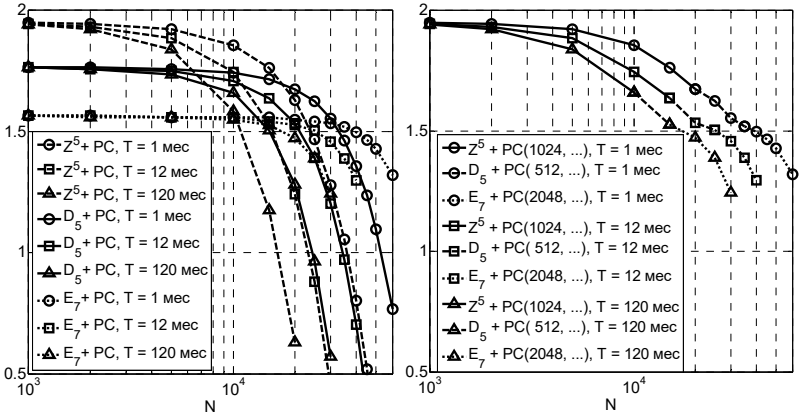


Рис. 6. Изменение скорости кодирования (плотности записи) R_d в зависимости от числа циклов перезаписи N при $P_b \le 10^{-12}$ для каскадных кодов (слева – все примеры, справа – лучшие варианты)

С ростом числа циклов перезаписи N и/или времени хранения данных T каскадная схема, использующая внутренний код, построенный на основе решетки D_5, обеспечивают большую плотность записи. В этом случае при условии, что время хранения данных $T = 1$мес, внешний код имеет параметры K_1 от 462 до 352. При дальнейшем ухудшении канала записи (то есть при увеличении значений T и/или N) лучшие характеристики обеспечивает схема с внутренним кодом, построенным на основе решетки E_7, и внешним кодом с длиной $N_1 = 2048$. Как следует из приведенных кривых, уменьшение скорости кодирования каскадной схемы позволяет сохранить надежность содержания данных на требуемом уровне. В рассмотренных примерах снижение скорости при адаптации составляет от 20%, при умеренном ухудшении параметров физического носителя, до примерно 30…35% для наиболее тяжелого сочетания числа циклов перезаписи и времени хранения данных. Технически такая адаптация может быть реализована в контроллере, реализующем процедуры кодирования/декодирования по мере обнаружения возрастающего числа исправляемых ошибок при декодировании внешнего кода. При

снижении плотности записи ниже заранее определенного уровня контроллер может сообщить, например, о необходимости замены запоминающего устройства и переносе данных на новый носитель, или применить какую-либо иную процедуру реагирования.

6. Заключение. В настоящей работе рассмотрена каскадная схема кодирования для многоуровневой флэш-памяти, внутренняя ступень которой представляет собой конечную подрешетку многоуровневой целочисленной решетки (lattice code), а в качестве внешней ступени используется код Рида — Соломона. Отличительными особенностями рассматриваемой конструкции являются: а) простая схема кодирования внутреннего кода; б) мягкое декодирование внутреннего кода; в) высокая степень гибкости, позволяющая обеспечить широкий диапазон вариантов кодирования в рамках обменного соотношения «плотность записи — вероятность ошибки»; г) возможность адаптации реализуемой плотности записи при увеличении времени хранения и/или числа циклов перезаписи. Введение внутреннего кодирования сопровождается некоторым снижением плотности записи, позволяя одновременно заметно повысить надежность хранения данных за счет декодирования слов внутреннего кода по максимуму правдоподобия с использованием мягких решений. Дальнейшее существенное повышение надежности хранения данных достигается за счет применения внешнего кодирования. В отличие от известных работ, связанных с помехоустойчивым кодированием для идеализированной модели флэш-памяти, в данной работе рассматривается более реалистичная модель ячейки флэш-памяти — с неравномерно расположенными целевыми уровнями напряжения в ячейке и дисперсией шума, зависящей от записанного значения (input-dependent additive Gaussian noise, ID-AGN).

В работе был предложен новый подход, позволяющий оценить вероятность ошибки декодирования. Этот подход развит применительно к модели канала с аддитивным гауссовским шумом с дисперсией, зависящей от записанного значения (input-dependent additive Gaussian noise). В нем для вычисления верхней границы вероятности ошибки декодировании слов внутреннего кода используются произведения характеристических функции соответствующих случайных величин. Показано, что вычисление этой верхней границы сводится к численному интегрированию по одной переменной некоторого подынтегрального выражения, зависящего от параметров канала и от структуры внутреннего кода.

С использованием этого подхода проанализирована кодовая конструкция, в которой внутренний код построен на основе решеток малой размерности, а на внешней ступени использовался кода Рида —

Соломона. В качестве конкретных примеров рассмотрены решетки $\mathbf{Z}^n$, D_n, $n=5$, и E_7, а внешний код представлял собой расширенный код Рида — Соломона длиной 1024, 512 и 2048 соответственно. В результате удалось построить гибкую схему каскадного кодирования с возможной адаптацией по скорости кода (плотности записи) в зависимости от ухудшения канала, возникающего с ростом числа циклов перезаписи и времени хранения данных. В рассмотренных примерах получено, что требуемый уровень вероятности ошибки на бит 10^{-12} достигается при скорости каскадного кода от 1.93 для сравнительно хороших условий записи до 1.30 бит/ячейка для наихудших сочетаний числа циклов перезаписи и времени хранении данных для кодов с длиной 128 и 512 соответственно. Дальнейшее увеличение плотности записи возможно при увеличении длины внутреннего и внешнего кодов.

Приложение. Формула для вероятности $\Pr[Z>0]$. Пусть Z — непрерывная вещественная случайная величина и $w_Z(\cdot)$ ее ф.п.в. Требуется вычислить вероятность $P=\Pr[Z>0]$. Очевидно, что:

$$P=\int_0^\infty w_Z(x)dx=\int_{-\infty}^{\infty} e(x)w_Z(x)dx=\overline{e(Z)}, \qquad (\text{П1})$$

где $e(x)$ — функция единичного скачка,

$$e(x)=\begin{cases}1, & x>0\\ 0, & x\le 0\end{cases},$$

а черта сверху означает усреднение. Обозначим $C_Z(\omega)$ характеристическую функцию случайной величины Z:

$$C_Z(\omega)=\overline{e^{j\omega Z}}=\int_{-\infty}^{\infty} e^{j\omega x}w_Z(x)dx. \qquad (\text{П2})$$

Далее рассмотрим подход, позволяющий вычислить вероятность P косвенным образом с использованием характеристической функции $C_Z(\omega)$. Этот подход близок к рассмотренному в [18].

Пусть $f(x)$ и $\varphi(x)$ — некоторые функции, такие что $f(x)\varphi(x)=e(x)$. Тогда можно записать, что:

$$P=\int_{-\infty}^{\infty} f(x)\varphi(x)w_Z(x)dx\,. \qquad (\text{П2})$$

Обозначим через $F(\alpha)$ преобразование Фурье функции $f(x)$, $F(\alpha)=\int_{-\infty}^{\infty} f(x)\exp(-j\alpha x)dx,$ и следовательно, $f(x)=(2\pi)^{-1}\int_{-\infty}^{\infty} F(\alpha)\exp(j\alpha x)d\alpha.$ Тогда выражение (П1) можно переписать в следующем виде:

$$P=\int_{-\infty}^{\infty} F(\alpha)\Phi(\alpha)d\alpha, \qquad (П3)$$

где $\Phi(\alpha)=(2\pi)^{-1}\int_{-\infty}^{\infty}\varphi(x)w_Z(x)e^{j\alpha x}dx.$ Равенство (П3) дает общее выражение для вероятности *P*. Чтобы привести его к удобному для использования виду, рассмотрим некоторый частный случай назначения функций $f(x)$ и $\varphi(x)$. Положим $\varphi(x)=\exp(\beta x)$, $\beta>0$, и

$$f(x)=\begin{cases} e^{-\beta x}, & x>0 \\ 0, & x\le 0 \end{cases}.$$

Очевидно, что $f(x)\varphi(x)=e(x).$ Тогда:

$$F(\alpha)=\int_{-\infty}^{\infty} f(x)e^{-j\alpha x}dx=\int_{0}^{\infty} e^{-\beta x}e^{-j\alpha x}dx=\frac{1}{\beta+j\alpha}, \qquad (П4)$$

а

$$\Phi(\alpha)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\beta x}e^{j\alpha x}w_Z(x)dx=\frac{1}{2\pi}\overline{e^{(\beta+j\alpha)Z}}=\frac{1}{2\pi}C_Z(\alpha-j\beta), \qquad (П5)$$

где $C_Z(\cdot)$ — характеристическая функция случайной величины *Z*, определенная равенством (П2). Комбинируя выражения (П3)-(П5), получаем, что:

$$P=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{C_Z(\alpha-j\beta)}{\beta+j\alpha}d\alpha. \qquad (П6)$$

Равенство (П6) справедливо для любых значений параметра β, $0<\beta<\beta_0$. Здесь первое неравенство обеспечивает сходимость

интеграла (П4), а значение β_0 во втором неравенстве — это максимальное значение, при котором сходится интеграл (П5).

Поскольку $P = \mathrm{Re}\, P$, то из (П6) следует, что:

$$P = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{Re} \frac{C_Z(\alpha - j\beta)}{\beta + j\alpha} d\alpha. \qquad (П7)$$

Нетрудно показать, что:

$$\mathrm{Re} \frac{C_Z(\alpha - j\beta)}{\beta + j\alpha} = \frac{\beta \overline{e^{\beta Z} \cos \alpha Z}}{\beta^2 + \alpha^2} + \frac{\alpha \overline{e^{\beta Z} \sin \alpha Z}}{\beta^2 + \alpha^2}$$

и представляет собой четную функцию от переменной α. Поэтому в (П7) можно заменить пределы интегрирования $(-\infty, +\infty)$ на $(0, +\infty)$ и удвоить результат, то есть записать, что:

$$P = \frac{1}{\pi} \int_{0}^{\infty} \mathrm{Re} \frac{C_Z(\alpha - j\beta)}{\beta + j\alpha} d\alpha, \quad 0 < \beta < \beta_0. \qquad (П8)$$

Вычисление вероятности P согласно выражению (П8) сводится к численному интегрированию по одной переменной и в ряде случаев оказывается более простым, чем прямое вычисление по формуле (П1).

Литература

1. *Micheloni R., Crippa L.* Multi-bit NAND flash memories for ultra-high density storage devices // Advances in Non-Volatile Memory and Storage Technology. 2014. pp. 75–119.
2. *Naso G. et al.* A 128 Gb 3b/cell NAND flash design using 20nm planar-cell technology // IEEE International Solid-State Circuits Conference Digest of Technical Papers. 2013. pp. 218–219.
3. *Im J.W. et al.* A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate // Proceedings of ISSCC. 2015. pp. 1–3.
4. *Aritome S.* NAND Flash Memory Technologies // John Wiley & Sons. 2016. 410 p.
5. *Yaakobi E. et al.* Error correcting coding for flash memories // Flash Memory Summit. 2009. vol. 31. pp. 4–10.
6. *Lou H.-L., Sundberg C.-E.* Coded modulation to increase storage capacity of multilevel memories // Proceedings of the IEEE Globecom (Globecom'1998). 1998. pp. 3379–3384.
7. *Sun F. et al.* Multilevel flash memory on-chip error correction based on trellis coded modulation // Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2006). 2006. pp. 1443–1446.
8. *Li S., Zhang T.* Improving multi-level NAND flash memory storage reliability using concatenated BCH-TCM coding // IEEE Trans. on VLSI Systems. 2010. vol. 18. no. 10. pp. 1412–1420.
9. *Xu Q., Gong P., Chen T.M.* Concatenated LDPC-TCM coding for reliable storage in multi-level flash memories // Proceedings of the 9th International Symposium On

Communication System, Networks & Digital Signal Processing (CSNDSP 2014). 2014. pp. 166–170.
10. *Kurkoski B.M.* Coded modulation using lattices and Reed-Solomon codes, with applications to flash memories // IEEE Trans. on Selected Areas in Communications. 2014. vol. 32. no. 5. pp. 900–908.
11. *Wang X., Dong G., Pan L., Zhou R.* Error correction codes and signal processing in flash memory // Flash Memories. URL: http://www.intechopen.com/books/flash-memories/error-correction-codes-and- signal-processing-in-flash-memory (дата обращения: 26.10.2014).
12. *Dong G. et al.* Estimating information- theoretical NAND flash memory storage capacity and its implication to memory system design space exploration // IEEE Trans. Very Large Scale Integration (VLSI) Systems. 2012. vol. 20. no. 9. pp. 1705–1714.
13. *Huang X. et al.* Multilevel Flash Memories: Channel modeling, Capacities and Optimal Coding Rates // International Journal on Advances in Systems and Measurement. 2013. vol. 6. no. 3, 4. pp. 364–373. URL: http://www.iariajournals.org/systems_and_measurements/sysmea_v6_n34_2013_paged.pdf (дата обращения: 26.10.2014).
14. *Sun F., Rose K., Zhang T.* On the Use of Strong BCH Codes for Improving Multilevel NAND Flash Memory Storage Capacity // URL: http://www.researchgate.net/publication/254376882_On_the_Use_of_Strong_BCHCodes for Improving_Multilevel NAND Flash_ Memory Storage Capacity (дата обращения: 26.10.2014).
15. *Трофимов А.Н., Таубин Ф.А.* Теоретико-информационный анализ многоуровневой flash памяти. Часть 1: Модель канала и границы случайного кодирования // Информационно-управляющие системы. 2016. Т. 81. № 2. С. 49–59.
16. *Forney G.D.* Coset Codes – Part 1: Introduction and geometrical classification // IEEE Trans. on Information Theory. 1988. vol. 34. no. 5. pp. 1123-1151.
17. *Bahl L.R., Cocke J., Jelinek F., Raviv J.* Optimal decoding of linear codes for minimum symbol error rate // IEEE Trans. on Information Theory. 1974. vol. 20. no. 2. pp. 284–287.
18. *Trofimov A.N.* Modified Chernoff bound and some applications / Krouk E., Semenov S. (Eds.) // Modulation and Coding Techniques in Wireless Communications. Chichester, West Sussex, UK: Wiley. 2011. pp. 206–220.

Таубин Феликс Алексаедрович — д-р техн. наук, профессор, профессор кафедры аэрокосмических компьютерных и программных систем института аэрокосмических приборов и систем, Санкт-Петербургский государственный университет аэрокосмического приборостроения (СПбГУАП). Область научных интересов: цифровые системы связи, методы помехоустойчивого кодирования, широкополосные системы, беспроводная связь. Число научных публикаций — 92. ftaubin@yahoo.com; ул. Большая Морская, 67, Санкт-Петербург, 190000; р.т.: +7(812)494-70-51, Факс: +7(812)494-70-51.

Трофимов Андрей Николаевич — к-т техн. наук, доцент, доцент кафедры инфокоммуникационных систем и кафедры безопасности информационных систем института информационных систем и защиты информации, Санкт-Петербургский государственный университет аэрокосмического приборостроения (СПбГУАП). Область научных интересов: теория цифровой связи, теория информации, теория корректирующего кодирования. Число научных публикаций — 67. andrei.trofimov@vu.spb.ru; ул. Большая Морская, 67, Санкт-Петербург, 190000; р.т.: +7(812)494-70-52, Факс: +7(812)494-70-52.

Поддержка исследований. Работа выполнена при поддержке Министерства образования и науки Российской Федерации при выполнении научно-исследовательской работы в рамках проектной части государственного задания в сфере научной деятельности по заданию № 2.2716.2014/К от 17.07.2014 и в рамках базовой части государственного задания в сфере научной деятельности на 2017-2019 гг. по заданию № 2.9214.2017/БЧ.

DOI 10.15622/sp.57.4

F.A.TAUBIN, A.N.TROFIMOV

CONCATENATED REED–SOLOMON/LATTICE CODING FOR MULTILEVEL FLASH MEMORY

Taubin F.A.,Trofimov A.N. **Concatenated Reed–Solomon/Lattice Coding for Multilevel Flash Memory.**

Abstract. The article considers concatenated coding scheme for multilevel flash memory. In this scheme the inner stage is a finite subset of a multidimensional lattice (lattice code) and the outer stage uses Reed–Solomon code.

Performance analysis is done for a model characterizing the basic physical features of a flash memory cell with non-uniform target voltage levels and noise variance dependent on the recorded value (input-dependent additive Gaussian noise, ID-AGN). For this model we develop a new approach to evaluating the error probability for the inner code. This approach is based on one-dimensional numerical integration of product of the characteristic functions of random variables used in the decoding process. It is shown how the parameters of the concatenated coding scheme can be adapted to keep the required error probability when the retention period and/or number of program-erasure cycles increase.

Keywords: concatenated code, multidimensional lattices, Reed-Solomon code, multilevel flash memory, decoding error probability.

Taubin Feliks Aleksandrovich — Dr. Sci., professor, professor of aerospace computer and software systems department of Institute of Aerospace Instruments and Systems, Saint Petersburg State University of Aerospace Instrumentation (SUAI). Research interests: communication theory, error-correcting coding, spread spectrum systems, wireless communication. The number of publications — 92. ftaubin@yahoo.com; 67, B. Morskaia St., 190000, St.-Petersburg, Russia; office phone: +7(812)494-70-51, Fax: +7(812)494-70-51.

Trofimov Andrey Nikolaevich — Ph.D., associate professor, associate professor of information and communication systems department and information systems security department of Institute of Information Systems and Information Security, Saint Petersburg State University of Aerospace Instrumentation (SUAI). Research interests: communication theory, error-correcting coding, information theory. The number of publications — 67. andrei.trofimov@vu.spb.ru; 67, B. Morskaia St., 190000, St.-Petersburg, Russia; office phone: +7(812)494-70-52, Fax: +7(812)494-70-52.

Acknowledgements. This research is supported by the Ministry of Education and Science of the Russian Federation in the framework of the project part of the government task in the field of scientific activity on assignment No. 2.2716.2014/K 17.07.2014, and within the framework of the government task in the field of scientific activity for 2017–2019 on assignment No. 2.9214.2017/БЧ.

References

1. Micheloni R., Crippa L. Multi-bit NAND flash memories for ultra-high density storage devices. Advances in Non-Volatile Memory and Storage Technology. 2014. pp. 75–119.
2. Naso G. et al. A 128Gb 3b/cell NAND flash design using 20nm planar-cell technology. IEEE International Solid-State Circuits Conference Digest of Technical Papers. 2013. pp. 218–219.

3. Im J.W. et al. A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate. Proceedings of ISSCC. 2015. pp. 1–3.
4. Aritome S. NAND Flash Memory Technologies. John Wiley & Sons. 2016. 410 p.
5. Yaakobi E. et al. Error correcting coding for flash memories. Flash Memory Summit. 2009. vol. 3. pp. 4–10.
6. Lou H.-L., Sundberg C.-E. Coded modulation to increase storage capacity of multilevel memories. Proceedings of the IEEE Globecom (Globecom'1998). 1998. pp. 3379–3384.
7. Sun F. et al. Multilevel flash memory on-chip error correction based on trellis coded modulation. Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2006). 2006. pp. 1443–1446.
8. Li S. and Zhang T. Improving multi-level NAND flash memory storage reliability using concatenated BCH-TCM coding. *IEEE Trans. on VLSI Systems*. 2010. vol. 18. no. 10. pp. 1412–1420.
9. Xu Q., Gong P., Chen T.M. Concatenated LDPC-TCM coding for reliable storage in multi-level flash memories. Proceedings of the 9th International Symposium on Communication System, Networks & Digital Signal Processing (CSNDSP). 2014. pp. 166–170.
10. Kurkoski B.M. Coded modulation using lattices and Reed-Solomon codes, with applications to flash memories. *IEEE Trans. on Selected Areas in Communications*. 2014. Vol. 32. no. 5. pp. 900–908.
11. Wang X., Dong G., Pan L., Zhou R. Error correction codes and cignal processing in flash memory. *Flash Memories*. Available at: http://www.intechopen.com/books/flash-memories/error-correction-codes-and-signal-processing-in-flash-memory (accessed: 26.10.2014).
12. Dong G. et al. Estimating information- theoretical NAND flash memory storage capacity and its implication to memory system design space exploration. *IEEE Trans. Very Large Scale Integration (VLSI) Systems.* 2012. vol. 20. no. 9. pp. 1705–1714.
13. Huang X. et al. Multilevel flash memories: Channel modeling, capacities and optimal coding rates. *International Journal on Advances in Systems and Measurement*. 2013. vol. 6. no. 3, 4. pp. 364–373. Available at: http://www.iariajournals.org/systems_and_measurements/sysmea_v6_n34_2013_paged.pdf (accessed: 26.10.2014).
14. Sun F., Rose K., Zhang T. On the Use of Strong BCH Codes for Improving Multilevel NAND Flash Memory Storage Capacity. Available at: http://www.researchgate.net/publication/254376882_On_the_Use_of_Strong_BCHCodes for Improving_Multilevel NAND Flash_ Memory Storage Capacity (accessed: 26.10.2014).
15. Trofimov A.N., Taubin F.A. [Information theory analysis of multilevel flash memory. Part 1: Channel model and random coding bounds]. *Informacionno-upravljajushhie sistemy – Information and control systems.* 2016. vol. 81. no. 2. pp. 49–59. (In Russ.).
16. Forney G.D. Coset Codes – Part 1: Introduction and geometrical classification. *IEEE Trans. on Information Theory*. 1988. vol. 34. no. 5. pp. 1123–1151.
17. Bahl L.R., Cocke J., Jelinek F., Raviv J. Optimal decoding of linear codes for minimum symbol error rate. *IEEE Trans. on Information Theory.* 1974. vol. 20. no. 2. pp. 284–287.
18. Trofimov A.N. Modified Chernoff bound and some applications. Krouk E., Semenov S. (Eds.). Modulation and Coding Techniques in Wireless Communications. Chichester, West Sussex, UK: Wiley. 2011. pp. 206–220.

УДК 519.246.8 **DOI** 10.15622/sp.57.5

А.В. Воробьев, Г.Р. Воробьева

ИНДУКТИВНЫЙ МЕТОД ВОССТАНОВЛЕНИЯ ВРЕМЕННЫХ РЯДОВ ГЕОМАГНИТНЫХ ДАННЫХ

Воробьев А.В., Воробьева Г.Р. **Индуктивный метод восстановления временных рядов геомагнитных данных.**

Аннотация. В настоящее время интенсивное развитие систем и технологий регистрации параметров магнитного поля Земли способствует экспоненциальному росту объемов геомагнитных данных, основным источником которых выступают постоянные магнитные станции. Несовершенство применяемой аппаратуры и задействованных каналов передачи информации обуславливает наличие пропусков во временных рядах зарегистрированных данных, что вместе с пространственной анизотропией создает серьезное препятствие для обработки геомагнитных данных при решении прикладных задач. Российские и зарубежные научные организации восстанавливают пропущенные геомагнитные данные методом линейной интерполяции, что обеспечивает приемлемые результаты в условиях спокойной магнитосферы, но значительно искажает временные ряды при изменении окружающей магнитной обстановки. В этой связи возникает актуальная научно-техническая задача разработки подхода к восстановлению геомагнитных данных в условиях возбужденной магнитосферы, обеспечивающего оптимальные метрики качества импутации временных рядов.

Авторами предложен метод восстановления временных рядов, основанный на индуктивном методе обучения алгоритмов. Согласно предлагаемому подходу, каждая магнитная станция оперирует собственной базой знаний, формируемой в ходе регистрации параметров геомагнитного поля и его вариаций. Комбинация значений ряда, предшествующих и следующих за пропуском, является признаковым описанием, применяемым для поиска прецедента в базе знаний магнитной станции. Результат содержит искомый фрагмент временного ряда и заменяет пропущенные значения его уровней. Сложность характера информационного сигнала, обусловленная неспокойной магнитной обстановкой, повышает точность поиска по прецедентам, эффективность которого тем выше, чем большей базой знаний располагает магнитная станция.

Проведенный анализ результатов восстановления пропусков временных рядов геомагнитных данных, зарегистрированных в условиях возбужденной магнитосферы, показал, что предложенный индуктивный метод импутации позволяет повысить точность восстановления пропущенных значений в среднем на 79.54 % по сравнению с используемыми в настоящее время методами, что позволит повысить эффективность обработки геомагнитных данных при решении прикладных задач.

Ключевые слова: геомагнитные данные, временные ряды, пропущенные значения, машинное обучение, обучение по прецедентам, импутация временных рядов.

1. Введение. Наблюдаемый в настоящее время рост объемов геомагнитных данных обусловлен интенсивным развитием наземных и спутниковых систем глобального мониторинга, обеспечивающих непрерывную регистрацию параметров магнитного поля Земли в режиме квазиреального времени [1]. Магнитные станции, аэромагнитные, гидромагнитные съемки, спутниковые и подземные скважинные исследования обеспечивают наблюдения, комплексный и своевременный анализ которых является основой для моделирования компонент геомаг-

нитного поля, понимания причин его эволюции и оценки активности, что особенно важно ввиду доказанной опасности геомагнитных вариаций для техногенных объектов и систем (спутников, коротковолновой радиосвязи, высокоточной магнитометрической аппаратуры, систем автоматики высокоширотных железных дорог и др.).

Открытый доступ к данным о непрерывных изменениях параметров магнитного поля Земли и объединение наземных высокотехнологичных магнитных станций в единую мировую информационную сеть ИНТЕРМАГНЕТ (INTERMAGNET — International Real-Time Magnetic Observatory Network) объясняют тот факт, что на сегодняшний день именно они являются наиболее распространенным, достоверным и доступным для большинства ученых и специалистов методом наблюдения параметров геомагнитного поля и его вариаций [2-4]. Регистрируемые станциями ежеминутные и ежесекундные данные наблюдений агрегированы в наборы односуточных текстовых файлов формата IAGA-2002 и доступны по протоколам FTP и HTTP конечным пользователям и приложениям.

Одной из важнейших задач магнитных станций ИНТЕРМАГНЕТ является обеспечение непрерывности регистрации данных об измеряемых параметрах геомагнитного поля и его вариаций [5]. Однако ввиду несовершенства используемых магнитными станциями аппаратуры и каналов данных исходные временные ряды геомагнитных данных содержат пропуски и «выбросы» за нормальное значение, которые являются необратимыми и могут привести к потере важной информации о геофизических явлениях и процессах [6].

В этой связи совершенствование методов и алгоритмов эффективной обработки больших объемов геомагнитных данных, включая способы восстановления пропущенных значений, входит в число первостепенных проблем современной геофизики.

В настоящее время сетью ИНТЕРМАГНЕТ задача заполнения пропусков геомагнитных данных решается простейшим, но малоэффективным методом — заменой пропусков на зарезервированные значения. Так, стандарт IAGA-2002 определяет последовательности «99999.00» и «88888.00» в качестве индикатора отсутствующего значения параметра геомагнитного поля, что при отсутствии предварительной обработки данных может существенно исказить результаты их интерпретации и анализа.

Другой известный и широко практикуемый в России и за рубежом подход [7-9] основан на линейной интерполяции временных рядов геомагнитных данных, содержащих пропущенные значения. Простота и вычислительная скорость данного метода являются его безусловным преимуществом, однако его эффективность при этом ограничена восста-

новлением только небольших пропущенных сегментов временного ряда геомагнитных данных при условии спокойной магнитной обстановки.

Ни один из представленных методов не решает выявленную проблему в достаточной мере. В этой связи авторами в данной работе предлагается новый подход к восстановлению временных рядов геомагнитных данных в условиях возбужденной магнитосферы, который позволит дополнить известные подходы к импутации пропусков (здесь и далее под импутацией понимается заполнение пропусков временных рядов) и повысит эффективность существующих методов и средств обработки данных наблюдений параметров геомагнитного поля и его вариаций.

2. Краткий анализ особенностей анализируемых временных рядов. Временные ряды геомагнитных данных, зарегистрированных магнитными станциями ИНТЕРМАГНЕТ, в общем виде обладают сходными характеристиками в контексте стационарности, тренда, регулярных и нерегулярных осцилляций. Краткий анализ этих особенностей в данной работе проводится авторами на примере типичного временного ряда геомагнитных данных.

В качестве экспериментального в работе рассматривается временной ряд, уровни которого представлены результатами поминутных измерений горизонтальной компоненты вектора геомагнитного поля, полученными магнитной станцией DOUrbes (50.1°N, 4.6° E) в 2016 году.

Для нивелирования влияния выбросов геомагнитные данные предлагается анализировать на основе медианы и интерквартильного размаха с визуализацией посредством диаграммы размаха, представленной на рисунке 1.

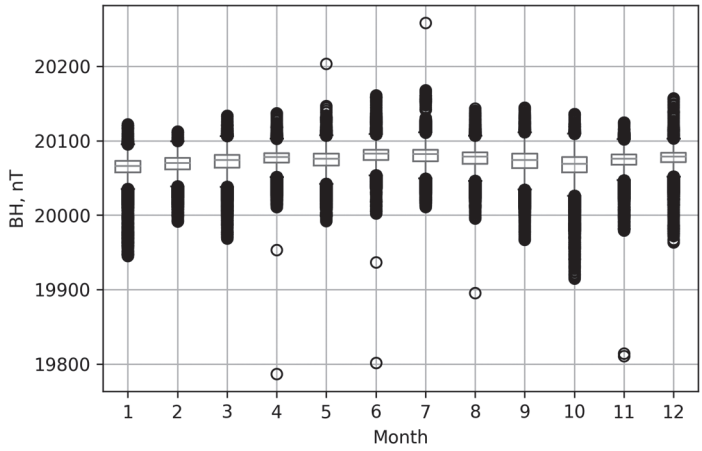


Рис. 1. Результаты анализа временного ряда геомагнитных данных обсерватории DOUrbes за 2016 год

Диаграмма размаха дает представление о медиане, разбросе и асимметрии в распределении значений исследуемого параметра геомагнитного поля для каждого календарного месяца анализируемого года наблюдений. Кроме того, данный способ визуализации является одним из индикаторов наличия выбросов во временном ряду. Так, диаграмма размаха, приведенная на рисунке 1*б*, показывает, что в анализируемом временном ряду на протяжении шести месяцев присутствовали выбросы как за минимальные, так и за максимальные значения. При этом сравнение параллельных графиков показывает, что в течение исследуемого года распределение параметра геомагнитного поля изменялось несущественно, сохраняя значение медианы исследуемого ряда примерно на одном и том же числовом уровне. Аналогичный разброс значений показывает и межквантильный размах: приведенные на диаграмме 25- и 75-процентные квантили относительно стабильны в течение всего рассматриваемого периода, в то время как минимальные и максимальные граничные значения отличаются более существенно (что в целом можно объяснить наличием выбросов в соответствующие календарные месяцы).

Сравнение асимметрии и эксцесса выбросов временного ряда геомагнитных данных с известными моментами нормального распределения посредством теста Харки — Бера [10] показало, что нулевая гипотеза о нормальности распределения выбросов анализируемого ряда отвергается с вероятностью, равной 1.16451551e-08. Это свидетельствует как о нормальном распределении ошибок наблюдения анализируемого геомагнитного параметра поля, так и об однородности всего временного ряда геомагнитных данных.

Кроме того, анализ исследуемого временного ряда геомагнитных данных показал наличие нелинейного тренда, сезонности и цикличности, что определяет аддитивный нестационарный характер его уровней (рисунок 2), при этом:

– нелинейный тренд представляет графическую интерпретацию зависимости уровней временного ряда геомагнитных данных от времени и имеет нейтральный характер флэта, поскольку не обладает выраженным восходящим или нисходящим развитием;

– сезонная компонента временного ряда с характерным для параметров геомагнитного поля 27-дневным периодом характеризует изменение значения горизонтальной составляющей вектора поля в соответствии с указанным циклом;

– циклическая составляющая временного ряда, выделенная в ходе анализа исследуемых уровней, характеризует только длину цикла, поскольку рассматривается в рамках одного календарного года.

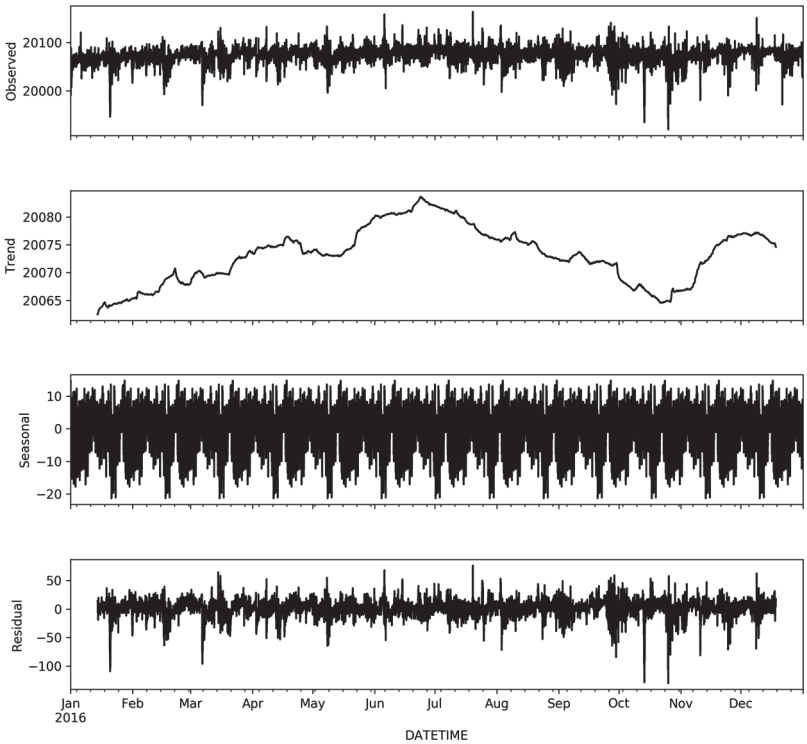


Рис. 2. Декомпозиция временного ряда геомагнитных данных обсерватории DOUrbes за 2016 год

Анализ и визуализация ряда большего временного диапазона подтвердили сделанное таким образом предположение относительно циклической составляющей исследуемого временного ряда.

Данное наблюдение также подтверждается результатами обобщенного теста Дикки — Фуллера (ADF) [11], выявившего наличие во временном ряду единичных корней, что, в свою очередь, свидетельствует о нестационарности уровней ряда.

Применение первой разности приводит временной ряд геомагнитных данных к стационарному виду, что подтверждается результатами того же теста Дикки — Фуллера, отвергающими нулевую гипотезу о наличии единичных корней. Данное наблюдение позволяет описать исследуемые уровни геомагнитных данных как интегрированный временной ряд первого порядка.

3. Классификация пропусков временного ряда геомагнитных данных. Предлагаемая авторами настоящей работы методика

восстановления временных рядов геомагнитных данных базируется на новом подходе к классификации пропущенных значений.

Пусть x_{iT} — уровни временного ряда геомагнитных данных за период T с наблюдениями $i = 1,...,n$. Значение ряда $x \in \{x_{iT} \mid \forall i,T\}$ описывается *измерительной мерой* m_{iT} (здесь и далее для наглядности наличие значения временного ряда наблюдений обозначается как определенное значение уровня ряда):

$$m_{iT} = \begin{cases} 0, \text{ значение } x_{iT} \text{ определено,} \\ 1, \text{ значение } x_{iT} \text{ заменено индикатором,} \\ 2, \text{ значение } x_{iT} \text{ отсуствует.} \end{cases}$$

Тогда временной ряд измерений параметров геомагнитного поля может быть определен как:

$$x_T = \{x_{\text{obs}}, x_{\text{mis}}\},$$

где x_{obs} — наблюдения с определенными значениями параметра $x_{iT} \in \{x_{\text{obs}} \mid \forall i,T\}$, x_{mis} — наблюдения с отсутствующими значениями параметра $x_{iT} \in \{x_{\text{mis}} \mid \forall i,T\}$.

При этом:

$$x_{\text{obs}} \equiv \{(x_{iT} \mid \forall i,T) \wedge (m_{iT} = 0)\};$$
$$x_{\text{mis}} \equiv \{(x_{iT} \mid \forall i,T) \wedge (m_{iT} > 0)\}.$$

Полным считается временной ряд x_T длиной n, для которого справедливо выражение:

$$x_T : \forall x_{iT} \in x_T \; m_{iT} = 0 \; (i = 1,...,n),$$
$$x_T = \{x_{\text{obs}}, x_{\text{mis}}\}, \; x_{\text{mis}} = \varnothing.$$

Если временной ряд x_T длиной n удовлетворяет следующим условиям:

$$x_T : \forall x_{iT} \in x_T \; m_{iT} \geq 0 \; (i = 1,...,n),$$
$$x_T = \{x_{\text{obs}}, x_{\text{mis}}\}, \; x_{\text{obs}} = \varnothing, \; , \; x_{\text{mis}} = \varnothing,$$

то он считается неполным.

Согласно принятой в теории анализа временных рядов классификации пропусков [12], отсутствующие значения геомагнитных данных относятся к типу MCAR (Missing Completely at Random), поскольку вероятность пропуска результата наблюдения одинакова для каждой записи набора зарегистрированных значений. Не предоставляя иной альтернативы (тип MAR или MNAR), данный факт не является основанием для выбора метода заполнения пропусков геомагнитных данных.

В этой связи в данной работе предлагается расширить принятую классификацию пропусков значений уровня временного ряда, введя ряд их дополнительных характеристик.

Предварительно для их определения вводится понятие сегмента временного ряда s, под которым понимается множество последовательных l значений уровня ряда:

$$s = \{s_{jT}\},\ j = 1,...,l;\ s_{jT} \in (x_{iT} \mid \forall i,T),\ i = 1,...,n,\ n > l,$$

где l — длина сегмента временного ряда.

Тогда дискретными будут называться пропущенные сегменты s временного ряда размерности n, длина которых строго равна одному:

$$s = \{s_{jT} : m_{jT} \geq 1\},\ j = 1;\ s_{jT} \in (x_{iT} \mid \forall i,T),\ i = 1,...,n,\ n > l.$$

Серийными будут обозначаться такие пропущенные сегменты s временного ряда размерности n, длина которых превышает один:

$$s = \{s_{jT} : m_{jT} \geq 1\},\ j = 1...,l; s_a \prec s_b,\ s_a,\ s_b \in s,\ a < b;$$
$$s_{jT} \in (x_{iT} \mid \forall i,T),\ i = 1,...,n,\ n > l;\ s \subset \{x_{iT}\},\ i = 1,...,l.$$

Если за пропуском длиной l следует сегмент определенных значений уровня, длина которого равна или превышает значение l, то такой пропуск будет называться интерполируемым:

$$s \subset \{x_{iT} : m_{iT} > 0\},\ i = 1,...,l;\ b \subset \{x_{iT} : m_{iT} = 0\},\ i = 1,...,n-l;\ \#(b) \geq \#(s).$$

В противном случае пропуск значений уровня временного ряда будет считаться экстраполируемым:

$$s \subset \{x_{iT} : m_{iT} > 0\},\ i = 1,...,l;\ b \subset \{x_{iT} : m_{iT} = 0\},\ i = 1,...,n-l;$$
$$\#(b) < \#(s) \text{ or } b = \varnothing.$$

С учетом приведенных характеристик, любой пропуск временного ряда геомагнитных данных может быть отнесен к одному из следующих типов: дискретный интерполируемый; дискретный экстраполируемый; серийный интерполируемый; серийный экстраполируемый.

Анализ экспериментальных данных наблюдений сети магнитных станций ИНТЕРМАГНЕТ показал следующее распределение типов пропусков геомагнитных данных за исследуемый 2016 год:

– дискретные интерполируемые — не обнаружено;

– дискретные экстраполируемые — не обнаружено;

– серийные интерполируемые — 10 пропусков (максимальная длина сегмента — 240 минут);

– серийный экстраполируемый — 27 пропусков (отсутствуют данные наблюдений за 27 непоследовательных суток).

Также для сравнения был проведен анализ временного ряда обсерватории AMS (Martinde Vivies, Amsterdam Island) с худшим по сравнению со станцией DOUrbes показателем относительной информационной эффективности [13]. Типы пропусков геомагнитных данных распределены следующим образом:

– дискретные интерполируемые — 3 пропуска;

– дискретные экстраполируемые — не обнаружено;

– серийные интерполируемые — 4 пропуска (максимальная длина сегмента — 1 108 минут);

– серийный экстраполируемый — 32 пропуска (из них отсутствуют данные 30 суток наблюдений).

Анализ подтверждает неравномерный характер распределения типов пропусков временного ряда геомагнитных данных для любой магнитной станции. Каждое из отсутствующих значений может быть восстановлено, но способ импутации напрямую зависит от характера пропуска. Неверный подбор метода восстановления пропуска может критически исказить весь временной ряд наблюдений и негативно отразиться на результатах его анализа. Поэтому целесообразным представляется подход, согласно которому определяется тип каждого пропуска временного ряда и в соответствии с этим применяется оптимальный метод его восстановления.

4. Анализ эффективности известных методов и моделей восстановления пропусков временного ряда. Основанием для выбора оптимального метода восстановления для каждого типа пропуска во временном ряду геомагнитных данных послужили результаты выполненного авторами анализа эффективности ряда известных моделей и методов импутации и прогнозирования данных, метрикой

качества которых послужило значение среднеквадратической ошибки восстановления ряда.

1) Упрощенный метод скользящей средней — частный случай способа сглаживания временных рядов. Данный подход реализован таким образом, что ширина N сегмента временного ряда фиксирована и равна 3, а пропущенное значение ряда рассчитывается как среднее арифметическое предшествующего и последующего замеров:

$$x_i = \frac{x_{i-1} + x_{i+1}}{2},\ i = 1,..., N,$$

где x_i — восстанавливаемое значение; x_{i-1} и x_{i+1} — предшествующее и последующее значения уровня временного ряда соответственно.

Поскольку характер изменения регистрируемого информационного сигнала исключает скачкообразные вариации (что в первую очередь обуславливается природой их происхождения), то данный метод требует минимальных затрат машинного времени и обеспечивает сопоставимую с другими методами метрику качества.

При этом следует оговорить ограничения, накладываемые на число и характер распределения пропущенных значений. Идеализированный вариант использования упрощенного метода скользящей средней предполагает единственное пропущенное значение между двумя известными геомагнитными измерениями. В действительности такая ситуация складывается крайне редко и реальные геомагнитные данные сопровождаются целой серией пропущенных значений, следующих во временном ряду последовательно друг за другом. В этом случае алгоритм предусматривает циклический поиск первого значимого замера (отличного от выброса / пропуска) и его подстановку в выражение расчета среднего арифметического значения. Очевидно, что чем дальше в ряду находится данное значение от пропущенного, тем больше величина среднеквадратического отклонения, возникающего при восстановлении временного ряда. Поэтому предпочтительно применение метода скользящей средней для восстановления дискретных интерполируемых значений с симметричными предшествующим и последующим сегментами временного ряда. В остальных случаях величина ошибки критически возрастает, что снижает целесообразность применения указанного метода.

2) Линейная интерполяция — метод, активно применяемый в настоящее время в геофизике для восстановления геомагнитных данных. Суть метода состоит в том, что крайние точки пропущенного сегмента временного ряда соединяются друг с другом прямой линией,

то есть составляется полином первой степени, поиск коэффициентов которого выполняется в ходе интерполяции:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1},\ x_1 \le x \le x_2,\ y = ax + b,$$

$$a = \frac{y_2 - y_1}{x_2 - x_1},\ b = y_1 - ax_1,$$

где x_1 и y_1 — первая крайняя точка пропуска и значение уровня в ней, x_2 и y_2 — вторая крайняя точка пропуска и значение уровня в ней, x и y — пропущенная точка и значение уровня в ней, a, b — коэффициенты построенной прямой.

Метод линейной интерполяции обеспечивает наибольшую эффективность при восстановлении дискретных пропусков обоих типов, поскольку интервал, в который попадает искомое значение, в данном случае минимален. Увеличение длины пропущенного сегмента приводит к пропорциональному увеличению значения среднеквадратической ошибки.

3) Метод кубической сплайн-интерполяции известен как более точный по сравнению с описанной выше линейной интерполяцией. Его отличительной особенностью является разбиение интервала интерполяции на отрезки, на каждом из которых функция задается полиномом третьей степени с коэффициентами, обеспечивающими как непрерывность функции, так и ее прохождение через заданные точки:

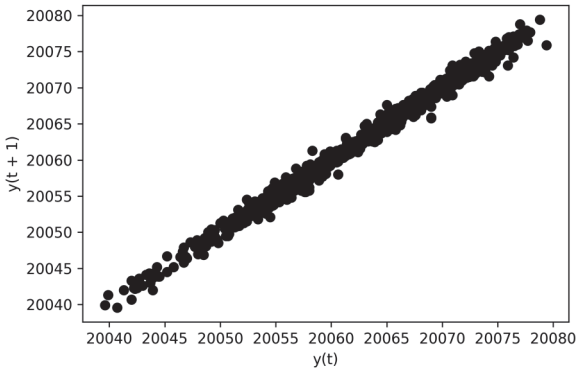
$$F_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3;$$
$$F = F_1 \text{ на интервале } [x_0, x_1],$$
$$F = F_2 \text{ на интервале } [x_1, x_2],$$
$$\ldots$$
$$F = F_N \text{ на интервале } [x_{N-1}, x_N],$$

где $F(x)$ — интерполирующая функция, $F_k(x)$ — кубический полином на отрезке $[x_{k-1}, x_k]$; a_k, b_k, c_k, d_k — коэффициенты полинома на отрезке $[x_{k-1}, x_k]$; $[x_0, x_N]$ — интерполируемый отрезок.

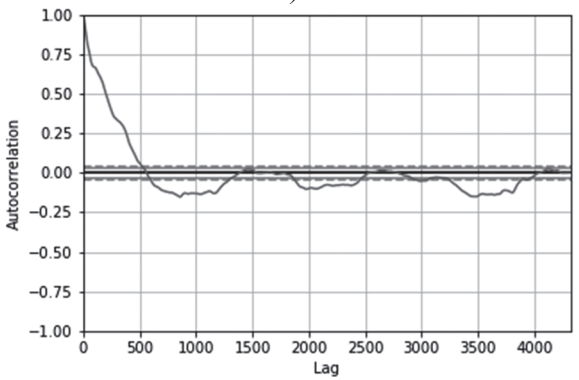
Специфика анализируемого явления позволяет нивелировать склонность кубических сплайнов к осцилляции в окрестностях точки, существенно отличающейся от своих соседей. В этой связи точность рассматриваемого метода повышается по сравнению с результатами, обеспечиваемыми методом кусочно-линейной интерполяции.

4) Модели и методы авторегрессии. Наличие тренда и сезонной / циклической составляющей временного ряда обуславливает корреляционную зависимость между его последовательными значениями, известную как автокорреляция уровней ряда.

Зависимость между измерениями из исследуемого ряда геомагнитных данных и их лагами приведена на рисунке 3*а*. Выбивающиеся из общего массива точек значения являются выбросами уровней ряда и могут быть нивелированы. Оставшиеся измерения визуально кластеризуются по диагонали от левого нижнего к правому верхнему краю диаграммы, что свидетельствует о положительной корреляционной связи между ними.



а)



б)

Рис. 3. Коррелограммы экспериментального временного ряда

Осциллирующая зависимость геомагнитных измерений и их лагов подтверждает детерминированный характер уровней анализируемого временного ряда (рисунок 3*б*). Из графика автокорреляции видно,

что коэффициент корреляции лагов исследуемого ряда отличен от нуля на протяжении всего временного интервала, что свидетельствует о прогнозируемости временного ряда на основе параметрических прогностических моделей и методов.

Суть предлагаемого подхода заключается в том, что известные уровни ряда, предшествующие пропущенному сегменту, выступают в роли обучающей выборки, на основании которой выполняется прогноз отсутствующего фрагмента искомой длины. Прогностические методы используют в своей основе модель авторегрессии, которую в общем виде можно представить как:

$$Y_i = a_0 + \sum_{i=1}^{p} a_i Y_{i-1} + \varepsilon_i,$$

где Y_i — целевая переменная (атомарное восстанавливаемое значение); p — порядок модели, a_0 — коэффициент, описывающий поведение модели при отсутствии внешних факторов, a_i — коэффициенты, описывающие влияние на поведение модели i внешних факторов; Y_{i-1} — прежние значения целевой переменной; ε_i — погрешность модели.

В рассматриваемом случае экспериментального временного ряда геомагнитных данных имеет место модель авторегрессии первого порядка, что позволяет оценивать изменение целевой переменной в зависимости от единственного фактора — ее собственного значения в прошлом периоде авторегрессии. Выбор такой модели авторегрессии обусловлен анализом коррелограммы (рисунок 3*б*), результат которого свидетельствует о том, что корреляция максимальна между двумя соседними значениями и непрерывно убывает по мере увеличения числа исследуемых лагов.

Та же закономерность прослеживается и при оценке точности прогнозирования пропущенных значений: чем больше восстанавливаемый сектор, тем большую погрешность обнаруживает метод и модель авторегрессии. В этой связи рассматриваемый метод авторегрессии используется итерационно для восстановления одного значения в пропущенном секторе, которое, в свою очередь, становится частью новой обучающей выборки для восстановления последующего элемента пропущенного сегмента и так далее.

Следующий прогностический метод, исследованный применительно к восстановлению геомагнитных данных, — интегрированная модель авторегрессии — скользящего среднего (ARIMA). Модель ха-

рактеризуется тремя параметрами: p — порядок авторегрессии, d — порядок интегрирования, q — порядок скользящего среднего [14-16]:

$$\left(\Delta^d X_t\right) = \sum_{i=1}^{p} \varphi_i \left(\Delta^d X_t\right) + \varepsilon_t + \sum_{j=1}^{q} \theta_j \left(\Delta^d \varepsilon_{t-j}\right),\ \varepsilon_t \sim N(0, \sigma_t^2),$$

где $\phi(\bullet)$, $\theta(\bullet)$ — полиномы степеней p и q, d — порядок взятия последовательной разности ($\Delta X_t = X_{t-1} - X_t = (1-B)X_t$, $\Delta^2 X_t = \Delta^2 X_{t+1} - -\Delta X_t = (1-B)^2 X_t, \ldots$), B — лаговый оператор ($B^j X_t = X_{t-1}$, $B^j \varepsilon_{j-1}$, $j = 0, \pm 1, \ldots$).

Итеративное исследование различных комбинаций перечисленных параметров модели ARIMA было выполнено с помощью «сетчатого поиска» и показало, что лучшее значение информационного критерия Акаике (AIC) достигается при $p = 1$, $d = 0$, $q = 1$. Результат подтверждает, что наилучшая корреляция наблюдается между двумя соседними значениями экспериментального ряда данных.

Анализ перечисленных методов восстановления данных был выполнен применительно к решению задачи импутации пропущенных значений экспериментального ряда геомагнитных данных обсерватории DOU за 2016 год (таблица 1).

Таблица 1. Среднеквадратическая ошибка восстановления основных типов пропусков известными методами импутации (нТл)

Тип пропуска \ Метод импутации	Скользящее среднее	Линейная интерполяция	Кубическая сплайн-интерполяция	Модель автокорреляции	ARIMA
Дискретный интерполируемый	0.110	0.153	0.153	0.475	0.482
Дискретный экстраполируемый	0.171	0.157	0.157	0.491	0.522
Серийный интерполируемый (получасовой пропуск при Kp < 3)	0.730	0.679	0.674	0.728	0.953
Серийный экстраполируемый (получасовой пропуск при Kp < 3)	8.402	13.738	12.694	8.508	10.670

Большинство известных методов анализа и восстановления временных рядов реализовано посредством программных средств, ориентированных на конечного пользователя и обладающие интуитивно понятным интерфейсов, и программных библиотек, ориентированных на разработчиков программных сценариев и продуктов [17-23]. Ярким представителем первого типа программных средств является инструментально-программная система «Гусеница» («Caterpillar»), доступная по URL: http://www.gistatgroup.com/gus/. Данная программа предоставляет конечному пользователю возможность загрузки временного ряда, его визуализации и аналитической обработки. Функция восстановления пропусков временного ряда не является здесь основной и реализована только для простейших и краткосрочных отсутствующих сегментов ряда. Основной используемый при этом метод — линейная рекуррентная формула — обеспечивает восстановления временного ряда геомагнитных данных со средним значением среднеквадратической ошибки 2.863 нТл, что существенно хуже методов, проанализированных выше (таблица 1).

Второй подход, ориентированный на разработчиков программных систем, — это прежде всего инструмент реализации известных методов восстановления временных рядов, в частности, перечисленных выше в таблице 1. Так, представленные результаты были получены авторами работы посредством языка программирования Python и его статистических библиотек. Вместе с тем собственными методами восстановления Python и ему подобные языки не располагают и тем более не учитывают специфики временных рядов геомагнитных данных, уровни которых зависят от текущего состояния Kp-индекса и, соответственно, окружающей магнитной обстановки в момент регистрации пропущенного сегмента ряда.

В целом анализ полученных результатов (таблица 1) показал следующее распределение методов восстановления данных по их эффективности. Метод скользящего среднего наиболее эффективен для импутации дискретных интерполируемых пропусков. Если дискретный пропуск является экстраполируемым, то для его восстановления целесообразнее применять метод кубической сплайн-интерполяции, равно как и для импутации серийного интерполируемого пропуска, длительность которого не превышает получасового периода. Серийный экстраполируемый пропуск был восстановлен с помощью метода скользящей средней, хотя значение среднеквадратической ошибки слишком велико и результат импутации недостоверен.

Важно отметить, что дискретные пропуски обоих типов восстанавливаются методами линейной интерполяции и кубической сплайн-интерполяции с равными показателями точности, поэтому могут в равной степени использоваться для восстановления пропущенных значений во временном ряду.

Кроме того, качество восстановления серийных пропусков определяется не только применяемым для импутации методом, но и внешними факторами, в частности, состоянием магнитосферы в соответствующий период времени. Так, указанные методы показывают достаточно хорошие метрики качества восстановления данных при условии спокойной магнитосферы (при значении индекса Kp, не превышающего 2). Однако дополнительно проведенные экспериментальные исследования показали, что среднеквадратическая ошибка восстановления данных увеличивается по мере возрастания значения коэффициента Kp и принимает значения порядка сотни нТл уже при Kp = 4.

Таким образом, результаты проведенного анализа известных методов восстановления временных рядов применительно к геомагнитным данным могут быть интерпретированы следующим образом:

1) дискретные интерполируемые пропуски должны быть восстановлены методом скользящей средней с шириной окна, равной 3. При этом внешние факторы в виде окружающей магнитной обстановки, количественно оцениваемой посредством Kp-индексов, не влияют на результаты импутации.

2) дискретные экстраполируемые пропуски должны быть восстановлены методом кубической сплайн-интерполяции. Как и в предыдущем случае значение Kp-индекса, отражающее состояние магнитосферы в момент регистрации пропуска, не влияет на результат импутации.

3) серийные интерполируемые пропуски в условиях спокойной магнитосферы (при Kp < 3) также должны быть восстановлены методом кубической сплайн-интерполяции, который в ходе проведенных экспериментов демонстрировал стабильно лучший результат;

4) серийные интерполируемые пропуски в условиях спокойной магнитосферы (при Kp < 3) должны быть восстановлены методом скользящей средней с шириной окна, равной 3, также показавшего лучший результат в серии экспериментов по восстановлению временных рядов геомагнитных данных в условиях спокойной магнитосферы.

Изменения уровней временного ряда в условиях неспокойной магнитной обстановки (при Kp > 2) сопровождаются характерными нерегулярными осцилляциями, затрудняющими восстановление пропусков в соответствующие временные интервалы. Сложное изменение характера информационного сигнала проявляется особенно на серийных сегментах ряда. При этом ни один из представленных методов не показал приемлемого результата по восстановлению временных рядов в таких условиях.

В этой связи возникает актуальная задача, заключающаяся в разработке и формализации нового метода восстановления временных рядов геомагнитных данных, обеспечивающего лучшие показатели качества при импутации данных в условиях неспокойной магнитной обстановки.

5. Метод прецедентного резервирования. Для восстановления временных рядов геомагнитных данных в условиях возбужденной магнитосферы в данной работе авторами предлагается метод, получивший название прецедентного резервирования. В его основе лежит концепция индуктивного обучения, заключающаяся в выявлении общих закономерностей по частным эмпирическим данным [17].

Ключевой идеей метода является предположение, что любому сегменту временного ряда можно с некоторой допустимой степенью точности поставить в соответствие один или несколько фрагментов предшествующих ему значений уровня (рисунок 4). В этом случае накопленные магнитной станцией статистические данные выступают в качестве базы прецедентов, где каждый сектор временного ряда заданной длины являет собой атомарный прецедент. Тем самым магнитная станция «резервирует» себя собственными ранее выполненными измерениями, которые при определенных ограничениях могут заменить пропущенные сегменты временного ряда.

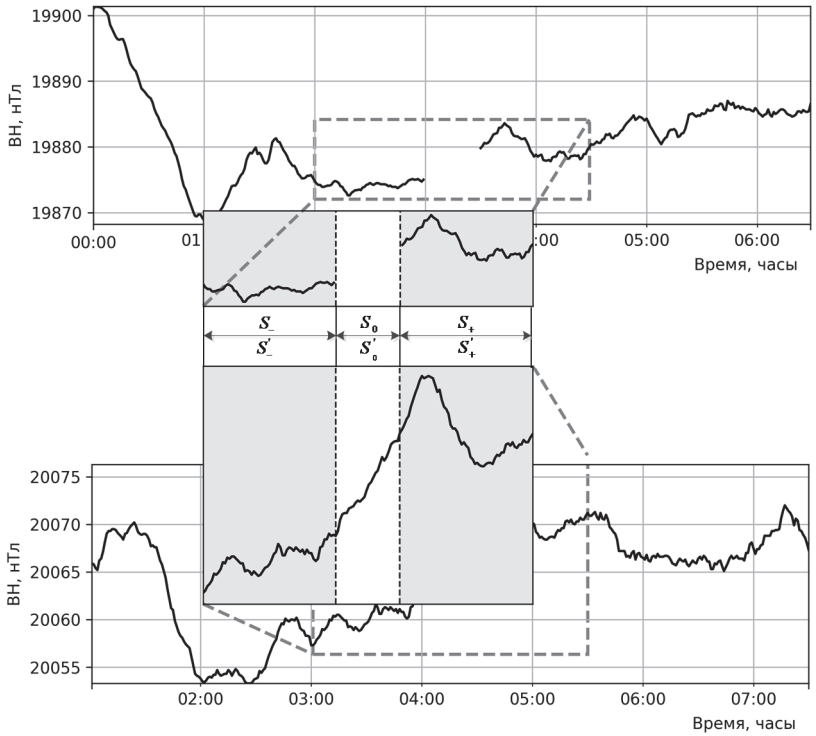


Рис. 4. Схема метода прецедентного резервирования

Пусть S — восстанавливаемая выборка, представленная тройкой из сегмента отсутствующих значений временного ряда, а также предшествующего и последующего за ним сегментов заданной длины:

$$S = \{S_-, S_0, S_+\},\ S_- = \{s_i\},\ i = 1,...,L,$$
$$S_0 = \{s_j\},\ j = L+1,...,M,\ S_+ = \{s_k\},\ k = M+1,...,N,$$

где S — восстанавливаемая выборка, S_0 — пропущенный сегмент, S_- — сегмент, предшествующий пропуску, S_+ — сегмент, следующий за пропуском.

Пусть S' — предшествующие восстанавливаемой выборке сегменты временного ряда, именуемые статистическими. Тогда паре предшествующего S'_- и следующего S'_+ за пропуском сегментов временного ряда можно поставить в соответствие пару статистических сегментов той же длины (S'_- и S'_+):

$$S' = \{s_k\},\ k = 1,...,N;$$
$$S'_- = \{s_i\},\ i = 1,...,|S_-|;\ S'_- \in S';$$
$$S'_+ = \{s_j\},\ j = 1,...,|S_+|;\ S'_+ \in S';$$
$$S'_- \to S_-,\ S'_+ \to S_+.$$

Разделяющие каждую пару сегменты считаются подобными и взаимозаменяемыми, что позволяет заполнить пропуски соответствующими значениями статистического сегмента временного ряда (с предварительной нормализацией данных):

$$S'_0 = \{s_n\},\ n = 1,...,|S_0|;\ S'_0 \in S';\ S'_0 \to S_0.$$

Мерой соответствия сегментов временного ряда выступает степень их линейной корреляции, что подтверждается выявленной сильной положительной автокорреляцией значений измерений геомагнитных параметров. При этом процедура поиска, заменяющего пропуск статистического сегмента, выполняется путем последовательного обхода элементов временного ряда в соответствии с жадным алгоритмом перебора значений, где размерность анализируемого фрагмента равна длине восстанавливаемой выборки и обрабатываемые сегменты пересекаются друг с другом.

Пусть шаблон поиска релевантных статистических сегментов представлен восстанавливаемой выборкой с вычтенным из нее сегментом отсутствующих значений, а обрабатываемая выборка — формируемой на каждой итерации перебора парой статистических сегментов. Тогда с учетом введенных обозначений с помощью коэффициента корреляции Пирсона [18] можно определить степень линейной зависимости между исследуемым сегментом и шаблоном поиска r_{xy} как:

$$r_{xy} \frac{\sum_{i=1}^{m}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_i-\overline{x}\right)^2\sum_{i=1}^{m}\left(y_i-\overline{y}\right)^2}},$$

где x_i — значения обрабатываемой выборки, y_i — значения восстанавливаемой выборки, $\overline{x}$ — среднее арифметическое обрабатываемой выборки; $\overline{y}$ — среднее арифметическое восстанавливаемой выборки.

Абсолютные величины расчетных значений коэффициентов корреляции Пирсона заносятся в предварительно выделенный пул, применяемый для определения наибольшего из значений и, как следствие, соответствующей ему обрабатываемой выборки. Максимальное значение коэффициента корреляции является основанием для подтверждения предположения о соответствии и взаимозаменяемости восстанавливаемой выборки и выделенной тройки статистических сегментов временного ряда.

Как и в большинстве задач данного рода, полученные оценки восстановления данных являются смещенными. В этой связи результаты, получаемые посредством указанного метода, должны быть нормализованы таким образом, чтобы результирующие значения были аппроксимированы относительно известных соседних пропуску значений уровней временного ряда. Соответствующая аппроксимация выполняется посредством метода наименьших квадратов, применение которого целесообразно ввиду единичного лага рассматриваемого временного ряда.

Задача аппроксимации в конечном счете сводится к определению значений коэффициентов линейной зависимости, при которых функция двух переменных a и b принимает наименьшее значение. Иными словами, при полученных a и b сумма квадратов отклонений экспериментальных данных от найденной прямой будет наименьшей:

$$F(a,b)=\sum_{i=1}^{n}\left(y_i-\left(ax_i+b\right)\right)^2.$$

Полученные методом наименьших квадратов коэффициенты a и b являются искомыми составляющими аппроксимации полученной выборки, необходимыми для нивелирования полученной в ходе применения метода смещенной оценки.

Для уменьшения уровней шума на завершающем этапе к полученным данным применяется метод медианного сглаживания, главным преимуществом которого является его устойчивость к выбросам. Для заданного медианного интервала временного ряда вычисляется сумма частот значений уровня, рассчитывается половина полученного значения и определяется, какое значение ряда на нее приходится:

$$M = x_0 + L\left(\frac{\frac{1}{2}\sum_{i=1}^{K} f_i - N_{\text{prev}}}{f_M}\right),$$

где M — медианное значение, x_0 — начальное значение медианного интервала, L — длина медианного интервала, K — длина временного ряда, $\sum_{i=1}^{K} f_i$ — сумма частот временного ряда, f_M — частота медианного интервала, N_{prev} — сумма частот интервалов, предшествующих медианному.

Для предотвращения коллизий начального и конечного значений уровней ряда применительно к временному ряду геомагнитных данных медианный интервал был определен в соответствии с процедурой Тьюки и принят $L = 3$ [19].

Важно отметить, что применение метода прецедентного резервирования в соответствии с принципом бритвы Оккама [20] сопряжено с оценкой минимального объема выборки, достаточного для восстановления данных за период упреждения. Иными словами, требуется определить длину сегментов S_+ и S_- восстанавливаемой выборки при известном числе значений уровня пропущенного сегмента в ней.

Очевидно, что простейшей будет являться модель временного ряда, в которой восстанавливаемая выборка несимметрична и охватывает все актуальные значения уровней, ограничивающие серию пропусков. Программная обработка выборки такой размерности требует значительных затрат вычислительных ресурсов и аппаратного времени, что в условиях восстановления большого числа пропущенных значений временных рядов геомагнитных данных неприемлемо.

Анализ и экспериментальное исследование метода прецедентного резервирования показали, что стратификацию временного ряда следует осуществлять по принципу эквивалентности длины подмножеств: сегменты S_+ и S_- восстанавливаемой выборки выбираются исходя из числа значений уровня в периоде упреждения. Так, к примеру, пропуск длиной в 30 значений формирует восстанавливаемую выборку из 90 последовательных значений уровня временного ряда геомагнитных данных со среднеквадратической ошибкой порядка 0.4 нТл.

Для оценки эффективности предложенного авторами метода была проведена серия экспериментов по восстановлению геомагнитных данных, пропущенных при различных значениях индекса Kp. Исследования проводились для серийных интерполируемых пропусков длиной в 30 значений (минут). Оценка была проведена применительно к двум методам: используемому в настоящее время методу линейной интерполяции и предложенному методу прецедентного резервирования. Полученные в ходе эксперимента результаты представлены на рисунке 5.

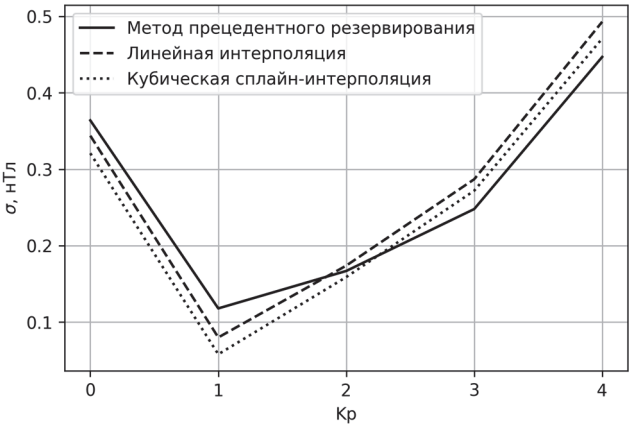


Рис. 5. Сравнительный анализ зависимости величины среднеквадратической ошибки от значения индекса Kp при восстановлении геомагнитных данных методами линейной интерполяции, кубической сплайн-интерполяции и прецедентного резервирования

Как видно из рисунка 5, в условиях спокойной магнитосферы (при $Kp \le 2$) метод линейной интерполяции обеспечивает меньшую среднеквадратическую ошибку, чем метод прецедентного резервирования. При приближении Kp к значению 2 величины среднеквадратической ошибки выравниваются, а с достижением индексом значения 2 метод прецедентного резервирования показывает лучший ре-

зультат по сравнению с методом линейной интерполяции. Следует отметить, что метод прецедентного резервирования обеспечивает приемлемую величину среднеквадратической ошибки только при $Kp \geq 2$ и $Kp \leq 4$. Далее значение ошибки импутации увеличивается в несколько раз и процедура восстановления данных теряет смысл.

6. Методика скользящего регенерирующего окна. Распространенный на сегодняшний день подход к восстановлению пропусков геомагнитных данных предполагает применение линейной интерполяции ко всему временному ряду независимо от типа пропуска и состояния магнитосферы в соответствующий временной период. Вместе с тем результаты проведенного авторами анализа показывают, что не все методы одинаково эффективны и значение среднеквадратической ошибки при их применении зависит от того, к какому типу относится восстанавливаемый пропуск и при каких внешних условиях он был зарегистрирован.

В этой связи в данной работе авторами предложен подход, получивший название скользящего регенерирующего окна. Его суть заключается в том, что временной ряд размерности n с m пропусками разбивается на m пересекающихся окон-сегментов, каждое из которых включает ровно один пропуск и определенные значения уровня, предшествующие и следующие за ним в ряду.

Размер скользящего окна определяется размерностью пропуска, при этом в случае интерполируемого пропуска длины сегментов, предшествующих и следующих за пропуском, равны между собой и составляют величину, равную длине отсутствующего сегмента. Если пропуск экстраполируемый, то длины сегментов, предшествующих и следующих за пропуском, не равны между собой и составляют доступную для каждого конкретного случая величину, меньшую или равную длине отсутствующего сегмента.

В ходе восстановления данных по предложенной методике каждое скользящее окно формирует независимый временной ряд, для восстановления пропуска в котором применяется тот метод, который наилучшим образом подходит для импутации отсутствующего значения соответствующего типа в определенной магнитной обстановке.

Так, к примеру, для восстановления дискретного интерполируемого пропуска подбирается скользящее окно, образующее временной ряд с тремя значениями уровня. Поскольку дискретные значения не зависят от магнитной обстановки, восстановление такого пропуска выполняется методом скользящей средней. Если во временном ряду пропущен сегмент из 30 значений, а последующий и предыдущий за ним сегменты того же размера содержат только определенные значения, то имеет место временной ряд данных, содержащий 90 значений

уровня. В случае, когда такой пропуск был зарегистрирован в условиях спокойной магнитосферы, он восстанавливается методом линейной интерполяции, иначе для импутации отсутствующих значений временного ряда используется метод прецедентного резервирования.

Очевидно, что в результате последовательного обхода временного ряда данных и разделения его на промежуточные временные ряды импутация отсутствующих значений обеспечивает лучшие показатели качества, чем при использовании одного метода, поскольку учитывает особенности каждого пропуска и условия его регистрации. Для подтверждения выдвинутого предположения об эффективности предложенной методики была проведена серия экспериментов, результаты одного из которых приводятся далее.

7. Верификация методики скользящего регенерирующего окна. Экспериментальный временной ряд представлен геомагнитными данными, полученными при измерении горизонтальной компоненты геомагнитного поля BH станцией DOU 1 января 2017 года (рисунок 6).

Оригинальный массив данных (рисунок 6*а*) не содержит пропусков и отражает суточные вариации параметров постоянного магнитного поля при изменении значения Kp-индекса от 2 от 4. Для анализа эффективности предложенной методики восстановления данных во временной ряд были введены 42 дискретных и 3 серийных интерполируемых пропуска. Дискретные пропуски были равномерно распределены по временному ряду, а серийные пропуски подобраны таким образом, чтобы индекс Kp в соответствующие периоды принимал значения 4, 3 и 2 соответственно.

Как показал анализ методов восстановления временных рядов, наименьшую среднеквадратическую ошибку при восстановлении дискретных интерполируемых пропусков геомагнитных данных обеспечивает метод скользящей средней с шириной окна, равной 3.

Выбор метода восстановления серийного пропуска зависит от значения Kp-индекса в период, когда был зафиксирован пропуск (рисунок 5). В рассматриваемом временном ряду определены 30-минутный пропуск при Kp = 4, 20-минутный пропуск при Kp = 3 и 10-минутный пропуск при Kp = 2. Согласно проведенным исследованиям, ожидается, что первые два пропуска могут быть восстановлены с минимальной среднеквадратической ошибкой с помощью метода прецедентного резервирования, а последний — посредством кубической сплайн-интерполяции.

Таким образом, в соответствии с методикой скользящего регенерирующего окна лучшую метрику качества восстановления экспериментального временного ряда обеспечивает последовательное применение следующих методов импутации пропусков:

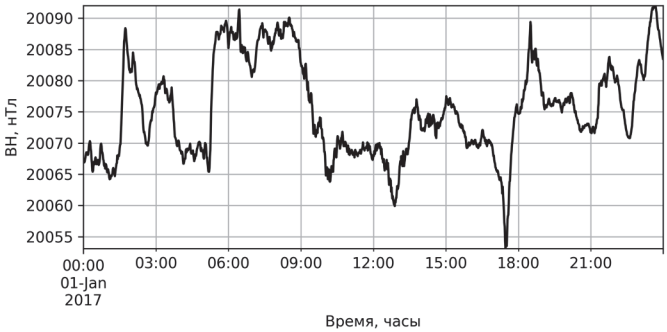
1) скользящая средняя для дискретных пропусков;

2) прецедентное резервирование для 30-минутного пропуска при Kp = 4;

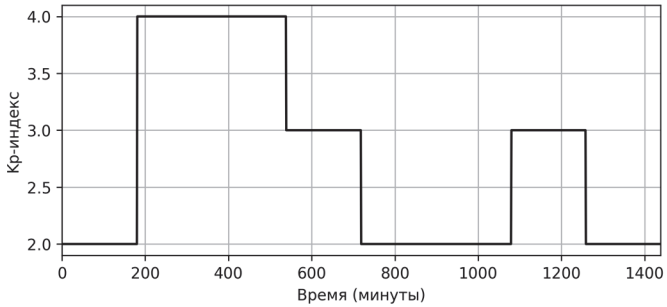
3) прецедентное резервирование для 20-минутного пропуска при Kp = 3;

4) кубическая сплайн-интерполяция для 10-минутного пропуска при Kp = 2.

а)



б)



в)

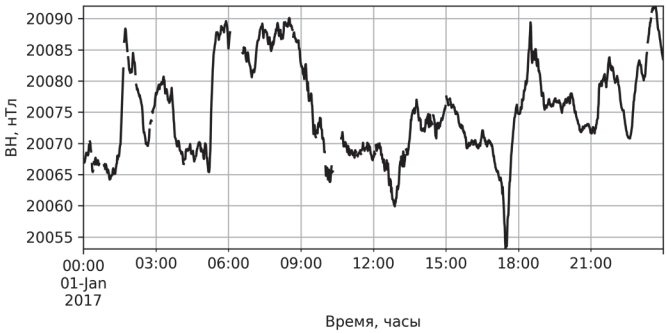


Рис. 6. Экспериментальный ряд геомагнитных данных, зарегистрированных станцией DOU 1 января 2017 года: *а* — оригинальный временной ряд; *б* — изменение Kp-индекса за 1 января 2017 года; *в* — временной ряд с искусственно введенными пропусками данных

Среднеквадратическая ошибка, полученная в результате восстановления экспериментального временного ряда геомагнитных данных посредством предложенных методики и метода, составила 0.219 нТл (рисунок 7). При этом 30-минутный пропуск при Kp = 4 и 20-минутный пропуск при Kp = 3 были восстановлены со значениями среднеквадратической ошибки 0.748 нТл и 0.445 нТл соответственно (рисунок 8).

Для сравнения тот же временной ряд был восстановлен с помощью двух известных подходов: используемого в настоящее время метода восстановления геомагнитных данных — линейной интерполяции, а также метода кубической сплайн-интерполяции, хорошо зарекомендовавшего себя при импутации небольших пропусков в условиях спокойной магнитосферы.

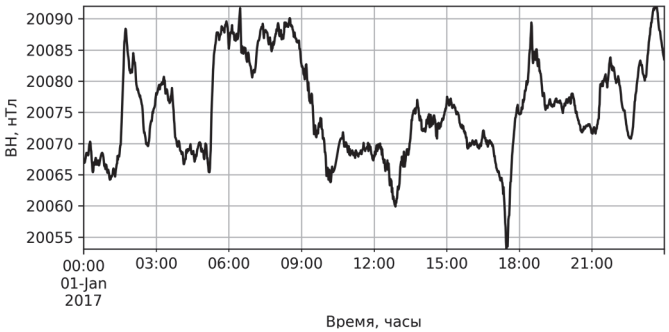


Рис. 7. Результат восстановления экспериментального временного ряда геомагнитных данных в соответствии с методикой скользящего регенерирующего окна

Импутация геомагнитных данных указанными методами показала худшую по сравнению с методикой скользящего регенерирующего окна метрику качества восстановления ряда, при этом общая среднеквадратическая ошибка составила 2.428 нТл и 2.315 нТл. Серийные интерполируемые 30-минутный пропуск при Kp = 4 и 20-минутный пропуск при Kp = 3 были восстановлены со значениями среднеквадратической ошибки соответственно 4.659 нТл и 3.491 нТл методом линейной интерполяции и 4.653 нТл и 3.487 нТл методом кубической сплайн-интерполяции (рисунок 8).

Таким образом, анализ результатов восстановления экспериментального временного ряда показал, что выбор оптимального способа импутации пропусков зависит от их типа и внешних факторов (значения индекса Kp). Так, дискретные пропуски могут быть восстановлены методом скользящей средней независимо от магнитной обстановки (и соответственно значения Kp-индекса). Серийные пропуски различной длительности рекомендуется восстанавливать методом кубической

сплайн-интерполяции (в случае спокойной магнитной обстановки при Kp < 3) или методом прецедентного резервирования (в случае неспокойной магнитной обстановки при Kp > 3). Применение единственного метода импутации без учета указанных факторов приводит к существенному искажению значений восстанавливаемого временного ряда, что независимо от формы анализируемого сигнала приводит к ошибочным выводам касательно локальной магнитной обстановки в отдельно взятой пространственной точке (в рамках магнитной обсерватории).

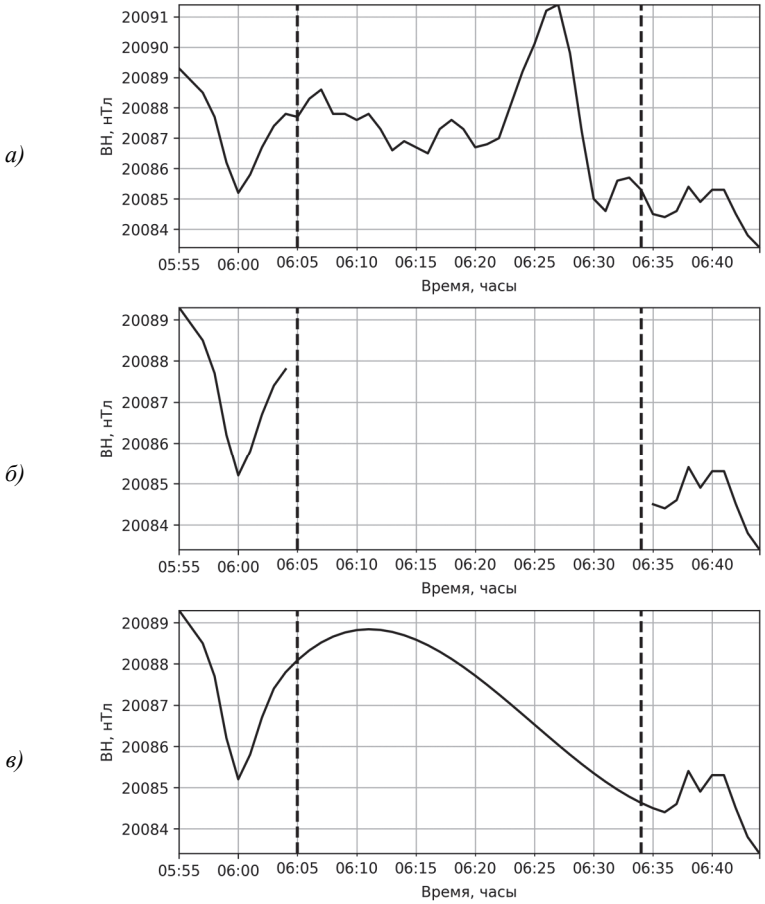


Рис. 8. Восстановление 30-минутного серийного интерполируемого пропуска геомагнитных данных при Kp = 4: *а* — оригинальные данные; *б* — данные с 30-минутным пропуском; *в* — результат восстановления методом кубической сплайн-интерполяции (среднеквадратическая ошибка = 4.653)

г)

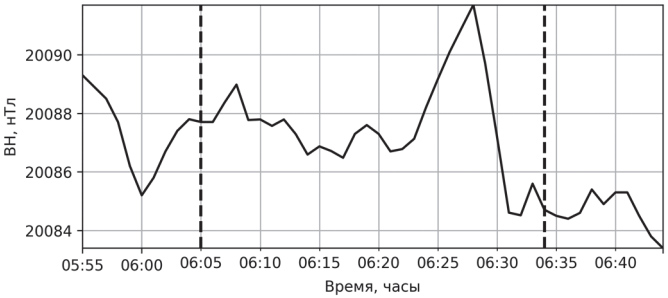


Рис. 8. Восстановление 30-минутного серийного интерполируемого пропуска геомагнитных данных при Kp = 4: *г* — результат восстановления методом прецедентного резервирования (среднеквадратическая ошибка = 0.748)

Основанная на данном принципе методика скользящего регенерирующего окна обеспечила лучшую метрику качества восстановления геомагнитных данных по сравнению с используемым для этих целей в настоящее время методом линейной интерполяции и методом, зарекомендовавшим себя в условиях спокойной магнитосферы — методом кубической сплайн-интерполяции.

8. Заключение. Одной из особенностей временных рядов геомагнитных данных является зависимость характера изменения их уровней от состояния магнитосферы в соответствующий момент времени. Сложность восстановления геомагнитных данных в условиях неспокойной магнитосферы обусловлена возникающими при этом вариациями параметров геомагнитного поля, которые приводят к сложным скачкообразным изменениям уровней временного ряда и разрыву линий тренда, нарушению их цикличности и периодичности.

Вместе с тем каждая магнитная станция располагает архивами выполненных ей поминутных измерений параметров геомагнитного поля, в совокупности составляющих базу ее прецедентов. Подтверждаемый коррелограммами линейный характер зависимости соседних значений уровня временного ряда позволяет сравнивать сегменты временного ряда по степени их корреляции и на основании этого считать их схожими и взаимозаменяемыми. Очевидно, что чем больше база прецедентов магнитной станции, тем больше корреляция между восстанавливаемым и замещающим сегментами и выше точность импутации значений временного ряда.

Данный подход положен в основу предложенного в настоящей работе метода прецедентного резервирования пропущенных значений временных рядов геомагнитных данных, базирующегося на индуктивном методе обучения по прецедентам и отличающимся тем, что в качестве признаков прецедентов используются данные, предшествую-

щие и последующие за пропуском во временном ряду. Эксперименты показали, что метод прецедентного резервирования позволяет в среднем на 79.54 % повысить точность восстановления временного ряда геомагнитных данных в условиях возбужденной магнитосферы по сравнению с известными методами импутации данных такого вида.

Кроме того, в настоящей работе авторами предложен и верифицирован подход к восстановлению пропущенных значений временного ряда геомагнитных данных, получивший название методики скользящего регенерирующего окна, суть которого сводится к декомпозиции временного ряда на множество пересекающихся сегментов, количество которых равно числу пропусков в ряду. К каждому выделенному сегменту применяется тот метод восстановления данных, который является наиболее эффективным применительно к выявленному типу пропуска и в условиях заданного значения Kp-индекса геомагнитной активности. Проведенные численные эксперименты показали, что восстановления временного ряда геомагнитных данных посредством данной методики позволяет повысить точность импутации по сравнению с подходом, при котором все пропуски восстанавливаются одним выбранным методом (например, используемым в настоящее время в геофизике методом линейной интерполяции).

По результатам проведенных исследований ожидается, что предлагаемые авторами метод и методика позволят повысить качество обработки и интерпретации геомагнитных данных, регистрируемых постоянными магнитными станциями, за счет более точного восстановления пропущенных при этом значений независимо от состояния магнитосферы в соответствующий период времени.

Литературы

1. *Мандрикова О.В., Жижикина Е.А.* Автоматический способ оценки состояния геомагнитного поля // Компьютерная оптика. 2015. Т. 39. № 3. С. 420–428.
2. INTERMAGNET technical reference manual. Version 4.6 / edited by S.-L. Benoît. Edinburgh: INTERMAGNET. BGS. 2012. 100 p.
3. *Love J.J., Chulliat A.* An international network of magnetic observatories // Eos, Transactions, American Geophysical Union. 2013. vol. 94(42). pp. 373–374.
4. *Macmillan S., Olsen N.* Observatory data and the Swarm mission // Earth, Planets and Space. 2013. vol. 65. no. 11. pp. 1355–1362.
5. *Гвишиани А.Д., Лукьянова Р.Ю.* Геоинформатика и наблюдения магнитного поля Земли: российский сегмент // Физика Земли. 2015. № 2. С. 3–20.
6. *Mandea M., Korte M.* Geomagnetic Observations and Models // Springer. 2011. 343 p.
7. *Рыбкина А.И. и др.* Интерполяция данных обсерваторских измерений и визуализация полной напряженности магнитного поля Земли // Вестник Отделения наук о Земле РАН. 2013. Т. 5. № 3002. С. 1–4.
8. *Gvishiani A. et al.* Survey of geomagnetic observations made in the northern sector of Russia and new methods for analysing them // Surveys in Geophysics. 2014. vol. 35(5). pp. 1123–1154.

9. *Soloviev A. et al.* Mathematical tools for geomagnetic data monitoring and the INTERMAGNET Russian segment // Data Science Journal. 2013. vol. 12. pp. WDS114-WDS119.
10. *Damodar N.* Gujarati. Basic Econometrics // The McGraw-Hill Companies. 2004. 1002 p.
11. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс // М.: Дело. 2007. 504 с.
12. *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов // М.: Финансы и статистика. 2003. 416 с.
13. *Vorobev A.V., Vorobeva G.R.* Web-oriented 2D/3D-visualization of geomagnetic field and its variations parameters // Scientific visualization. 2017. vol. 9. no.2. pp. 94–101.
14. *De Gooijer J.G.* Elements of nonlinear time series analysis and forecasting // Springer. 2017. 618 p.
15. *Jain E., Mallick D.* A Study of time series models ARIMA and ETS // International Journal of Modern Education and Computer Science (IJMECS). 2017. vol. 9. no.4. pp. 57–63.
16. *Pfaff B.* Analysis of integrated and cointegrated time series with R // Springer. 2008. 190 p.
17. *Чучуева И.А.* Модель прогнозирования временных рядов по выборке максимального подобия: диссертация канд. тех. наук // М.: Московский государственный технический университет им. Н.Э. Баумана. 2012. 154 с.
18. *Box G. et al*. Time series analysis: forecasting and control // New York: John Wiley & Sons. 2017. 712 p.
19. *Кобзарь А. И.* Прикладная математическая статистика // М.: Физматлит. 2006. 403 с.
20. *Langford J.* Quantitatively tight sample complexity bounds // Carnegie Mellon Thesis. 2002. 130 p.
21. *Moritz S., Sard´a A., Bartz-Beielstein T.* Comparison of different Methods for Univariate Time Series Imputation in R // arXiv preprint arXiv:1510.03924. 2015.
22. *Литтл Р.Дж.А., Рубин Д.Б.* Статистический анализ данных с пропусками // М. 1991. 336 с.
23. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies. 2002. vol. 6. no. 1. pp. 51–61.

Воробьев Андрей Владимирович — к-т техн. наук, доцент, доцент кафедры геоинформационных систем факультета информатики и робототехники, ФГБОУ ВО Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: геоинформационные технологии, цифровая обработка сигналов. Число научных публикаций — 138. geomagnet@list.ru, http://www.geomagnet.ru; ул. К. Маркса, 12, Уфа, 450008; р.т.: +7(917)345-2299.

Воробьева Гульнара Равилевна — к-т техн. наук, доцент кафедры вычислительной математики и кибернетики факультета информатики и робототехники, ФГБОУ ВО Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки информации. Число научных публикаций — 114. gulnara.vorobeva@gmail.com, http://www.geomagnet.ru; ул. К. Маркса, 12, Уфа, 450008; р.т.: +7(917)417-4111.

DOI 10.15622/sp.57.5

A.V. VOROBEV, G.R. VOROBEVA

INDUCTIVE METHOD OF GEOMAGNETIC DATA TIME SERIES RECOVERING

Vorobev A.V., Vorobeva G.R. **Inductive Method of Geomagnetic Data Time Series Recovering.**

Abstract. Today intensive development of systems and technologies for registration of the Earth's magnetic field parameters causes an exponential increase of geomagnetic data quantity, mainly collected by the ground magnetic stations. Imperfection of applied equipment and enabled channels of information transfer leads to the presence of omissions in the registered data time series. Along with spatial anisotropy it creates a serious obstacle to the processing of geomagnetic data. Russian and foreign scientific organizations are used to recover missing geomagnetic data by the linear interpolation. The approach provides admissible results in conditions of a quiet magnetosphere, but significantly distorts time series when changing the surrounding magnetic environment. This fact causes a scientific and technical problem, which is concerned with the development of new approach to recovering geomagnetic data registered in unquiet magnetosphere with acceptable time series imputation quality metrics.

The authors suggest the approach for time series recovering based on inductive method of machine learning. According to the approach each magnetic station operates its own knowledge base, which is formed during the registration of geomagnetic field and its variations parameters. The combination of the values of the series preceding and following the gap is supposed to be a characteristic description, which is used for searching the precedent in the magnetic station knowledge base. The result contains the required fragment of the time series, which replaces the missing values. The complexity of the information signal, caused by an unquiet magnetic environment, increases the accuracy of search by precedents. The greater the knowledge base of the magnetic station, the higher the effectiveness of the search.

Analysis of the results obtained during gap recovering in geomagnetic data time series (registered in conditions of unquiet magnetosphere) demonstrated that the suggested inductive method of imputation allows increasing the accuracy of the missing values recovery by an average of 79.54% compared with the methods currently used. The approach will enhance the efficiency of geomagnetic data processing for solving applied problems.

Keywords: geomagnetic data, time series, missing values, machine learning, learning by precedents, time series imputation.

Vorobev Andrei Vladimirovich — Ph.D., associate professor, associate professor of geoinformation systems department of computer science and robotics faculty, Ufa State Aviation Technical University (USATU). Research interests: geoinformation technologies, digital signal processing. The number of publications — 138. geomagnet@list.ru, http://www.geomagnet.ru; 12, K. Marx St., Ufa, 450008, Russia; office phone: +7(917)345-2299.

Vorobeva Gulnara Ravilevna — Ph.D., associate professor of computational mathematics and cybernetics department of computer science and robotics faculty, Ufa State Aviation Technical University (USATU). Research interests: geoinformation and web technologies, systems of information storing and processing. The number of publications — 114. gulnara.vorobeva@gmail.com, http://www.geomagnet.ru; 12, K. Marx St., Ufa, 450008, Russia; office phone: +7(917)417-4111.

References

1. Mandrikova O.V., Zhizhikina E.A. [Automated approach to estimate geomagnetic field state]. *Komp'juternaja optika – Computer Optics*. 2015. vol. 39. no. 3. pp. 420–428. (In Russ.).

2. INTERMAGNET technical reference manual. Version 4.6. Edited by S.-L. Benoît. Edinburgh: INTERMAGNET. BGS. 2012. 100 p.
3. Love J.J., Chulliat A. An international network of magnetic observatories. *Eos, Transactions, American Geophysical Union*. 2013. vol. 94(42). pp. 373–374.
4. Macmillan S., Olsen N. Observatory data and the Swarm mission. *Earth, Planets and Space*. 2013. vol. 65. no. 11. pp. 1355–1362.
5. Gvishiani A.D., Luk'janova R.Ju. [Geoinformatics and the Earth's magnetic field observations: Russian segment]. *Fizika Zemli – Physics of the Earth*. 2015. vol. 2. pp. 3–20. (In Russ.).
6. Mandea M., Korte M. Geomagnetic Observations and Models. Springer. 2011. 343 p.
7. Rybkina A.I. et al. [Interpolation of observatory measurements data and visualization of the Earth's magnetic field intensity]. *Vestnik Otdelenija nauk o Zemle RAN – Bulletin of the Department of Earth Sciences of the Russian Academy of Sciences*. 2013. vol. 5. no. 3002. pp. 1–4. (In Russ.).
8. Gvishiani A. et al. Survey of geomagnetic observations made in the northern sector of Russia and new methods for analysing them. *Surveys in Geophysics*. 2014. vol. 35(5). pp. 1123–1154.
9. Soloviev A. et al. Mathematical tools for geomagnetic data monitoring and the INTERMAGNET Russian segment. *Data Science Journal*. 2013. vol. 12. pp. WDS114-WDS119.
10. Livshic M. *Sluchajnye processy – ot teorii k praktike* [Random processes – from theory to practice]. M.: Lan'. 2016. 320 p.
11. Aue A., Dubart D., Hörmann S. On the prediction of stationary functional time series. *Journal of the American Statistical Association*. 2015. vol. 110(509). pp. 378–392.
12. Lukashin Ju. P. *Adaptivnye metody kratkosrochnogo prognozirovanija vremennyh rjadov* [Adaptive methods of short-term forecasting of time series]. Moscow: Finance and Statistics. 2003. 416 p.
13. Vorobev A.V., Vorobeva G.R. Web-oriented 2D/3D-visualization of geomagnetic field and its variations parameters. *Scientific visualization*. 2017. vol. 9. no. 2. pp. 94–101.
14. De Gooijer J.G. Elements of nonlinear time series analysis and forecasting. Springer. 2017. 618 p.
15. Jain E., Mallick D. A Study of time series models ARIMA and ETS. *International Journal of Modern Education and Computer Science (IJMECS)*. 2017. vol. 9. no.4. pp. 57–63.
16. Pfaff B. Analysis of integrated and cointegrated time series with R. Springer. 2008. 190 p.
17. Chuchueva I.A. *Model' prognozirovanija vremennyh rjadov po vyborke maksimal'nogo podobija. Dissertacija kand. teh. nauk* [Model of forecasting of time series on a sample of the maximum similarity. Ph.D. Thesis]. Moscow. 2012. 154 p. (In Russ.).
18. Box G. et al. Time series analysis: forecasting and control. New York: John Wiley & Sons. 2017. 712 p.
19. Kobzar A.I. *Prikladnaja matematicheskaja statistika* [Applied Mathematical Statistics]. Moscow: Physmatlit. 2006. 403 p. (In Russ.).
20. Langford J. Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. 2002. 130 p.
21. Moritz S., Sard´a A., Bartz-Beielstein T. Comparison of different Methods for Univariate Time Series Imputation in R. arXiv preprint arXiv:1510.03924. 2015.
22. Little R.J.A., Rubin D.B. *Statistical analysis with missing data*. John Wiley & Sons, 1987. 278 p. (Russ. ed.: Littl R.Dzh.A., Rubin D.B. *Statisticheskij analiz dannyh s propusk*ami. Moscow. 1990. 336 p.).
23. Zloba E., Yackiv I. [Statistical methods for recovering missing data]. *Computer Modelling & New Technologies*. 2002. vol. 6. no. 1. pp. 51–61.

УДК 004.896 **DOI** 10.15622/sp.57.6

В.П. Андреев, П.Ф. Плетенев

МЕТОД ИНФОРМАЦИОННОГО ВЗАИМОДЕЙСТВИЯ ДЛЯ СИСТЕМ РАСПРЕДЕЛЕННОГО УПРАВЛЕНИЯ В РОБОТАХ С МОДУЛЬНОЙ АРХИТЕКТУРОЙ

Андреев В.П., Плетенев П.Ф. **Метод информационного взаимодействия для систем распределенного управления в роботах с модульной архитектурой.**

Аннотация. В статье приведено решение проблемы информационного взаимодействия между встраиваемыми вычислительными устройствами при реализации *распределенного управления* в информационно-измерительной и управляющей системе (ИИУС) роботов с модульной архитектурой. Распределенное управление реализуется за счет проектирования каждого модуля как устройства со своей собственной ИИУС, содержащей все необходимые для выполнения своего функционала компоненты, включая вычислительные устройства. Вследствие такой функциональной завершенности модулей происходит *распараллеливание вычислительного процесса* функционирования робота как единой мехатронной системы. В результате существенно снижаются требования к мощности вычислительных устройств ИИУС модулей, в качестве которых оказывается возможным использовать недорогие микроконтроллеры и одноплатные ЭВМ — встраиваемые вычислительные устройства.

Предложена сетевая организация структуры ИИУС робота, что позволило перенести свойство *реконфигурируемости* сети на структуру модульного робота. Анализ различных топологий сети показал, что топология типа «звезда» имеет ряд преимуществ по сравнению с топологией типа «шина» для применения в гетерогенных модульных роботах.

Показано, что использование Robot Operating System (ROS) для реализации информационного взаимодействия между встраиваемыми вычислительными устройствами либо невозможно, либо существенно затруднено. Предложена спецификация, предназначенная для создания соответствующих программных интерфейсов и языка межмодульного взаимодействия, обеспечивающих включение модулей сторонних производителей в режиме «plug and play». Спецификация основана на принципах ROS, но позволяет реализовать ПО на встраиваемых вычислительных устройствах. На основе многокритериальной оптимизации по Парето получены рекомендации для выбора соответствующих аппаратно-программных средств.

Работоспособность предложенного решения была доказана в ходе экспериментов на установке, состав которой приближен к условиям работы ИИУС гетерогенного модульного робота. Эксперименты показали, что совместная работа программной и аппаратной частей удовлетворяет всем обозначенным требованиям и применима для передачи сообщений исполнительного уровня с частотой до 100 Гц при любой нагрузке на сеть.

Ключевые слова: модульный робот, гетерогенный робот, мобильный робот, реконфигурируемость, система управления, распределенное управление.

1. Введение. Мобильные роботы (МР), предназначенные для работы в условиях, когда заранее невозможно определить вид предполагаемой работы, должны допускать быстрое изменение своей структуры непосредственно на месте проведения работ и во время самих работ. В работе А.В. Лопоты и Е.И. Юревича [1] отмечено, что «Именно в экстремальных ситуациях различных катастроф, аварий и активных противодействий имеет место предельная априорная

неопределенность как условий предстоящих работ, так и самого их перечня. Что делает особо актуальной возможность компоновать состав роботов непосредственно на месте работы и корректировать его в ходе самих работ».

Аналогичные требования предъявляются и к роботам, предназначенным для космических исследований. Так, в обзоре [2] робототехники, необходимой для создания лунной базы, приводится перечисление роботов различного назначения: от роботов-бульдозеров до роботов-строителей и роботов-ремонтников. Такое большое разнообразие объясняется тем, что «у лунной робототехники есть важная особенность. Обычно под роботом понимается машина, способная функционировать в заранее «недоопределенной» среде …».

Такой широкий спектр роботов экономически невыгодно завозить на Луну или в зону с экстремальными условиями по отдельности. Достаточно создать набор функциональных модулей и механизм автоматической реконфигурации (средствами самих модулей робота или внешней робототехнической или роботизированной системы). Тогда появляется возможность на месте автоматически собирать из этого набора роботы необходимого назначения.

Следовательно, для выполнения таких работ должны использоваться роботы с переменной структурой, то есть *реконфигурируемые* модульные мобильные роботы.

К реконфигурируемым модульным мобильным роботам относятся *гомогенные* и *гетерогенные* модульные роботы.

Гомогенные роботы состоят из множества одинаковых модулей. Каждый модуль содержит в себе все необходимые для работы компоненты — датчики, двигатели, движители, аккумуляторы, системы управления и так далее. По сути, гомогенный модульный робот — это суперпозиция множества отдельных одинаковых роботов. В качестве примера можно привести модульный робот M-TRAN [3]. Несмотря на простоту выполняемых каждым из модулей движений, множество модулей, построившись в некую структуру, способны выполнять сложные согласованные движения. M-TRAN, как и другие гомогенные роботы, например, ATRON [4] и TRANSMOTE [5], имеют общую особенность — несмотря на то, что каждый из модулей независим, управление работой всех модулей осуществляется модулем-супервизором на *исполнительном уровне*. В результате оказывается необходимым разрабатывать дополнительное достаточно сложное программное обеспечение (ПО), которое переводит команды модуля-супервизора в последовательности исполнительных команд для каждого из модулей-исполнителей и обеспечивает их согласованную работу. Это, в свою очередь, приводит к необходимости устанавливать на модуль-супервизор «мощный» компьютер.

Гетерогенные роботы состоят из конечного набора модулей различной функциональности, которые соединяются друг с другом через унифицированные механические, электрические и программные интерфейсы. В отличие от гомогенных, модули гетерогенных роботов, как правило, не могут нормально функционировать друг без друга. Первым из известных МР, который является гетерогенным и полностью модульным, подразумевающим единство интерфейсов, можно считать робот CEBOT [6]. К классу гетерогенных также относятся роботы SMART [7], Thor [8] и модульный робот для космических исследований [9]. Каждый из данных роботов — это синергетическое объединение различных модулей и субмодулей (отдельных механических компонент модулей) в единую мехатронную систему. Робот Thor является прямым наследником гомогенного модульного робота ODIN [10], включая системы взаимодействия — в обоих роботах используется шина RS-485. Робот SMART несколько отличается — один из его модулей, подключенный по беспроводному каналу сети Bluetooth к супервизору, становится ведущим на шине CAN, через которую он управляет другими модулями, транслируя полученные от супервизора команды, а полученные от остальных модулей данные отсылает через радиоканал супервизору. Иными словами, управление работой всех модулей осуществляется модулем-супервизором также на *исполнительном уровне*, что, как и в случае гомогенных роботов, приводит к необходимости использовать центральный вычислитель большой производительности.

С позиции создания мобильных роботов с переменной структурой, то есть *реконфигурируемых* модульных роботов, ориентированных на использование в случае предельной априорной неопределенности условий и видов предстоящих работ, считаем, что на сегодняшний день *преимуществами обладают гетерогенные роботы*. Также полагаем, что такие роботы должны обладать свойством автоматической перестройки элементов информационно-измерительной и управляющей системы (ИИУС) робота под новые состав и конфигурацию.

2. Система распределенного управления для гетерогенного робота. Одна из основных трудностей создания реконфигурируемых систем заключается в разработке систем управления, способных работать в режиме «plug and play». В таких роботах должна осуществляться автоматическая реорганизация общей структуры системы управления в соответствии с быстро меняющимся составом и конфигурацией роботизированного устройства. Данное требование приводит к значительному увеличению сложности системы скоординированного управления модулями гетерогенного робота как единой мехатронной системы.

Упростить такую систему управления позволяет использование принципа полной функциональности в конструкции и программном обеспечении модулей МР, представляющих собой мехатронные устройства. *Полная функциональность мехатронного устройства — это способность выполнять свою целевую функцию, используя только собственные средства для выполнения команд от внешней системы управления* [11].

Гетерогенный модульный робот — это конструкция из отдельных функциональных устройств — модулей. Каждый модуль в таком роботе — это мехатронное или электронное устройство. Такие устройства должны иметь универсальные механические, электрические и программные интерфейсы. Лучшей взаимозаменяемости модулей и гибкости структуры МР можно достичь, если каждый из модулей разработан для выполнения одной специфической функции и выполняет ее максимально хорошо, то есть является *полнофункциональным устройством*.

Полная функциональность модулей робота достигается путем проектирования каждого модуля как устройства со своей собственной ИИУС, содержащей все необходимые компоненты, включая вычислительные устройства. Такая функциональная завершенность является *основным отличием* нашего решения от рассмотренных гетерогенных роботов, где модуль-супервизор управляет работой всех модулей-исполнителей на исполнительном уровне. В нашем решении модуль-супервизор формирует лишь цель управления, которую передает в модуль-исполнитель, и проверяет только результат ее достижения исполнительным модулем, но не управляет процессом выполнения задачи этим модулем. Как следствие такого подхода, реализуется *распределенное управление*, что существенно снижает требование к вычислительной мощности компьютеров как модуля-супервизора, так и модулей-исполнителей. Такое распределенное управление позволяет распараллеливать вычислительный процесс реализации целевой функции робота за счет разделения процесса на функциональные подзадачи и распределения их между вычислительными устройствами ИИУС модулей. При этом в качестве таких вычислительных устройств должны выступать относительно простые и дешевые малогабаритные микроконтроллеры и/или одноплатные ЭВМ — встраиваемые вычислительные устройства.

Такая распределенная система управления модульного робота нуждается в специфическом механизме информационного взаимодействия, в котором каждый полнофункциональный модуль должен иметь возможность обмениваться информацией (включая

команды управления) *непосредственно с любым другим модулем*. Данное условие обеспечивает распараллеливание и независимость информационных потоков в коммуникационном канале. Такой механизм (framework) уже существует и называется Robot Operating System (ROS) [12]. Эта система считается одной из наиболее развитых для использования в мобильной робототехнике. Однако ее невозможно применить для работы на микроконтроллерах, а использование на многих одноплатных ЭВМ имеет существенные ограничения.

3. Метод информационного межмодульного взаимодействия. Предлагаемый метод предназначен для организации межмодульного информационного взаимодействия в системе распределенного управления гетерогенных модульных роботов с функционально-модульной структурой, работа которой основана на следующих принципах [13]:

1. Информационно-измерительная и управляющая система робота строится как локальная вычислительная сеть, узлами которой являются встраиваемые вычислительные устройства ИИУС модулей.
2. Каждый модуль является полнофункциональным электронным или мехатронным устройством со своей собственной ИИУС.
3. Включение любых мехатронных или электронных устройств в сетевую структуру робота осуществляется с помощью специального программного обеспечения — драйверов.

Сетевая структура ИИУС модульного робота придает роботу свойство *реконфигурируемости* и масштабируемости за счет использования хорошо проработанных сетевых протоколов и развитых библиотек. *Функциональная завершенность модулей* означает конструктивную и функциональную независимость, что позволяет использовать модули сторонних производителей, идентичные по своим функциям. *«Концепция драйверов»* заключается в том, что для каждого электронного или мехатронного устройства создается унифицированный сетевой управляющий протокол (программный интерфейс) на базе существующих низкоуровневых программных интерфейсов взаимодействия. Тогда производитель может сравнительно просто интегрировать свое устройство в ИИУС модульного робота, создав лишь соответствующий драйвер, реализующий детерминированные протоколом программные инструкции управления и протоколы сетевого взаимодействия. При появлении в сети нового устройства ПО системы управления каждого из модулей в процессе взаимодействия с драйвером нового модуля автоматически включает это устройство в общую сеть — реализуется режим «plug and play».

Для реализации предлагаемого метода межмодульного информационного взаимодействия необходимо было решить следующие задачи:

– выбрать тип и топологию коммуникационной сети;

– создать спецификацию — документ, определяющий требования и рекомендации к реализации механизма взаимодействия;

– определить требования к аппаратно-программному обеспечению модулей;

– разработать систему команд для каждого модуля МР.

4. Коммуникационные сети: выбор типа и топологии. В современных модульных роботах, как гомогенных, так и гетерогенных, в связи с миниатюризацией самих модулей и минимизацией количества контактов на механических интерфейсах используются различные последовательные шины:

– I2C на модулях ЦНИИ РТК;

– CAN — в роботах M-TRAN, SMART;

– IrDA (по сути RS-232) — в роботе ATRON;

– RS-485 — в роботах Thor и Odin.

Последовательная шина, в широком смысле, используется и на роботе TRANSMOTE (используется беспроводной интерфейс ZigBee).

Чаще всего коммуникация в известных модульных роботах строится на топологии типа «шина» с локальными ответвлениями (рисунок 1*а*), реже — по топологии типа «звезда» (рисунок 1*б*). Топология «шина» наиболее полно отражает структуру модульного робота — каждый модуль имеет нескольких «соседей», через которых он взаимодействует со всей остальной системой. Недостатком таких шин является их низкая пропускная способность. Но в перечисленных конструкциях пропускной способности «шины» достаточно, поскольку у таких роботов есть, по сути, два варианта работы:

– выполнять набор простых правил, которые разработчики поместили в ПЗУ микроконтроллера в каждом из модулей;

– выполнять команды внешнего компьютера, подаваемые модулям через какой-либо один модуль, к шине которого (а значит, и к общей шине соединенных модулей) подключен компьютер.

Это означает, что роль системы межмодульного взаимодействия сводится только к наиболее быстрой и эффективной передаче команд исполнительного или тактического уровней и показаний датчиков. Тогда информационный поток в процессе межмодульного взаимодействия невелик, поскольку на каждом модуле либо нет датчиков, либо их установлено не так много и, фактически, необходимо передавать лишь команды *исполнительного уровня*.

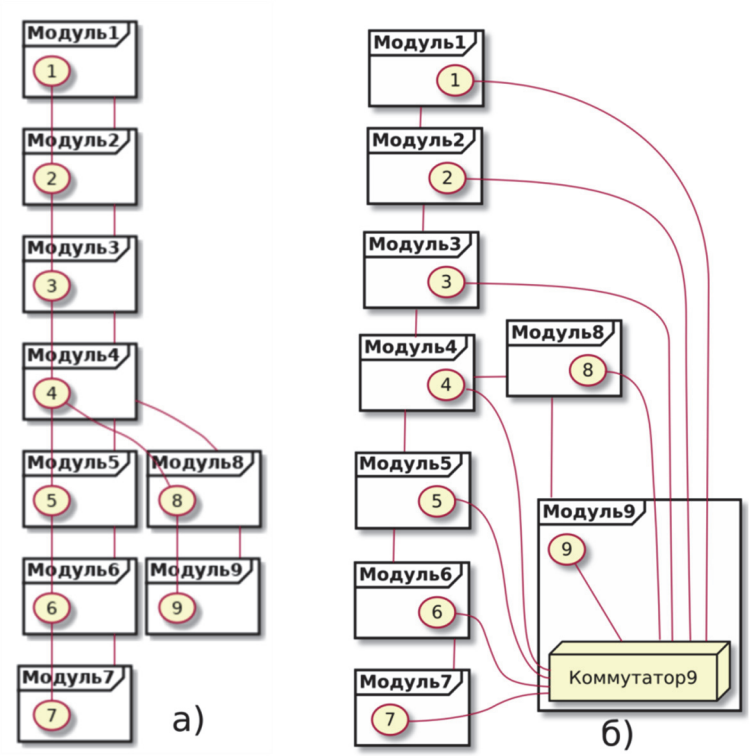


Рис. 1. Топологии сети Ethernet: а) «шина»; б) «звезда»

Многие компании, разрабатывающие средства автоматизации, в настоящее время переходят к использованию коммуникационных сетей, основанных на стандарте Ethernet или его модификациях, ориентированных для работы с детерминированными временами доставки сообщений. Такая популярность Ethernet в среде промышленных сетей является следствием универсальности и открытости стандарта Ethernet. Примерами доработки стандарта являются EtherCAT [14] и Ethernet Powerlink [15]. Оба этих стандарта переопределяют формат кадра, передаваемого по сети, что позволяет добиться значительно меньших времен задержек — 50-100 наносекунд в случае EtherCAT и единиц микросекунд в случае Ethernet Powerlink.

Сравнение различных существующих сетей приведено в таблице 1.

Таблица 1. Свойства различных сетей

Характеристики	Реальное время	Основные топологии	Макс. скорость, Мбит/с	Механизм борьбы с коллизиями	Служебная инф. и макс. длина сообщения, бит	Распространенность оборудования	Цены на оборудование
RS-485	Жесткое[1]	Шина	10 (на 10 м)	Не определен	80[7] / 2088[10]	Высокая	От низких до средних
CAN	Жесткое	Шина	1 (на 40 м)	CR[4]	44 или 64[8] / 64	Средняя	От средних до высоких
Ethernet	Мягкое	Точка-точка, разные[2]	100/1000 (на 100 м)	CD[4]	464 или 336 / 11536 или 11664[9]	Очень высокая	От низких до средних
EtherCAT	Жесткое	Точка-точка, шина, разные[3]	100 (на 100 м)	Master-Slave[5]	160 / 11888	Низкая	Высокие
Powerlink	Жесткое	Шина	100/1000 (на 100 м)	Master-Slave[6]	128 / 11920	Высокая	От низких до средних

Примечания к таблице 1: (1) — зависит от реализации, стандарт не определяет этот параметр; (2) — топология может быть как «шина», так и «звезда» в зависимости от дополнительного оборудования; (3) — может иметь любую топологию, однако большинство устройств поддерживают топологию «шина»; (4) — CAN является синхронной шиной с типом доступа Collision Resolving (CR, разрешение коллизии), который, в отличие от Collision Detect (CD, обнаружение коллизии) в сетях Ethernet, детерминировано обеспечивает доступ на передачу сообщения; (5) — использует специальный формат кадра, который формирует ведущее устройство, а ведомые читают обращенные к ним части кадра и записывают в кадр свои значения, тем самым достигается контроль занятия шины в каждый момент времени; (6) — часы на всех устройствах синхронизируются, ведущий подает сигнал начала обмена и получает или передает данные ведомым устройствам; (7) — для реализации MODBUS; (8) – для обычного и расширенного формата кадра; (9) — для протоколов TCP и UDP соответственно; (10) — для реализации MODBUS длина сообщений строго фиксирована.

Использование принципа функциональной завершенности модулей снимает необходимость в передаче команд *исполнительного уровня* – передаются только команды верхних (тактического и стратегического) уровней. В этом случае допустимы более «мягкие» требования к времени доставки сообщений и использование интерфейсов с «мягким реальным временем». В нашем случае мы планируем использовать в сенсорной системе робота большое количество разнообразных датчиков и многокамерную систему технического зрения (многопотоковое видео). Следовательно, *межмодульный коммуникационный канал должен быть широкополосным*.

Согласно данным таблицы 1, Ethernet является наиболее скоростным интерфейсом, а широкая распространенность соответствующего оборудования обеспечивает его низкую стоимость. В случае использования EtherCAT или Powerlink все межмодульное взаимодействие обязано происходить через «ведущего» шины, что снижает эффективность прямого межмодульного взаимодействия (не реализуется распараллеливание информационных потоков). Дополнительным препятствием для использования сетей EtherCAT и Powerlink является необходимость создания сложных встраиваемых устройств «с нуля» для реализации систем управления отдельными модулями. В случае Powerlink возможно использование готовых одноплатных компьютеров, однако стоимость таких компьютеров на 1-2 порядка больше стоимости решений на основе микроконтроллеров.

Сеть Ethernet изначально не является шиной — в ней используются соединения типа «один к одному». Возможно преобразовать ее в «шину», установив в узлы сети коммутаторы (switch), что позволит создавать соединения (топологию) типа «один ко многим». Такое решение обеспечивает оба варианта топологий сети: «шина» и «звезда».

На рисунках 1*а* и 1*б* приведены топологии сети «шина» и «звезда», наложенные на топологию модульного робота. Сравнение этих топологий приведено в таблице 2. Как видно из таблицы, обе топологии имеют недостатки — «шина» требует либо использовать большое число коммутаторов, либо встраивать в каждое устройство несколько интерфейсов Ethernet (по аналогии с EtherCAT), а «звезда» заставляет прокладывать значительно большее число проводов. Последнее в нашем случае не имеет существенного значения, поскольку проводная система локальной вычислительной сети располагается в пределах небольших размеров конструкции робота. Поэтому, мы предпочли топологию типа «звезда».

Таблица 2. Свойства различных топологий сети

Характеристики	«Шина»	«Звезда»
Централизованность	Децентрализованная	Централизованная
Количество коммутаторов	n коммутаторов с $s+m_i$ интерфейсами	1 коммутатор с $\sum_{i=1}^{n} m_i$ интерфейсами
Число линий Ethernet, проходящих через каждый модуль	$s+m_i$ проводов от коммутатора до физических интерфейсов	В худшем случае $(\sum_{i=1, i\neq i_c}^{n} m_i) - m_k$ проводов, где i_c — модуль с коммутатором
Дополнительные ограничения	Максимум 4 коммутатора на пути соединения «один к одному» для минимизации коллизий	Нет

Описание переменных, используемых в таблице 2: n — число модулей в модульном мобильном роботе; m_i — число Ethernet-подключений в каждом модуле (чаще всего $m_i = 1$); s — число возможных соседей модуля (чаще всего 2 или 6).

При выборе топологии сети для модульного робота следует, по-видимому, руководствоваться сопоставлением целого ряда параметров: помехоустойчивость, желаемая частота обмена сообщениями, наличие необходимых сетевых протоколов и соответствующих библиотек, программного и аппаратного обеспечения, требования к сборочными и экономическим показателям, ограничения функционирования сети той или иной топологии и тому подобное. Например, для роботов змеевидной конструкции, скорее всего, подойдет шинная топология, а для антропоморфных роботов — «звезда». Но этот вопрос требует подробного изучения и выработки соответствующих рекомендаций, что планируется сделать в дальнейших исследованиях.

5. Спецификация. Для совместимости программного обеспечения модулей робота необходимо создать *спецификацию* — документ, определяющий требования и рекомендации по реализации метода межмодульного взаимодействия. Подобные спецификации уже достаточно давно выпускает организация, создающая стандарты для работы Интернета, такие как, например, стандарт работы сети Ethernet [16]. Созданы также системы для описания жизненного цикла открытой спецификации — от разработки до «взросления» и «устаревания» [17]. Одним из требований к создаваемой спецификации является возможность использования готовых библиотек и устройств. Конечная цель — создание на основе такой спецификации специального ПО — драйверов и языка межмодульного взаимодействия, то есть создание соответствующих программных интерфейсов, обеспечивающих включение модулей сторонних производителей в режиме «plug and play».

Как было отмечено ранее, одним из самых известных методов межмодульного информационного взаимодействия является программный каркас (framework) ROS. Для ROS разработано большое количество средств визуализации, симуляции и отладки. Однако ROS не ориентирован для работы непосредственно на встраиваемых системах (микропроцессорах): необходим как минимум компьютер с полноценной Linux-подобной ОС, для которой существует порт этой системы; для работы с ним необходимо устанавливать на вычислительные устройства модулей довольно сложное и объемное ПО; требуется высокий уровень входа для его использования (для конфигурации, например, используются диалекты языка XML).

Наш предыдущий подход к реализации межмодульного информационного взаимодействия, предложенный в работе [18], основывался на библиотеке zmq_robot [19]. Она построена на основе

библиотеки ZMQ [20] и позволяет создавать как ПО для исполнительных модулей, так и управляющие программы для модуля-супервизора. Однако эта библиотека имеет тот же недостаток, что и ROS — для работы с ней необходимо устанавливать на устройства сложное ПО, которое работает только на полноценных ОС и не работает на встраиваемых системах.

Поэтому предлагается относительно простая система, которая совместима с ROS и при этом реализуема на встраиваемых системах. Под совместимостью с ROS здесь понимается возможность создания специального ПО — «моста» между создаваемой системой и узлами ROS. Это, в свою очередь, позволяет прозрачно передавать сообщения между системами и воспользоваться обширной базой существующих библиотек и систем, разработанных для ROS.

Сущность предлагаемого решения заключается в использовании библиотеки ZMQ, работающей по протоколу TCP и дополненной протоколом UDP с широковещательными сообщениями, который дает возможность реализовать спецификацию для встраиваемых систем. В отличие от ROS, здесь каждый модуль имеет один общий для всех модулей интерфейс, через который предоставляет общую информацию о себе: имя модуля, его описание, функциональные интерфейсы и прочее. Подробное описание спецификации привести в данной статье не представляется возможным; полный текст спецификации можно найти на интернет-сайте [21].

6. Требования к аппаратно-программной реализации системы распределенного управления. Программно-аппаратное обеспечение включает в себя аппаратно-программное средство (АПС) и соответствующее программное обеспечение. АПС — это электронное вычислительное устройство или набор таких устройств, работающих совместно и создающих как сетевой интерфейс, так и систему управления с помощью соответствующего программного обеспечения.

Аппаратно-программное средство для систем управления модулей робота должно иметь в минимальном исполнении встраиваемое вычислительное устройство, способное работать в сети Ethernet по протоколу UDP с производительностью, достаточной для реализации функционального назначения модуля. АПС также должно иметь минимальные массогабаритные параметры и минимально возможную стоимость. За счет распределения вычислений между микропроцессорами модулей система управления модульного робота в целом имеет более высокую производительность по сравнению с системами управления, которые построены на одном вычислителе.

Для выбора АПС, на базе которого можно реализовать предложенный механизм межмодульного информационного взаимодействия, был выполнен анализ доступных моделей АПС различных производителей:

- Arduino Uno + Ethernet Shield,
- Arduino Mega + Ethernet Shield,
- Arduino Due + Ethernet Shield,
- Taijiuino Due Pro R3 + Ethernet PHY,
- Arduino Yun,
- Seeeduino Cloud,
- Arduino Uno + Yun Shield,
- Arduino Mega + Yun Shield,
- Intel Galileo gen 2,
- Intel Edison,
- Raspberry Pi B Plus,
- Orange Pi Zero,
- 86duino ZERO,
- 86duino One.

По перечисленным АПС была проведена дискретная многокритериальная оптимизация с использованием критерия Парето [22]. Все критерии качества АПС оценивались по описаниям и техническим характеристикам рассмотренных устройств. Ниже приведены эти критерии, часть из которых являются объективными, часть — субъективными:

- $K1$: Возможность реализации спецификации межмодульного взаимодействия; критерий является суммой подкритериев:
  - $K1.1$: Наличие реализации библиотеки ZeroMQ для АПС (0 – реализации не существует; 0.5 — реализация не завершена, имеются серьезные недоработки; 1 — реализация стабильна).
  - $K1.2$: Наличие ограничений на число одновременных подключений к интерфейсу (-1 — ограничения есть и они существенны; 0 — информации об ограничениях нет; 1 — ограничений нет).
  - $K1.3$: Наличие ограничений на число одновременно работающих интерфейсов (-1 — ограничения есть и они существенны; 0 – информации об ограничениях нет; 1 — ограничений нет).
- $K2$: Возможность создавать приложения, работающие в жестком реальном времени; критерий является суммой подкритериев:

– $K2.1$: Наличие операционной системы реального времени (0.5 — ОС существует, но сложна в установке/удалении, 1 — ОС существует и легка в установке и использовании).

– $K2.2$: Уровень входа в программирование в реальном времени (-1 — высокий; 0 — средний; 1 — низкий).

– $K3$: Наличие аппаратного ускорения операций над вещественными числами (0 — нет, 0.5 — ускорения нет, но имеется хорошая программная реализация, 1 — есть).

– $K4$: Количество входов и выходов общего назначения (0, 1, 2, …).

– $K5$: Разрядность таймеров для генерации ШИМ-сигнала (0, 1, 2, …; обычно — 8, 16, 32).

– $K6$: Количество каналов генерации ШИМ-сигнала (0, 1, 2, …).

– $K7$: Количество каналов прерываний по уровню (0, 1, 2, …).

– $K8$: Количество полноценных каналов работы с энкодерами (0, 1, 2, …).

– $K9$: Количество каналов АЦП (0, 1, 2, …).

– $K10$: Стоимость АПС в рублях в России; если АПС состоит из нескольких частей, то их стоимость складывается. Так как стоимость АПС необходимо минимизировать, то этот критерий умножается на -1.

– $K11$: Стоимость АПС в рублях при заказе из Китая с учетом доставки; если АПС состоит из нескольких частей, то их стоимость складывается. Так как стоимость АПС необходимо минимизировать, то этот критерий умножается на -1.

Сравнительный анализ показал, что из 14 исследованных оптимальными являются 7:

– Arduino Uno + Ethernet Shield

$K_{1...10} = \{-1.5; 2; 0; 10; 8; 6; 2; 1; 6; -2347; -2347\}$,

– Arduino Mega + Ethernet Shield

$K_{1...10} = \{-1.5; 2; 0; 50; 8; 15; 6; 3; 16; -3547; -3547\}$,

– Taijiuino Due Pro R3

$K_{1...10} = \{0.5; 1; 0,5; 54; 12; 12; 54; 27; 12; -2699; -2699\}$,

– Seeeduino Cloud $K_{1...10} = \{3; 2; 0; 18; 8; 7; 2; 1; 12; -2850; -2850\}$,

– 86duino ZERO $K_{1...10} = \{3; 1; 1; 14; 32; 7; 14; 4; 6; -3690; -3641\}$,

– 86duino One $K_{1...10} = \{3; 1; 1; 45; 32; 11; 45; 4; 7; -5250; -5611\}$,

– Raspberry Pi B+ $K_{1...10} = \{3; -0.5; 1; 27; 0; 0; 0; 0; 0; -3100; -3100\}$.

7. Система команд модулей. Система команд для каждого модуля определяется, в первую очередь, его функционалом. По принципу полной функциональности можно выделить следующие функции модулей: транспортную, энергетическую, информационную, коммуникационную, технологическую, вспомогательную функции и функцию общего управления. В рамках разработанной спецификации к настоящему времени разработаны системы команд для следующих функциональных модулей лабораторного макета гетерогенного мобильного робота (рисунок 2):

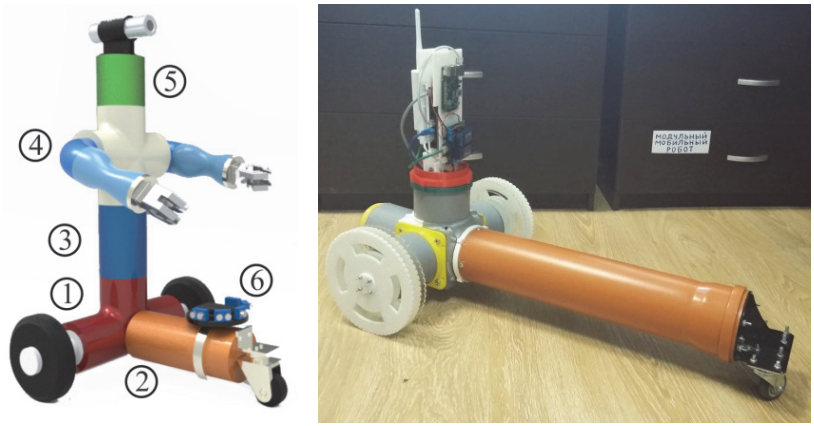


Рис. 2. Компьютерная модель и фотография лабораторного макета гетерогенного модульного мобильного робота: 1 — транспортный модуль, 2 — силовой модуль, 3 — модуль супервизорного управления, 4 — модуль манипуляторов, 5 — сенсорный модуль дальнего радиуса действия, 6 — сенсорный модуль ближнего радиуса действия

1. Транспортный модуль — обеспечивает перемещение робота (транспортная функция).

2. Силовой модуль — обеспечивает энергоснабжение электронных и электромеханических компонент робота и их безопасное включение и отключение (энергетическая функция).

3. Сенсорная система: модули датчиков ближнего и дальнего радиуса действия обеспечивают обнаружение препятствий и объектов манипулирования на разных дистанциях и предоставляют сенсорную информацию другим модулям (информационная функция).

4. Модуль супервизорного управления — обеспечивает управление роботом в целом, синхронизирует информационное взаимодействие всех модулей, выполняет обработку информации и постановку задач, поступающих от внешнего супервизора робота,

формирует и распределяет задания между модулями, являясь, таким образом, для них супервизором, создает и контролирует работу радиоканала (коммуникационная функция и функция общей системы управления). В перспективе коммуникационная функция будет реализована в виде отдельного модуля.

8. Эксперименты. Работоспособность предложенного сетевого и аппаратно-программного решения, предназначенного для реализации метода межмодульного информационного взаимодействия, была проверена в ходе поставленных экспериментов. В программной части использована библиотека ZMQ, дополненная протоколом UDP с широковещательными сообщениями. Была собрана установка, состоящая из следующих компонент:

- два АПС Arduino Uno с Ethernet Shield;
- АПС Arduino Yun с ОС OpenWRT 15.05;
- АПС Orange Pi Zero с ОС Armbian;
- управляющий ноутбук с ОС Arch Linux;
- сетевой коммутатор D-Link 1005C;
- сетевые кабели длиной ~ 1 м;
- шестнадцатиканальный логический анализатор Saleae Logic;
- провода питания.

Состав экспериментальной установки приближен к условиям работы ИИУС гетерогенного модульного робота: сеть построена по топологии типа «звезда», коммуникации выполнены кабелем типа UTP и имеют длину не более метра, узлы сети — рекомендованные к применению АПС (см. раздел 6). Несмотря на то, что следствием реализации спецификации для межмодульного информационного взаимодействия является низкая нагрузка на сеть, возможны ситуации, когда из-за ряда обстоятельств, например, таких как одновременная передача нескольких видеопотоков, возможна большая нагрузка на сеть. Это может повлиять на качество передачи сообщений. Следовательно, необходимо проверить работу сети *под нагрузкой и без нагрузки*.

Для эмуляции нагрузки на сеть использована утилита *iperf*. Она состоит из двух частей — «серверной» и «клиентской». «Серверная» часть принимает трафик и выбрасывает его, замеряя скорость приема, «клиентская» часть генерирует трафик, замеряя скорость передачи. Эмуляция передачи 2-х видеопотоков от 2-х модулей выполняется посредством запуска «серверной» части *iperf* на АПС Orange Pi Zero и «клиентских» частей на управляющем ноутбуке и АПС Arduino Yun.

Проведены следующие эксперименты:

1. Проверка предположения: сообщения, передаваемые по сети, не теряются и передаются последовательно без потерь.

2. Определение максимальной частоты синхронного обмена сообщениями.

3. Определение синхронности получения сообщений.

4. Измерение времени передачи и приема сообщений.

Эксперимент 1. Проверка потерь и перемешивания пакетов в сети. Цель эксперимента — проверить тот факт, что при разных условиях загрузки сети пакеты не теряются и их последовательность не перемешивается (хотя это и не гарантируется протоколом UDP).

Загрузим в АПС Arduino Uno «прошивку», посылающую раз в 10 миллисекунд широковещательное сообщение:

```
{ "data": "digital", "n": "0x000000386f" }
```

Листинг 1. Формат широковещательного сообщения

где ключ «n» содержит номер сообщения в виде шестнадцатеричного числа, записанного в виде строки фиксированной длины; тем самым все сообщение получается фиксированной длины в 43 символа. Будем получать на управляющем ноутбуке широковещательные сообщения от АПС Arduino Uno и расшифровывать их, сравнивая полученный номер с предыдущим. Если предыдущий номер отличается от текущего не на -1, выводим соответствующее сообщение. Проведем тесты в условиях сети без нагрузки и под нагрузкой.

Эксперимент показал, что вне зависимости от нагрузки на сеть, *пакеты не теряются и их последовательность не меняется.*

Эксперимент 2. Определение максимальной частоты синхронного обмена сообщениями. Цель эксперимента — замерить максимальную частоту синхронного обмена сообщениями и влияние на эту величину нагрузки на сеть. Для этого будем посылать с управляющего компьютера на метаинтерфейс одного из АПС Arduino Uno команду «ping» и принимать ответ на нее, замеряя время от начала передачи команды до получения и расшифровки ответа. Проверим для двух условий — без нагрузки и под нагрузкой, замерив 10000 посылок.

Определим минимальное, максимальное, медианное и среднее время задержки (Δt). Рассчитаем также среднеквадратическое отклонение среднего времени (σ) и среднюю частоту передачи сообщений ($\overline{f}$). Результаты экспериментов представлены ниже.

Отправка и получение сообщений без нагрузки:

$$\Delta t_{min} = 1.5 \text{ мс},$$

$$\Delta t_{max} = 17.1 \text{ мс},$$

$$\Delta t_{med} = 1.6 \text{ мс},$$

$$\overline{\Delta t} \pm \sigma = 2.0 \pm 0.9 \text{ мс},$$

$$\overline{f} = 499 \text{ Гц}.$$

Отправка и получение сообщений под нагрузкой:

$$\Delta t_{min} = 1.6 \text{ мс},$$

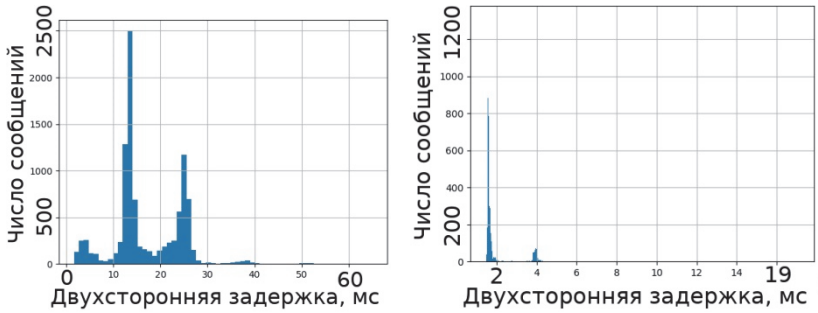
$$\Delta t_{max} = 65.0 \text{ мс},$$

$$\Delta t_{med} = 14.0 \text{ мс},$$

$$\overline{\Delta t} \pm \sigma = 16.9 \pm 7.3 \text{ мс},$$

$$\overline{f} = 59 \text{ Гц}.$$

Результаты эксперимента приведены на рисунке 3. Видно, что скорость передачи падает почти на порядок — с 500 до 60 Гц. Нижняя граница достаточна для управления человеком, *но почти неприменима для таких систем исполнительного уровня, где необходимы частоты порядка 500 Гц и выше.*



а) *б)*

Рис. 3. Гистограмма значений двухсторонней задержки в эксперименте: а) без нагрузки и б) под нагрузкой

Однако такое изменение не скажется на качестве передачи сообщений — представленная ранее спецификация межмодульного взаимодействия рекомендует использовать «синхронные» сокеты только для передачи команд тактического уровня, частота работы которого может быть на один-два порядка ниже, чем полученная в эксперименте частота в 60 Гц.

Эксперимент 3. Определение синхронности получения сообщений. Эксперименты с получением широковещательных сообщений от «асинхронного по UDP» сокета. Цель эксперимента — определить влияние загрузки сети на синхронность получения сообщений. АПС Arduino Uno каждые 10 мс отправляет описанное ранее сообщение (см. листинг 1), а программа на управляющем

ноутбуке замеряет время между получением каждого сообщения. Результаты экспериментов для двух условий — без нагрузки и с нагрузкой представлены ниже:

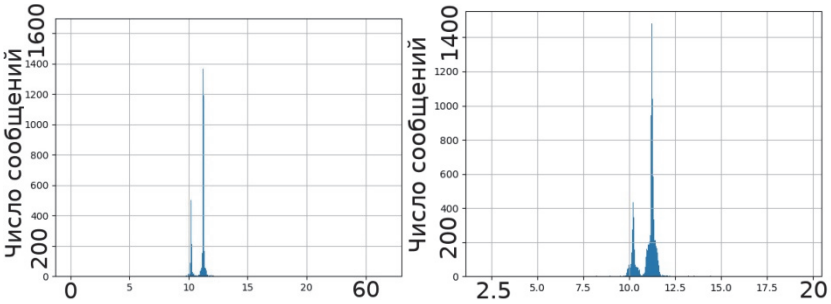
Прием широковещательных сообщений без нагрузки:

$$\Delta t_{min} = 6.5 \text{ мс},$$
$$\Delta t_{max} = 18.5 \text{ мс},$$
$$\Delta t_{med} = 13.3 \text{ мс},$$
$$\overline{\Delta t} \pm \sigma = 13.1 \pm 0.5 \text{ мс}.$$

Прием широковещательных сообщений под нагрузкой

$$\Delta t_{min} = 2.1 \text{ мс},$$
$$\Delta t_{max} = 28.8 \text{ мс},$$
$$\Delta t_{med} = 13.3 \text{ мс},$$
$$\overline{\Delta t} \pm \sigma = 11.2 \pm 0.5 \text{ мс}.$$

Результаты приведены на рисунке 4. Значение Δt_{min} в обоих экспериментах получено как время от запуска программы на управляющем компьютере до момента получения первого сообщения от АПС. Запуск программы на компьютере специально не синхронизовался с передачей сообщений от АПС, и потому Δt_{min} сильно отличается от среднего.



а) *б)*

Рис. 4. Гистограмма значений задержки между приемом сообщений в эксперименте: а) без нагрузки и б) под нагрузкой

По результатам эксперимента можно утверждать, что *нагрузка на сеть незначительно влияет на синхронность передачи широковещательных сообщений с частотой передачи до 100 Гц.*

Эксперимент 4. Измерение времени доставки сообщений. Целью эксперимента является измерение влияния загруженности сети

на время доставки сообщений. Для этого составим две разные программы для двух одинаковых АПС Arduino Uno. Первая программа (сервер) каждые 10 мс меняет состояние одного из выводов АПС на противоположное и отсылает его по сети в виде широковещательного сообщения. Вторая программа (клиент) ожидает прихода сообщения, получая и распаковывая его, устанавливает один из своих выводов в состояние, идентичное полученному по сети. Оба АПС Arduino подключаются к логическому анализатору, который, в свою очередь, подключается к управляющему ноутбуку. Для записи состояний выводов использовалась программа *Saleae Logic* версии 1.2.11. Запишем состояния выводов двух АПС на, соответственно, нулевой и первый канал логического анализатора в двух режимах работы сети — без нагрузки и под нагрузкой. Записи из ПО логического анализатора были помещены в файл и обработаны. В результате были получены следующие характеристики.

Передача сообщений без нагрузки на сеть. Произведена запись в течение 60 секунд; было получено 2725 фронта. Результаты:

$$\Delta t_{min} = 0.71 \text{ мс},$$
$$\Delta t_{max} = 0.80 \text{ мс},$$
$$\Delta t_{med} = 0.76 \text{ мс},$$
$$\overline{\Delta t} \pm \sigma = 0.76 \pm 0.02 \text{ мс}.$$

Передача сообщений при нагрузке на сеть. Произведена запись в течение 60 секунд; было получено 2725 фронта. Результаты:

$$\Delta t_{min} = 0.00 \text{ мс},$$
$$\Delta t_{max} = 1.20 \text{ мс},$$
$$\Delta t_{med} = 0.76 \text{ мс},$$
$$\overline{\Delta t} \pm \sigma = 0.76 \pm 0.03 \text{ мс}.$$

Как видно из результатов эксперимента, нагрузка на сеть *несущественно влияет на время доставки сообщений.*

По итогам приведенных выше экспериментов можно сделать *общий вывод*: выбранный *протокол UDP применим для передачи сообщений исполнительного уровня с частотой обмена сообщениями между устройствами не более 100 Гц при любой нагрузке на сеть. При снижении нагрузки на сеть эта частота может быть увеличена.*

Разработанный механизм межмодульного информационного взаимодействия был реализован для транспортного, силового и частично для модуля сетевого управления в составе лабораторного макета гетерогенного модульного мобильного робота (см. рисунок 2).

Для вычислительных устройств ИИУС модулей были созданы управляющие программы, работающие на основе представленной выше спецификации. Для силового модуля, как наименее производительной части системы, удалось добиться частоты обмена сообщениями до 500 Гц. Эксперименты показали эффективность предложенных решений.

8. Заключение. В качестве основного результата проведенной работы следует считать *решение проблемы информационного взаимодействия* между встраиваемыми вычислительными устройствами при реализации распределенного управления в информационно-измерительной и управляющей системе (ИИУС) роботов с модульной архитектурой.

Предложенное решение основывается на функциональной завершенности модулей робота, которая достигается за счет проектирования каждого модуля как мехатронного или электронного устройства со своей информационно-измерительной и управляющей системой, содержащей все необходимые для реализации своего функционала компоненты. Как следствие такого подхода, реализуется *распределенное управление*, что существенно снижает требование к вычислительной мощности вычислительных устройств отдельных модулей. Это, в свою очередь, позволяет использовать в качестве таких вычислительных устройств недорогие малогабаритные микроконтроллеры и одноплатные ЭВМ.

Предложена сетевая организация структуры ИИУС робота, что позволило перенести такие свойства локальной вычислительной сети как *реконфигурируемость* и масштабируемость на структуру модульного робота. Анализ различных топологий сети показал, что топология типа «звезда» имеет ряд преимуществ по сравнению с топологией типа «шина» для применения в гетерогенных модульных роботах.

Показано, что реализация распределенного управления нуждается в таком механизме информационного взаимодействия, в котором каждый модуль имеет возможность обмениваться информацией *непосредственно с любым другим модулем*. Сравнительный анализ существующих сетей показал, что данному требованию удовлетворяет лишь сеть Ethernet, которая имеет преимущество также по целому ряду других показателей, включая пропускную способность, распространенность оборудования и его стоимость.

Для совместимости программного обеспечения модулей разработана спецификация, предназначенная для создания специального ПО (драйверов) и языка межмодульного информационного взаимодействия, то есть соответствующих

программных интерфейсов, обеспечивающих включение модулей сторонних производителей в режиме «plug and play». Спецификация основана на принципах широко известного программного каркаса (framework) ROS и позволяет реализовать ПО на встраиваемых вычислительных устройствах, в отличие от ROS, который невозможно использовать из-за высоких требований к вычислительной мощности АПС. Спецификация построена на основе библиотеки обмена сообщениями по протоколу TCP — библиотека ZeroMQ, и в дополнение используются сокеты по протоколу UDP для прямого взаимодействия со встраиваемыми системами.

На основе многокритериальной оптимизации по Парето получены рекомендации для выбора аппаратно-программных средств, на базе которых можно реализовать разработанный механизм межмодульного информационного взаимодействия. Соответствующим требованиям удовлетворяют 7 доступных АПС, что обеспечивает широкий выбор для разработчиков модулей.

Работоспособность предложенного метода межмодульного информационного взаимодействия, реализованного на базе сетевого (ЛВС типа Ethernet) и аппаратно-программного (встраиваемые АПС) решения, была проверена в ходе экспериментов на установке, состав которой приближен к условиям работы ИИУС гетерогенного модульного мобильного робота. В программной части использована библиотека ZMQ, дополненная протоколом UDP с широковещательными сообщениями, использованные АПС взяты из рекомендованных к применению. Выполнена проверка работы сети под нагрузкой и без нагрузки. Эксперименты показали, что совместная работа программной и аппаратной частей удовлетворяет всем обозначенным требованиям и *применима для передачи сообщений исполнительного уровня с частотой до 100 Гц, которая может быть увеличена при снижении нагрузки на сеть.*

Предлагаемый метод межмодульного информационного взаимодействия был реализован для транспортного, силового и частично для модуля-супервизора в составе лабораторного макета гетерогенного модульного мобильного робота. Библиотеки функций, реализующих разработанную спецификацию, созданы для трех встраиваемых устройств (Arduino Yun, Arduino Mega2560, Raspberry Pi B+) и использованы для реализации соответствующих драйверов. Разработана система команд для перечисленных модулей.

В дальнейшем необходимо создать реализации спецификации для других архитектур встраиваемых устройств (таких как ARM или Xtensa). Необходимо разработать системы команд для таких модулей,

как модуль манипулятора, сенсорный модуль дальнего радиуса действия и других.

Также предполагается разработать алгоритмы супервизорного управления движением робота в недетерминированной среде и алгоритмы супервизорного управления манипулированием различными объектами.

Литература

1. *Лопота А.В., Юревич Е.И.* Этапы и перспективы развития модульного принципа построения робототехнических систем // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. СПб: Изд-во Политехнического ун-та. 2013. №1(164). С. 98–103.
2. *Платонов А.К.* Робототехника лунной базы // XXXIV Чтения по космонавтике. URL: http://www.keldysh.ru/section5/report.xhtm?src=section5.xml&filter=3. (дата обращения: 10.06.2017).
3. *Murata S. et al.* M-TRAN: self-reconfigurable modular robotic system // IEEE/ASME Transactions on Mechatronics. 2002. vol. 7(4). pp. 432–441.
4. *Østergaard E.H., Kassow K., Beck R., Lund H.H.* Design of the ATRON lattice-based self-reconfigurable robot // Autonomous Robots. 2006. vol. 21(2). pp. 165–183.
5. Qiao G. et al. Design of Transmote: a Modular Self-Reconfigurable Robot with Versatile Transformation Capabilities // Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics. 2012. pp. 1331–1336.
6. *Fukuda T., Ueyama T., Kawauchi Y., Arai F.* Concept of cellular robotic system (CEBOT) and basic strategies for its realization // Computers Elect Engng. 1992. vol. 18. no. 1. pp. 11–39.
7. *Baca J., Ferre M., Aracil R.* A heterogeneous modular robotic design for fast response to a diversity of tasks // Robotics and Autonomous Systems. 2012. vol. 60. no. 4. pp. 522–531.
8. *Lyder A.H. et al.* On sub-modularization and morphological heterogeneity in modular robotics // Intelligent Autonomous Systems of Advances in Intelligent Systems and Computing. 2013. vol. 193. no. 12. pp. 649–661.
9. *Hancher M.D., Hornby G.S.* A modular robotic system with applications to space exploration // 2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT'06). 2006. pp. 132–140.
10. *Garcia R.F.M., Lyder A., Christensen D.J., Stoy K.* Reusable Electronics and Adaptable Communication as Implemented in the Odin Modular Robot // IEEE International Conference on Robotics and Automation. 2009. pp. 1152–1158.
11. *Andreev V., Kim V., Pletenev P.* The principle of full functionality – the basis for rapid reconfiguration in heterogeneous modular mobile robots // Proceedings of the 28th DAAAM International Symposium. 2017. pp. 0023–0028.
12. *Quigley M. et al.* ROS: an open-source Robot Operating System // ICRA workshop on open source software. 2009. vol. 3. no. 3.2. pp. 5.
13. *Андреев В.П., Ким В.Л., Подураев Ю.В.* Сетевые решения в архитектуре гетерогенных модульных мобильных роботов // Робототехника и техническая кибернетика. 2016. № 3(12). С. 23–29.
14. EtherCAT Technology Group, Industrial Ethernet Technologies. URL: https://www.ethercat.org/download/documents/Industrial_Ethernet_Technologies.pdf. (дата обращения: 17.03.2016).
15. Ethernet POWERLINK Communication Profile Specification Version 1.2.0. URL: http://www.ethernet-powerlink.org/en/downloads/technical-documents/action/open-

download/download/epsg-ds-301-v120-communication-profile-specification/element/5158/?no_cache=1 (дата обращения: 17.03.2016).
16. RFC: 791 Internet Protocol. URL:https://tools.ietf.org/html/rfc791 (дата обращения: 17.05.2017).
17. Digital Standards Organization COSS – Consensus Oriented Specification System. URL: https://web.archive.org/web/20161002092144/http://www.digistan.org/spec:1/coss (дата обращения: 17.05.2017).
18. *Андреев В.П., Плетенев П.Ф.* Разработка технологии межмодульного общения в гетерогенном модульном мобильном роботе // Труды международной научно-технической конференции «Экстремальная робототехника». 2016. С. 245–255.
19. *Kirsanov K.* Software architecture of control system for heterogeneous group of mobile robots // Procedia Engineering. 2015. vol. 100. pp. 216–221.
20. *Hintjens P.* "0MQ – The Guide". URL: http://zguide.zeromq.org/page:all (дата обращения: 17.03.2016).
21. *Плетенев П.Ф.* 1/ПММВ – Протокол взаимодействия в гетерогенном модульном мобильном роботе. URL: https://asmfreak.github.io/modular_robots_rfc/1/ПММВ/ (дата обращения: 20.01.2017).
22. *Глухих И.Н.* Оптимизация векторного критерия. Парето-оптимальные решения. URL: http://systematy.ru/articles/44_optimizatsiya_vektornogo_kriteriya__pareto-optimalnyie_resheniya (дата обращения: 17.03.2017).

Андреев Виктор Павлович — к-т физ.-мат. наук, д-р техн. наук, профессор кафедры сенсорных и управляющих систем, Московский государственный технологический университет «Станкин» (МГТУ "СТАНКИН"). Область научных интересов: мехатроника, робототехника, информационно-измерительные и управляющие системы, системы технического зрения. Число научных публикаций — 127. andreevvipa@yandex.ru; Вадковский пер., 1, Москва, 127055; р.т.: +7(965)210-7951.

Плетенев Павел Филиппович — аспирант кафедры робототехники и мехатроники, Московский государственный технологический университет «Станкин» (МГТУ "СТАНКИН"). Область научных интересов: робототехника, мехатроника, распределенные системы управления, самоорганизующиеся системы управления. Число научных публикаций — 7. cpp.create@gmail.com; Вадковский пер., 1, Москва, 127055; р.т.: +7(903)2547693.

Поддержка исследований. Работа выполняется при финансовой поддержке РФФИ (гранты 16-07-00811а и 16-07-01264а).

DOI 10.15622/sp.57.6

V.P. ANDREEV, P.F. PLETENEV

METHOD OF INFORMATION INTERACTION FOR DISTRIBUTED CONTROL SYSTEMS OF ROBOTS WITH MODULAR ARCHITECTURE

Andreev V.P., Pletenev P.F. **Method of Information Interaction for Distributed Control Systems of Robots with Modular Architecture.**

Abstract. The paper presents a solution to the problem of information interaction between embedded computing devices in the implementation of *distributed control* in the information-measuring and control system (IMCS) of robots with modular architecture. Distributed control is realized by designing each module as a device with its own IMCS containing all the components necessary to perform its functionality, including computing devices. As a consequence of such a functional completion of the modules, the computations needed to control the robot as a single mechatronic system are *parallelized*. As a result, the requirements to the power of computing devices of modules' IMCS are significantly reduced, as it is possible to use inexpensive microcontrollers and single-board computers — embedded computing devices.

The network organization of the IMCS structure of the robot is proposed, which allowed the transfer of the *reconfigurability* property of the network to the structure of the modular robot. Analysis of various network topologies has shown that the topology of the "star" type has several advantages over the topology of the "bus" type for use in heterogeneous modular robots.

It is shown that the use of the Robot Operating System (ROS) for the implementation of information interaction between embedded computing devices is either impossible or significantly hampered. A specification is proposed for the creation of appropriate programming interfaces and a language for inter-module communication, which enable the inclusion of third-party modules in the plug and play mode. The specification is based on the principles of ROS, but allows one to implement software on embedded computing devices. On the basis of Pareto's multi-criteria optimization, recommendations were obtained for selecting the appropriate hardware and software.

The efficiency of the proposed solution was proved in the course of experiments on an installation which composition is close to the working conditions of the IMCS of a heterogeneous modular robot. The experiments showed that the joint operation of the software and hardware parts meets all the specified requirements and is applicable for the transmission of messages of the executive level with a frequency of up to 100 Hz under any load on the network.

Keywords: modular robot, heterogeneous robot, mobile robot, reconfigurability, control system, distributed control.

Andreev Victor Pavlovich — Dr. Sci., professor of sensory and drive systems department, Moscow State University of Technology «STANKIN» (MSTU «STANKIN»). Research interests: mechatronics, distributed control systems, self-organizing control systems. The number of publications — 127. andreevvipa@yandex.ru; 1, Vadkovsky per., Moscow, 127055, Russia; office phone: +7(965)210-7951.

Pletenev Pavel Filippovich — Ph.D. student of robotics and mechatronics department, Moscow State University of Technology «STANKIN» (MSTU «STANKIN»). Research interests: robotics, mechatronics, distributed control systems, self-organizing control systems. The number of publications — 7. cpp.create@gmail.com; 1, Vadkovsky per., Moscow, 127055, Russia; office phone: +7(903)2547693.

Acknowledgements. This research is supported by RFBR (grants 16-07-00811a and 16-07-01264a).

References

1. Lopota A.V., Jurevich E.I. [Stages and development prospects of robotic systems design modular principle]. *Nauchno-tehnicheskie vedomosti SPbGPU. Informatika. Telekommunikacii. Upravlenie – St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunication and Control Systems*. 2013. vol. 1(164). pp. 98–103. (In Russ.).
2. Platonov A.K. [Robotics of a moon base]. *XXXIV Chtenija po kosmonavtike – XXXIV Readings on Cosmonautics*. Available at: http://www.keldysh.ru/section5/report.xhtm?src=section5.xml&filter=3 (accessed 10.06.2017). (In Russ.).
3. Murata S. et al. M-TRAN: self-reconfigurable modular robotic system. *IEEE/ASME Transactions on Mechatronics*. 2002. vol. 7(4). pp. 432–441.
4. Østergaard E.H., Kassow K., Beck R., Lund H.H. Design of the ATRON lattice-based self-reconfigurable robot. *Autonomous Robots*. 2006. vol. 21(2). pp. 165–183.
5. Qiao G. et al. Design of Transmote: a Modular Self-Reconfigurable Robot with Versatile Transformation Capabilities. Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics. 2012. pp. 1331–1336.
6. Fukuda T., Ueyama T., Kawauchi Y., Arai F. Concept of cellular robotic system (CEBOT) and basic strategies for its realization. *Computers Elect Engng*. 1992. vol. 18. no. 1. pp. 11–39.
7. Baca J., Ferre M., Aracil R. A heterogeneous modular robotic design for fast response to a diversity of tasks. *Robotics and Autonomous Systems*. 2012. vol. 60. no. 4. pp. 522–531.
8. Lyder A.H. et al. On sub-modularization and morphological heterogeneity in modular robotics. Intelligent Autonomous Systems of Advances in Intelligent Systems and Computing. 2013. vol. 193. no. 12. pp. 649–661.
9. Hancher M.D., Hornby G.S. A modular robotic system with applications to space exploration. 2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT'06). 2006. pp. 132–140.
10. Garcia R.F.M., Lyder A., Christensen D.J., Stoy K. Reusable Electronics and Adaptable Communication as Implemented in the Odin Modular Robot. IEEE International Conference on Robotics and Automation. 2009. pp. 1152–1158.
11. Andreev V., Kim V., Pletenev P. The principle of full functionality – the basis for rapid reconfiguration in heterogeneous modular mobile robots. Proceedings of the 28th DAAAM International Symposium. 2017. pp. 0023–0028.
12. Quigley M. et al. ROS: an open-source Robot Operating System. ICRA workshop on open source software. 2009. vol. 3. no. 3.2. pp. 5.
13. Andreev V.P., Kim V.L., Poduraev Yu.V. [Network–based design of heterogeneous modular mobile robotic systems] *Robototehnika i tehnicheskaja kibernetika – Robotics and Technical Cybernetics*. 2016. vol. 3(12). pp. 23–29. (In Russ.).
14. EtherCAT Technology Group, Industrial Ethernet Technologies. Available at: https://www.ethercat.org/download/documents/Industrial_Ethernet_Technologies.pdf. (accessed: 17.03.2016).
15. Ethernet POWERLINK Communication Profile Specification Version 1.2.0. Available at: http://www.ethernet-powerlink.org/en/downloads/technical-documents/action/open-download/download/epsg-ds-301-v120-communication-profile-specification/element/ 5158/?no_cache=1 (accessed: 17.03.2016).
16. RFC: 791 Internet Protocol. Available at: https://tools.ietf.org/html/rfc791 (accessed: 17.05.2017).
17. Digital Standards Organization COSS – Consensus Oriented Specification System. Available at: https://web.archive.org/web/20161002092144/http://www.digistan.org/spec:1/coss (accessed: 17.05.2017).

18. Andreev V.P., Pletenev P.F. [Developing inter-modular communication for heterogeneous mobile robot]. *Trudy mezhdunarodnoj nauchno-tehnicheskoj konferencii "Jekstremal'naja robototehnika"* [Proceedings of the International Scientific and Technological Conference "Extreme robotics"]. 2016. pp. 245–255. (In Russ.).
19. Kirsanov K. Software architecture of control system for heterogeneous group of mobile robots. *Procedia Engineering*. 2015. vol. 100. pp. 216–221.
20. Hintjens P. "0MQ – The Guide". Available at: http://zguide.zeromq.org/page:all (accessed: 17.03.2016).
21. Pletenev P.F. [1/PIMI – Protocol of intermodular interaction]. Available at: https://asmfreak.github.io/modular_robots_rfc/1/ПММВ/ (accessed: 20.01.2017). (In Russ.).
22. Gluhih I.N. [Optimization of a vector criteria. Pareto-optimal solutions]. Available at: http://systematy.ru/articles/44_optimizatsiya_vektornogo_kriteriya__pareto-optimalnyie_resheniya (accessed: 17.03.2017). (In Russ.).

УДК 004.896 **DOI** 10.15622/sp.57.7

Г.А. Киселёв, А.И. Панов

ЗНАКОВЫЙ ПОДХОД К ЗАДАЧЕ РАСПРЕДЕЛЕНИЯ РОЛЕЙ В КОАЛИЦИИ КОГНИТИВНЫХ АГЕНТОВ

Киселёв Г.А., Панов А.И. **Знаковый подход к задаче распределения ролей в коалиции когнитивных агентов.**

Аннотация. В настоящей работе рассмотрена задача распределения ролей при составлении общего плана действий в коалиции когнитивных агентов. Когнитивные агенты реализуют основные функции интеллектуального агента с использованием моделей когнитивных функций человека, к которым относятся применяемые в данной работе функции обучения концептуальным знаниям и планирования коллективного поведения. В работе представлен оригинальный метод распределения ролей — алгоритм MultiMAP, основанный на знаковом способе планирования поведения агента. Представлены основные особенности описываемого подхода, включающие способы представления знаний агента о себе и о других агентах, способы знаковой коммуникации и сохранения опыта кооперации с другими агентами. Описаны модельные эксперименты, демонстрирующие основные преимущества представленного подхода и некоторые недостатки, на устранение которых направлена будущая работа в данном направлении.

Ключевые слова: когнитивные агенты, коалиции агентов, планирование, распределение ролей, знак, знаковый подход, теория деятельности.

1. Введение. В настоящей работе представлен оригинальный подход к синтезу общего плана поведения для группы интеллектуальных агентов. В качестве примера агентов, для которых задача распределения ролей в общем плане особенно актуальна, можно рассмотреть робототехнические системы, обладающие различными типами манипуляторов, предназначенных для взаимодействия с объектами внешней среды, которые обладают дополняющими деятельность других агентов характеристиками. Например, одни агенты могут оперировать исключительно крупными объектами, которые можно поднять только мощным манипулятором, а другие — лишь небольшими объектами, требующими более точного и менее мощного инструмента. Некоторые агенты способны передвигаться между объектами инфраструктуры в пределах одного города, а другие совершать междугородние перелеты. Такое различие в функциональности приводит к тому, что задачи, которые не мог решить один интеллектуальный агент, могут быть решены группой кооперирующихся агентов, которые учитывают специфику способностей других агентов при составлении общего плана.

Для представления знаний агента о процессах и объектах внешней среды, знаний о себе и о других агентах используется модель знаковой картины мира [1, 2], которая является формализацией качественной теории целенаправленного поведения человека в социуме — психологической теории деятельности [3]. На основе знаковой картины мира уда-

ется построить психологически правдоподобные модели ряда когнитивных функций человека: планирования поведения, целеполагания, категоризации и обобщения [2], которые, в свою очередь, применяются для построения когнитивных агентов, обладающих более высокой степенью автономности по сравнению с существующими интеллектуальными агентами [4]. В дальнейшем будут использоваться принятые в математической теории картины мира термины и понятия, заимствованные из психологии, но имеющие точные формальные определения.

Базовым элементом картины мира является четырехкомпонентная структура — знак, который сочетает в себе как декларативные, так и процедурные знания об объекте, процессе или субъекте деятельности (см. формальное определение в [2]). Каждая компонента знака: образ, значение и личностный смысл — отвечает за определенный тип информации, относящейся к представляемому объекту или процессу. Особенностью предлагаемого подхода является то, что агент хранит представление знаний о себе и о своих способностях, а также о способностях других агентов в тех же структурах, что и информацию об объектах внешней среды — в знаках «Я», «Агент 1», «Агент 2» и так далее. Как будет продемонстрировано ниже, это позволяет составлять коллективный план с использованием той же процедуры, с помощью которой производилось индивидуальное планирование.

Статья посвящена описанию оригинального метода коллективного планирования MultiMAP, который включает в себя этап распределения ролей и является расширением ранее предложенного алгоритма MAP [5] индивидуального планирования поведения в знаковой картине мира. Предложенный алгоритм коллективного планирования является итерационным и иерархическим. В нем используется информация о способностях как самого агента, так и других членов коалиции, а также предыдущий опыт составления и согласования планов. В работе описана процедура согласования планов, использующая оригинальный протокол коммуникации и сообщения, вырабатываемые на основе знаковой картины мира агента. В качестве иллюстрации особенностей предложенного алгоритма рассмотрен ряд экспериментальных задач, в которых используются следующие утверждения: а) все агенты обладают актуальной информацией о способностях других агентов; б) состав коалиции и способности агентов не меняются в процессе планирования; в) способности агентов могут частично совпадать, то есть имеются действия, которые могут совершить одновременно несколько агентов; г) все агенты имеют общую цель и не ставят свои собственные подцели; д) план согласуется централизованным способом; е) этап выполнения плана не рассматривается, поэтому в качестве опыта со-

храняется история планирования. В работе даны примеры вводимых понятий и проведены модельные эксперименты с использованием ряда известных в классическом планировании задач (доменов), таких как «Мир кубиков» [6], «Логистика» и так далее. В заключении проведено обсуждение основных преимуществ предлагаемого подхода и перспективных направлений развития представленных работ.

Статья организована следующим образом. В разделе 2 приводится краткий обзор наиболее значимых работ по распределению ролей в многоагентных системах (МАС). В разделе 3 даны базовые принципы знакового подхода к представлению знаний и введены основные понятия. В разделе 4 представлен алгоритм планирования в группе агентов, которые используют знания о других агентах для распределения ролей. В разделе 5 даны результаты модельных экспериментов и их обсуждение.

2. Обзор работ по планированию в группе агентов. Для решения задач, которые невозможно выполнить одним агентом, используются различные виды многоагентных систем (МАС), которые в зависимости от сложности организации называются группами, роями, стаями. Основными причинами объединения агентов в МАС являются: отсутствие достаточного количества ресурсов для самостоятельного достижения целевой ситуации, наличие общих целей и функциональное дополнение друг друга. Методы многоагентного планирования применяются для большого спектра прикладных задач, в которые входят задачи логистики, моделирования предприятий, политических ситуаций и другие. Для осуществления распределения действий в МАС используются алгоритмы планирования и алгоритмы, реализующие протоколы коммуникаций между участниками кооперации [7, 8]. Рассмотрим основные проблемы, возникающие в многоагентном планировании и включающие проблему динамического перепланирования при наличии факторов, не позволяющих агенту совершить требуемое действие (поломка агента, наличие внешних факторов), проблему использования проверенных агентов (тех, с кем имеются прецеденты действия) и проблему избыточности, распространяемой агентами информации. Данные проблемы нашли решение в предлагаемом в статье подходе и далее будут рассмотрены некоторое аналогичные работы, связанные с планированием.

В рамках построения многоагентного плана действий, планирующий агент должен иметь представление о возможностях всех агентов МАС. Процесс построения коллективного плана занимает большое количество времени, поэтому отказ от выполнения предписанных агенту-участнику действий ведет к замедлению работы всей системы. Наличие возможности выполнить одно и то же действие с помощью

различных агентов позволяет динамически перестроить многоагентный план без кардинального изменения планов других участников МАС. В качестве примера способов динамического перепланирования в процессе взаимодействий агентов рассмотрим работу [9], в которой агенты могут создавать свои индивидуальные планы в рамках совместной работы над общим планом. Для построения совместного плана агентов авторы приводят ряд обязательных условий: во-первых, все агенты должны быть согласны с набором соблюдаемых параметров для достижения целевой ситуации. Примером параметра может служить часть карты, в которой агенты совершают действия и за которую не следует выходить. Выработка общего набора параметров может осуществляться с помощью различных средств достижения агентами соглашений, например с помощью аукционов [10], контрактных сетей [11], моделей социальных соглашений [12] и так далее. Во-вторых, группы агентов должны составлять план действий только из тех действий, которые агенты в группах способны выполнить. В-третьих, групповой план может иметь в качестве компонентов как планы индивидуальных агентов, так и групповые планы входящих в нее подгрупп. Планы каждого из агентов состоят из набора действий по достижению цели, где каждое комплексное действие является подпланом по достижению подцели задачи и имеет несколько абстрактных уровней заполненности готовыми действиями. Здесь под заполненностью подразумевается различная степень выполнения поддействий комплексного действия. После окончания процесса планирования группа агентов приступает к реализации плана. По мере взаимодействия каждого агента группы с окружающей средой и группами других агентов происходит процесс заполнения общего плана действий, который заполняется законченными действиями других агентов. Учитывая динамику окружающей среды, какое-то действие может оказаться невыполнимым в текущих условиях, поэтому агент должен иметь возможность вернуться к предыдущему уровню описания плана, на котором это действие отмечено как невыполненное. Каждый агент имеет дерево соблюдаемых параметров для выполнения определенного группового действия, которое состоит из множеств ограничений и поддействий, по мере взаимодействия с другими агентами и окружающей средой, агент пополняет свое дерево соблюдаемых параметров. Каждый агент решает вступать ли ему в кооперацию с остальными участниками коллектива на основе вероятностной шкалы осуществления действий другими агентами. Это представление создается при помощи опытного взаимодействия с другими членами кооперации, основываясь на прецедентах действий исследуемого агента с другими агентами. Например, иссле-

дуемый агент может продолжить совершение коллективного плана даже при изменении его отношения к ранее согласованным параметрам действий плана, выразив тем самым свою надежность в роли партнера, либо агент отказывается от следования ранее составленному плану во время процесса его выполнения, заставляя всех агентов-участников перепланировать свои действия. В процессе совершения действий в рамках реализации группового плана работ агент выстраивает шкалу доверия, по которой оценивается вероятность будущих повторных взаимодействий с другими агентами.

Понятия доверия и надежности агента рассматривается в модели Beliefe-Desire-Intention (BDI), которая описана в статье [13]. В модели BDI используются так называемые графы репутации [14], отражающие вероятностную модель межагентных взаимодействий внутри коллектива. Согласно принципам модели К. Рейнольдса [15], агенты взаимодействуют при выполнении локальных задач без управляющего внешнего воздействия, согласно своим представлениям о вовлеченности других агентов коллектива в свою стратегию достижения цели и о взаимозаменяемости агентов. Эти представления возникают из-за наличия определенного набора доступных действий у каждого из агентов, часть которых может быть выполнена другими агентами. Так как агенты взаимозаменяемы, возникает проблема распределения ролей при составлении многоагентного плана действования. Наличие собственной стратегии достижения цели у каждого из агентов, в которой субъектами действования могут выступать другие агенты, способствует увеличению времени на построение общего плана действий. Для уменьшения комбинаторной сложности задачи выбора субъекта действия, в модели BDI используются графы репутации, которые составляются на основе опыта планирующего агента, полученного при взаимодействии с другими участниками коллектива. Помимо использования собственного опыта планирования, агент может выстраивать графы репутаций, основываясь на опыте других участников кооперации [16], запрашивая их мнение относительно остальных агентов планирования. Для построения графа репутации агент использует два основных протокола сообщений: разведывательный протокол и протокол запросов [17]. Разведывательный протокол позволяет опрашивать окружающих агентов по поводу заведомо знакомых агенту вещей, что позволяет дать оценку правдивости опрошенного агента. После завершения процедуры оценивания правдивости окружающих агентов, агент использует протокол запросов и опрашивает мнение агентов, подтвердивших свою правдивость. Множество гетерогенных агентов с архитектурой, описанной в статье, имеют проблемы с устойчивостью

протоколов коммуникаций и с подбором необходимого протокола для каждого агента персонально.

Многие протоколы коммуникаций агентов в MAC распространяют избыточную информацию об агенте. В статье [18] представлен алгоритм многоагентного планирования MA-STRIPS, в котором реализован протокол коммуникаций агентов без распространения избыточной информации. На первом шаге алгоритма происходит считывание задачи планирования, описанной на многоагентной версии языка PDDL [19]. На следующем шаге агенты разбивают факты задачи на три подгруппы: публичные, внутренние и смешанные. Факт считается публичным, если он задействован в действиях минимум двух агентов; факт является внутренним, если он не публичен, но задействован хотя бы в одном действии планирующего агента; факт является смешанным, если он публичный или внутренний для агента. Затем осуществляется процесс планирования, который заключается в построении каждым агентом недетерминированного конечного автомата (НКА), представляющего все доступные агенту варианты планов. Планировочный конечный автомат для STRIPS [20] задачи $\Pi=\langle P,A,I,G\rangle$ — это НКА $\Gamma=\langle A,S,I,\delta,F\rangle$, в котором A — алфавит действий Π, состояния из S — это подмножества P (множества условий задачи), δ это функция перехода на множестве состояний, а $F \subseteq S$ содержит все состояния, которые удовлетворяют цели G. Чтобы избежать путаницы, авторы предлагают, что алфавит A содержит уникальные идентификаторы действий, а не полные действия. Публичная версия недетерминированного конечного автомата, состоящая только из публичных фактов, рассылается другим субъектам планирования по цепочке. После получения сообщения со всеми возможными планами другого агента, агенты коалиции пытаются расширить получившийся автомат до достижения конечной ситуации планирования. Такой подход создает большое количество возможных вариантов достижения целевой ситуации, что приводит к увеличению времени на поиск оптимального для всех агентов плана.

Рассмотренные подходы используют психологическую составляющую деятельности агентов для описания понятий доверия, надежности (BDI), избыточности информации (MA-STRIPS) и других факторов, и одновременно не учитывают единый способ представления знаний [2] всеми агентами коалиции. Унифицированный способ представления знаний всеми членами группы является немаловажным фактором при выборе способа коммуникаций агентов [21]. Протокол коммуникации агентов зависит от состава исполняемых ими ролей, процесс распределения которых является трудоемкой задачей. От решения задачи распределения ролей зависит достижение поставленных целей

коалиции [22]. Когнитивные агенты способны осуществлять деятельность, раскрывающую возможности не только одной выполняемой роли, но сразу нескольких. Так как качество выполняемых агентами действий может быть различным, процесс распределения ролей служит для создания наиболее подходящих составов групп агентов [23], которые имели опыт выполнения требуемых действий.

Далее в статье изложен психологически правдоподобный метод решения конфликтных ситуаций при распределении ролей между агентами, представления знаний о других агентах [24, 25] и построения коллективного плана действий агентов.

3. Знаковый подход к синтезу поведения агента. В настоящей работе в качестве основного способа представления знаний используется модель знаковой картины мира, базовым элементом которой является четырехкомпонентная структура, называемая знаком [2, 4]. Знак может представлять как статический объект, так и действие. Знак задается именем и содержит компоненты образа, значения, личностного смысла. Компонента образа содержит характерные признаки представляемого объекта или процесса. Компонента значения представляет доступные коллективу агентов обобщенные сценарии использования объектов. Компонента личностного смысла знака несет информацию о личном значении объекта для агента, то есть о текущей ситуации или ситуации, имевшей место в прошлом, составной частью которой является данный объект. Личностные смыслы знака формируются в процессе деятельности субъекта и являются конкретизацией сценариев из значений этого знака. Они задают предпочтения субъекта деятельности и формируют опыт выполнения действий или планирования.

Компоненты знака составляют специальные семантические (каузальные) сети, в узлах которых располагаются так называемые каузальные матрицы. Каждая каузальная матрица представляет собой последовательность списков (столбцов) признаков данного компонента знака. Признаками являются либо элементарные данные с сенсоров, либо ссылки на соответствующие знаки. Казуальный матрицы бывают двух типов: объектные, столбцы которых равнозначны (представляют описание статических объектов), и процедурные, столбцы которых следуют в определенном порядке, моделируя следование причины за следствием (представляют описание действий и процессов). Например, каузальная матрица знака «Находиться на» состоит из двух столбцов: в левый столбец входит знак блока X, а в правый — знак блока Y, что означает, что блок X находится на блоке Y. Другим примером каузальной матрицы можно рассмотреть процедурную каузальную матрицу действия «Поднять» из задачи «Мир блоков». В левую часть матрицы входят

условия действия: столбец, в который включены ссылка на знак блока, который требуется поднять, ссылка на знак размера блока, например «Большой», и ссылка на связанный с ним знак «Тип-блока». В следующий столбец входят ссылки на знак «Пустой», обозначающего отсутствие блоков на целевом блоке и ссылка на знак целевого блока. В третий столбец включены ссылки на знак «На столе», обозначающего наличие целевого блока на столе и ссылка на знак целевого блока. Еще один столбец включает ссылку на знак «Актуатор пуст», обозначающего отсутствие блоков в манипуляторе агента и ссылку на знак агента. В правую часть матрицы входит столбец с ссылкой на знак «Актуатор не пуст», обозначающий наличие целевого блока в манипуляторе агента, ссылка на знак агента и ссылка на знак целевого блока. Следующий столбец эффектов включает ссылку на знак целевого блока и связанные с ним знаки «Тип блока» и «Большой» (рисунок 1).

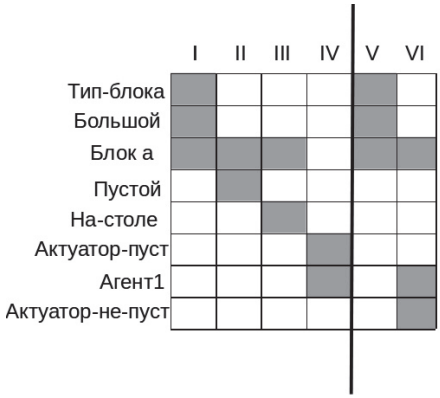


Рис. 1. Каузальная матрица знака действия «Поднять» в задаче «Мир блоков». Матрица является битовой матрицей, единицы которой обозначены серым, а нули – белыми клетками. Столбцы I, II, III, IV является столбцами условий, V и VI – эффектов

Другим примером каузальной матрицы является матрица знака ситуации, в левую часть которой входят столбцы с ссылками на объектные и процедурные знаки (рисунок 2). Например, каузальная матрица начальной ситуации задачи «Логистика» будет включать столбцы со ссылками на знаки перевозимых объектов «Объект1» ,..., «ОбъектN», ссылками на связанные с объектными знаками, знаки типов объектов «Тяжелый», «Легкий» и так далее ссылки на объектные знаки «В», «Находится», «Дислокация» и другие.

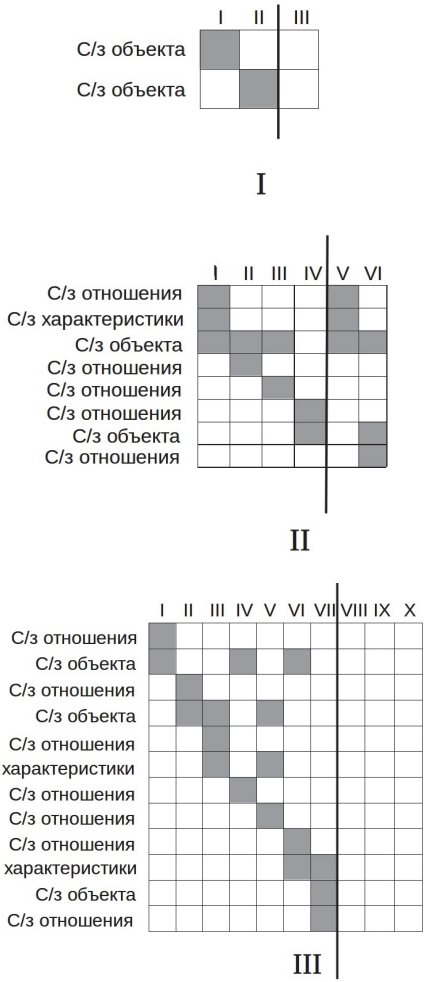


Рис. 2. Пример: I) объектная каузальная матрица; II) процедурная каузальная матрица; III) каузальная матрица ситуации (с/з – ссылка на знак)

Для каждой из трех каузальных сетей описан ряд семантических отношений на множестве знаков [2]. Так среди отношений на множестве знаков на сети значений есть отношение класс-подкласс, когда для одного знака, обозначающего какую-либо роль, может быть несколько играющих эту роль знаков, формирующих его каузальную матрицу на сети значений.

Формально, знаком *s* называют кортеж из четырех компонент: *<n, p, m, a>*, где n — имя знака, *p* — образ знака, соответствующий узлу $w_p(s)$ каузальной сети на образах, *m* — значение знака, соответствующий узлу $w_m(s)$ каузальной сети значений, *a* — личностный смысл знака, соответствующий узлу $w_a(s)$ каузальной сети на смыслах. R_n — отношения на множестве знаков, а Θ— операции на множестве знаков, полученные на основе фрагментов каузальных сетей, к которым принадлежат соответствующие компоненты знаков. Кортеж из пяти элементов $\langle W_p, W_m, W_a, R_n, \Theta \rangle$ является моделью картины мира агента.

Модель знаковой картины мира в настоящей работе используется в качестве базового способа представления знаний каждым агентом для построения как индивидуального, так и коллективного планов. В рамках процесса по нахождению плана осуществляется обратный процесс планирования (от целевой ситуации), подробно описанный в разделе 4.2. Агенты совершают различные действия, исходя из своих личностных смыслов, и пытаются достичь целевой ситуации. Знания о возможностях планирующего агента и других агентов представлены каузальными матрицами на сетях личностных смыслов знака «Я» и знаков «Агент1», «Агент2», … для каждого агента соответственно. Агенты создают планы по достижению целевой ситуации, распределяя роли агентов, осуществляющих деятельность исходя из критериев выполнимости действия различными агентами. Как и любой знак, знак «Я» состоит из личностных смыслов агента, осуществляющего процесс планирования, его образа и значений. Образ самого агента и других участников группы представляет основные характеристики агента, важные для распознавания других агентов и объектов по данным сенсоров, поэтому в настоящей работе этот компонент опущен. Значение знака «Я» и знаков других агентов — это обобщенные сценарии (действия), в которых агент может выступать субъектом либо непосредственно, либо через свои классы. Все действия, которые может совершить агент, представлены в его личностных смыслах и являются частично конкретизированными значениям, роли субъектов и объектов, в которых предзаполняются в соответствии с блоком его ограничений, представленных в задаче планирования. Пример знака «Я» в задаче «Логистика» приведен на рисунке 3. Планирующим агентом является агент-«грузовик», сеть личностных смыслов которого включает матрицы действий «загрузить», «разгрузить» и «вести грузовик». Матрицы всех действий частично конкретизируются, оставляя не конкретизированными часть знаков ролей объектов.

Знаки других агентов связаны отношением класс-подкласс с абстрактным знаком «Они». В знаки агентов входят представления агента об остальных агентах, полученные из общего описания задачи пла-

нирования. В каузальную матрицу на сети значений знака «Они» входят ссылки на знаки других агентов.

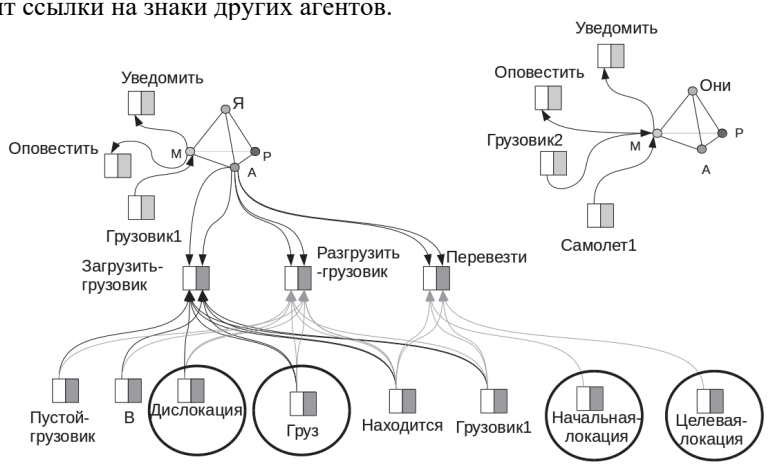


Рис. 3. Знак «Я» (слева) и знак «Они» (справа) в задаче «Логистика». Направление стрелок на рисунках обозначает отношения класс-подкласс на сети значений и тип-подтип на сети личностных смыслов. Знаки ролей обведены чёрными овалами

Примером знака другого агента в задаче «Логистика» является знак «Самолёт1», в личностные смыслы которого входят каузальные матрицы действий «Загрузить-самолёт», «Разгрузить-самолёт» и действия «Перелёт-самолёта» (рисунок 4), а в значения знаки ролей, характеризующие агента «Самолёт1» как объект, средство передвижения и самолет.

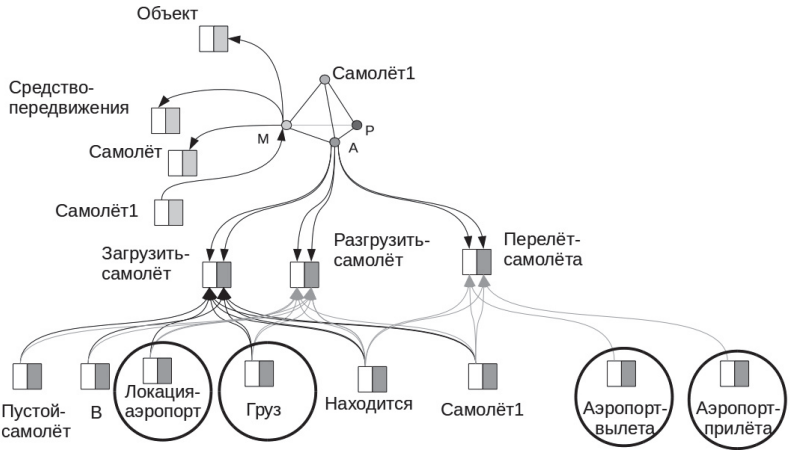


Рис. 4. Знак агента «Самолёт» из задачи «Логистика»

4. Распределение ролей в процессе планирования. В настоящей работе под интеллектуальным агентом будет пониматься аппаратная или программная система, обладающая свойствами автономности, реактивности, активности и коммутативности. Эти свойства позволяют агенту взаимодействовать со средой, которая включает в себя различные типы объектов. Планом P агента A_k будем называть последовательность действий $P(A_k)=<a_1(A_{i1}), a_2(A_{i2}),\ldots>$, полученных в результате работы алгоритма планирования, где A_{ij} — агент, выполняющий действие a_j. План формируется агентом на основе цели, информации о текущем состоянии окружающей среды и динамике ее изменения. Описание среды и задачи представлено на языке PDDL3 или ML-PDDL [26]. В домен задачи $D=<V,TO,A>$ включены описания предикатов V, типов объектов TO и действия агентов A. В описание действия $a=<n, Cond, Eff>$ входит имя действия n, список его предусловий $Cond$ и эффектов выполнения Eff. В список предусловий входят предикаты, формирующие условие применения этого действия, а в эффект — предикаты, значение которых стало истинной после применения действия. Типы объектов объединены в иерархию класс-подкласс. Предикаты описывают некоторое утверждение об объекте (например, предикат <размер блока>: большой блок). В домен планирования входят общие предикаты, которые описывают вид отношений между объектом некоторого типа и конкретным его свойством. В задачу планирования T входят специфицированные предикаты, где на место абстрактного типа подставлены конкретные значения. В задаче планирования в предикатной форме описаны начальное Sit_{start} и конечное Sit_{goal} состояние среды. Также, в задачу планирования входит блок ограничений на действия агентов $C(ag)$, состоящий из утверждений, истинных для конкретных агентов.

В процессе составления плана, учитывающего действия других агентов, наибольший интерес представляет этап выбора не только подходящего действия, но и подходящего агента. Включение процесса распределения ролей в алгоритм индивидуального планирования в знаковой картине мира привело к созданию алгоритма multiMAP, который состоит из четырех основных этапов: этап означивания, этап индивидуального планирования, этап согласования планов и этап сохранения опыта.

4.1. Этап означивания. На этапе означивания агент заполняет собственную картину мира знаками объектов, предикатов и действий, полученными из описаний домена и задачи планирования. Создаются объектные и процедурные каузальные матрицы на сетях значений и личностных смыслов. Создаются знаки «Я» и «Они», которые служат представлением о возможностях других агентов в картине мира агента.

Этап означивания начинается с процесса получения задачи планирования в предикатном виде. Домен и задача планирования описаны на языке ML-PDDL. Задача планирования состоит из набора фактов, описывающих объекты среды (в поле :объекты), начального и конечного состояний агентов (:начало и :цель) и ограничений на действия

агентов (поле :ограничения). В домен задачи включены описания предикатов, типов объектов и действия агентов. После получения домена и задачи планирования создается проблема планирования, в которой формируются знаки агентов (шаги 2-10 листинга 1), в значения знака «Я» добавляется каузальная матрица знака планирующего агента (шаг 8), в сеть значений знака «Они» добавляются каузальные матрицы остальных агентов планирования (шаг 9). **function** $GROUND(T)$:

1. $agent = T.agent$
2. $I_{sign} = Sign(I)$
3. $.They_{sign} = Sign(They)$.
4. **for** sj **in** subjects:
5. $sj = Sign(sj)$
6. **if** $agent = sj$ **then**
7. $I_{sign}.addfeature(sj)$
8. **else**: $They_{sign}.connect(sj)$
9. $z_{sj} = Z_m(sj)$
10. **for** ob **in** objects:
11. $ob = Sign(ob)$
12. $z_{ob} = Z_m(ob)$
13. **for** role **in** roles:
14. $role = Sign(role)$
15. $z_r = Z_m(r)$
16. $z_r.addfeature(z_{ob})$
17. **for** pr **in** predicates:
18. $pr = Sign(pr)$
19. $z_{pr} = Z(pr)$
20. $z_{pr}.addfeature(z_r)$
21. **for** A **in** actions:
22. $A = Sign(A)$
23. $z_A = Z(A)$
24. $z_A.addfeature(z_{pr})$
25. $z^a_{start}, z^a_{goal} := Sit(start, goal)$

Листинг 1. Формирование знаков в задаче планирования

Добавление каузальных матриц в сети знаков происходит с помощью функции «add feature», которая создает новый столбец в каузальной матрице знака с ссылкой на добавляемый знак. Далее формируются знаки и каузальные матрицы на сети значений для объектов задачи с помощью описания объектов в домене планирования (шаги 11-13). На следующем шаге алгоритм означивания создает знаки ролей объектов и связывает значения знаков ролей с соответствующими знаками объектов (шаги 14-17). После этого формируются знаки предикатов, в каузальные матрицы значений которых включаются ссылки на знаки ролей объектов, участвующих в них (шаги 18-21).

Заключительным шагом является создание знаков и процедурных каузальных матриц действий задачи (шаги 22-25). Ссылки на знаки ролей в процедурных каузальных матрицах частично заменяются на ссылки на знаки объектов и субъектов действования, что позволяет заранее создавать соответствующие блоку ограничений действия. Последним шагом алгоритма означивания является создание знаков и каузальных матриц на сети личностных смыслов начальной $z^a_{start}, z^a_{goal} := Sit(start, goal)$ и конечной ситуаций z^a_{goal} планирования.

4.2. Этап индивидуального планирования. Следующим этапом является этап индивидуального планирования, в котором агент создает все возможные планы по достижению целевой ситуации с помощью действий, включенных в личностные смыслы знаков «Я» и «Они». На шаге 6 (листинг 2) происходит назначение каузальной матрицы z_{cur} узлом сети личностных смыслов, который описывает целевую ситуацию, на шаге 7 происходит назначение каузальной матрицы z_{start} узлом сети личностных смыслов, который описывает первичную ситуацию. После этого происходит вызов рекурсивной функции MAP_ITERATION, которая возвращает список всех построенных планов.

Функция рекурсивного поиска планов начинается со сравнения шага рекурсии с максимальным числом итераций i, которое является внешним параметров и служит для ограничения глубины рекурсии. Если максимальное число итераций i_{max} было выполнено, а конечная ситуация не была достигнута, то поиск останавливается и множество планов формируется из уже найденных планов. Далее происходит поиск прецедентов достижения текущей ситуации z_{cur} из целевой z_{start}. Если была найдена процедурная каузальная матрица, в эффекты которой входят знаки текущей ситуации, а в условия знаки целевой ситуации, то она добавляется в список прецедентов действий $A_{precedences}$.

1. **for** agent **in** agents:
2. $T_{agent} = T, D$
3. $W_{agent} := GROUND\left(T_{agent}\right)$
4. $Plan := MAPSEARCH\left(W_{agent}\right)$
5. **function** $MAPSEARCH\left(W_{agent}\right)$
6. $z_{cur} := z^{a}_{goal}$
7. $z_{start} := z^{a}_{start}$
8. $Plans := MAPITERATION\left(z_{cur}, z_{start}, \varnothing, 0\right)$
9. $\left\{Plan_0, Plan_1, ...\right\} = SORT\left(Plans\right)$
10. return $Plan_0$
11. **function** $MAPITERATION\left(z_{cur}, z_{start}, Plan_{cur}, i\right)$
12. **if** $i > i_{max}$ **then**
13. **return** $\varnothing$
14. $A_{precedences} = Z\left(W_{agent}\right)$
15. $Act_{chains} = getsitsigns\left(z_{cur}\right)$
16. **for** chain **in** Act_{chains}:
17. $A_{signif} = getactions\left(chain\right)$
18. **for** z_{signif} **in** A_{signif}:
19. $Ch = openaction\left(z_{signif}\right)$
20. $P = generatemeanings\left(Ch, agents\right)$
21. $checked = activity\left(z_{cur}, P, A_{precedences}\right)$
22. $A_{cand} = metaactivity\left(checked\right)$
23. $exp_{cand} = exp\left(agents, A_{cand}\right)$
24. **for** A, ag **in** $.exp_{cand}.$:
25. $z_{cur+1} = Sit\left(z_{cur}, A\right)$
26. $plan.append\left(A, ag\right)$
27. **if** $z_{cur+1} \in z_{start}$ **then**
28. $F_{plans}.append\left(plan\right)$
29. **else**: $Plans := MAPITERATION\left(z_{cur+1}, z_{start}, plan, i+1\right)$

Листинг 2. Основные функции алгоритма планирования

На шаге 17 листинга 2 происходит получение всех процедурных каузальных матриц A_{signif}. На 18-22 шагах происходит создание (шаги 18-19) и означивание (шаг 20) возможных процедурных каузальных матриц P на сети личностных смыслов агента, среди которых выбираются применимые в текущей ситуации A_{cand}. На шаге 23 отбираются каузальные матрицы exp_{cand}, которые присутствуют в опыте планирования. На 24-28 шагах происходит создание следующей ситуации z_{cur+1} для каждой из оставшихся процедурных матриц. Если следующая ситуация является целевой ситуацией z_{start}, то полученный план действий добавляется в список конечных планов F_{plans}, если нет, то происходит рекурсивный вызов функции поиска планов в шаге 29.

4.3. Этап согласования. Процесс согласования планов агентами коалиции происходит в три шага. Первым является шаг выбора плана каждым из агентов, в результате у агента остается только один план из множества всех доступных планов, за который он будет голосовать. Далее следует шаг выбора способа коммуникации между агентами и создание сообщения для других агентов, на этом шаге агент преобразует оставшийся план в текстовое сообщение. Последним является шаг проведения аукциона, по итогу которого агенты получают финальный план действий. Аукцион производится на сервере коммуникации агентов, который является централизованной системой связи всех агентов.

На первом шаге происходят следующие действия:

1. $shortestplans = min\left(len\left(plans\right)\right)$
2. $longest = max\left(followers\left(shortestplans\right)\right)$
3. $smallestagents = min\left(agents\left(longest\right)\right)$
4. $goodplans = includeI\left(shortestplans\right)$
5. $thebestplans = random\left(goodplans\right)$

Действие 1 позволяет выбрать среди всех планов планы наименьшей длины, действие 2 выбирает те планы, в которых последовательность подряд идущих действий агентов максимальна. Применение действия 2 обосновано уменьшением нагрузки на коммуникационную сеть при информировании следующего агента плана о потребности начать предписанные ему действия. Действие 3 позволяет выбрать те планы, в которых фигурирует наименьшее количество агентов, действие 4 выбирает из оставшихся планов те планы, в которых фигурирует планирующий агент. Действие 5 применяется только в тех случаях, когда остался не единственный план, полученный после выполнения действия 4. После этого происходит выбор типа соединения с другими агентами кооперации, вы-

бирается способ соединения с помощью процедурной каузальной матрицы знака «Проинформировать» (рисунок 5).

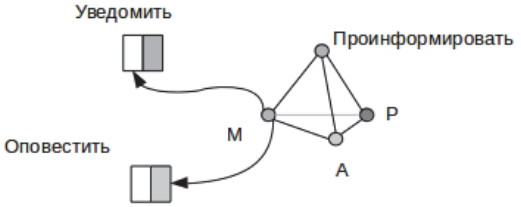


Рис. 5. Знак «Проинформировать»

4.4. Этап сохранения опыта. Следующим этапом является сохранения опыта агентов. Агент удаляет все матрицы ситуаций, кроме матриц начальной z^a_{start} и конечной ситуаций z_{start}, удаляет все процедурные каузальные матрицы, не вошедшие в план, например матрицы действий, которые включены в другие планы агента и отвергнуты на этапе согласования планов. Далее агент создает знак действия, чья каузальная матрица описывает возможность достижения конечной ситуации из начальной (рисунок 6) и сохраняет оставшиеся знаки. В образы знака действия, обозначающего опыт агента, входит каузальная матрица, содержащая ссылки на знаки всех действий, которые составляют конечный план. В смыслы знака действия входит каузальная матрица, чьи условия содержат ссылку на z^a_{start}, а эффекты ссылку на z_{start}.

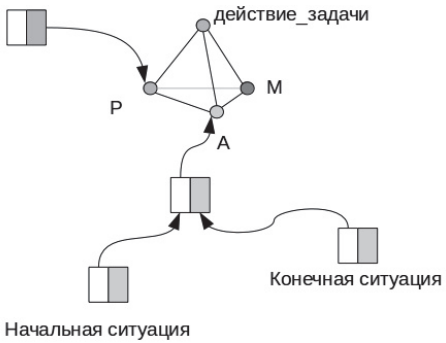


Рис. 6. Знак сохраненного плана выполнения задачи

При решении последующих схожих задач агент способен использовать опыт предыдущих прецедентов a_{exp} в качестве действия, в таком случае план будет выглядеть:

$$P(I) \coloneqq a_1(I), a_2(агент_2), a_{exp}(\varnothing),$$

a_1, a_2 — действия агентов 1 и $агент_2$ соответственно, пустое множество обозначает отсутствие конкретного агента-исполнителя для действия, обозначающего прецедент планирования.

5. Экспериментальное исследование. В рамках демонстрации особенностей применения знакового подхода к решению задачи распределения ролей были проведены экспериментальные исследования, в которых рассмотрены решения задач планирования «Мир блоков» и «Логистика». На каждом шаге итерации процесса планирования возникают варианты выбора агента, который будет совершать требуемое действие, что приводит к значительному увеличению числа построенных планов. Далее представлен пример решения проблемы выбора общего плана действий в задаче «Мир блоков» в экспериментах 1-3 и задаче «Логистика» в экспериментах 4-6.

Домен планирования задачи «Мир блоков» состоит из описания действий: «Поднять», «Положить», «Состыковать», «Снять» и описания предикатов «На», «На столе», «Пустой», «Актуатор пуст», «Тип блока». Задача планирования состоит из описания блоков, размеров блоков, агентов, начального и конечного положения блоков и актуаторов, ограничений на деятельность агентов. Ограничения состоят из предикатов, истинных для конкретных агентов планирования: Агент1: <размер блока>: большой блок, <размер блока>: средний блок, Агент2: <размер блока>: средний блок, <размер блока>: маленький блок и так далее.

В эксперименте 1 участвуют три агента планирования, каждый из которых может взаимодействовать с двумя блоками. Всего используется три вида блоков, в количестве двух блоков каждого вида. Начальная ситуация планирования включает 6 блоков, лежащих на столе, а конечная ситуация является описанием башни из блоков, которую необходимо построить агентам.

В процессе планирования каждый из агентов находит все 32 возможных плана с наименьшей длиной равной 10. Среди построенных планов выбираются планы наименьшей длины и планы, в которых агенты совершают наибольшее количество последовательных действий. Следующим шагом производится сортировка планов по количеству задействованных агентов и выбираются планы с минимальным числом агентов. Среди оставшихся планов агенты выбирают планы, в которых они являются субъектами деятельности. Далее агенты используют каузальную матрицу знака «Проинформировать» для получения матрицы знака «Оповестить» либо матрицы знака «Уведомить». Исследуемая задача планирования поддерживает работу трех агентов, из-за чего выбирается матрица знака «Оповестить» и добавляется к конечному плану каждого из агентов. Выбор матрицы «Оповестить» запускает процедуру формирования текстового сообщения для уведомления всех агентов о наличии построенного плана действий. Далее это сообщение отсылается на сервер коммуникации агентов и включается в словарь построенных планов, в котором ключами к словарю являются планы агентов, а значе-

ниями — число агентов, построивших такой же план. План, который удовлетворял наибольшее число агентов, признается окончательным.

Каждый из агентов сохраняет прецедент достижения целевой ситуации в виде знака действия, применимого в задачах с такими же ограничениями (рисунок 7). Каузальная матрица на сети личностных смыслов созданного знака действия содержит ссылку на каузальную матрицу знака целевой ситуации в левой части матрицы и ссылку на каузальную матрицу знака первичной ситуации в правой части матрицы. Каузальная матрица целевой ситуации задачи эксперимента 1 состоит из ссылок на знак «На-столе», который связан со всеми знаками блоков задачи, ссылки на знаки размеров блоков, знака «Актуатор-пуст», связанного со знаками агентов задачи и знака «Пустой», связанного со знаками блоков задачи. Каузальная матрица первичной ситуации отличается от каузальной матрицы целевой наличием знака «На», связанного со всеми блоками задачи и несущего смысловую нагрузку наличия башни из блоков, в которой с верхним блоком связан знак «Пустой», а с нижним - знак «На-столе». В образы знака действия входит каузальная матрица, в столбец условий которой входят ссылки на знаки всех действий, выполненных агентами по достижению целевой ситуации планирования.

В эксперименте 2 используется опыт построения планов агентами из эксперимента 1 при повторном решении задачи эксперимента 1. Каждый из агентов строит план длиною в два действия, использовав действие, обозначающее опыт нахождения плана первого эксперимента и добавив действие согласования построенных планов.

$$P(A_1) := \left\langle a_{\exp}(\varnothing), notify(A_1) \right\rangle$$

$$P(A_1) := \left\langle a_{\exp}(\varnothing), notify(A_1) \right\rangle$$

$$P\left(агент_1\right) := a_{exp}\left(\varnothing\right), Оповестить\left(агент_1\right);$$

$$P\left(агент_2\right) := a_{exp}\left(\varnothing\right), Оповестить\left(агент_2\right);$$

$$P\left(агент_3\right) := a_{exp}\left(\varnothing\right), Оповестить\left(агент_3\right).$$

В эксперименте 3 задача эксперимента 1 расширяется с помощью добавления двух больших блоков в вершину башни и изменяется порядок блоков при построении башни. На первой итерации планирования агенты активируют матрицы означенных действий, сохраненных в рамках первого эксперимента. Полученные матрицы используются агентами при разрешении неоднозначности выбора действия, позволяя агенту выбирать те действия, которые составляют сохраненные прецеденты. При решении задачи в эксперименте 1 агент совершал действия «Поднять блок g» и «Состыковать блок g и блок e», в эксперименте 3 возникла неопределенность выбора агента для манипуляций с блоком «g», которая была решена в пользу агента уже совершавшего действия с этим блоком в эксперименте 1.

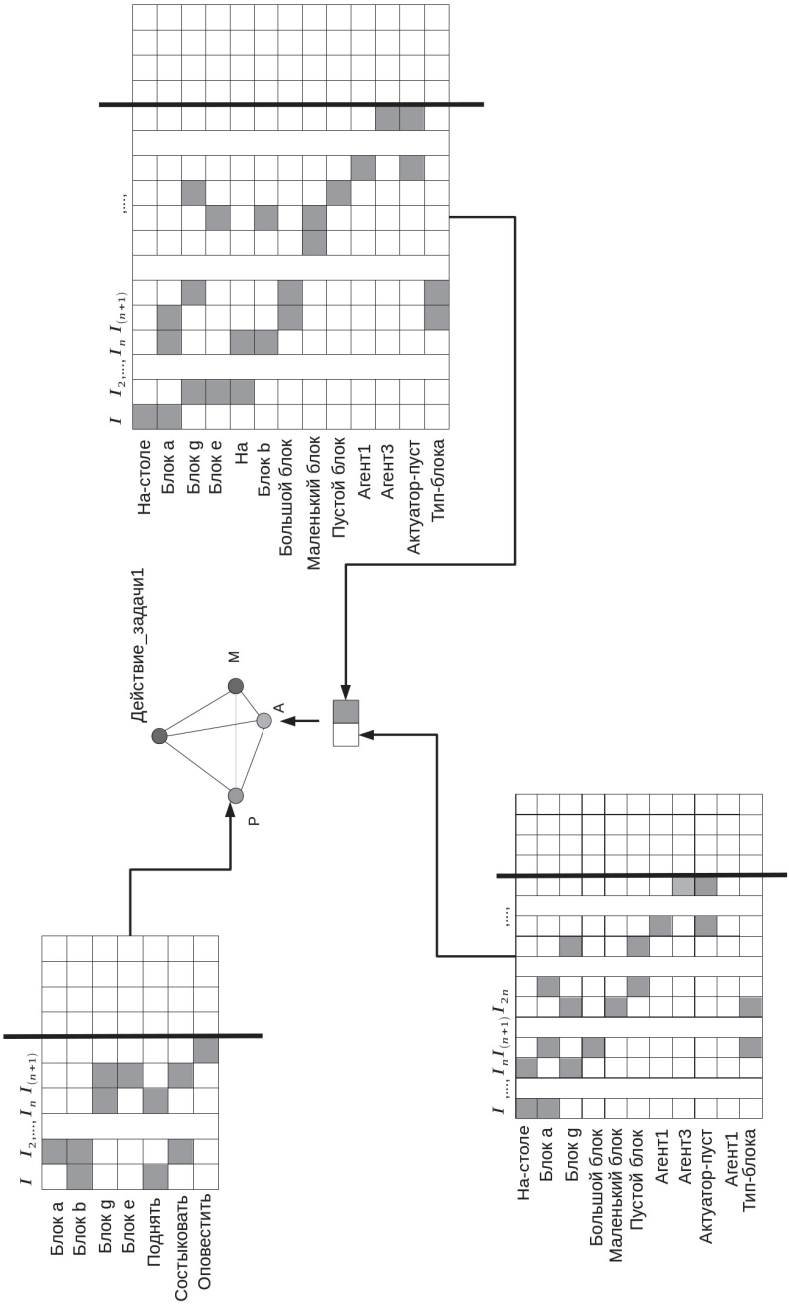


Рис. 7. Знак опыта планирования в задаче «Мир блоков»

В таблице 1 представлены условия экспериментов 1 — 3. КА — количество агентов, КТ — количество типов блоков, КТА — количество типов блоков, с которыми может взаимодействовать агент, КБ — количество блоков. В таблице 2 описаны результаты экспериментов 1-3. ДП — длина согласованного плана агентов, КЗ — количество знаков в картинах мира агентов. ВП — время, затраченное агентами для поиска всех планов достижения целевой ситуации, ИО — использование опыта агентами, КП — количество памяти, затраченное агентами на решение задачи.

Эксперименты 4-6 выполнены с задачей «Логистика», домен и задача планирования которой имеет схожее описание с задачей «Мир блоков». Отличием является наличие различных типов агентов (грузовик и самолет), каждый из которых имеет свой набор доступных действий, недоступный агентам другого типа. Домен планирования задачи «Логистика» состоит из описания действий: «Загрузить грузовик», «Разгрузить грузовик», «Перевезти грузовик», «Загрузить самолёт», «Разгрузить самолёт», «Перелёт самолёта» и описания предикатов «Аэропорт-прилёта», «Аэропорт-вылета», «Содержит», «Пустой», «Груз», «Дислокация», «Целевая-локация», «Начальная локация» и т.д. Задача планирования описывает первичные и целевые местоположения грузов и средств передвижения, а также ограничения на вместительность средств передвижения.

Таблица 1. Условия экспериментов 1-3

#эксп.	КА	КТ	КТА	КБ
1	3	3	2	6
2	3	3	2	6
3	3	3	2	8

Таблица 2. Результаты экспериментов 1-3

#эксп.	ДП	КЗ	ВП	ИО		КП
1	11	35	8940	нет		38,6
2	2	35	17	да		33,3
3	15	40	6400	да		39,9

В эксперименте 4 рассматривается задача перемещения двумя грузовиками разной грузоподъемности четырех различных грузов в аэропорт. После перемещения грузов в аэропорт агент-самолет перевозит их в аэропорт другого города. В эксперименте 5 рассматривается задача, в которой используется два вида грузов, которые находятся в различных местах двух разных городов. В каждом городе находятся грузовики различной грузоподъемности, которые должны доставлять грузы до аэропорта или до точек дислокации грузов. Между городами осуществлено воздушное сообщение. В эксперименте 6 решается задача доставки грузов трех видов

в аэропорт грузовиками, каждый из которых может быть загружен только двумя видами грузов. С помощью алгоритма multiMAP были найдены планы по перемещению грузов в целевое место назначения.

В таблице 3 представлены условия экспериментов 4-6. КАГ — количество агентов типа грузовик, КАС — количество агентов типа самолет, КГ — количество грузов, КТГ — количество типов грузов.

В таблице 4 описаны результаты экспериментов 4-6. ДП — длина согласованного плана агентов, ВП — время, затраченное агентами для поиска всех планов достижения целевой ситуации, КП — количество памяти, затраченное агентами на решение задачи.

На основе полученных данных были построены графики на рисунке 8.

Таблица 3. Условия экспериментов 4-6

#эксп.	КАГ	КАС	КГ	КТГ
1	2	1	4	2
2	4	1	4	2
3	3	0	6	3

Таблица 4. Результаты экспериментов 4-6

#эксп.	ДП	ВП	КП
1	40	4040	62,16
2	184	16000	180,4
3	48	5020	69,5

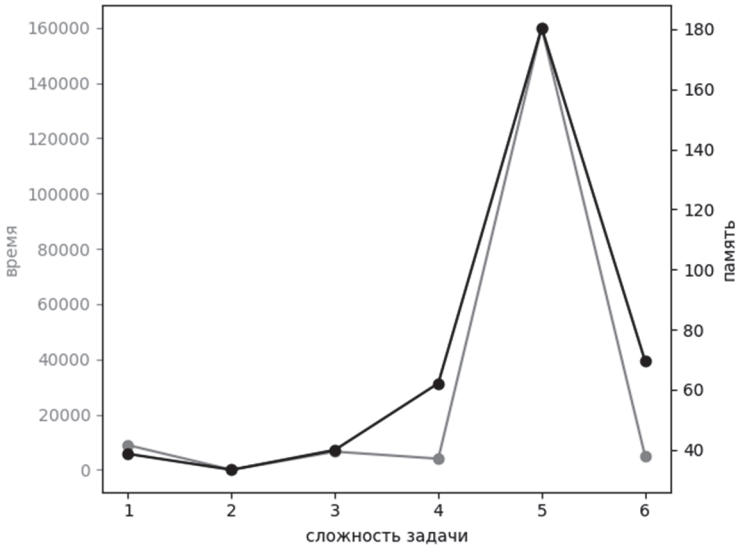


Рис. 8. Графики зависимости используемой памяти и времени подсчета планов относительно сложности задачи

Согласно данным, представленным на рисунке 8, можно сделать вывод, что с увеличением числа субъектов и объектов деятельности в алгоритме планирования наблюдается значительный рост потребления ресурсов, что усложняет процесс когнитивной оценки ситуации. Несмотря на длительность процесса построения оптимального плана решения в многоагентных задачах, алгоритм, основанный на знаковом подходе, предоставляет возможность заранее распределить роли агентов и предотвратить ситуации, в которых агенты одновременно пытаются выполнить действия с одним объектом. Агенты со знаковой картиной мира используют накопленный опыт для разработки планов действий. Подход с использованием опыта планирования позволяет значительно уменьшить потребление вычислительных ресурсов агента и ускорить процесс построения планов.

6. Заключение. В работе представлен знаковый подход к решению задачи коллективного планирования поведения и распределения ролей в коалиции. Рассмотренные эксперименты показывают, что проблема распределения ролей в коллективе является основополагающей проблемой многоагентного планирования, решение которой требует от агента серьезных временных и вычислительных затрат. Знаковый подход к планированию позволяет использовать опыт агентов, что значительно уменьшает ресурсоемкость задачи, одновременно позволяя унифицировать алгоритм добавления агентов в коллектив при наличии информации о возможностях агента. В будущих этапах разработки подхода планируется экспериментально описать проблему распределения ролей агентов с помощью взаимодействия физических роботов в реальном мире и в среде эмуляции Gazebo.

Литература

1. *Осипов Г.С., Панов А.И., Чудова Н.В.* Управление поведением как функция сознания. I. Картина мира и целеполагание // Известия Российской академии наук. Теория и системы управления. 2014. № 4. С. 49–62.
2. *Осипов Г.С., Панов А.И., Чудова Н.В.* Управление поведением как функция сознания. II. Синтез плана поведения // Известия Российский академии наук. Теория и системы управления. 2015. № 6. С. 47–61.
3. *Леонтьев А.Н.* Деятельность. Сознание. Личность: изд. 2-е. //М.: Политиздат. 1977. 304 с.
4. *Макаров Д.А., Панов А.И., Яковлев К.С.* Архитектура многоуровневой интеллектуальной системы управления беспилотными летательными аппаратами // Искусственный интеллект и принятие решений. 2015. № 3. С. 18–33.
5. *Panov A.I.* Behavior Planning of Intelligent Agent with Sign World Model // Biol. Inspired Cogn. Archit. 2017. vol. 19. pp. 21–31.
6. *Slaney J., Thiébaux S.* Blocks World revisited // Artificial Intelligence. 2001. vol. 125. no. 1–2. pp. 119–153.
7. The Foundation for Intelligent Physical Agents – FIPA. URL: http://www.fipa.org (дата обращения: 10.12.2017).

8. *Городецкий В.И. и др.* Прикладные многоагентные системы группового управления // Искусственный интеллект и принятие решений. 2009. № 2. С. 3–24.
9. *Gal Y. et al.* Agent decision-making in open mixed networks // Artificial Intelligence. 2010. vol. 174. no. 18. pp. 1460–1480.
10. *Primeau N. et al.* Improving task allocation in risk-aware robotic sensor networks via auction protocol selection // 2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES). 2016. pp. 21–26.
11. *Холодкова А.В.* Применение агентов в модели договорных сетей // Системы обработки информации. 2012. № 4(102). С. 142-145.
12. *Lorini E., Verdicchio M.* Towards a Logical Model of Social Agreement for Agent Societies // Coordination, Organizations, Institutions and Norms in Agent Systems V. 2010. pp. 147–162.
13. *Sabater J., Sierra C.* Review on Computational Trust and Reputation Models // Artificial intelligence review. 2005. vol. 24. no. 1. pp. 33–60.
14. *Huynh T.D., Jennings N.R., Shadbolt N.R.* An integrated trust and reputation model for open multi-agent systems // Autonomous Agents and Multiagent Systems. 2006. vol. 13. no. 2. pp. 119–154.
15. *Бурдун И.Е., Бубин А.Р.* Исследования и разработки в области мобильной робототехники стайного применения (краткий технический обзор зарубежных публикаций) // Морские информационно-управляющие системы. 2012. № 1. С. 46–56.
16. *Granatyr J. et al.* Trust and Reputation Models for Multiagent Systems // ACM Comput. Surv. 2015. vol. 48. no. 2. pp. 1–42.
17. *Serrano E., Rovatsos M., Botia J.* A qualitative reputation system for multiagent systems with protocol-based communication // Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. 2012. vol. 1. pp. 307–314.
18. *Jakubův J., Tožička J., Komenda A.* Multiagent Planning by Plan Set Intersection and Plan Verification // Proceedings of the International Conference on Agents and Artificial Intelligence. SCITEPRESS - Science and Technology Publications. 2015. pp. 173–182.
19. *Kovacs D.L.* A Multi-Agent Extension of PDDL3 // ICAPS 2012 Proceedings of the 3rd Workshop on the International Planning Competition (WS-IPC 2012). 2012. pp. 19–37.
20. *Fikes R.E., Nilsson N.J.* STRIPS: A new approach to the application of theorem proving to problem solving // Artificial Intelligence. 1971. vol. 2. no. 3–4. pp. 189–208.
21. *Городецкий В.И., Карсаев О.В., Самойлов В.В., Серебряков С.В.* Открытые сети агентов // Труды СПИИРАН. 2007. № 4. С. 11–35.
22. *Городецкий В.И., Карасев О.В.* Самоорганизация группового поведения кластера малых спутников распределенной системы наблюдения // Известия ЮФУ. Технические науки. 2017. Т. 187. № 2. С. 234–247.
23. *Kiselev G.A., Panov A.I.* Synthesis of the Behavior Plan for Group of Robots with Sign Based World Model // International Conference on Interactive Collaborative Robotics. 2017. pp. 83–94.
24. *Bazenkov N., Korepanov V.* Double Best Response as a Network Stability Concept // 7th International Conference on ETwork Games, COntrol and OPtimization (NetGCoop). 2014. pp. 201–207.
25. *Корепанов В.О., Новиков Д.А.* Метод рефлексивных разбиений в моделях группового поведения и управления // Проблемы управления. 2011. № 1. С. 21–32.
26. *Gerevini A.E., Long D.* Plan Constraints and Preferences in PDDL3 // Technical Report. 2005. 12 p.

Киселёв Глеб Андреевич — программист лаборатории динамических интеллектуальных систем института системного анализа, Федеральный исследовательский центр "Информатика и управление" Российской академии наук (ФИЦ ИУ РАН). Область научных

интересов: многоагентное планирование, целеполагание, когнитивные способы представления знаний. Число научных публикаций — 3. kiselev@isa.ru; пр-т 60-летия Октября, 9, Москва, 117312; р.т.: +7(906)799-3329.

Панов Александр Игоревич — к.т.н., старший научный сотрудник лаборатории динамических интеллектуальных систем института системного анализа, Федеральный исследовательский центр "Информатика и управление" Российской академии наук (ФИЦ ИУ РАН), доцент базовой кафедры интеллектуальных технологий системного анализа и управления федерального исследовательского центра "Информатика и управление" Российской академии наук (ФИЦ ИУ РАН) факультета компьютерных наук, Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ). Область научных интересов: искусственный интеллект, групповая робототехника, планирование, семиотика, моделирование когнитивных функций. Число научных публикаций — 51. pan@isa.ru; пр-т 60-летия Октября, 9, Москва, 117312; р.т.: +79161445255.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 16-37-60055-мол_а_дк).

DOI 10.15622/sp.57.7

G.A. KISELEV, A.I. PANOV

SIGN-BASED APPROACH TO THE TASK OF ROLE DISTRIBUTION IN THE COALITION OF COGNITIVE AGENTS

Kiselev G.A., Panov A.I. **Sign-based Approach to the Task of Role Distribution in the Coalition of Cognitive Agents.**

Abstract. In this paper we consider the problem of the role distribution during the construction of a general plan of actions in the coalition of cognitive agents. Cognitive agents realize the basic functions of an intelligent agent using models of human cognitive functions. As a psychological basis for constructing models of cognitive functions, the theory of activity and the formalization of sign-based world model were used. The paper presents an original method for roles distribution - the MultiMAP algorithm, based on the sign-based method of agent's behavior planning. The main features of the described approach are presented, including ways of representing the agent's knowledge of himself and other agents, methods of sign communication and preserving the experience of cooperation with other agents. Model experiments are described that demonstrate the main advantages of the approach presented and some of the shortcomings to be eliminated in future work.

Keywords: cognitive agents, coalition, planning, role distribution, sign, sign-based world model, activity theory.

Kiselev Gleb Andreevich — programmer of dynamic intelligent systems laboratory of the Institute of System Analysis Federal research center "Computer science and control" of Russian academy of science (FRC CSC RAS). Research interests: multi-agent planning, goal-setting, cognitive types of knowledge representation. The number of publications — 3. kiselev@isa.ru; 9, pr. 60-letiya Oktyabrya, Moscow, 117312, Russia; office phone: +7(906)799-3329.

Panov Aleksandr Igorevich — Ph.D., senior researcher of dynamic intelligent systems laboratory of the Institute of System Analysis Federal research center "Computer science and control" of Russian academy of science (FRC CSC RAS), associate professor of intelligent technologies in system analysis and management department with Federal Research Center of Computer Science and Control of Russian Academy of Sciences of faculty of Computer Science, National Research University Higher School of Economics (NRU HSE). Research interests: machine learning, pattern recognition, cognitive computer modeling, multiagent systems. The number of publications — 51. pan@isa.ru; 9, pr. 60-letiya Oktyabrya, Moscow, 117312, Russia; office phone: +79161445255.

Acknowledgements. This research is supported by RFBR (grant 16-37-60055-mol_a_dk).

References

1. Osipov G.S., Panov A.I., Chudova N.V. [Behavior control as a function of consciousness. I. World model and goal setting]. *Izvestija Rossijskij akademii nauk. Teorija i sistemy upravlenija – Journal of Computer and Systems Sciences International*. 2014. vol. 4. pp. 517–529. (In Russ.).
2. Osipov G.S., Panov A.I., Chudova N.V. [Behavior Control as a Function of Consciousness. II. Synthesis of a Behavior Plan]. *Izvestija Rossijskij akademii nauk. Teorija i sistemy upravlenija – Journal of Computer and Systems Sciences International.* 2015. vol. 6. pp. 882–896. (In Russ.).
3. Leont'ev A.N. *Dejatel'nost'. Soznanie. Lichnost': Izd. 2-e.* [The Development of Mind: 2nd ed.]. M.: Politizdat. 1977. 304 p. (In Russ.).
4. Emel'yanov S. et al. [Multilayer cognitive architecture for UAV control]. *Iskusstvennyj intellekt i prinjatie reshenij – Artificial Intelligence and Decision Making*. 2016. vol. 39. pp. 58–72. (In Russ.).
5. Panov A.I. Behavior Planning of Intelligent Agent with Sign World Model *Biol. Inspired Cogn. Archit.* 2017. vol. 19. pp. 21–31.

6. Slaney J., Thiébaux S. Blocks World revisited. *Artificial Intelligence*. 2001. vol. 125. no. 1–2. pp. 119–153.
7. The Foundation for Intelligent Physical Agents – FIPA. Available at: http://www.fipa.org (access: 10.12.2017).
8. Gorodetsky V.I. et al. [Applied multiagent sytems for group control]. *Iskusstvennyj intellekt i prinjatie reshenij – Artificial Intelligence and Decision Making*. 2009. vol. 2. pp. 3–34. (In Russ.).
9. *Gal Y. et al.* Agent decision-making in open mixed networks. *Artificial Intelligence*. 2010. vol. 174. no. 18. pp. 1460–1480.
10. Primeau N. et al. Improving task allocation in risk-aware robotic sensor networks via auction protocol selection. 2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES). 2016. pp. 21–26.
11. Holodkova A.V. [Application of agents is in model of contractual network]. *Sistemy obrabotki informacii – Information Processing Systems*. 2012. № 4(102). pp. 142–145. (In Russ.).
12. Lorini E., Verdicchio M. Towards a Logical Model of Social Agreement for Agent Societies. Coordination, Organizations, Institutions and Norms in Agent Systems V. 2010. pp. 147–162.
13. Sabater J., Sierra C. Review on Computational Trust and Reputation Models. *Artificial intelligence review*. 2005. vol. 24. no. 1. pp. 33–60.
14. Huynh T.D., Jennings N.R., Shadbolt N.R. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multiagent Systems*. 2006. vol. 13. no. 2. pp. 119–154.
15. Burdun I.Y., Bubin A.R. [Research and Development in the Field of Mobile Robotics for Swarming Applications (a brief review of publications)]. *Morskie informacionno-upravljajushhie sistemy – Marine Information and Control Systems*. 2012. vol. 1. pp. 46–56. (In Russ.).
16. Granatyr J. et al. Trust and Reputation Models for Multiagent Systems. *ACM Comput. Surv*. 2015. vol. 48. no. 2. pp. 1–42.
17. Serrano E., Rovatsos M., Botia J. A qualitative reputation system for multiagent systems with protocol-based communication. Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. 2012. vol. 1. pp. 307–314.
18. Jakubův J., Tožička J., Komenda A. Multiagent Planning by Plan Set Intersection and Plan Verification. Proceedings of the International Conference on Agents and Artificial Intelligence. 2015. pp. 173–182.
19. Kovacs D.L. A Multi-Agent Extension of PDDL3. ICAPS 2012 Proceedings of the 3rd Workshop on the International Planning Competition (WS-IPC 2012). 2012. pp. 19–37.
20. Fikes R.E., Nilsson N.J. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*. 1971. vol. 2. no. 3-4. pp. 189–208.
21. Gorodetsky V.I., Karsaev O.V., Samoilov V.V., Serebryakov S.V. [Open networks of agents]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2007. vol. 4. pp. 11–35. (In Russ.).
22. Gorodetsky V.I., Karsaev O.V. [Distributed Surveillance System Based on Self-Organized Collective Behavior of Small Satellite Cluster]. *Izvestija JuFU. Tehnicheskie nauki – Izvestia SFU. Technical sciences*. 2017. vol. 187. no. 2. pp. 234–247. (In Russ.).
23. Kiselev G.A., Panov A.I. Synthesis of the Behavior Plan for Group of Robots with Sign Based World Model. International Conference on Interactive Collaborative Robotics. 2017. pp. 83–94.
24. Bazenkov N., Korepanov V. Double Best Response as a Network Stability Concept. 2014 7th International Conference on NETwork Games, COntrol and OPtimization (NetGCoop). pp. 201–207.
25. Korepanov V.O., Novikov D.A. [Method of reflexive partitions in models of group behavior and control]. *Problemy upravlenija – Control problems*. 2011. vol. 1. pp. 21–32. (In Russ.).
26. Gerevini A.E., Long D. Plan Constraints and Preferences in PDDL3. Technical Report. 2005. 12 p.

УДК 621.3 **DOI** 10.15622/sp.57.8

О.С. Авсентьев, И.Г. Дровникова, И.И. Застрожнов, А.Д. Попов, Е.А. Рогозин

МЕТОДИКА УПРАВЛЕНИЯ ЗАЩИТОЙ ИНФОРМАЦИОННОГО РЕСУРСА СИСТЕМЫ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

Авсентьев О.С., Дровникова И.Г., Застрожнов И.И., Попов А.Д., Рогозин Е.А. **Методика управления защитой информационного ресурса системы электронного документооборота.**

Аннотация. В статье рассматриваются методологические основы организационно-технологического управления (ОТУ) защитой информационного ресурса (ЗИР) систем электронного документооборота (СЭД) на базе программных средств (ПСр) защиты информации. Разработана концептуальная модель управления ЗИР СЭД на основе концептуальной проработки аспектов формирования методологии ОТУ ЗИР СЭД на базе ПСр ЗИР, обладающая широкими возможностями по ее использованию для разработки способов решения управленческих задач. Представлена методика управления эффективностью функционирования подсистемы защиты информационного ресурса (ПЗИР) в СЭД, предполагающая оптимизацию управляемых параметров подсистемы, обеспечивающих максимизацию интегрального показателя эффективности функционирования ПЗИР, и соответственно, выполнение требований, предъявляемых к подсистеме. Приведен алгоритм определения оптимальных значений управляемых параметров ПЗИР и оптимального значения интегрального показателя эффективности функционирования подсистемы, обеспечивающий возможность создания конкретных подсистем автоматизированного управления эффективностью функционирования ПЗИР в СЭД. Анализируются результаты расчетов по исследованию показателя временной неконфликтности функционирования ПЗИР.

Ключевые слова: организационно-технологическое управление, система электронного документооборота, защита информации, информационный ресурс, эффективность функционирования системы, управление эффективностью.

1. Введение. Значительное увеличение потока информации в жизнедеятельности общества и повсеместная компьютеризация различных сфер деятельности человека способствовало широкому внедрению систем электронного документооборота (СЭД) в структуру различных организаций. Особенностью функционирования СЭД является работа в многопользовательском режиме с информацией разного уровня конфиденциальности. При этом пользователи СЭД имеют различные полномочия по доступу к информации, циркулирующей в ней. Поэтому проблема защиты информационного ресурса (ЗИР) в СЭД от угроз несанкционированного доступа (НСД) с целью обеспечения информационной безопасности (ИБ) СЭД является актуальной. Для решения задач ЗИР СЭД от НСД создается специализированная подсистема ЗИР.

Одно из основных требований стандарта по ИБ [1] — организация управления процессом ЗИР автоматизированных систем (АС). Поэтому при организации ЗИР СЭД на основе ПЗИР необходимо обеспе-

чить непрерывное управление ЗИР. Важный элемент процесса управления сложными системами, включая и управление ЗИР СЭД — процесс принятия решений (ПР) [2-4]. При этом управляющие решения целесообразно принимать с учетом оценки эффективности функционирования АС, рассматриваемой как объект управления (ОУ). Поэтому перспективным направлением организации управления ПЗИР является реализация ОТУ защитой информации в СЭД на базе оценки эффективности функционирования подсистемы. Исходя из этого, актуальна проблема формирования методологии ОТУ ЗИР от НСД в СЭД на основе комплексной оценки эффективности ПЗИР.

Разработанная методика ОТУ ЗИР от НСД в СЭД позволяет автоматизировать процесс принятия управленческих решений по ЗИР СЭД.

2. Концептуальная модель ОТУ ЗИР в СЭД. Под ОТУ ЗИР в СЭД с использованием ПСр ЗИР понимают меры и мероприятия, установленные инструкциями организации, эксплуатирующей СЭД, а также способы управления на базе ПСр управления ЗИР СЭД, которые позволяют автоматизировать процедуру ПР или обеспечить компьютерную поддержку для ПР [4]. Вопросам управления ЗИР уделяется значительное внимание в стандартах по информационной безопасности (ИБ) [1, 5-9]. В основных стандартах по ИБ [1, 9] функциональные требования содержат классы требований управления безопасностью, регламентирующие аспекты управления средствами защиты, параметрами средств защиты, атрибутами безопасности и конфигурацией механизмов защиты. Во всех разделах функциональных требований стандартов [1, 9] существует пункт «Управляемые параметры», обеспечивающий управление параметрами ПСр ЗИР. В этом пункте приводятся параметры, которые позволяют осуществлять управление ПСр ЗИР для реализации требований соответствующего раздела.

Анализ литературы по ИБ и управлению АС [1-4, 12-14, 16, 17] позволяет определить структуру задач ОТУ ЗИР СЭД на базе ПСр ЗИР. Защита информации в СЭД реализуется комплексом программных средств защиты (КПСЗ) (входящим в состав ПЗИР), включающим систему разграничения доступа (СРД) и следующие основные подсистемы: обеспечения целостности; управления доступом; криптографическая; регистрации и учета. Подсистема управления доступом осуществляет идентификацию и аутентификацию пользователей при их доступе в СЭД. Подсистема обеспечения целостности осуществляет контроль целостности модулей ПЗИР, а также файлов и каталогов пользователя. Подсистема регистрации и учета выполняет регистрацию событий, связанных с работой ПЗИР, регистрацию запуска и завершения программ, а также сигнализацию попыток нарушения ЗИР. Криптографическая подсистема осуществляет шифрование защищаемой информации при ее хранении или передаче по открытым каналам. СРД реализует полномо-

чия пользователей СЭД по доступу к файлам, дискам, ПСр и так далее. Исходя из этого, управление КПСЗ должно включать в себя управление СРД и управление вышеназванными подсистемами.

Процесс управления АС в общем случае включает следующие этапы [3, 4, 12, 17]:

– сбор информации о параметрах функционирования ОУ (получение данных об информационном процессе, передача этих данных для обработки);

– обработка информации и ПР (анализ накопленной, справочной и поступающей информации; ПР по результатам анализа);

– исполнение принятого решения (создание управляющего сигнала и выполнение воздействия на ОУ).

ОТУ ЗИР обеспечивается подсистемой управления ЗИР (ПУЗР). ПУЗР — это функциональная подсистема СЭД, включающая программные средства и организационные мероприятия, предусмотренные для осуществления ОТУ ЗИР СЭД. Данная подсистема управления ЗИР реализует два взаимосвязанных вида управления: управление КПСЗ (осуществляет управление отдельными ПСр ПЗИР) и управление ПЗИР (осуществляет управление организацией ЗИР СЭД на базе ПСр ЗИР). В каждом из этих видов управления существуют два взаимодействующих блока — ОУ и управляющая система. Данные о функционировании ОУ поступают в управляющую систему, где производится их обработка и анализ, по результатам которых формируется управляющее воздействие на ОУ. Для управления КПСЗ ОУ является КПСЗ, а управляющей системой является подсистема управления эффективностью функционирования ПЗИР, функционально относящаяся и к ПЗИР, и к ПУЗР. При управлении ПЗИР ОУ — ПЗИР, а в качестве управляющей системы используется подсистема управления ПЗИР, функционально относящаяся к ПУЗР.

При управлении КПСЗ его ПСр, с одной стороны, играют роль исполнительных органов, получающих управляющие сигналы от подсистемы управления ПЗИР (набор целесообразных для использования ПСр ЗИР и конфигурация ПЗИР), с другой стороны — управляющих органов, осуществляющих управление параметрами КПСЗ. Управление КПСЗ представляет собой управление некоторыми ПСр ЗИР и соответствующими организационными мероприятиями. Организация управления КПСЗ основана на возможности изменения значений некоторых параметров ПСр ЗИР. Поэтому процесс управления КПСЗ включает в себя процедуры анализа состояния ПЗИР, принятие управляющего решения и на его основе осуществление соответствующего воздействия на ПСр ЗИР путем изменения значений их управляемых параметров.

При управлении ПЗИР управляющими воздействиями являются набор целесообразных для использования ПСр ЗИР, конфигурация ПЗИР и связанные с этим управлением организационные мероприятия.

Управление структурой ПЗИР осуществляется на основе принципа блочной архитектуры, учитывая специфику задач ЗИР СЭД [1-4, 10, 15, 16]. Учет этого принципа при разработке ПЗИР позволяет использовать унифицированные стандартные ПСр ЗИР, что дает возможность упростить разработку, отладку, контроль и верификацию алгоритмов и программ, модернизацию ПЗИР, обеспечить простоту и удобство эксплуатации. Используя данный принцип, можно сформировать ядро защиты ПЗИР, обеспечивающее минимально допустимый уровень защищенности СЭД, а при необходимости повысить уровень защиты путем инсталляции дополнительных ПСр ЗИР.

Обоснованные предложения, принципы организации управления сложными системами позволили разработать структурную схему концептуальной модели ОТУ ПЗИР, представленную на рисунке 1.

Рис. 1. Структурная схема концептуальной модели ОТУ ПЗИР

3. Оптимальное управление ЗИР СЭД. Задача управления ЗИР СЭД с помощью ПСр ЗИР при управлении КПСЗ представляет собой задачу оптимального управления, решение которой обеспечивает поддержание экстремального значения целевой функции. Данная функция зависит, во-первых, от группы регулируемых параметров, значения

которых могут изменяться с помощью управляющего сигнала подсистемы управления эффективностью функционирования ПЗИР; во-вторых, от группы внешних параметров, нерегулируемых данной подсистемой.

При управлении процессами ЗИР в СЭД необходимо учесть противоречивость требований ИБ СЭД требованиям к СЭД по ее назначению (производительности, удобству эксплуатации и т.д.) [2, 4]. Учет этого противоречия позволяет определить принцип методологии оценивания эффективности функционирования ПЗИР, заключающийся в том, что эффективность функционирования СЭД по прямому назначению неразрывно связана с эффективностью функционирования ПЗИР СЭД.

Эффективное функционирование ПЗИР обеспечивает возможность эффективного функционирования защищаемой СЭД. Управление ЗИР СЭД заключается в решении задачи оптимизации эффективности функционирования ПЗИР в СЭД, позволяющей, с одной стороны, обеспечить ЗИР СЭД, с другой — оказать минимальное негативное влияние реализации функций ЗИР на функционирование СЭД по назначению. Оптимальное управление ЗИР целесообразно осуществлять по интегральному показателю эффективности функционирования ПЗИР, поэтому для реализации данного управления необходимо решать многокритериальную задачу. Данная задача управления является трудной, так как необходимо решать достаточно сложные задачи формализации процессов ЗИР и ПР. Задачу ПР при оптимальном управлении функционированием ПЗИР СЭД при управлении КПСЗ можно формализовано представить как задачу оптимизации: необходимо выбрать альтернативу $\vec{a}$ (комплект значений управляемых параметров ПЗИР) из множества A ($\vec{a} \in A$) (всех возможных комплектов), при которой интегральный показатель эффективности функционирования ПЗИР $\vec{E}_u$ имел бы максимальное значение:

$$\vec{E}_u(a) \rightarrow max. \qquad (1)$$

Под максимизацией векторного показателя $\vec{E}_u$ (1) понимается увеличение наименьшего значения элементов вектора данного показателя.

4. Система показателей эффективности функционирования ПЗИР. Оценку эффективности реализации функций ЗИР в АС проводят с помощью показателей, которые достаточно полно характеризуют эффективность ЗИР в АС [4, 18-23]. Для оценки эффективности функционирования ПЗИР и осуществления на ее базе ОТУ ЗИР в СЭД целесообразно использовать систему показателей, содержащую интегральный показатель $\vec{E}_u$, который агрегирует 6 элементарных показателей: $E_ф$ — показатель функциональности ПЗИР; $\vec{E}_{аф}$ — показатель

адекватности функционирования ПЗИР; $E_{вн}$ — показатель временной неконфликтности функционирования ПЗИР; $E_{рн}$ — показатель ресурсной неконфликтности функционирования ПЗИР; $E_{фн}$ — показатель функциональной неконфликтности функционирования ПЗИР; $E_{уи}$ — показатель удобства использования ПЗИР.

Показатель адекватности функционирования ПЗИР отражает соответствие подсистемы требованиям по ЗИР, характеризующим эффективность реализации защитных функций ПЗИР. Оценку показателя $\vec{E}_{аф}$ осуществляют путем анализа параметров управляемых программных средств (УПСр) ЗИР подсистемы, которые характеризуют эффективность выполнения этими средствами своих функций. Показатель адекватности функционирования ПЗИР представляется в виде вектора частных показателей адекватности УПСр ЗИР ($E_{аф\,i}$) и оценивается по таблицам соответствия значений управляемых параметров ПЗИР значениям частных показателей адекватности функционирования УПСр ЗИР [4]. Измерение частных показателей осуществляется при помощи качественной шкалы, предполагающей балльную оценку.

Показатели $E_{ф}, E_{фн}, E_{рн}$ и $E_{уи}$, отражают соответствие ПЗИР требованиям по полноте реализуемого набора защитных функций ЗИР, функциональной и ресурсной неконфликтности функционирования ПЗИР в СЭД (неконфликтность взаимодействия ПЗИР с другими подсистемами и ПСр СЭД) и удобству использования ПЗИР в процессе эксплуатации СЭД соответственно. Оценку данных показателей проводят путем определения соответствия ПЗИР предъявляемым к ней требованиям на основе анализа ее программной документации [4, 23, 24]. Измерение этих показателей осуществляют по качественной шкале, имеющей значения «допустимо» и «недопустимо», что позволяет использовать булеву переменную. Таким образом, элементарные показатели $E_{ф}$, $\overrightarrow{E_{аф}}$, $E_{фн}, E_{рн}$ и $E_{уи}$ являются качественными показателями эффективности функционирования ПЗИР.

Количественный показатель временной неконфликтности функционирования ПЗИР отражает вероятностно-временные свойства динамики функционирования ПЗИР, влияющие на эффективность функционирования СЭД. Реализация функций ПЗИР приводит к увеличению продолжительности решения функциональных задач по назначению СЭД, так как часть процессорного времени СЭД тратится на решение задач ЗИР. При этом ИБ СЭД может быть обеспечена только при своевременной реализации защитных функций ПЗИР. Поэтому временная неконфликтность функционирования ПЗИР определяется как вероятность своевременной реализации функций ЗИР:

$$E_{вн} = P(\tau \le \tau_{max}), \quad (2)$$

где τ — время выполнения подсистемой функций ЗИР, τ_{max} — максимально допустимое время выполнения подсистемой функций ЗИР. Оценка данного показателя осуществляется с помощью полумарковской модели на основе представления динамики функционирования ПЗИР в виде конечного полумарковского процесса [4, 10, 11].

Комплексная оценка эффективности функционирования ПЗИР проводится с использованием интегрального показателя эффективности функционирования ПЗИР, агрегирующего рассмотренные выше элементарные показатели. Показатели $E_ф, E_{фн}, E_{рн}$ и $E_{уи}$ можно использовать только в ограничениях, так как они являются булевыми функциями. Учитывая, что показатель временной неконфликтности функционирования ПЗИР отражает временные ограничения функционирования ПЗИР в СЭД (2), этот показатель также необходимо использовать в ограничениях. Напротив, показатель адекватности ПЗИР характеризует эффективность реализации этой подсистемой своих функций, поэтому данный показатель необходимо рассматривать как интегральный показатель эффективности функционирования ПЗИР, учитывая выполнение вышеназванных ограничений.

С учетом выше изложенного, оценивание интегрального показателя эффективности функционирования ПЗИР предлагается выполнять с помощью выражения:

$$\overrightarrow{E_и} = \begin{cases} \overrightarrow{E_{аф}}, & \textit{если} \quad \left(E_{вн} \ge E_{\min\ вн}\right),\ E_ф \wedge E_{фн} \wedge E_{рн} \wedge E_{уи} = 1, \\ 0, & \textit{иначе} \end{cases} \quad (3)$$

где $E_{\min вн}$ — минимально допустимое значение временной неконфликтности функционирования ПЗИР, заданное документацией на СЭД [4, 11].

5. Организационно-технологическое управление эффективностью функционирования ПЗИР. Управление эффективностью функционирования ПЗИР реализуется с помощью управляемых параметров, позволяющих регулировать эффективность функционирования ПЗИР путем изменения их значений при воздействии сигналов управления. В качестве управляемых параметров ПЗИР взяты параметры УПСр ЗИР подсистемы, которые оказывают влияние на динамику функционирования ПЗИР. К управляемым параметрам функционирования ПЗИР, выявленным на основе анализа динамики ее функционирования [4], можно отнести: $l_{аут}$ — количество символов пароля, вводимого пользователем вручную для его аутентификации; $l_{доп\ аут}$ — количество символов пароля, вводимого пользователем вручную для его дополнительной аутентификации в процессе обращения к наиболее важному ресурсу; $p_{оп}$ — вероятность приме-

нения специальных преобразований для файлов; $p_{кц}$ — вероятность старта теста проверки целостности рабочей среды СЭД. Значения этих управляемых параметров определяют значения частных показателей адекватности УПСр ЗИР $E_{аф\ аут}, E_{аф\ доп.\ аут}, E_{аф\ сп}, E_{аф\ кц}$ отражающих эффективность выполнения ПЗИР возложенных на нее защитных функций по основной и дополнительной аутентификации пользователей СЭД, специальным преобразованиям информации и контролю целостности рабочей среды СЭД соответственно.

Оптимальное управление АС в случае, когда нерегулируемые параметры в системе в рассматриваемом периоде времени неизменны, сводится к установлению таких значений регулируемых параметров, при которых обеспечивается максимизация (или минимизация) критерия оптимального управления [3, 4, 12, 25, 26]. Для управления эффективностью функционирования ПЗИР необходимо определить такой вектор значений управляемых параметров ПЗИР, который обеспечивает максимизацию интегрального показателя эффективности функционирования ПЗИР (3). В данном случае задачу ПР при управлении эффективностью функционирования ПЗИР можно формализовано представить как задачу математического программирования [4, 12] — необходимо выбрать такую альтернативу из множества альтернатив, чтобы выполнялись условия:

$$\overrightarrow{E_{аф}} \to \max, \tag{4}$$

$$E_{вн} \geq E_{\min вн}, \tag{5}$$

$$E_{ф} \wedge E_{фн} \wedge E_{рн} \wedge E_{уи} = 1. \tag{6}$$

Исходя из предположения о равнозначности УПСр подсистемы с точки зрения ЗИР, максимизация векторного показателя $\overrightarrow{E_{аф}}$ (4) заключается в последовательном увеличении значения частных показателей УПСр ЗИР, имеющих наименьшее значение. Выражения (5) и (6) в данной системе являются ограничениями, отражающими требования к ПЗИР в СЭД. Выполнение данных ограничений обеспечивает достаточную полноту набора функций ПЗИР, своевременность выполнения данных функций подсистемой, функциональную и ресурсную неконфликтность ПЗИР в СЭД и удобное использование ПЗИР СЭД пользователями и обслуживание персоналом.

Анализ функционирования ПЗИР [2, 22, 25, 26] показал, что увеличение времени, отводимого на выполнение функций ЗИР, дает возможность увеличения эффективности функционирования УПСр ЗИР, измеряемой набором частных показателей адекватности данных средств ЗИР, которые определяют показатель адекватности ПЗИР. Исходя из этого задача оптимального управления эффективностью функционирования ПЗИР сводится

к задаче выбора оптимальных значений управляемых параметров, обеспечивающих выполнение ограничений (5), (6) и выражения:

$$E_{вн} - E_{min\ вн} \rightarrow min. \quad (7)$$

При этом, учитывая критичность СЭД к обеспечению ИБ, в качестве критерия оптимальности при выборе значений управляемых параметров предлагается использовать критерий максимизации наименьшего из значений частных показателей адекватности УПСр ЗИР.

При условии $E_{вн} < E_{min\ вн}$ управление эффективностью функционирования ПЗИР целесообразно осуществлять более чувствительным параметром из вышеназванной совокупности управляемых параметров, который обеспечивает выполнение ограничения (5) за счет минимального снижения эффективности функционирования УПСр ЗИР в данной ситуации. В этом случае в управлении эффективностью функционирования ПЗИР не используются управляемые параметры:

– имеющие значения, соответствующие минимальным значениям, заданным эксплуатационной документацией на СЭД;

– УПСр ЗИР которых имеют наименьшее значение частных показателей адекватности (исключая случай, когда все УПСр подсистемы имеют одинаковые значения частных показателей).

В случае, если при управлении ПЗИР все управляемые параметры достигли минимальных значений, заданных эксплуатационной документацией на СЭД, а ограничение (5) не выполнено, то данную ПЗИР заменяют другой.

Если $E_{вн} < E_{min\ вн} + \delta$, где δ — заданная величина, то управление эффективностью функционирования ПЗИР целесообразно осуществлять управляемым параметром средств ЗИР, имеющим наименьшее значение оценки частного показателя адекватности. При наличии нескольких УПСр ЗИР, имеющих наименьшее значение частного показателя, управление осуществляют менее чувствительным управляемым параметром, который обеспечивает выполнение выражений (5) и (7) при максимальном увеличении эффективности функционирования ПЗИР.

Структурно-функциональная модель управления эффективностью функционирования ПЗИР СЭД представлена на рисунке 2. ОТУ эффективностью функционирования ПЗИР предлагается осуществлять с помощью подсистемы автоматизированного управления эффективностью функционирования ПЗИР в СЭД. Данная подсистема реализует приведенную выше модель оптимального управления эффективностью функционирования ПЗИР через управляемые параметры подсистемы.

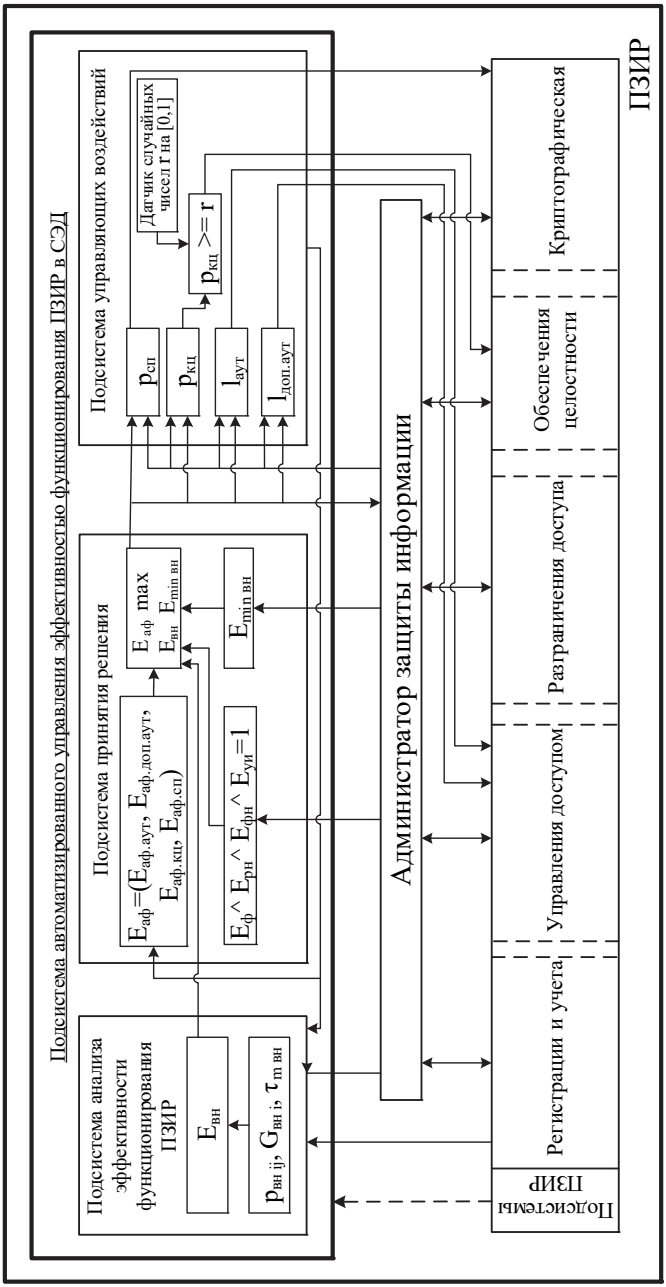


Рис. 2. Структурно-функциональная модель управления эффективностью функционирования ПЗИР СЭД

Подсистема автоматизированного управления эффективностью функционирования ПЗИР включает в свой состав подсистемы: анализа эффективности функционирования ПЗИР, ПР и управляющих воздействий.

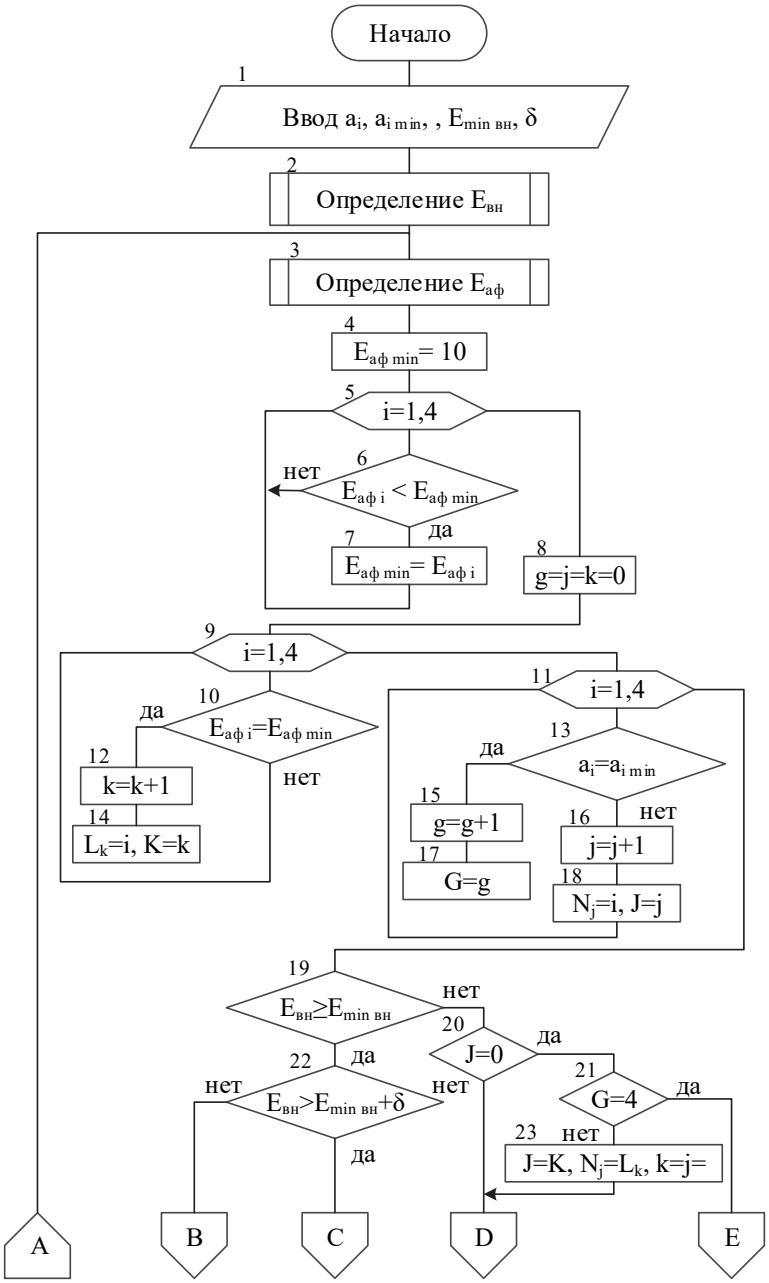
Подсистема анализа эффективности функционирования ПЗИР осуществляет оценку показателя $E_{вн}$ на основе данных о выполнении ПЗИР своих функций в СЭД. Эти данные поступают от подсистемы регистрации и учета. Эффективность функционирования ПЗИР оценивается в этом случае для обеспечения обратной связи в процессе управления эффективностью функционирования ПЗИР.

Подсистема ПР реализует функцию ПР по оптимальному управлению ЗИР в СЭД [18-21, 23]. Принятие решения осуществляется на основе комплексной оценки эффективности функционирования ПЗИР для обеспечения и поддержания разумного компромисса между уровнем защищенности информации в СЭД и эффективностью функционирования СЭД по прямому назначению. В результате принятия управленческого решения выбирается такой набор значений управляемых параметров функционирования ПЗИР, который обеспечивает максимальное значение интегрального показателя.

Подсистема управляющих воздействий формирует управляющее воздействие на ПЗИР в соответствии с ПР (набором значений управляемых параметров), которое обеспечивает выполнение условий (5), (7).

6. Алгоритмизация управления ЗИР СЭД. Алгоритмизация процесса управления на основе комплексной оценки эффективности функционирования ПЗИР заключается в разработке алгоритма оптимизации управляемых параметров при ОТУ эффективностью функционирования ПЗИР АС. Алгоритм определения оптимальных значений управляемых параметров ПЗИР и оптимального значения интегрального показателя эффективности функционирования ПЗИР при управлении ЗИР представлен на рисунке 3.

В рассматриваемом алгоритме вначале выполняются процедуры определения показателей $E_{вн}$ (блок 2) и $\overrightarrow{E_{аф}}$ (блок 3), алгоритмы которых приведены в [4]. Далее проводится начальная установка значения $E_{аф\,\min}$, соответствующее максимально возможному значению частных показателей $E_{аф\,i}$ (блок 4), и реализуется цикл определения минимального значения частных показателей адекватности функционирования УПСр ЗИР (блоки 5-7). Затем последовательно выполняются циклы определения частных показателей адекватности функционирования УПСр ЗИР (блоки 9, 10, 12, 14) и текущих управляемых параметров, имеющих минимальные значения (блоки 11, 13, 15-18).


Начало
1
Ввод a_i, $a_{i\,min}$, , $E_{min\ вн}$, δ
2
Определение $E_{вн}$
3
Определение $E_{аф}$
4
$E_{аф\,min}$= 10
5
i=1,4
6
нет
$E_{аф\,i} < E_{аф\,min}$
да
7
$E_{аф\,min}$= $E_{аф\,i}$
8
g=j=k=0
9
i=1,4
10
да
$E_{аф\,i}$=$E_{аф\,min}$
нет
12
k=k+1
14
L_k=i, K=k
11
i=1,4
13
да
a_i=$a_{i\,min}$
15
g=g+1
16
нет
j=j+1
17
G=g
18
N_j=i, J=j
19
$E_{вн}$≥$E_{min\ вн}$
нет
20
J=0
да
21
G=4
да
22
да
нет
$E_{вн}$>$E_{min\ вн}$+δ
нет
23
нет
J=K, N_j=L_k, k=j=
да
A
B
C
D
E

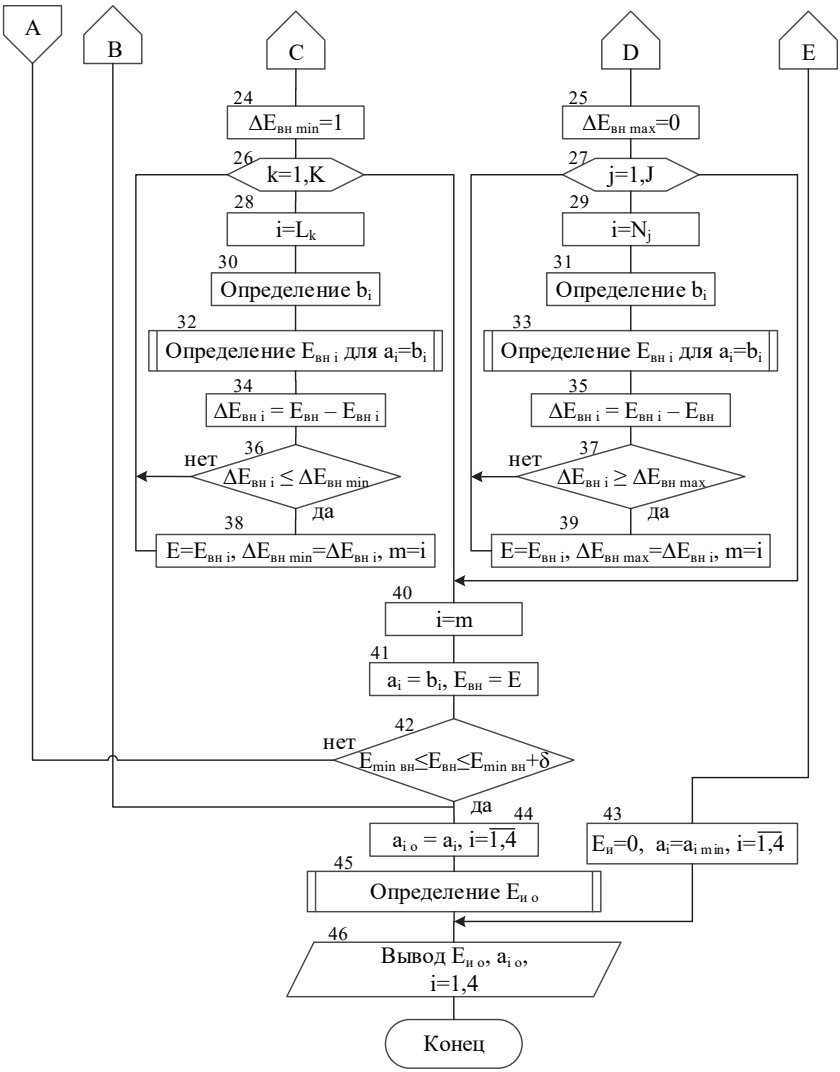


Рис. 3. Алгоритм определения оптимальных значений управляемых параметров и интегрального показателя эффективности функционирования ПЗИР при управлении ЗИР

Исходя из результата выполнения условий блоков 19 и 22 реализуются циклы выбора и увеличения (блоки 26, 28, 30, 32, 34, 36, 38) или уменьшения значения (блоки 27, 29, 31, 33, 35, 37, 39) управляемого параметра ПЗИР для выполнения выражения (7) или (5) соответ-

ственно. При выполнении условий блока 42 текущие значения управляемых параметров фиксируются как оптимальные (блок 44) и для них определяется значение интегрального показателя эффективности функционирования ПЗИР (блок 45). Иначе процедура регулирования управляемых параметров продолжается.

7. Исследование показателей эффективности функционирования ПЗИР. Построение и исследование графических зависимостей показателей эффективности функционирования ПЗИР от управляемых параметров для разных значений внешних параметров подсистемы являются важными при выявлении и изучении закономерностей ОТУ эффективностью функционирования ПЗИР в процессе эксплуатации. Оценка качественных показателей не требует проведения вычислений, и исследования зависимостей этих показателей от варьируемых параметров не представляют интереса. Значительный интерес вызывают зависимости показателя временной неконфликтности функционирования ПЗИР от варьируемых параметров.

Оценка показателя временной неконфликтности функционирования ПЗИР, активно используемая при управлении эффективностью функционирования ПЗИР, производится с использованием математической модели, созданной на базе графовой формализации динамики функционирования ПЗИР [4, 10]. Математическая модель оценки показателя временной неконфликтности функционирования ПЗИР как ОУ приведена в [4].

Комплекс программ (КП), реализующий математическую модель комплексной оценки эффективности функционирования ПЗИР, разработан на базе алгоритмов оценки показателей эффективности функционирования ПЗИР. Структурная схема КП для комплексной оценки эффективности функционирования ПЗИР в СЭД приведена на рисунке 4.

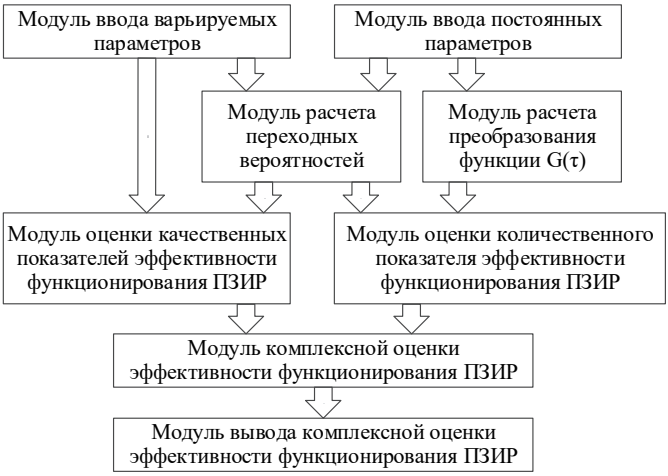


Рис. 4. Структурная схема КП для комплексной оценки эффективности функционирования ПЗИР в СЭД

Кроме управляемых параметров в качестве варьируемых параметров использовались следующие внешние параметры: $p_{да}$ — вероятность применения дополнительной аутентификации пользователя при его обращении к наиболее важному ресурсу; $p_{сб}$ — вероятность применения системной дискеты; $p_{б}$ — вероятность блокировки монитора и клавиатуры в случае не допустимых действий пользователя; $p_{пи}$ — вероятность применения преобразования информации; $p_{рв}$ — вероятность ручного восстановления вычислительной среды; $\tau_{т\,вн}$ — среднее значение максимально допустимого времени выполнения ПЗИР защитных функций [25, 26]. Для более полного анализа зависимостей $E_{вн}(a_i)$ исследования графических зависимостей осуществлялись при предельных значениях внешних параметров. На рисунке 5 приведены некоторые зависимости показателя временной неконфликтности функционирования ПЗИР от управляемых параметров.

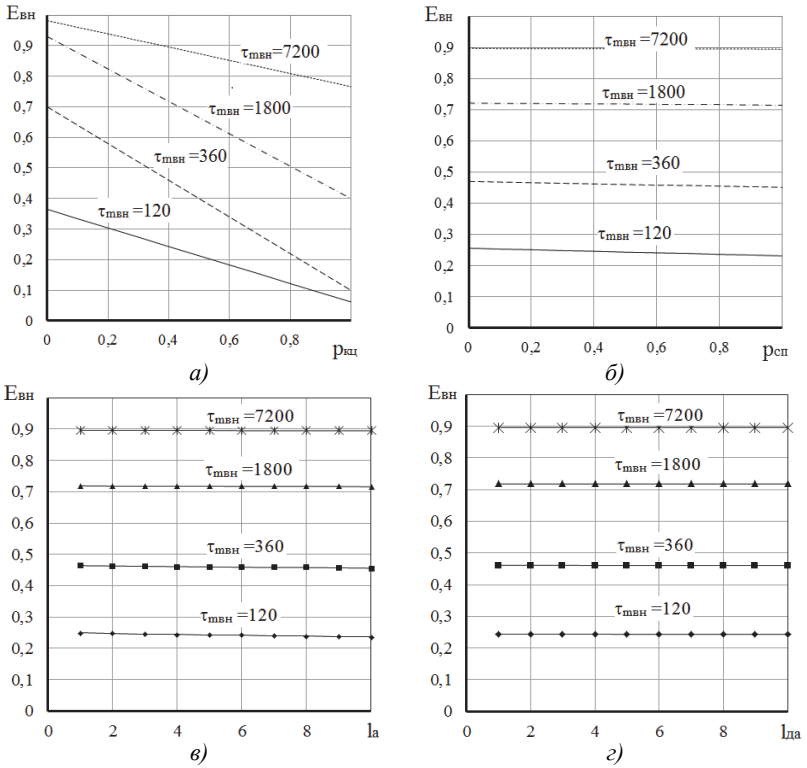


Рис. 5. Зависимости показателя временной неконфликтности функционирования ПЗИР от управляемых параметров

Зависимости даны для значений $p_{да} = 0,1$; $p_{сб} = 0,05$; =0,05; $p_{рв} = 0,01$; $p_{б} = 0,03$; $p_{пи} = 0,8$. Значения управляемых параметров, не участвующих в определении конкретной графической зависимости, были фиксированы: $p_{сп} = 0,5$; $p_{кц} = 0,4$; $l_{аут} = 5$; $l_{доп\ аут} = 5$ и приняты в качестве типовых. Если зависимости $E_{вн}(a_i)$ с ростом управляемого параметра возрастают, то это обозначает повышение эффективности функционирования ПЗИР по данному показателю, а если убывают — снижение ее эффективности.

Исследования и анализ зависимостей показателя временной неконфликтности функционирования рассматриваемой ПЗИР [4], позволяют сделать следующие выводы.

1. Зависимости $E_{вн}(a_i)$ при варьировании внешних параметров сохраняют характер своих изменений. Зависимость $E_{вн}(p_{кц})$ является линейной, убывающей при изменении $p_{кц}$ от 0 до 1. Зависимости $E_{вн}(l_{аут})$, $E_{вн}(l_{доп\ аут})$, $E_{вн}(p_{сп})$ являются линейными, значения которых при увеличении соответствующих управляемых параметров не возрастают.

2. Значения показателя временной неконфликтности функционирования исследуемой ПЗИР существенно зависят от изменений $p_{кц}$, а от изменений остальных управляемых параметров зависят слабо (диапазон изменений — единицы процентов). Это связано с тем, что временные затраты на ввод пароля в процессе стандартной и дополнительной аутентификации пользователя и использование специальных преобразований отдельных файлов незначительны для исследуемой ПЗИР.

3. При изменении $\tau_{т\,вн}$ закономерность изменения зависимостей $E_{вн}(a_i)$ сохраняется при варьировании параметров динамики функционирования ПЗИР. Значение показателя $E_{вн}$ возрастает при увеличении $\tau_{т\,вн}$. Для зависимости $E_{вн}(p_{кц})$ при увеличении $\tau_{т\,вн}$ интенсивность изменения максимального и минимального значений различна. Для малых значений $\tau_{т\,вн}$, возрастание данного параметра вызывает более интенсивное возрастание максимального значения $E_{вн}(p_{кц})$, увеличивая наклон этой зависимости. Рост $\tau_{т\,вн}$, при больших его значениях, наоборот, вызывает более интенсивное возрастание минимального значения $E_{вн}(p_{кц})$, уменьшая наклон данной зависимости.

Для других управляемых параметров зависимости $E_{вн}$ возрастают при увеличении $\tau_{т\,вн}$, уменьшая наклон этих кривых до горизонтального положения. Это связано с тем, что повышается эффективность функционирования ПЗИР при снижении требований по ЗИР СЭД.

На основе проведенного анализа можно сделать вывод, что для исследуемой ПЗИР наиболее эффективным параметром управления ее эффективностью является $p_{кц}$.

8. Заключение. Предложенная модель управления процессами ЗИР воплощает в себе результаты проработки аспектов создания методологии ОТУ ЗИР СЭД на базе ПСр ЗИР и обладает широкими возможностями по ее применению при разработке методов решения задач управления ЗИР. Концептуальная модель ОТУ ЗИР в СЭД представляет два взаимосвязанных вида управления: модель управления КПСЗ и модель управления ПЗИР. Управление КПСЗ реализуется путем оптимального управления эффективностью функционирования ПЗИР, обеспечивающее определение набора значений управляемых параметров подсистемы, позволяющего максимизировать уровень ЗИР при малом отрицательном воздействии ПЗИР на эффективность функционирования СЭД по назначению.

Разработанная методика ОТУ эффективностью функционирования ПЗИР реализуется с помощью управляемых параметров, позволяющих регулировать эффективность функционирования ПЗИР путем изменения их значений при воздействии сигналов управления. Данное управление является оптимальным, при котором определяются значения регулируемых параметров, обеспечивающие максимизация интегрального показателя эффективности функционирования ПЗИР и, соответственно, выполнение требований, предъявляемых к подсистеме. Предложенный метод организации управления позволяет осуществлять эффективное управление ПЗИР с целью повышения безопасности информационного ресурса в СЭД. Представленный алгоритм дает широкие возможности для применения при разработке ПК подсистемы автоматизированного управления эффективностью функционирования ПЗИР в СЭД.

Проведенное исследование показателя временной неконфликтности функционирования ПЗИР позволило определить некоторые закономерности, имеющие место при ОТУ эффективностью функционирования ПЗИР. Полученные результаты исследований по оценке эффективности функционирования ПЗИР как ОУ не противоречат известным данным, а также показывают широкие возможности данного метода при управлении эффективностью функционирования ПЗИР.

Разработанные модели и алгоритм ОТУ, приведенные в статье, в дальнейшем могут быть использованы для разработки предложений по совершенствованию как существующих, так и разрабатываемых СЗИ от НСД с целью повышения ИБ АС.

Литература

1. ГОСТ Р ИСО/МЭК 15408-2013. Информационная технология. Методы и средства обеспечения безопасности. Критерии оценки безопасности информационных технологий // М.: Издательство стандартов. 2013. 267 с.
2. *Шаньгин В.Ф.* Защита информации в компьютерных системах и сетях // М.: ДМК Пресс. 2012. 592 с.
3. *Кузьмин А.В.* Теория систем автоматического управления // М.: ООО «ТНТ». 2012. 224 с.
4. *Застрожнов И.И.* Методологические основы безопасности использования информационных технологий в системах электронного документооборота // Воронеж: ИПЦ «Научная книга». 2011. 252 с.
5. *Qiu L. et al.* Trusted computer system evaluation criteria // National Computer Security Center. 1985. 116 p.
6. Information Technology Security Evaluation Criteria. Harmonized Criteria of France-Germany-Netherlands-United Kingdom // Department of Trade and Industry. 1991. 163 p.
7. Federal Criteria for Information Technology Security // National Institute of Standards and Technology & National Security Agency. Version 1.0. 1992.
8. The Canadian Trusted Computer Product Evaluation Criteria // Canadian System Security Center. Version 3.0e. 1993.
9. Common Criteria for Information Technology Security Evaluation // Common Criteria Project Sponsoring Organisations. Version 2.1. 1999.
10. *Львович Я.Е. и др.* Формализация функционирования перспективной программной системы защиты информации автоматизированных систем // Телекоммуникации. 2004. № 1. С. 38–43.
11. *Муратов А.В., Рогозин Е.А., Застрожнов И.И., Дубровин А.С.* Оценка качества функционирования перспективной программной системы защиты информации автоматизированных систем при проектировании радиоэлектронных средств // Проектирование и технология электронных средств. 2004. № 2. С. 2–5.
12. *Табак Д., Куо Б.* Оптимальное управление и математическое программирование // М.: Главная редакция физико-математической литературы издательства "Наука". 1975. 280 с.
13. *Yun L., Sheng-Peng L., Li L., Yuan-Yuan M.* Effectiveness Evaluation on Cyberspace Security Defense System // International Conference on Network and Information Systems for Computers. 2015. pp. 576–579.
14. *Xin Z., Shaojie M., Fang Z.* Research on effectiveness evaluation of the mission-critical system // Proceedings of 2013 2nd International Conference on Measurement, Information and Control. 2013. pp. 869–873.
15. *Ptitsyn P.S., Radko D.V., Lankin O.V.* Designing architecture of software framework for building security infrastructure of global distributed computing systems // ARPN Journal of Engineering and Applied Sciences. 2016. vol. 11. no. 19. pp. 11599–11610.
16. *Заряев А.В. и др.* Методическое обеспечение управления доступом пользователей к рабочей среде автоматизированных систем // Телекоммуникации. 2004. № 2. С. 39–44.
17. *Ощепков А.Ю.* Системы автоматического управления. Теория, применение, моделирование в MATLAB // СПб.: Лань. 2013. 208 с.

18. *Скрыль С.В., Окрачков А.А.* Метод количественной оценки показателей эффективности систем защиты информации от несанкционированного доступа // Вестник Воронежского института МВД России. 2013. № 3. С. 78–83.
19. *Мещерякова Т.В.* Показатели для оценки эффективности информационных процессов в условиях обеспечения их защищенности в автоматизированных информационных системах органов внутренних дел // Вестник Воронежского института МВД России. 2014. № 1. С. 141–150.
20. *Скрыль С.В. и др.* Показатели эффективности информационной деятельности в условиях комплексного технического контроля обеспечения защищенности речевой информации // Приборы и системы. Управление, контроль, диагностика. 2016. № 2. С. 28–34.
21. *Скрыль С.В., Голубков Д.А., Половинкин В.А.* Показатели эффективности информационных процессов в интегрированных системах безопасности в условиях обеспечения антивирусной защиты // Вестник Воронежского института МВД России. 2014. № 4. С. 212–220.
22. *Змеев С.А., Селютин И.Н., Скрыль Е.Б., Никитин А.А.* Рациональный выбор средств защиты при структурном синтезе программных систем защиты информации в системах электронного документооборота // Вестник Воронежского института ФСИН России. 2013. № 2. С. 55–59.
23. *Родионова Н.С., Белокуров С.В., Скрыль С.В.* Показатели эффективности управления информационными процессами в интегрированных системах безопасности // Вестник Воронежского государственного университета инженерных технологий. 2014. № 4. С. 79–84.
24. *Мещерякова Т.В., Фирюлин М.Е., Хворов Р.А.* Аналитические модели показателей состояния защищенности информации в центрах обработки данных органов внутренних дел// Вестник Воронежского института МВД России. 2015. № 3. С. 104–113.
25. *Ланкин О.В., Мещерякова Т.В., Селютин И.Н.* Обеспечение целостности информационных ресурсов подсистемы безопасности распределенных информационно-вычислительных систем // Вестник Воронежского института МВД России. 2017. № 1. С. 35–42.
26. *Скрыль С.В. и др.* Оценка характеристик компонент защиты информации от несанкционированного доступа для реализации функций обеспечения целостности и доступности информации // Приборы и системы. Управление, контроль, диагностика. 2014. № 7. С. 21–25.

Авсентьев Олег Сергеевич — д-р техн. наук, профессор, профессор кафедры информационной безопасности, Воронежский институт Министерства внутренних дел России. Область научных интересов: информационная безопасность, защита информации, моделирование систем защиты информации. Число научных публикаций — 87. osaos@mail.ru; пр. Патриотов, 53, Воронеж, 394065; р.т.: +7(473)200-52-36.

Дровникова Ирина Григорьевна — д-р техн. наук, доцент, профессор кафедры автоматизированных информационных систем органов внутренних дел, Воронежский институт Министерства внутренних дел России. Область научных интересов: системы защиты информации, эволюционное моделирование, автоматизированные информационные системы, теория вероятности, управление в социально-экономических системах. Число научных публикаций — 210. idrovnikova@mail.ru; пр. Патриотов, 53, Воронеж, 394086; р.т.: +7(472)200-51-88.

Застрожнов Игорь Иванович — к-т техн. наук, доцент кафедры эксплуатации авиационного оборудования, Военный учебно-научный центр ВВС «Военно-воздушная академия имени профессора Н.Е. Жуковского и Ю.А. Гагарина». Область научных интересов: защита информации от НСД в АИС, проектирование и управление процессами защиты

информации на основе количественной оценки СЗИ от НСД в АИС. Число научных публикаций — 112. zasigor@yandex.ru; ул. Краснознаменная, 153, Воронеж, 394052; р.т.: +7(473)200-51-88.

Попов Антон Дмитриевич — адъюнкт кафедры автоматизированных информационных систем органов внутренних дел, Воронежский институт МВД России. Область научных интересов: защита информации от НСД в АИС, проектирование и управление процессами защиты информации на основе количественной оценки СЗИ от НСД в АИС, тестирование и анализ СЗИ, разработка АИС, прикладная информатика. Число научных публикаций — 40. anton.holmes@mail.ru; пр. Патриотов, 53, Воронеж, 394086; р.т.: +7(473)200-51-80.

Рогозин Евгений Алексеевич — д-р техн. наук, профессор, профессор кафедры автоматизированных информационных систем органов внутренних дел, Воронежский институт Министерства внутренних дел России. Область научных интересов: защита информации от НСД в АИС, проектирование и управление процессами защиты информации на основе количественной оценки СЗИ от НСД в АИС, прикладная информатика. Число научных публикаций — 240. evgenirogozin@yandex.ru; пр. Патриотов, 53, Воронеж, 394086; р.т.: +7(473)200-51-88.

DOI 10.15622/sp.57.8

O.S. AVSENTEV, I.G. DROVNIKOVA, I.I. ZASTROZHNOV, A.D. POPOV, E.A. ROGOZIN

CONTROL TECHNIQUES OF INFORMATION RESOURCE PROTECTION OF ELECTRONIC DOCUMENT MANAGEMENT SYSTEM

Avsentev O.S., Drovnikova I.G., Zastrozhnov I.I., Popov A.D., Rogozin E.A. **Control Techniques of Information Resource Protection of Electronic Document Management System.**

Abstract. The article discusses methodological bases for the organizational and technological control (OTC) of the protection of an information resource (PIR) of electronic document management systems (EDMS) based on software (SW) of information security. The authors developed a conceptual model of control of PIR of EDMS on the basis of conceptual study of the aspects of the formation of OTC PIR EDMS methodology on the basis of the SW of PIR, which has ample opportunities to be used for developing methods of administrative tasks solution. The paper presents a technique for efficiency management of functioning of the information resource protection subsystem (IRPS) in EDMS, assuming optimization of the subsystem controlled parameters that maximize an integral index of efficiency of IRPS functioning and respectively execution of requirements imposed to the subsystem. The algorithm for determining best values of the IRPS controlled parameters and best value of an integral index of efficiency of the subsystem, providing a possibility of creating specific subsystems of IRPS automated management efficiency in EDMS, is given. Results of calculations for a research of an index of temporal non-conflictness of IRPS functioning are analyzed.

Keywords: organizational and technological control, protection of an information resource, electronic document management system, subsystem of protection of an information resource, efficiency of a subsystem, control of efficiency.

Avsentev Oleg Sergeevich — Ph.D., Dr. Sci., professor, professor of information security department, Voronezh Institute of the Ministry of Interior. Research interests: information security, information protection, modeling of information security systems. The number of publications — 87. osaos@mail.ru; 53, pr. Patriotov, Voronezh, 394086, Russia; office phone: +7(473)200-52-36.

Drovnikova Irina Grigorevna — Ph.D., Dr. Sci., associate professor, professor of automated information systems in interior affairs department, Voronezh Institute of the Ministry of Interior. Research interests: information systems security, evolutionary modeling, automated information systems, probability theory, social-and-economic system management. The number of publications — 210. idrovnikova@mail.ru; 53, Prospekt Patriotov, Voronezh, 394086, Russia; office phone: +7(472)200-51-88.

Zastrozhnov Igor Ivanovich — Ph.D., associate professor of aeronautical equipment department, Military educational scientific center of the Air Force "Air-force academy of a name of professor N.E. Zhukovsky and Yu. A. Gagarin". Research interests: information security from the unauthorized access in AIS; design and management of information security processes based on the quantitative assessment of ISS security from the unauthorized access in AIS. The number of publications — 112. zasigor@yandex.ru; 153, Krasnoznamennaya St., Voronezh, 394052, Russia; office phone: +7(473)200-51-88.

Popov Anton Dmitrievich — Ph.D. student of automated information systems in interior affairs department, Voronezh Institute of the Ministry of Interior. Research interests: infor-

mation protection against unauthorized access in AIS, design and management of information security processes based on quantitative assessment of ISS, testing and analysis of ISS, AIS development, applied computer science. The number of publications — 40. anton.holmes@mail.ru; 53, Prospekt Patriotov, Voronezh, 394086, Russia; office phone: +7(473)200-51-80.

Rogozin Evgeniy Alekseevich — Ph.D., Dr. Sci., professor, academician of RANS, professor of automated information systems in interior affairs department, Voronezh Institute of the Ministry of Interior. Research interests: information security from the unauthorized access in AIS, design and management of information security processes based on the quantitative assessment of ISS security from the unauthorized access in AIS, applied informatics. The number of publications — 240. evgenirogozin@yandex.ru; 53, Prospekt Patriotov, Voronezh, 394086, Russia; office phone: +7(473)200-51-88.

References

1. GOST P 15408-2013. [Methods and security protections. Criteria for evaluation of safety of information technologies]. M.: Izdatel'stvo standartov. 2013. 267 p. (In Russ.).
2. Shangin V.F. *Zashhita informacii v komp'juternyh sistemah i setjah* [Information security in computer systems and networks: studies. manual]. M.: DMK Press. 2012. 592 p. (In Russ.).
3. Kuzmin A.V., Skhirtladze A.G. *Teorija sistem avtomaticheskogo upravlenija* [Theory of systems of automatic control]. M.: LLC TNT. 2012. 224 p. (In Russ.).
4. Zastrozhnov I.I. Rogozin E.A., Bagayev M.A. *Metodologicheskie osnovy bezopasnosti ispol'zovanija informacionnyh tehnologij v sistemah jelektronnogo dokumentooborota* [Methodological bases of safety of use of information technologies in electronic document management systems: monograph]. Voronezh: IPTs "Scientific Book". 2011. 252 p. (In Russ.).
5. Qiu L. et al. Trusted computer system evaluation criteria. National Computer Security Center. 1985. 116 p.
6. Information Technology Security Evaluation Criteria. Harmonized Criteria of France-Germany-Netherlands-United Kingdom. Department of Trade and Industry. 1991. 163 p.
7. Federal Criteria for Information Technology Security. National Institute of Standards and Technology & National Security Agency. Version 1.0. 1992.
8. Canadian Trusted Computer Product Evaluation Criteria. Canadian System Security Center. Version 3.0e. 1993.
9. Common Criteria for Information Technology Security Evaluation. Common Criteria Project Sponsoring Organisations. Version 2.1. 1999.
10. Lvovich Ya.E. et al. [Formalization of functioning of perspective program system of information security of automated systems]. *Telekommunikacii – Telecommunications.* 2004. vol. 1. pp. 38–43. (In Russ.).
11. Muratov A.V., Rogozin E.A., Zastrozhnov I.I., Dubrovin A.S. [Otsenka of quality of functioning of perspective program system of information security of automated systems in case of design of radio-electronic means]. *Proektirovanie i tehnologija jelektronnyh sredstv – Design and technology of electronic means*. 2004. vol. 2. pp. 2–5. (In Russ.).
12. Tabak D., Kuo B.C. *Optimal Control by Mathematical Programming.* Prentice-Hall. 1971. (Russ. ed.: Tabak D., Kuo B. *Optimal'noe upravlenie i matematicheskoe programmirovanie*. M.: Glavnaja redakcija fiziko-matematicheskoj literatury izdatel'stva "Nauka". 1975. 280 p.).
13. Yun L., Sheng-Peng L., Li L., Yuan-Yuan M. Effectiveness Evaluation on Cyberspace Security Defense System. International Conference on Network and Information Systems for Computers. 2015. pp. 576–579.

14. Xin Z., Shaojie M., Fang Z. Research on effectiveness evaluation of the mission-critical system.Proceedings of 2013 2nd International Conference on Measurement, Information and Control. 2013. pp. 869–873.
15. Ptitsyn P.S., Radko D.V., Lankin O.V. Designing architecture of software framework for building security infrastructure of global distributed computing systems. *ARPN Journal of Engineering and Applied Sciences*. 2016. vol. 11. no. 19. pp. 11599–11610.
16. Zaryaev A.V. et al. [Methodical support of access control of users to a work environment of automated systems]. *Telekommunikacii – Telecommunications.* 2004. vol. 2. pp. 39–44. (In Russ.).
17. Oshchepkov A.Yu. [Systems of automatic control. The theory, application, simulation in MATLAB]. SPb.: Fallow deer. 2013. 208 p. (In Russ.).
18. Skryl' S.V., Okrachkov A.A. [The method of quantifying performance information protection from unauthorized access] *Vestnik Voronezhskogo gosudarstvennogo universiteta inzhenernyh tehnologij - Bulletin of Voronezh State University of Engineering Technology*. 2013. vol. 3. pp. 78–83. (In Russ.).
19. Mesherjakova T.V. [The method of quantifying performance information protection from unauthorized access of the Law Enforces Agencies]. *Vestnik Voronezhskogo gosudarstvennogo universiteta inzhenernyh tehnologij - Bulletin of Voronezh State University of Engineering Technology.* 2014. vol. 1. pp. 141–150. (In Russ.).
20. Skryl' S.V. et al. [Indicators of effectiveness of information activity in the conditions of complex technical control provide voice information security]. *Pribory i system. Upravlenie, kontrol', diagnostika - Instruments and systems. Management, monitoring, diagnostics*. 2016. vol. 2. pp. 28–34. (In Russ.).
21. Skryl' S.V., Golubkov D.A., Polovinkin V.A. [Performance indicators of information processes in integrated security systems in the conditions of anti-virus protection software]. *Vestnik Voronezhskogo instituta MVD Rossii - Bulletin of the Voronezh Institute of the Ministry of the Interior of Russia*. 2014. vol. 4. pp. 212–220. (In Russ.).
22. Zmeev S.A., Seljutin I.N., Skryl' E.B., Nikitin A.A. [Rational choice of means of protection in the structural synthesis of software systems for information protection in electronic document management systems]. *Vestnik Voronezhskogo instituta FSIN Rossii - Bulletin of the Voronezh Institute of the Federal Penitentiary Service of Russia*. 2013. vol. 1. pp. 82–85. (In Russ.).
23. Rodionova N.S., Belokurov S.V., Skryl' S.V. [Performance information processes management in integrated security systems]. *Vestnik Voronezhskogo gosudarstvennogo universiteta inzhenernyh tehnologij - Bulletin of Voronezh State University of Engineering Technology.* 2014. vol. 4. pp. 79–84. (In Russ.).
24. Meshherjakova T.V., Firjulin M.E., Hvorov R.A. [Analytical models of information security status indicators in the data processing centers of internal affairs bodies]. *Vestnik Voronezhskogo instituta MVD Rossii - Bulletin of the Voronezh Institute of the Ministry of the Interior of Russia*. 2015. vol. 3. pp. 104–113. (In Russ.).
25. Lankin O.V., Meshherjakova T.V., Seljutin I.N. [Ensuring the integrity of information resources of the security subsystem of distributed information and computing systems]. *Vestnik Voronezhskogo instituta MVD Rossii - Bulletin of the Voronezh Institute of the Ministry of the Interior of Russia*. 2017. vol. 1. pp. 35–42. (In Russ).
26. Skryl' S.V. et al. [Sizintsev Evaluation of information security components of the characteristics of the unauthorized access to the realization of functions to ensure the integrity and availability of information]. *Pribory i system. Upravlenie, kontrol', diagnostika - Instruments and systems. Management, monitoring, diagnostics*. 2014. vol. 7. pp. 21–25. (In Russ).

УДК 004.056 **DOI** 10.15622/sp.57.9

Е.В. Дойникова, И.В. Котенко

СОВЕРШЕНСТВОВАНИЕ ГРАФОВ АТАК ДЛЯ МОНИТОРИНГА КИБЕРБЕЗОПАСНОСТИ: ОПЕРИРОВАНИЕ НЕТОЧНОСТЯМИ, ОБРАБОТКА ЦИКЛОВ, ОТОБРАЖЕНИЕ ИНЦИДЕНТОВ И АВТОМАТИЧЕСКИЙ ВЫБОР ЗАЩИТНЫХ МЕР

Дойникова Е.В., Котенко И.В. **Совершенствование графов атак для мониторинга кибербезопасности: оперирование неточностями, обработка циклов, отображение инцидентов и автоматический выбор защитных мер.**

Аннотация. Своевременность и адекватность реагирования на инциденты компьютерной безопасности, а также потери организаций от компьютерных атак, зависят от точности определения ситуации при мониторинге кибербезопасности. Статья посвящена совершенствованию моделей атак в виде графов для задач мониторинга кибербезопасности. Рассматривается ряд актуальных проблем, связанных с использованием графов атак, и способов их решения, в том числе оперирование неточностями при определении пред- и постусловий выполнения атакующих действий, обработка циклов при использовании байесовского вывода для анализа графа атак, отображение инцидентов на графе атак, а также автоматический выбор защитных мер в случае высокого уровня риска. Представлен реализованный ранее и модифицированный с учетом предложенных изменений программный прототип компонента системы мониторинга кибербезопасности и результаты экспериментов. Влияние изменений на результаты мониторинга кибербезопасности показано на примере оценки защищенности фрагмента компьютерной сети.

Ключевые слова: граф атак, вероятность атаки, мониторинг кибербезопасности, компьютерные сети, оценка защищенности, показатели защищенности, реагирование на атаки, оценивание уязвимостей.

1. Введение. В современном мире, когда компьютерные сети являются важнейшей частью инфраструктуры большинства организаций, наблюдается повышенный интерес к проведению кибератак, использующих уязвимости компьютерных сетей. Успешная реализация таких атак может привести к серьезному ущербу для деятельности организаций, напрямую зависящих от информационных технологий. Поскольку задача интегрированной киберзащиты организаций и устранения уязвимостей является трудоемкой, затратной по стоимости и не всегда оправданной, особенно важно проводить проактивный мониторинг кибербезопасности [1, 2] для своевременного выявления кибератак, распознавания целей и квалификации атакующих, эффективного предотвращения развития атак в системе, выявления текущего состояния защищенности и слабых мест системы, а также решения задач компьютерной криминалистики (форензики).

Для выявления возможных путей атак в компьютерных сетях широко используются графы атак. Они представляют собой множество

возможных атакующих действий и переходов между ними. В существующих исследованиях описываются различные виды графов атак [3-18]. Данное исследование основано на работах [19, 20], в которых каждое атакующее действие представляет собой эксплуатацию уязвимости сети. Переходы между действиями задаются пред- и постусловиями эксплуатации уязвимостей, определяемыми на основе индексов (показателей) системы оценивания уязвимостей Common Vulnerability Scoring System (CVSS) версии 2.0 [21]. На основе графа атак можно проследить путь атаки в системе от источника до цели атаки, определить, какие ресурсы сети находятся под угрозой, определить текущее состояние атаки на основе информации об инцидентах безопасности и сделать выводы о характеристиках атакующего.

Тем не менее практическое применение графов атак для заявленных целей мониторинга кибербезопасности затрудняется такими аспектами, как неопределенность исходных данных, сложность построения и анализа графов, и отсутствие удовлетворительных программных решений в области их построения и анализа.

В работе развивается метод построения и анализа модели атак в виде графа, предложенный в [19, 20, 22, 23]. Метод усовершенствован для удовлетворения целей мониторинга кибербезопасности и выбора защитных мер. Основной особенностью исходного метода является автоматизированное оперативное построение и анализ графов атак с применением открытых стандартов представления данных по безопасности и открытых баз уязвимостей. При этом к методу предъявляются следующие требования: учет всех возможных последовательностей атакующих действий; адекватная оценка защищенности анализируемой системы за счет вычисления уровня риска на основе вероятностей успешной реализации атакующих действий и тщательного учета ущерба, наносимого в результате успешной реализации атакующих действий.

Основными недостатками с точки зрения поставленных целей и требований являются:

– неточность определения значений индексов CVSS версии 2.0, что ведет к неточностям при построении графа (под неточностью в данном случае понимается такое определение значений, индексов, которое допускает неоднозначное толкование) и нарушает требование учета всех возможных последовательностей атакующих действий;

– отсутствие учета атак, не использующих уязвимости, что нарушает требование учета всех возможных последовательностей атакующих действий и создает сложности при отображении инцидентов на граф атакующих действий;

– удаление циклических связей между узлами графа в процесс анализа графа, что также нарушает требование учета всех возможных последовательностей атакующих действий;

– отсутствие выделения классов угроз на графе, что препятствует тщательному учету ущерба, наносимого в результате успешной реализации атакующих действий, и затрудняет автоматический выбор контрмер (в текущей реализации угроза определяется как последовательность эксплуатации уязвимостей, ущерб от их эксплуатации может отличаться, соответственно, различные классы угроз не разделяются).

Цель исследования — разработка эффективного подхода к построению и анализу графов атак, который позволит осуществлять адекватный мониторинг кибербезопасности и выбирать рациональные защитные меры. В данной работе в рамках поставленной цели решаются задачи по модификации метода построения и анализа графа атак для устранения перечисленных выше недочетов, а именно:

– для устранения неточностей, связанных с применением CVSS версии 2.0, предложены модификации подхода к построению и анализу графа атак на основе использования новой версии CVSS версии 3.0 [24];

– для учета атак, не использующих уязвимости, предложены модификации подхода на основе использования шаблонов атак CAPEC [25];

– определены модификации подхода для обработки циклов на графе атак для последующего анализа;

– узлы графа атак переопределены путем разбиения уязвимостей на новые группы, определенные в зависимости от причиняемого ущерба, для выделения различных классов угроз и эффективного выбора контрмер.

В статье проведен анализ влияния предложенных изменений на процесс мониторинга кибербезопасности и выбора контрмер. Таким образом, основной вклад данной статьи состоит в совершенствовании подхода к построению и анализу графа атак (на основе CVSS версии 3.0 и CAPEC, выделения и обработки различных типов циклов графа, а также выделения различных классов угроз в зависимости от причиняемого ущерба), необходимого для мониторинга кибербезопасности и выбора контрмер. Статья организована следующим образом. Во втором разделе рассмотрены релевантные исследования в области генерации графов атак, анализа защищенности и мониторинга кибербезопасности. В третьем разделе приведено описание подхода к построению и анализу графа атак, взятого за

основу, и предложения по его изменению. В четвертом разделе приведен пример применения подхода, кратко описаны результаты экспериментов и дискуссия. В заключении сделаны выводы по результатам работы и описаны будущие направления исследований.

2. Релевантные исследования. Для решения задач мониторинга кибербезопасности были разработаны различные классы систем мониторинга и управления инцидентами (SIEM). Тем не менее они, как правило, не реализуют функции детальной оценки рисков, моделирования и прогнозирования атак. Шаги в этом направлении были сделаны, например, в системе MaxPatrol SIEM компании Positive Technologies [26] и продукте OSSIM компании AlientVault [27], а также в рамках проекта MASSIF FP7 путем внедрения компонента моделирования атак на основе графов и оценки защищенности [28].

В то же время существует большое количество исследований в области построения и анализа графов атак [3-20]. В [3, 4] предлагаются подходы к повышению защищенности компьютерной сети с использованием графов атак. В [9] рассматривается анализ защищенности с использованием деревьев атак. Подходы к оценке риска на основе графов атак сформулированы в [13, 15]. В [14] для повышения защищенности компьютерной сети используются графы атак совместно с теорией игр. В [16] анализа графов атак применяется приближенный вывод. В [17] защищенность компьютерной сети оценивается на основе графов зависимостей эксплойтов. Основным недостатком графов, предложенных в данных работах, является сложность построения. В [5, 6, 20] рассматривается решение проблемы оперативного построения графов.

В ряде работ предлагается использовать вероятностные графы атак для оценки риска [13, 29], анализа защищенности с учетом характеристик атакующего [8], повышения защищенности динамических сетей [18], реагирования на вторжения [30, 31]. В других работах применяются байесовские графы атак. В том числе для оценки защищенности [11, 32, 33], оценки защищенности с учетом характеристик атакующего [34], динамической оценки защищенности [36].

Одной из основных сложностей при формировании байесовских графов атак является обработка циклов. Этот вопрос рассматривается в [35-37].

Кроме того, существуют инструменты, реализующие методики моделирования атак и оценки защищенности. К ним относятся система построения и анализа графов NetSPA [4], CAULDRON [38], SecurITree [39] и другие.

Несмотря на большое количество исследований, в них в недостаточной степени учитываются аспекты комплексного применения открытых стандартов, таких как система оценивания уязвимостей (CVSS) [21, 24] и классификация шаблонов атак (CAPEC) [25], для аналитического моделирования атак и контрмер. В данной работе эта задача рассматривается с учетом некоторых недостатков моделирования атак, которые необходимо решить для эффективного мониторинга кибербезопасности и выбора защитных мер. Предлагается модифицированный подход к построению и анализу графа, который учитывает особенности CVSS версии 3.0 и шаблоны CAPEC, позволяет обрабатывать циклы графа для вычисления показателей защищенности, а также выделять различные классы угроз для последующего выбора защитных мер путем переопределения узлов графа за счет разбиения уязвимостей на группы в зависимости от причиняемого ущерба.

3. Модель атак для мониторинга кибербезопасности и поддержки принятия решений. Развиваемый авторами подход к мониторингу кибербезопасности и поддержке принятия решений базируется на аналитическом моделировании. Основой для последующих модификаций является модель атак в виде графа, предложенная и подробно описанная в [20]. Входными данными для построения модели являются данные об анализируемой компьютерной сети, данные об уязвимостях ее программно-аппаратного обеспечения, и характеристики уязвимостей, полученные из открытых источников. Результатом является множество возможных последовательностей атакующих действий. Методика анализа защищенности и выбора контрмер на базе данного графа представлена в [22, 23].

В основе подхода к построению и анализу графа, а также выбору контрмер лежит применение открытых стандартов, позволяющих формализовать исходные данные, (CVSS — для определения связей в графе и вычисления показателей защищенности [21], Common Platform Enumeration (CPE) [40] — для представления программно-аппаратного обеспечения и Common Configuration Enumeration (CCE) [41] — для представления уязвимых конфигураций) и данных по безопасности из открытых источников (база NVD [42]) для автоматизации процесса анализа защищенности. В процессе исследований был выявлен ряд недочетов, уже освященных в предыдущих разделах. Модификации процесса генерации графа атак для устранения этих недочетов описаны ниже.

3.1. Модель атак с учетом CVSS версии 3.0. Граф атак задается как множество взаимосвязанных атакующих действий [20]. Каждое атакующее действие определяется как эксплуатация уязвимости некоторой группы. Для связи атакующих действий в последовательности атак, группы выделяются в соответствии с предусловиями эксплуатации уязвимостей на основе индексов CVSS версии 2.0 (*AccessVector* (*AV*) — вектор доступа к уязвимости) и постусловиями их эксплуатации (*priv* — полученные привилегии и/или *CIA* — ущерб конфиденциальности, целостности и доступности) (таблица 1).

Таблица 1. Группы уязвимостей, выделенные на основе CVSS версии 2.0

Группа	Индексы CVSS		
	AV	*priv*	*CIA*
Группа 1	N/A (сетевой доступ или доступ из смежной сети)	user/other (привилегии пользователя или другое)	any (любой ущерб)
Группа 2	N/A	admin (привилегии администратора)	any
Группа 3	N/A	none (не дает привилегий)	P/C (частичный или полный ущерб)
Группа 4	L (локальный доступ)	admin	any
Группа 5	L	user/other	$CIA > CIA_{группа1}$ (учитываются только те уязвимости, чья эксплуатация ведет к большему ущербу, чем при эксплуатации уязвимостей группы 1)
Группа 6	L	none	$CIA > CIA_{группа1}$

На рисунке 1*а* представлены связи между атакующими действиями, использующими уязвимости соответствующей группы в рамках одного хоста (узла сети).

Однако определение индексов CVSS версии 2.0 имеет ряд неопределенностей, что ведет к неточностям при формировании графа атакующих действий, в том числе:

– индекс *AV* принимает значение L (локальный доступ) как в случае физического, так и в случае логического доступа к компьютеру;

– индексы *Impact* не учитывают область, на которую распространяется ущерб от эксплуатации уязвимости.

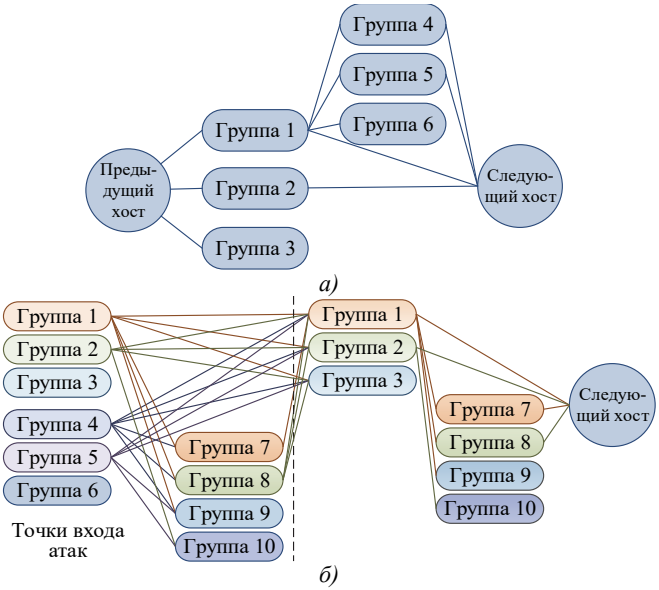


Рис. 1. Связи между группами уязвимостей для CVSS 2.0 (а) и CVSS 3.0 (б)

Кроме того, ряд неопределенностей ведет к неточностям при анализе графа атакующих действий, в том числе:

– индекс *AccessComplexity* (*AC*), характеризующий сложность эксплуатации уязвимости, не позволяет отдельно учитывать необходимость взаимодействия с пользователем;

– индекс *Authentication* (*Au*) определяет, требуется ли прохождение дополнительной процедуры аутентификации для эксплуатации уязвимости, но не определяет требуемый уровень привилегий.

Спецификация CVSS версии 3.0 была сформирована, чтобы устранить проблемы CVSS версии 2.0. При этом изменились определения и возможные значения ряда индексов CVSS:

– индекс *AV* переименован в *AttackVector* (обозначим его AV_{v3}, чтобы избежать путаницы с индексом *AV* CVSS версии 2.0), его значения разделены на «локальный» (L) и «физический» (P), чтобы выделить физический и логический тип доступа к компьютеру;

– для определения области, на которую распространяется ущерб от эксплуатации уязвимости, введен новый индекс *Scope* (*S*);

– индекс *AC* переименован в *AttackComplexity* (обозначим его AC_{v3}, чтобы избежать путаницы с индексом *AC* CVSS версии 2.0), и больше не включает взаимодействие с пользователем. Для учета необходимости взаимодействия с пользователем введен отдельный индекс *UserInteraction*;

– индекс *Au* заменен на индекс *PrivilegesRequired* (*PR*), определяющий требуемый для эксплуатации уязвимостей уровень привилегий.

На основе новых индексов CVSS версии 3.0 и их значений авторами данной работы были сформированы новые группы уязвимостей (таблица 2). На рисунке 1*б* представлены связи между атакующими действиями, использующими уязвимости соответствующей группы (с учетом входных точек атаки: Группа 1 — Группа 3 — у атакующего есть сетевой (удаленный) доступ к хосту в терминах сетевого уровня модели OSI, Группа 7 — Группа 10 — у атакующего есть локальный доступ к хосту, то есть для эксплуатации уязвимости атакующий должен быть авторизован на хосте или дождаться определенных действий авторизованного пользователя, и Группа 4 — Группа 6 — у атакующего есть физический доступ к хосту).

Таблица 2. Группы уязвимостей, выделенные на основе CVSS версии 3.0

Группа	Индексы CVSS			
	AV_{v3}	*PR*	*priv*	*CIA*
Группа 1	N/A	N (доступ к файлам и настройкам не требуется)	user/ other	any
Группа 2	N/A	N	admin	any
Группа 3	N/A	N	none	P/C
Группа 4	P (физический доступ)	N	user/ other	any
Группа 5	P	N	admin	any
Группа 6	P	N	none	P/C
Группа 7	L	L/N (L — атакующий зарегистрирован в системе с привилегиями, предоставляющими базовые возможности пользователя)	admin	any
Группа 8	L	L/N	user/ other	CIA>($CIA_{группа1}$) ИЛИ ($CIA_{группа4}$)
Группа 9	L	L/N	none	CIA>($CIA_{группа1}$) ИЛИ ($CIA_{группа4}$) ИЛИ ($CIA_{группа5}$)
Группа 10	L	H (для эксплуатации уязвимости требуются привилегии администратора)	any	CIA> ($CIA_{группа2}$) ИЛИ ($CIA_{группа5}$)

Как видно из рисунка 1*б*, схема связей между группами уязвимостей усложнилась, однако она позволяет сформировать возможные пути атак точнее, чем предложенная ранее схема, представленная на рисунке 1*а*.

Тем не менее в данном случае все еще не учитываются атакующие действия, не использующие уязвимости программно-аппаратного обеспечения, например разведывательные действия.

3.2. Модель атак с учетом шаблонов атак CAPEC. Для учета атакующих действий, не использующих уязвимости программно-аппаратного обеспечения, предлагается применять шаблоны атак CAPEC [25].

Шаблоны атак можно классифицировать по целям атаки. Для этого предлагается использовать значение поля «Цель» (“Purpose”) шаблона атаки. Данное поле может принимать значения: разведка, проникновение, эксплуатация, обфускация [25]. Например, шаблон CAPEC-169 «Footrinting» относится к шаблонам с целью «разведка», а шаблон CAPEC-100 «Overflow Buffers» — к шаблонам с целями «проникновение» и «эксплуатация».

При моделировании атак для учета данных шаблонов применяется следующий алгоритм:

1. Для атакующего выбираются доступные шаблоны атак с целью «разведка» с учетом уровня навыков атакующего (на основе значения поля шаблона «Attacker Skills or Knowledge Required») и слабых мест из базы CWE (Common Weakness Enumeration [43, 44]) доступных хостов (в случае, если не существует шаблонов с целью «разведка», соответствующих слабым местам доступных хостов, данный узел добавлен не будет). Указанные шаблоны добавляются между хостами как узел графа «Разведка» (рисунок 2*а*).

2. Узел «Разведка» соединяется с группами уязвимостей доступных хостов (рисунок 2*а*).

3. Узел графа «Обфускация» добавляется после группы уязвимостей в случае, если существуют шаблоны атак с целью «обфускация», соответствующие слабым местам CWE уязвимостей группы. Данный узел соединяется с соответствующей группой уязвимостей и следующей доступной группой уязвимостей, либо с узлом графа «Разведка» (см. шаг 4, рисунок 2*б*).

4. Узел графа «Разведка» добавляется после группы уязвимостей внутри хоста (либо после узла графа «Обфускация»), в случае, если существуют шаблоны атак с целью «разведка», соответствующие слабым местам CWE уязвимостей доступных групп (рисунок 2*в*).

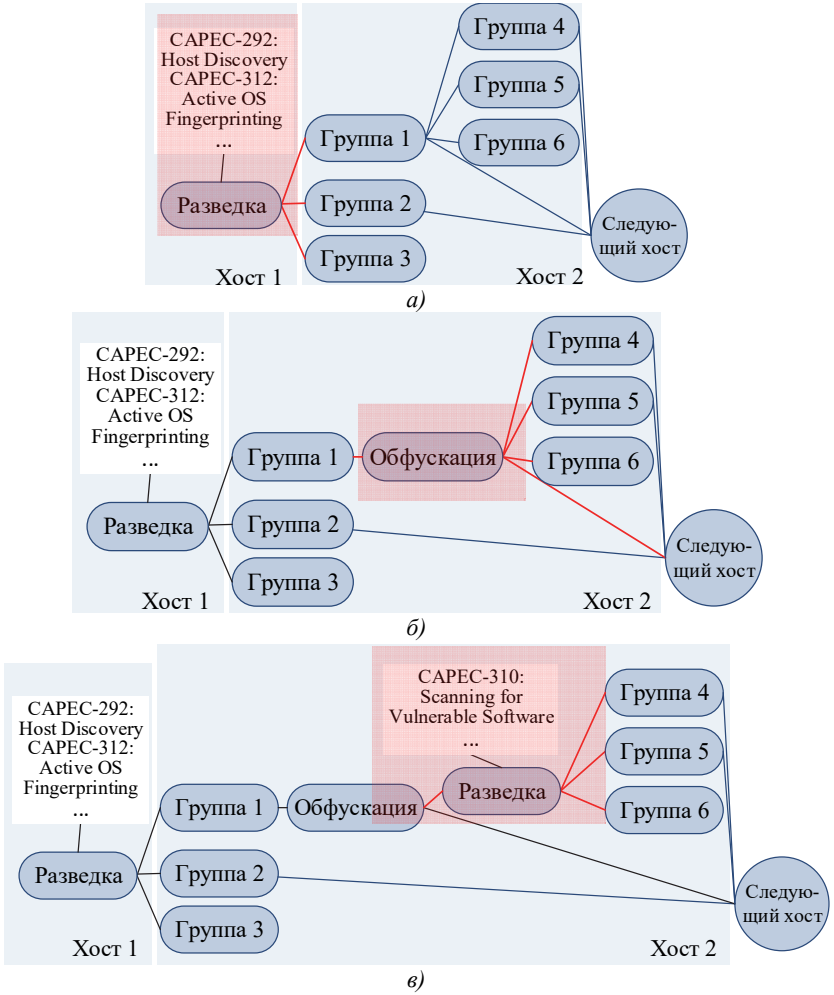


Рис. 2. Добавление шаблонов атак CAPEC на граф атакующих действий

Вышеперечисленные шаги выполняются для всех хостов компьютерной сети одновременно с формированием графа атак на основе уязвимостей. Добавление узлов, соответствующих шаблонам атак, позволяет сопоставить им обнаруженные инциденты разведывательной деятельности или заметания следов, информация о которых поступает от модуля корреляции событий безопасности SIEM-системы. Сопоставление осуществляется компонентом программы для целей анализа защищенности и выбора контрмер на

основе информации об узле сети, на котором зафиксирован инцидент, и последствий обнаруженного инцидента.

3.3. Обработка циклов для формирования вероятностного графа атак. Для анализа защищенности компьютерной сети при мониторинге кибербезопасности был выбран подход на основе байесовского вывода. Для преобразования исходного графа в байесовский граф атак каждому узлу графа были поставлены в соответствие следующие параметры: состояние атакующего действия *St* (вводится для последующего учета динамического характера атак, *St*={True, False}, где True означает, что узел скомпрометирован), локальная вероятность успешной реализации атакующего действия (вероятность, что *St*=True, без учета вероятности компрометации связанных узлов), условная вероятность успешной реализации атакующего действия (вероятность того, что *St*=True в случае различных состояний связанных узлов), и полная вероятность успешной реализации атакующего действия (или того, что действие находится в состоянии *St* (Pr∈[0,1]), с учетом всех возможных состояний связанных узлов). Кроме того, для графа атак определены два типа отношений между связями: И — для перехода в скомпрометированное состояние необходимо, чтобы все узлы-предки, связанные данным отношением, были скомпрометированы (цепочка последовательно связанных узлов графа); ИЛИ — для перехода в скомпрометированное состояние необходимо, чтобы хотя бы один из узлов-предков, связанных данным отношением, был скомпрометирован (узлы графа, находящиеся на одном уровне) [33, 36].

Данная модель была выбрана, так как байесовские графы атак позволяют учитывать влияние событий на состояние системы и в соответствии с этим делать предположения о предыдущих шагах атаки и прогнозировать развитие атаки в будущем. Кроме того, они позволяют делать выводы об атаке на основе субъективных знаний при отсутствии статистических данных об успешном использовании уязвимостей сети. Байесовский вывод применим только для графов, не содержащих циклов, поэтому для его использования необходимо обработать циклы исходного графа.

Исходный алгоритм формирования графа атак включает два основных шага:

1 Определение возможных атакующих действий на объекты компьютерной сети.

2 Формирование связей между ними.

На шаге 2 могут образоваться циклы. Типы циклов, которые возникают в процессе формирования связей, представлены на рисунке 3.

Предлагаются следующие методы обработки циклов:

– циклы *типа 1* (рисунок 3*а*) — узлы графа атак находятся на одном уровне структуры графа) могут быть удалены, поскольку вероятность попасть в вершину напрямую, а не через соседнюю вершину графа, будет выше (то есть вероятность попасть в вершину «Атакующее действие 2» напрямую из вершины «Атакующее действие 1» выше, чем вероятность попасть в вершину «Атакующее действие 2» из вершины «Атакующее действие 1» через вершину «Атакующее действие 3», так как во втором случае добавляется дополнительный элемент цикла, а любой дополнительный элемент уменьшает вероятность);

– циклы *типа 2* (рисунок 3*б*) — целевой узел расположен на более высоком уровне структуры графа (уровни структуры графа могут отличаться от структуры анализируемой компьютерной сети), чем исходный узел, и связан с ним путем на графе) могут быть удалены, поскольку для атакующего не имеет смысла возвращаться назад;

– циклы *типа 3* (рисунок 3*в*) сохраняются, но их связи помечаются как несуществующие и обрабатываются отдельно при анализе графа атак (входящая связь обрабатывается с учетом предположения, что исходящая связь не существует, и наоборот).

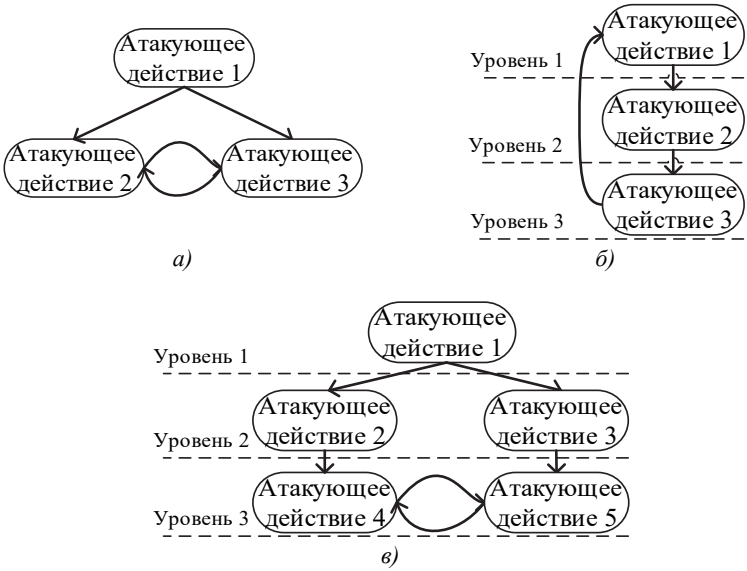


Рис. 3. Типы циклов для графа атакующих действий

На рисунке 4*а* приведен пример фрагмента компьютерной сети, граф атак для которой содержит цикл типа 1. Граф атак, содержащий цикл типа 1 представлен на рисунке 4*б*. На рисунке 4*в* представлен граф после обработки цикла.

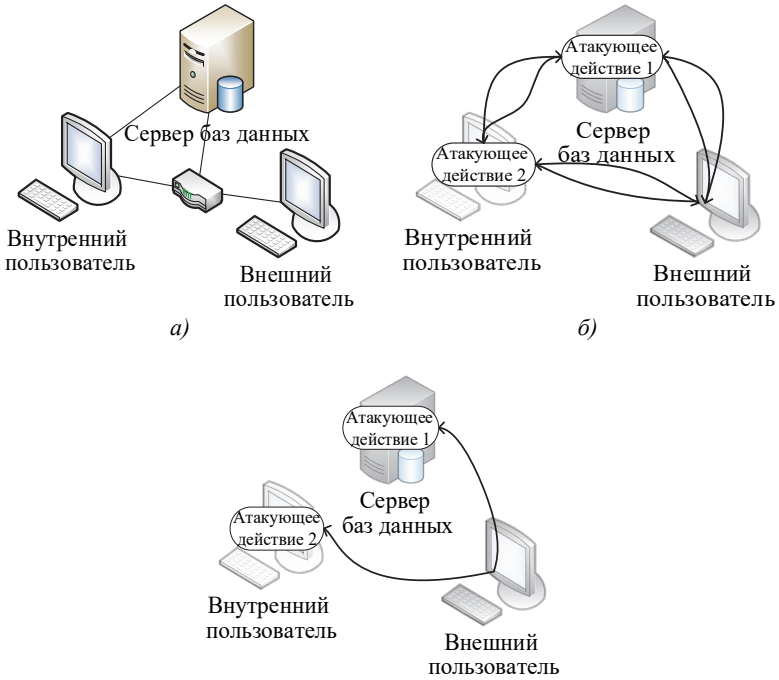


Рис. 4. Пример удаления циклов типа 1 для графа атакующих действий

На рисунке 5а приведен пример фрагмента компьютерной сети, граф атак для которой содержит цикл типа 2. Граф атак, содержащий цикл типа 2, представлен на рисунке 5*б*. При этом предполагается, что циклы типа 1 уже удалены. На рисунке 5*в* представлен граф после обработки цикла.

На рисунке 6а приведен пример фрагмента компьютерной сети, граф атак для которой содержит цикл типа 3. Граф атак, содержащий цикл типа 3, представлен на рисунке 6*б*. При этом предполагается, что циклы типа 1 и 2 уже удалены. На рисунке 6*в* представлен граф после обработки цикла.

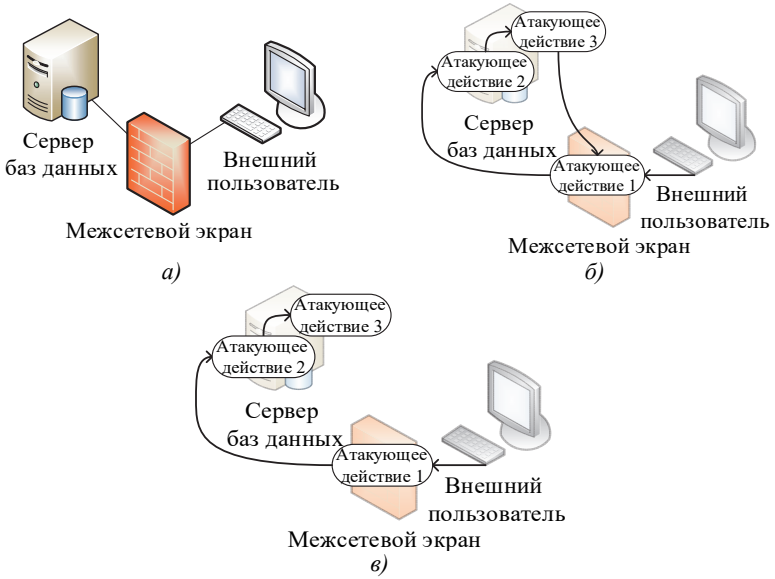


Рис. 5. Пример удаления циклов типа 2 для графа атакующих действий

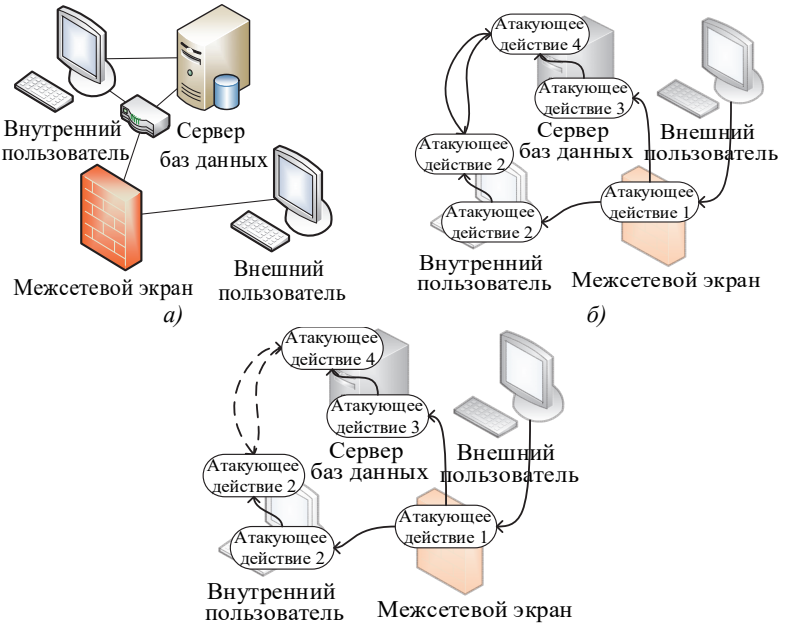


Рис. 6. Пример удаления циклов типа 3 для графа атакующих действий

Данное решение позволяет вычислять вероятность атаки для узлов и путей графа атак и выбирать контрмеры для узлов с неприемлемым уровнем риска, но вводит ограничения на идентификацию пути атакующего на графе атак после возникновения инцидента (некоторые пути графа атак, соответствующие выполненным атакующим действиям, могут быть удалены с графа атак).

3.4. Переопределение модели атак для локализации угроз. Еще одно ограничение исходного графа было выявлено при отображении различных инцидентов безопасности на граф атак для последующего выбора защитных мер. В рамках введенной модели атак под путем атаки понимается последовательность атакующих действий (узлов графа атак). Каждый путь соответствует угрозе для некоторого сетевого актива. Инциденты безопасности отображаются на граф согласно нанесенному ущербу и активу, на который они повлияли. Однако при текущей реализации, когда атакующие действия определены на основе групп (таблица 1 и таблица 2), не учитывающих различные виды ущерба, они могут отображаться сразу на несколько узлов.

С другой стороны, защитные действия ограничиваются добавлением или удалением путей на граф атак (например, закрытие доступа к хосту, удаление уязвимостей) и не учитывают тип угроз. При этом не очевидно, каким образом защитные меры, действующие против нарушения конфиденциальности и целостности, влияют на граф атак.

Поэтому предлагается ввести новое разделение уязвимостей на группы по типам угроз для точного отображения инцидента на узлы графа атак в зависимости от ущерба, нанесенного атакой.

Согласно системе CVSS, выделяются показатели ущерба конфиденциальности (C), целостности (I) и доступности (A). Можно выделить следующие сочетания данных показателей: CIA, C, I, A, CI, CA, IA, и none (нет ущерба). В соответствии с этим каждая группа в таблицах 1 и 2 будет поделена на 8 подгрупп по полю *CIA*.

Для каждого вида ущерба должны быть определены соответствующие защитные меры. Например, в таблице 3 выделены угрозы и возможные контрмеры согласно стандарту ГОСТ Р ИСО/МЭК 27005-2010 [23, 45] (где C обозначает ущерб конфиденциальности, I — ущерб целостности, A — ущерб доступности).

Угрозы и защитные меры определяются индивидуально для каждой сети. Каждый класс защитных мер в таблице 3 включает конкретные меры, имеющиеся в наличии в анализируемой системе для статического режима (т.е. режима проектирования) и динамического режима (т.е. режима эксплуатации) [23].

Таблица 3. Классы угроз и защитные меры

Угрозы	Свойство безопасности	Контрмеры							
		Предохранение от вредоносного кода	Идентификация и аутентификация	Логическое управление и аудит доступа	Управление безопасностью сети	Криптография	Обнаружение и предотвращение вторжений	Резервные копии	Управление персоналом
Злонамеренный код	C	SD					D		
	I	SD					D	SD	
	A	SD					D		
Подмена личности пользователя	C	SD	SD	SD	SD	S			
	I	SD	SD	SD	SD	S		SD	
	A	SD	SD	SD	SD	S		SD	
Ложная маршрутизация/перенаправление сообщений	C				SD	S			
	I				SD	S			
	A				SD	S			
Несанкционированный доступ к компьютерам, данным, сервисам и приложениям	C		SD	SD	SD	S			
	I		SD	SD	SD	S		SD	
	A		SD	SD	SD	S			
Разрушительная атака									
	A		SD	SD				SD	S
Неправильное использование ресурсов									
	A		SD	SD	SD				S
Перегрузка трафика									
	A				SD			SD	

Например, мера «Идентификация и аутентификация» может быть реализована с использованием программного токена в статическом режиме работы системы (обозначение S в таблице 3). В динамическом режиме работы системы (обозначение D в таблице 3), при поступлении информации об инциденте безопасности, соответствующем несанкционированному доступу, программный токен используется для активации многофакторной аутентификации. Контрмеры отображаются на соответствующие свойства безопасности, что позволяет отобразить их на узлы графа атак.

3.5. Оценка защищенности с использованием модифицированного графа атак. Мониторинг кибербезопасности с использованием описанного графа происходит на основе ряда показателей защищенности [22, 23]. Предложенные модификации методики формирования графа атак влияют на процесс оценки защищенности. Переход на CVSS версии 3.0 привел к модификации уравнений вычисления показателей. Обработка циклов графа позволила применить байесовский метод для вычисления показателей вероятности успешной реализации атаки. Однако при этом необходимо выполнить дополнительные действия при вычислении вероятностей, что повысило сложность обработки графа атак. Новое определение групп уязвимостей не повлияло на уравнения для вычисления показателей.

Ниже рассмотрены некоторые показатели защищенности, используемые для мониторинга кибербезопасности и выбора контрмер, на которые повлияли предложенные в предыдущих разделах изменения модели атак, в том числе:

– сложность атакующего действия (вычисляется на основе индекса CVSS *AccessComplexity*);

– ущерб для свойств безопасности от атакующего действия (вычисляется на основе индексов CVSS ущерб конфиденциальности, целостности и доступности);

– ущерб от атакующего действия (с учетом критичности активов);

– вероятность атаки (вычисляется на основе индексов CVSS *AccessVector*, *Authentication* и *AccessComplexity*).

При мониторинге показатель сложности атакующего действия *a* *AttackComplexity*(*a*) используется для определения уровня навыков атакующего. Более точное определение данного показателя ведет в свою очередь к более корректным выводам о навыках атакующего. Переход к CVSS версии 3.0 привел к изменению формулы для вычисления данного показателя. Для CVSS версии 2.0:

$$AttackComplexity(a) = AccessComplexity,$$

где *AccessComplexity* — максимальная сложность эксплуатации уязвимостей группы, соответствующей атакующему действию *a*, согласно CVSS версии 2.0, $AccessComplexity \in \{\text{High}, \text{Medium}, \text{Low}\}$.

Для CVSS версии 3.0 уравнение меняется следующим образом:

$$AttackComplexity(a) = AttackComplexity,$$

где *AttackComplexity* — максимальная сложность атаки для уязвимостей группы, соответствующей атакующему действию *a*, согласно CVSS версии 3.0, $AttackComplexity \in \{\text{High}, \text{Low}\}$.

Ущерб от атакующего действия *a* используется при расчете уровней риска узлов графа атак для выявления слабых мест сети и для последующего выбора защитных мер. Точное определение данного показателя позволяет корректно определить уровень возможных потерь организации в случае успешной атаки. Переход к CVSS версии 3.0 привел к изменению формулы для вычисления данного показателя. При использовании CVSS версии 2.0 показатель определялся следующим образом:

$$AttackImpact(a) = Criticality \times [CI \quad II \quad AI],$$

где *Criticality* — критичность актива, против которого направлено атакующее действие *a*, *CI*, *II*, *AI* — ущерб конфиденциальности, целостности и доступности согласно CVSS версии 2.0 соответственно.

При переходе на CVSS версии 3.0 показатели ущерба CVSS версии 2.0 заменяются на соответствующие индексы CVSS версии 3.0. Качественные значения индексов меняются с {None, Partial, Complete} на {None, Low, High}. Новая шкала интуитивно понятнее. Количественные значения индексов CVSS версии 3.0 немного ниже, чем значения индексов CVSS версии 2.0. Это незначительно уменьшает ущерб, создаваемый эксплуатацией уязвимостей.

Показатель вероятности атаки при мониторинге используется, с одной стороны, для расчета уровня риска, для выявления слабых мест сети и для последующего выбора защитных мер, а с другой стороны, для локализации пути атаки, ее источника и целей. Соответственно, более точное определение данного показателя ведет к более точному определению характеристик атаки и эффективному проактивному выбору защитных мер.

Для вычисления вероятности успешного выполнения атакующего действия (полных вероятностей для узлов графа атак) используется формула определения полной вероятности. При обходе графа в случае наличия циклов третьего типа (рисунок 3в), полная вероятность рассчитывается без учета исходящей дуги (образующей цикл), но с учетом входящей дуги. При этом вероятность связанного узла рассчитывается без учета исходящей дуги, но с учетом входящей дуги.

Для определения дискретных локальных распределений условных вероятностей *Pc* используются формулы из [33]. В случае связей типа «И» между узлами предками (для успешной компрометации узла потомка *S* необходимо, чтобы все узлы предки *Pa*(*S*) были скомпрометированы) применяется уравнение:

$$Pc(S \mid Pa(S)) = \begin{cases} 0, & \exists S \in Pa(S) \mid S = 0 \\ p(S), & \text{иначе} \end{cases}, \quad (1)$$

где $p(S)$ — локальная вероятность, соответствующая узлу *S*.

В случае связей типа «ИЛИ» между узлами предками (для успешной компрометации узла потомка необходимо, чтобы хотя бы один узел предок был скомпрометирован) применяется уравнение:

$$Pc(S \mid Pa(S)) = \begin{cases} 0, & \forall S \in Pa(S) \mid S = 0 \\ p(S), & \text{иначе} \end{cases}. \quad (2)$$

Локальную вероятность $p(S)$ атакующего действия *a*, соответствующую узлу *S*, предлагается определять на основе индексов CVSS. Поэтому формулы вычисления локальной вероятности меняются при переходе на CVSS версии 3.0. При использовании CVSS версии 2.0 этот показатель определялся следующим образом:

$$p(S) = \begin{cases} 2 \times AV \times AC \times Au, & \text{если } S \in S_r \\ 2 \times AC \times Au, & \text{иначе} \end{cases}, \quad (3)$$

где S_r — множество корневых (входных) узлов графа; *AV* — показатель, характеризующий доступ, необходимый для эксплуатации уязвимости по CVSS версии 2.0; *AC* — сложность доступа к уязвимости по CVSS версии 2.0; *Au* — аутентификация, требуемая для эксплуатации уязвимости по CVSS версии 2.0 [22, 23].

При использовании CVSS версии 3.0 указанный показатель предлагается определять на основе CVSS уравнения для показателя эксплуатации уязвимости следующим образом:

$$p(S) = \begin{cases} (8.22 \times AV \times AC \times PR \times UI - 0.2) \times 2.7 / 10, & \text{если } S \in S_r \\ (8.22 \times AC \times PR \times UI - 0.2) \times 2.7 / 10, & \text{иначе} \end{cases}, \quad (4)$$

где коэффициенты введены для нормализации значения вероятности от 0 до 1; *AV* — показатель, характеризующий доступ, необходимый для эксплуатации уязвимости (*AttackVector* по CVSS версии 3.0); *AC* — показатель, характеризующий сложность эксплуатации уязвимости (*AttackComplexity* по CVSS версии 3.0); *PR* — привилегии,

требуемые для эксплуатации уязвимости по CVSS версии 3.0; и *UI* — показатель, определяющий, требуется ли взаимодействие с пользователем для эксплуатации уязвимости по CVSS версии 3.0.

В следующих разделах анализируются результаты применения заданных уравнений.

4. Пример применения, эксперименты и дискуссия. Программный прототип, реализующий предложенный подход к мониторингу кибербезопасности и выбору контрмер, разработан с использованием Java на Microsoft Windows Intel Core i7 ЦПУ и 12 Гб ОЗУ [19, 20, 22, 23].

Для проведения экспериментов прототип был модифицирован с учетом предложенных изменений модели атак и методики анализа. На примере фрагмента тестовой сети рассмотрим результаты предложенных изменений (рисунок 7).

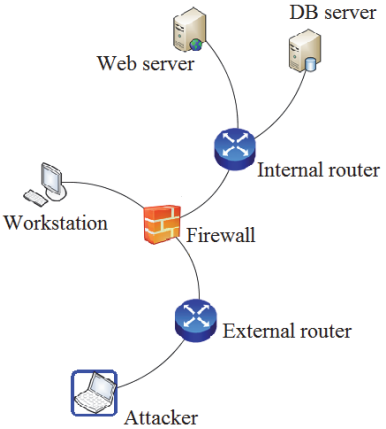


Рис. 7. Фрагмент тестовой сети

Фрагмент тестовой сети содержит:

– веб сервер *Web server* (с Windows Server 2008 R2 SP1 (64 бит), JBoss AS 5.0.1, ApacheStruts2 framework);

– сервер баз данных *DB server* (с Windows Server 2008 R2 (64 бит), Microsoft SQL Server 2008 R2 (64 бит), CA Spectrum 9.2, EMC Unisphere for VMAX 8.1);

– межсетевой экран *Firewall* (с Novell SUSE Linux Enterprise Server 11.0 SP3 Long Term Service Pack Support, Netfilter);

– рабочую станцию *Workstation* (с Microsoft Windows 7 64-bit, Apple ITunes 9.0.3, Microsoft Office 2007 SP1, Microsoft Internet Explorer 7).

На рисунке 7 представлены связи между объектами сети (на сетевом уровне).

Итоговый ациклический граф, сгенерированный с использованием CVSS версии 2.0, представлен на рисунке 8.

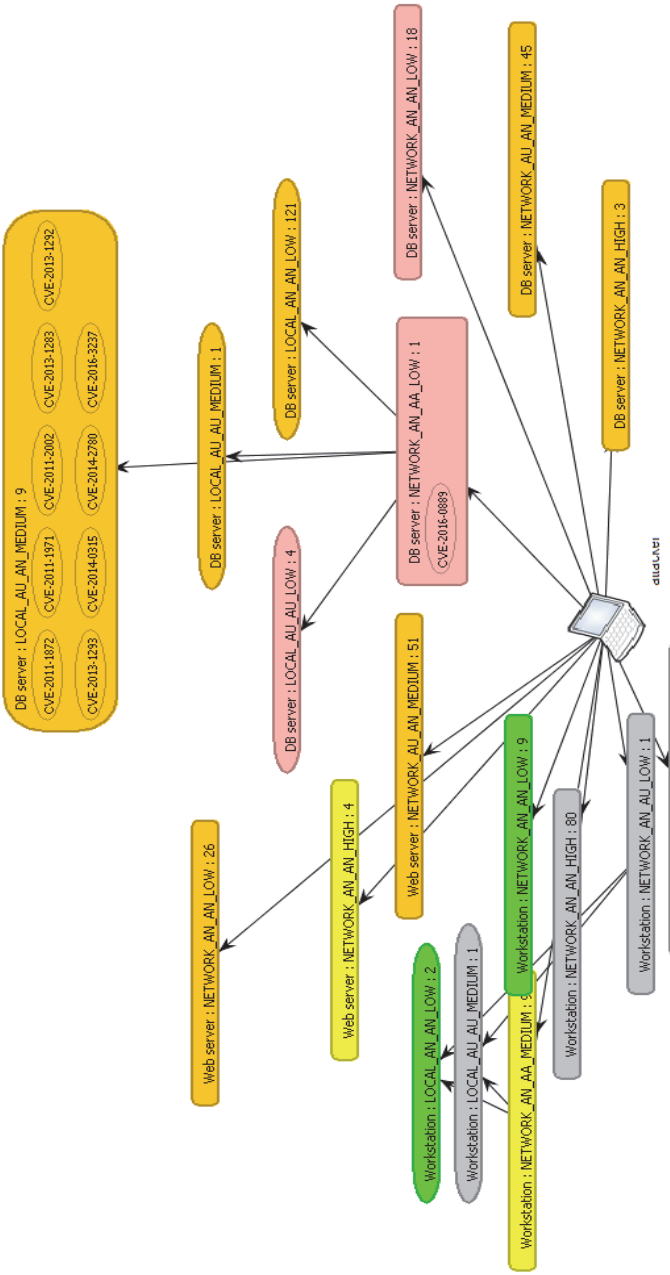


Рис. 8. Исходный граф атак для фрагмента тестовой сети

Узлы графа в рамках программного прототипа раскрашены в соответствии с уровнем риска:

- зеленый цвет — низкий уровень риска;
- желтый цвет — средний уровень риска;
- оранжевый цвет — высокий уровень риска;
- красный цвет — критичный уровень риска.

Каждый узел, соответствующий группе уязвимостей CVSS версии 2.0 обозначается с использованием показателей CVSS версии 2.0 в следующем формате:

“Host_name : AccessVector_Authentication_ GainedPrivileges_
AccessComplexity : number_of_vulnerabilities”,

где Host_name — имя хоста; AccessVector — показатель, характеризующий доступ, необходимый для эксплуатации уязвимости по CVSS версии 2.0; Authentication — аутентификация, требуемая для эксплуатации уязвимости по CVSS версии 2.0; GainedPrivileges — привилегии, полученные в результате эксплуатации уязвимости; AccessComplexity — показатель, характеризующий сложность эксплуатации уязвимости по CVSS версии 2.0; number_of_vulnerabilities — количество уязвимостей в группе.

Каждый узел содержит уязвимости соответствующей группы.

Граф атак с учетом модификаций в результате применения CVSS версии 3.0 представлен на рисунке 9.

Уязвимости, перемещенные в другие группы, определенные с использованием CVSS версии 3.0, выделены красными прямоугольниками. Красные стрелки, идущие от прямоугольников, показывают перемещение уязвимостей в новые группы. Каждый узел, соответствующей группе, выделенной с использованием CVSS версии 3.0, обозначается следующим образом:

“Host_name : AttackVector_PrivilegesRequired_
GainedPrivileges_AttackComplexity”,

где Host_name — имя хоста; AttackVector — показатель, характеризующий доступ, необходимый для эксплуатации уязвимости по CVSS версии 3.0; PrivilegesRequired — привилегии, требуемые для эксплуатации уязвимости по CVSS версии 3.0; GainedPrivileges — привилегии, полученные в результате эксплуатации уязвимости; AttackComplexity — сложность эксплуатации уязвимости по CVSS версии 3.0.

Поскольку в настоящий момент открытые базы уязвимостей содержат оценки CVSS версии 3.0 не для всех уязвимостей, полностью перейти на CVSS версии 3.0 невозможно. Поэтому в рамках прототипа две системы оценки используются совместно.

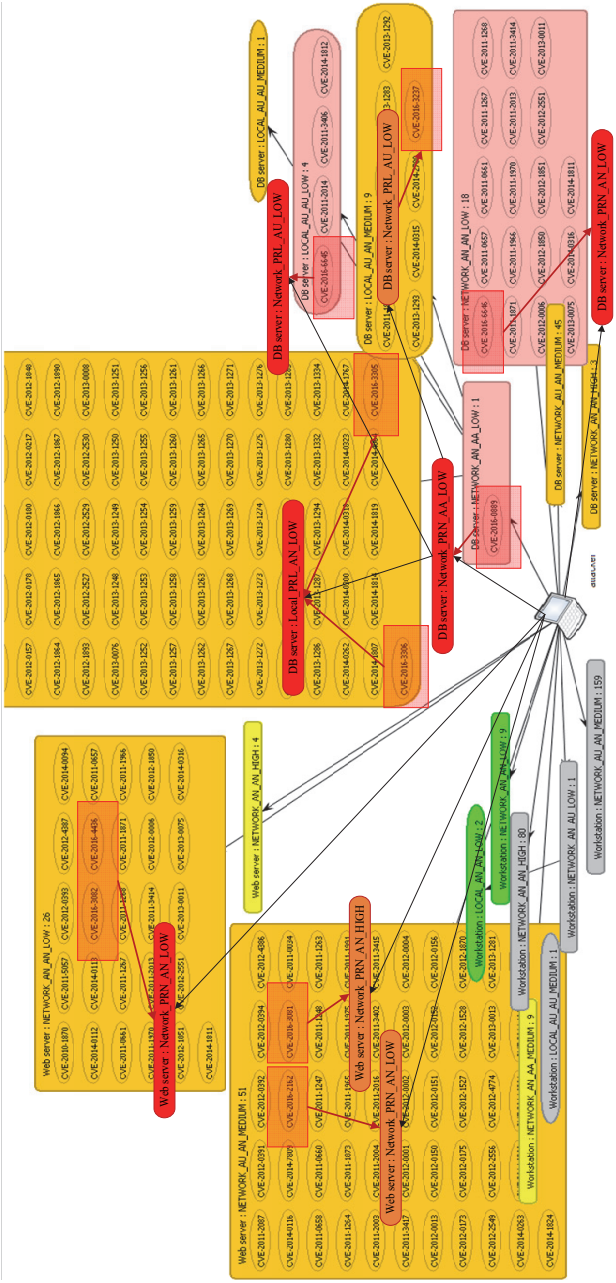


Рис. 9. Граф атак с учетом модификаций в результате применения CVSS версии 3.0

Из рисунка 9 видно, что структура графа в целом совпадает со структурой исходного графа (рисунок 8), за исключением новых узлов, соответствующих уязвимостям, оцененным с использованием CVSS версии 3.0. Большинство групп CVSS версии 3.0 имеют тот же уровень риска, что и соответствующие группы CVSS версии 2.0 (из которых уязвимости были перемещены). Но для хоста «Web server» есть две уязвимости, изменившие свои оценки риска с «Высокий» на «Критичный», то есть общий уровень риска для данного хоста вырос. Таким образом, новые оценки могут значительно изменить распределение уровней риска в компьютерных сетях. Для анализа возможных изменений продолжаются эксперименты с различными сетевыми топологиями. Другие эксперименты были проведены для оценки выбора контрмер с использованием новых групп, сгенерированных с учетом различных видов ущерба.

Все еще остается ряд аспектов, которые необходимо исследовать, например: сложность анализа графа атак; некоторые показатели CVSS, которые все еще не учитываются при оценке (“User Interaction” и “Scope”); дальнейшее развитие модели атак с применением шаблонов атак и показателей CAPEC. Их планируется исследовать в будущем.

5. Заключение. В работе рассмотрены вопросы модификации процесса генерации модели атак и методики его анализа для более адекватного мониторинга кибербезопасности и выбора защитных мер. Выделены существующие ограничения, в том числе неточность графа атакующих действий, обусловленная неоднозначностью индексов CVSS, лежащих в его основе, отсутствием учета атак, не использующих уязвимости, пренебрежением циклическими связями на графе, а также допущения при определении ущерба от атак вследствие неоднозначного определения области воздействия атаки. Приведены предложения по устранению указанных ограничений за счет использования новой спецификации CVSS версии 3.0, использования шаблонов атак CAPEC и обработки циклических связей графа. Проведен анализ влияния введенных изменений на результаты анализа защищенности, показавший изменение распределения уровней риска в тестовой компьютерной сети.

Изменения реализованы в рамках ранее разработанного программного прототипа для проведения экспериментов. На примере показано влияние предложенных изменений на результаты оценки защищенности. Кратко описаны проведенные эксперименты, показавшие качественное улучшение процессов оценки защищенности и выбора контрмер с точки зрения корректности прогнозирования атак и рациональности выбора контрмер.

В будущем планируется дополнительно провести ряд экспериментов для тщательной проверки адекватности предложенной модели и получения количественных оценок улучшения процессов оценки защищенности и выбора контрмер, развить модель атак с учетом сложности анализа графа атак, проанализировать еще неучтенные показатели системы CVSS, а также показатели и шаблоны атак CAPEC.

Литература

1. *Котенко И.В., Саенко И.Б.* Построение системы интеллектуальных сервисов для защиты информации в условиях кибернетического противоборства // Труды СПИИРАН. 2012. Вып. 3(22). С. 84–100.
2. *Котенко И.В., Саенко И.Б.* Архитектура системы интеллектуальных сервисов защиты информации в критически важных инфраструктурах // Труды СПИИРАН. 2013. Вып. 1(24). С. 21–40.
3. *Artz M.* NetSPA, a network security planning architecture. Master's thesis // Massachusetts Institute of Technology. 2002. 96 p.
4. *Lippmann R.P. et al.* Validating and restoring defense in depth using attack graphs // Proceedings of the 2007 IEEE Military Communications Conference. 2006. pp. 1–10.
5. *Ingols K., Lippmann R., Piwowarski K.* Practical attack graph generation for network defense // Proceedings of the 22nd Annual Conference on the Computer Security Applications. 2006. pp. 121–130.
6. *Singhal A., Ou X.* Security risk analysis of enterprise networks using probabilistic attack graphs // Network Security Metrics. 2017. pp. 53–73.
7. *Abraham S., Nair S.* A predictive framework for cyber security analytics using attack graphs // International Journal of Computer Networks & Communications (IJCNC). 2015. vol. 7. no.1. pp. 1–17.
8. *Janse van Rensburg A., Nurse J.R.C., Goldsmith M.* Attacker-Parametrised Attack Graphs // Proceedings of the Tenth International Conference on Emerging Security Information, Systems and Technologies. 2016. pp. 316–319.
9. *Kordy B., Pièvre-Cambacédès L., Schweitzer P.* DAG-based attack and defense modeling: don't miss the forest for the attack trees // Computer Science Review. 2014. vol. 13–14. pp. 1–38. URL: http: // www.sciencedirect.com/science/article/pii/S1574013714000100 (дата обращения: 17.07.2017).
10. *Shandilya V., Simmons C.B., Shiva S.* Use of attack graphs in security systems // Journal of Computer Networks and Communications. 2014. vol. 2014. 13 p. URL: http: // dx.doi.org/10.1155/2014/818957 (дата обращения: 17.07.2017).
11. *Muñoz-González L., Sgandurra D., Barrère M., Lupu E.C.* Exact inference techniques for the analysis of Bayesian attack graphs // IEEE Transactions on Dependable and Secure Computing. 2017. pp. 1–14.
12. *Noel S., Jajodia S.* Metrics suite for network attack graph analytics // Proceedings of the 9th Cyber and Information Security Research Conference. 2014. pp. 5–8.
13. *Alhomidi M., Reed M.* Attack graph-based risk assessment and optimisation approach // International Journal of Network Security & Its Applications (IJNSA). 2014. vol. 6. no. 3. pp. 31–43.
14. *Durkota K., Lisy V., Bošansky B., Kiekintveld C.* Optimal network security hardening using attack graph games // Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015. pp. 526–532.
15. *Sembiring J., Ramadhan M., Gondokaryono Y.S., Arman A.A.* Network security risk analysis using improved MulVAL Bayesian attack graphs // International Journal on Electrical Engineering and Informatics. 2015. vol. 7. no. 4. pp. 735–753.

16. *Muñoz-Gonzalez L., Sgandurra D., Paudice A., Lupu E.C.* Efficient attack graph analysis through approximate inference // ACM Transactions on Privacy and Security. 2017. vol. 20. no. 3. 30 p.
17. *Bhattacharya P. Ghosh S.K.* Analytical framework for measuring network security using exploit dependency graph // IET Information Security. 2012. vol. 6. no. 4. pp. 264–270.
18. *Almohri H.M.J., Watson L.T., Yao D., Ou X.* Security optimization of dynamic networks with probabilistic graph modeling and linear programming // IEEE Transactions on Dependable and Secure Computing. 2016. vol. 13. no. 4. pp. 474–487.
19. *Kotenko I., Stepashkin M.* Attack graph based evaluation of network security // Proceedings of the Communications and Multimedia Security (CMS 2006). 2006. LNCS 4237. pp. 216–227.
20. *Kotenko I., Chechulin A.* Attack Modeling and Security Evaluation in SIEM Systems // International Transactions on Systems Science and Applications. 2012. vol. 8. pp. 129–147.
21. *Mell P., Scarforne K., Romanosky S.* A Complete Guide to the Common Vulnerability Scoring System (CVSS) Version 2.0. 2007. URL: https://www.first.org/cvss/v2/guide (дата обращения: 17.07.2017).
22. *Kotenko I., Doynikova E.* Dynamical calculation of security metrics for countermeasure selection in computer networks // Proceedings of the 24th Euromicro International Conference on Parallel, Distributed and network-based Processing. 2016. pp. 558–565.
23. *Doynikova E., Kotenko I.* Countermeasure selection based on the attack and service dependency graphs for security incident management // Proceedings of the 10th International Conference on Risks and Security of Internet and Systems. 2016. LNCS 9572. pp. 107–124.
24. FIRST Org. Inc, Common Vulnerability Scoring System v3.0: Specification Document. 2015. URL: https://www.first.org/cvss/specification-document (дата обращения: 17.07.2017).
25. *Barnum S.* Common Attack Pattern Enumeration and Classification (CAPEC). Schema Description. 2008. 26 p.
26. Positive Technologies web site. URL: https:/www.ptsecurity.com/ww-en (дата обращения: 17.07.2017).
27. *Lorenzo J.M.* Alienvault users manual. Version 1.0. Alienvault LC. 2011. 225 p.
28. CORDIS website. URL: http://cordis.europa.eu/project/rcn/95310_en.html (дата обращения: 17.07.2017).
29. *Man D. et al.* A quantitative evaluation model for network security // Proceedings of the 2007 Intern. Conference on Computational Intelligence and Security. 2007. pp. 773–777.
30. *Wu Y.-S. et al.* Automated adaptive intrusion containment in systems of interacting services // Computer Networks: The International Journal of Computer and Telecommunications Networking. 2007. vol. 51. pp. 1334–1360.
31. *Stakhanova N., Basu S., Wong J.* A cost-sensitive model for preemptive intrusion response systems // Proceedings of the 21st International Conference on Advanced Networking and Applications. 2007. pp. 1–8.
32. *Liu Y., Man H.* Network vulnerability assessment using Bayesian networks // SPIE. 2005. vol. 5812. pp. 61–71.
33. *Frigault M., Wang L., Singhal A., Jajodia S.* Measuring network security using dynamic Bayesian network // Proceedings of the ACM Workshop on Quality of Protection. 2008. pp. 23–30.
34. *Dantu R., Kolan P., Cangussu J.* Network risk management using attacker profiling // Security and Communication Networks. 2009. vol. 2. no. 1. pp. 83–96.
35. *Wang L. et al.* An attack graph-based probabilistic security metric // Proceedings of the 22nd annual IFIP WG 11.3 working conference on Data and Applications Security. 2008. pp. 283–296.
36. *Poolsappasit N., Dewri R., Ray I.* Dynamic security risk management using Bayesian attack graphs // Proceedings of the IEEE Transactions on Dependable and Security Computing. 2012. vol. 9. no. 1. pp. 61–74.

37. *Dacier M., Deswarte Y., Kaâniche M.* Quantitative assessment of operational security: Models and tools // Information Systems Security. 1996. pp. 179–86.
38. CyVision website. CAULDRON tool. URL: https://www.benvenisti.net/cauldron/ (дата обращения: 17.07.2017).
39. SecurITree. Amenaza Technologies Limited. URL: http://www.amenaza.com (дата обращения: 17.07.2017).
40. Common Platform Enumeration (CPE). NVD website. URL: https://nvd.nist.gov/cpe.cfm (дата обращения: 17.07.2017).
41. Common Configuration Enumeration (CCE). NVD website. URL: https://nvd.nist.gov/cce/index.cfm (дата обращения: 17.07.2017).
42. NVD website. URL: https://nvd.nist.gov (дата обращения: 17.07.2017).
43. Common Weakness Enumeration (CWE). MITRE website. URL: https://cwe.mitre.org/ (дата обращения: 15.08.2017).
44. *Wu Y., Yesha Y., Bojanova I.* They Know Your Weaknesses – Do You?: Reintroducing Common Weakness Enumeration // CrossTalk: The Journal of Defense Software Engineering. 2016. vol. 29. no. 3. pp. 19–24.
45. ГОСТ Р ИСО/МЭК 27005-2010. Информационная технология. Методы и средства обеспечения безопасности. Менеджмент риска информационной безопасности // М.: Стандартинформ. 2011. 47 с.

Дойникова Елена Владимировна — научный сотрудник лаборатории проблем компьютерной безопасности, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), научный сотрудник международной лаборатории информационной безопасности киберфизических систем, ФГАОУ ВО "Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики" (Университет ИТМО). Область научных интересов: безопасность компьютерных сетей, методы анализа рисков компьютерных сетей, управление информационными рисками. Число научных публикаций — 71. elenadoynikova@mail.ru; 14-я линия В.О., 39, Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-7181, Факс: +7(812)328-4450.

Котенко Игорь Витальевич — д-р техн. наук, профессор, заведующий лабораторией проблем компьютерной безопасности, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), руководитель международной лаборатории информационной безопасности киберфизических систем, ФГАОУ ВО "Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики" (Университет ИТМО). Область научных интересов: безопасность компьютерных сетей, в том числе управление политиками безопасности, разграничение доступа, аутентификация, анализ защищенности, обнаружение компьютерных атак, межсетевые экраны, защита от вирусов и сетевых червей, анализ и верификация протоколов безопасности и систем защиты информации, защита программного обеспечения от взлома и управление цифровыми правами, технологии моделирования и визуализации для противодействия кибер–терроризму. Число научных публикаций — 500. ivkote@comsec.spb.ru, http://www.comsec.spb.ru; 14-я линия В.О., 39, Санкт-Петербург, 199178; р.т.: +7-(812)328-7181, Факс: +7(812)328–4450.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проекты 16-37-00338, 16-29-09482 и 18-07-01488), гранта Президента РФ № МК-314.2017.9, стипендии Президента РФ № СП-751.2018.5, при частичной поддержке бюджетных тем № 0073-2015-0004 и 0073-2015-0007, а также при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

DOI 10.15622/sp.57.9

E.V. DOYNIKOVA, I.V. KOTENKO

IMPROVEMENT OF ATTACK GRAPHS FOR CYBERSECURITY MONITORING: HANDLING OF INACCURACIES, PROCESSING OF CYCLES, MAPPING OF INCIDENTS AND AUTOMATIC COUNTERMEASURE SELECTION

Doynikova E.V., Kotenko I.V. **Improvement of Attack Graphs for Cybersecurity Monitoring: Handling of Inaccuracies, Processing of Cycles, Mapping of Incidents and Automatic Countermeasure Selection.**

Abstract. Both timely and adequate response on the computer security incidents and organization losses from the computer attacks depend on the accuracy of situation recognition under the cybersecurity monitoring. The paper is devoted to the enhancement of the attack models in the form of attack graphs for the cybersecurity monitoring tasks. A number of important issues related to the application of attack graphs and their solutions are considered. They include inaccuracies in the definition of the pre- and post-conditions of attack actions, the processing of attack graph cycles for the application of Bayesian inference for the attack graph analysis, the mapping of security incidents on an attack graph, the automatic countermeasure selection in case of a high security risk level. The paper demonstrates a software prototype of the security monitoring system component which was earlier implemented and modified considering the suggested enhancements. The results of experiments are described. The influence of the modifications on the cybersecurity monitoring results is shown on a case study.

Keywords: attack graph, attack probability, cybersecurity monitoring, computer networks, security assessment, security metrics, attack response, vulnerability assessment.

Doynikova Elena Vladimirovna — researcher of computer security problems laboratory, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), researcher of information security of cyber-physical systems international laboratory , ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics). Research interests: computer network security, risk analysis methods for computer networks, information security risk management. The number of publications — 71. elenadoynikova@mail.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-7181, Fax: +7(812)328-4450.

Kotenko Igor Vitalievich — Ph.D., Dr. Sci., professor, head of computer security problems laboratory, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), head of information security of cyber-physical systems international laboratory , ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics). Research interests: computer network security, including security policy management, access control, authentication, network security analysis, intrusion detection, firewalls, deception systems, malware protection, verification of security systems, digital right management, modeling, simulation and visualization technologies for counteraction to cyber terrorism. The number of publications — 500. ivkote@comsec.spb.ru, http://www.comsec.spb.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7-(812)328-7181, Fax: +7(812)328–4450.

Acknowledgements. This research is supported by RFBR (projects No. 16-37-00338, 16-29-09482 and 18-07-01488), Grants of the President of the Russian Federation No. MK-314.2017.9, SP-751.2018.5, by the budget (projects No. 0073-2015-0004 and 0073-2015-0007), and by Government of the Russian Federation, Grant 074-U01.

References

1. Kotenko I.V., Saenko I.B. [Developing the system of intelligent services to protect information in cyber warfare]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. vol. 3(22). pp. 84–100. (In Russ.).

2. Kotenko I.V., Saenko I.B. [Architecture of the system of intelligent services to protect information in cyber warfare]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2013. vol. 1(24). pp. 21–40. (In Russ.).
3. Artz M. NetSPA, a network security planning architecture. Master's thesis. Massachusetts Institute of Technology. 2002. 96 p.
4. Lippmann R.P. et al. Validating and restoring defense in depth using attack graphs. Proceedings of the 2007 IEEE Military Communications Conference. 2006. pp. 1–10.
5. Ingols K., Lippmann R., Piwowarski K. Practical attack graph generation for network defense. Proceedings of the 22nd Annual Conference on the Computer Security Applications. 2006. pp. 121–130.
6. Singhal A., Ou X. Security risk analysis of enterprise networks using probabilistic attack graphs. Network Security Metrics. 2017. pp. 53–73.
7. Abraham S., Nair S. A predictive framework for cyber security analytics using attack graphs. *International Journal of Computer Networks & Communications (IJCNC)*. 2015. vol. 7. no.1. pp. 1–17.
8. Janse van Rensburg A., Nurse J.R.C., Goldsmith M. Attacker-Parametrised Attack Graphs. Proceedings of the Tenth International Conference on Emerging Security Information, Systems and Technologies. 2016. pp. 316–319.
9. Kordy B., Pièrre-Cambacédès L., Schweitzer P. DAG-based attack and defense modeling: don't miss the forest for the attack trees. Computer Science Review. 2014. vol. 13–14. pp. 1–38. Available at: http: // www.sciencedirect.com/science/article/pii/S1574013714000100 (access: 17.07.2017).
10. Shandilya V., Simmons C.B., Shiva S. Use of attack graphs in security systems. Journal of Computer Networks and Communications. 2014. vol. 2014. 13 p. Available at: http://dx.doi.org/10.1155/2014/818957 (access: 17.07.2017).
11. Muñoz-González L., Sgandurra D., Barrère M., Lupu E.C. Exact inference techniques for the analysis of Bayesian attack graphs. *IEEE Transactions on Dependable and Secure Computing*. 2017. pp. 1–14.
12. Noel S., Jajodia S. Metrics suite for network attack graph analytics. Proceedings of the 9th Cyber and Information Security Research Conference. 2014. pp. 5–8.
13. Alhomidi M., Reed M. Attack graph-based risk assessment and optimisation approach. *International Journal of Network Security & Its Applications (IJNSA)*. 2014. vol. 6. no. 3. pp. 31–43.
14. Durkota K., Lisy V., Bošansky B., Kiekintveld C. Optimal network security hardening using attack graph games. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015. pp. 526–532.
15. Sembiring J., Ramadhan M., Gondokaryono Y.S., Arman A.A. Network security risk analysis using improved MulVAL Bayesian attack graphs. *International Journal on Electrical Engineering and Informatics*. 2015. vol. 7. no. 4. pp. 735–753.
16. Muñoz-Gonzalez L., Sgandurra D., Paudice A., Lupu E.C. Efficient attack graph analysis through approximate inference. *ACM Transactions on Privacy and Security*. 2017. vol. 20. no. 3. 30 p.
17. Bhattacharya P. Ghosh S.K. Analytical framework for measuring network security using exploit dependency graph. IET Information Security. 2012. vol. 6. no. 4. pp. 264–270.
18. Almohri H.M.J., Watson L.T., Yao D., Ou X. Security optimization of dynamic networks with probabilistic graph modeling and linear programming. *IEEE Transactions on Dependable and Secure Computing*. 2016. vol. 13. no. 4. pp. 474–487.
19. Kotenko I., Stepashkin M. Attack graph based evaluation of network security. Proceedings of the Communications and Multimedia Security (CMS 2006). 2006. LNCS 4237. pp. 216–227
20. Kotenko I., Chechulin A. Attack Modeling and Security Evaluation in SIEM Systems. *International Transactions on Systems Science and Applications*. 2012. vol. 8. pp. 129–147.
21. Mell P., Scarforne K., Romanosky S. A Complete Guide to the Common Vulnerability Scoring System (CVSS) Version 2.0. 2007. Available at: https://www.first.org/cvss/v2/guide (accessed: 17.07.2017).

22. Kotenko I., Doynikova E. Dynamical calculation of security metrics for countermeasure selection in computer networks. Proceedings of the 24th Euromicro International Conference on Parallel, Distributed and network-based Processing. 2016. pp. 558–565.
23. Doynikova E., Kotenko I. Countermeasure selection based on the attack and service dependency graphs for security incident management. Proceedings of the 10th International Conference on Risks and Security of Internet and Systems. 2016. LNCS 9572. pp. 107–124.
24. FIRST Org. Inc, Common Vulnerability Scoring System v3.0: Specification Document. 2015. Available at: https://www.first.org/cvss/specification-document (accessed: 17.07.2017).
25. Barnum S. Common Attack Pattern Enumeration and Classification (CAPEC). Schema Description. 2008. 26 p.
26. Positive Technologies web site. Available at: https:/www.ptsecurity.com/ww-en (accessed: 17.07.2017).
27. Lorenzo J.M. Alienvault users manual. Version 1.0. Alienvault LC. 2011. 225 p.
28. CORDIS website. Available at: http://cordis.europa.eu/project/rcn/95310_en.html (accessed: 17.07.2017).
29. Man D. et al. A quantitative evaluation model for network security. Proceedings of the 2007 Intern. Conference on Computational Intelligence and Security. 2007. pp. 773–777.
30. Wu Y.-S. et al. Automated adaptive intrusion containment in systems of interacting services. Computer Networks: *The International Journal of Computer and Telecommunications Networking*. 2007. vol. 51. pp. 1334–1360.
31. Stakhanova N., Basu S., Wong J. A cost-sensitive model for preemptive intrusion response systems. Proceedings of the 21st International Conference on Advanced Networking and Applications. 2007. pp. 1–8.
32. Liu Y., Man H. Network vulnerability assessment using Bayesian networks. SPIE. 2005. vol. 5812. pp. 61–71.
33. Frigault M., Wang L., Singhal A., Jajodia S. Measuring network security using dynamic Bayesian network. Proceedings of the ACM Workshop on Quality of Protection. 2008. pp. 23–30.
34. Dantu R., Kolan P., Cangussu J. Network risk management using attacker profiling. Security and Communication Networks. 2009. vol. 2. no. 1. pp. 83–96.
35. Wang L. et al. An attack graph-based probabilistic security metric. Proceedings of the 22nd annual IFIP WG 11.3 working conference on Data and Applications Security. 2008. pp. 283–296.
36. Poolsappasit N., Dewri R., Ray I. Dynamic security risk management using Bayesian attack graphs. Proceedings of the IEEE Transactions on Dependable and Security Computing. 2012. vol. 9. no. 1. pp. 61–74.
37. Dacier M., Deswarte Y. et al. Quantitative Assessment of Operational Security: Models and Tools. *Information Systems Security*. 1996. pp. 179–86.
38. CyVision website. CAULDRON tool. Available at: https://www.benvenisti.net/cauldron/ (accessed: 17.07.2017).
39. SecurITree. Amenaza Technologies Limited. Available at: http://www.amenaza.com (accessed: 17.07.2017).
40. Common Platform Enumeration (CPE). NVD website. Available at: https://nvd.nist.gov/cpe.cfm (accessed: 17.07.2017).
41. Common Configuration Enumeration (CCE). NVD website. Available at: https://nvd.nist.gov/cce/index.cfm (accessed: 17.07.2017).
42. NVD website. Available at: https://nvd.nist.gov (accessed 17.07.2017).
43. Common Weakness Enumeration (CWE). MITRE website. Available at: https://cwe.mitre.org/ (accessed: 15.08.2017).
44. Wu Y., Yesha Y., Bojanova I. They Know Your Weaknesses – Do You?: Reintroducing Common Weakness Enumeration. *CrossTalk: The Journal of Defense Software Engineering*. 2016. vol. 29. no. 3. pp. 19 –24.
45. *GOST R ISO/IEC 27005-2010* [Information technology. Security techniques. Information security risk management]. M.: Standartinform. 2011. 47 p. (In Russ.).

Signed to print 20.03.2018
Printed in Publishing center GUAP, 67, B. Morskaya, St. Petersburg, 190000, Russia

The journal is registered in Russian Federal Agency for Communications
and Mass-Media Supervision, certificate ПИ № ФС77-41695 dated August 19, 2010 г.
Subscription Index П5513, Russian Post Catalog

Подписано к печати 20.03.2018. Формат 60×90 1/16. Усл. печ. л. 15,12. Заказ № 111.
Тираж 150 экз., цена свободная.
Отпечатано в Редакционно-издательском центре ГУАП, 190000, Санкт-Петербург, Б. Морская, д. 67

Журнал зарегистрирован Федеральной службой по надзору в сфере связи
и массовых коммуникаций,
свидетельство ПИ № ФС77-41695 от 19 августа 2010 г.

Подписной индекс П5513 по каталогу «Почта России»

РУКОВОДСТВО ДЛЯ АВТОРОВ

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Труды СПИИРАН» http://www.proceedings.spiiras.nw.ru. При регистрации авторам рекомендуется заполнить все предложенные поля данных.

Подготовка статьи ведется с помощью текстовых редакторов MS Word 2007 и выше. Объем основного текста – от 15 до 25 страниц включительно. Формат страницы документа – A5 (148 мм ширина, 210 мм высота); ориентация – портретная; все поля – 20 мм. Верхний и нижний колонтитулы страницы – пустые. Основной шрифт документа – Times New Roman, основной кегль (размер) шрифта – 10 pt. Переносы разрешены. Абзацный отступ устанавливается размером в 10 мм. Межстрочный интервал – одинарный. Номера страниц не проставляются.

В основную часть допускается помещать рисунки, таблицы, листинги и формулы. Правила их оформления подробно рассмотрены на нашем сайте в разделе «Руководство для авторов».

AUTHOR GUIDELINES

Interaction between each potential author and the Editorial board is realized through the personal account on the website of the journal "Proceedings of SPIIRAS" http://www.proceedings.spiiras.nw.ru. At the registration the authors are requested to fill out all data fields in the proposed form.

The submissions should be prepared using MS Word 2007 text editor or higher versions, at that, only manuscripts in *.docx format will be considered. The text of the paper in the main part of it should be from 15 – 25 pages of A5 size that is 210 X 148 mm; orientation – portrait; all margins – 20 mm. The font of the main paper text is Times New Roman of 10 pt font size. The pages' headers and footers should be empty; indention – 10 mm; line spacing – single; pages are not numbered; hyphenations are allowed.

Certain figures, tables, listings and formulas are allowed in the main section, and their typography is considered by the paper template in more detail in journal web.

ISSN 2078-9181

9 772078 918785 >

Вып. 2(57)

ТРУДЫ СПИИРАН
proceedings.spiiras.nw.ru

С-Пб
2018