ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 174

Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Problems of Automation of the Aggression Analysis in Socio-Cyberphysical Environment

Irina Kulagina
Surgut State University
Surgut, Russia
cogito@ngs.ru

Andrey Iskhakov*
V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russia
iay@ipu.ru

***Abstract*—The article presents the results of the carried out interdisciplinary research that allowed creating algorithmic software to determine the structural composition of investigated cases of electronic communication that could be interpreted as aggressive. Such interpretation is offered to be carried out by means of the system analysis of the hypertext according to the following metrics: the intensity of discussion, density and reference. The research allowed us to define sociometric models of electronic communication accompanied by aggression; the primary detection of these models is the basis for carrying out the procedure of a detailed linguistic analysis of the text of the communication. Such an approach promotes the reduction of expenses for the management of aggression processes in the socio-cyberphysical environment. The received results prove the possibility to automate the analysis of information content and to raise the interpretability of phenomena in the socio-cyberphysical environment. We examined the aspects of crawling and parsing of modern web-platforms, focused data gathering, automatic and automated extraction of comments with the application of various technologies. The architecture of a system of multithreaded data gathering is offered.**

Keywords—management automation, sociometrics, electronic communication, aggression, socio-cyberphysical environment

I. Introduction

The development of socio-cyberphysical environment and active usage of electronic communication leads to the demand to research the manifestation of aggression and practice of social pressure [1-4]. It requires an attentive analysis of modern informational processes in socio-cyberphysical environment and solutions of the problem of control system automation [5-7]. The society expects high speed and quality of processing the information containing aggression and psychological pressure from modern software management of electronic communication to react promptly. The methods of research of electronic texts traditionally include linguistic, cognitive, culturological ones etc. [8-10]. Among the main problems of the application of such methods, it is possible to notice a constant expansion of ways of realization of practice of electronic communication, manifestation of aggression and social pressure that are falling out of the limits of linguistic forms and expressed in new forms of information transmission – likes, photo and video components, links etc. It compels to expand the approaches to the analysis of electronic communication and to use methods of software detection of aggression to control the processes in the socio-cyberphysical environment [3]. One of the problems is the impossibility to know in advance, what type the practices of electronic communication would contain aggression. Aggression itself can be expressed in new forms, out of the linguistic markers of the conflict, and its post factum detection decelerates the decision-making process.

On the basis of the leading sociological researches of 32 cases of debate practices in electronic communication, we put forward a hypothesis about the possibility of algorithmic software detection of the structural composition of communication content; on the basis of this hypothesis, a qualitative evaluation of the content on whether it contains aggression can be carried out. We assume that such a possibility arises on the basis of system analysis of hypertext according to a number of criteria such as the intensity of discussion, density and reference.

The given metrics were defined during a number of investigations carried out on the basis of quantitative methods on the array of cases of electronic communication (N=32) from September to December 2019 in the most popular resources for Internet communication: Instagram, VKontakte, Twitter, Odnoklassniki, Facebook, YouTube.

The research objective was to build the criteria for the development of a search algorithm for aggression at the development of an automated system of processing of electronic communication texts.

Among the primary goals set in the research, we note the search and formation of analysis mechanisms of the modern electronic communication processes and formation of hypotheses on distinctive features of such structures in the cases containing signs of aggression.

II. Results of the Analysis of Comment Threads

During the research, we obtained the results that allowed us to divide the discussion practices in electronic communication on 2 types: short-term (the metrics describing the number of comments a day decrease, from 80 to 100 % of statements were made in the first 5 days of discussion, the discussion subjects are localized on one or several topics of discussion) and long-term (the density of comments is rather evenly distributed and is not exhausted during a long time, is in the nature of a multidirectional forum). It is necessary to note that the investigation of electronic communication texts has a number of features, unlike real communication. First of all, this is the user's capability to independently edit the texts; correction, removals or blockings of comments by the resource administration, and also complete removal of the discussion by the author or the resource administration. These features also influence the capability to analyze the data of electronic communication and prove the need for quick removal of the information on the course of communicative processes in the information society to make administrative decisions.

The distribution of the total number of comments in short-term discussions was examined in the given research. According to the data of the cases, the general volume of comments appears during the first six hours from the start of the discussion (67.2 % of all comments), and in distribution, from all studied time of duration of the discussion, this volume slightly decreases to 60.3 % (see Fig.1, Fig.2).

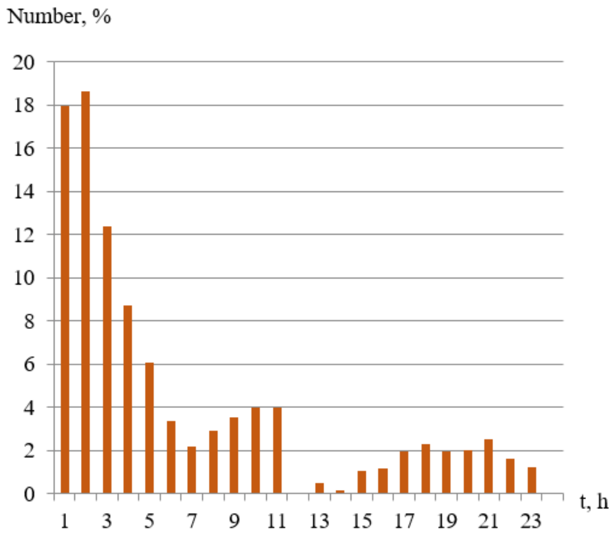

Fig. 1. Time distribution of comments in % in the investigated practices of electronic communication (N=32).

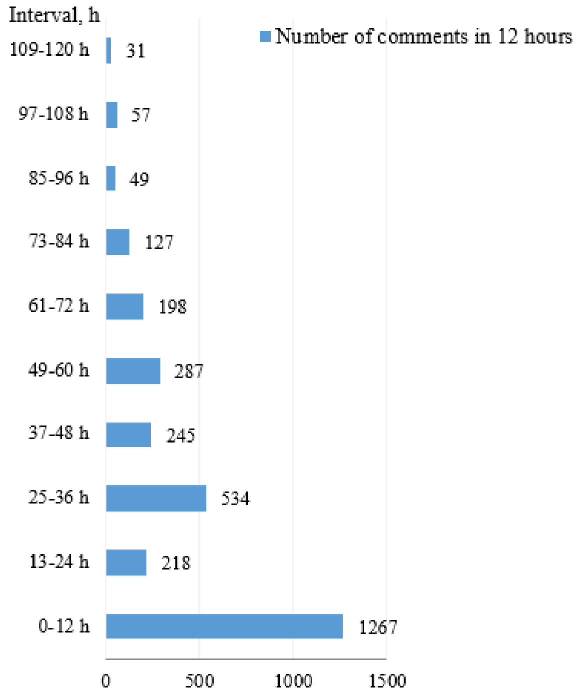

Fig. 2. Distribution of comments in the first five days from the start discussion in the investigated practices of electronic communication (N=32).

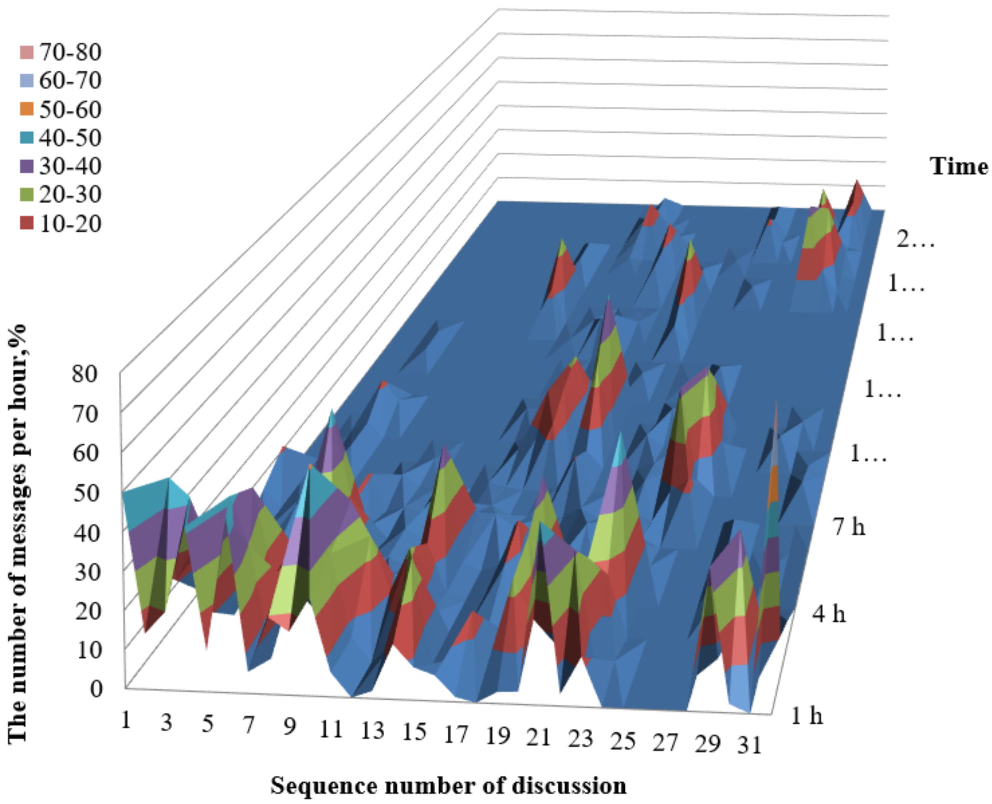

Fig. 3. The diagram of message density in % from the total amount in discussion within the first 24 hours in the investigated practices of electronic communication (N=32)

The further content-analysis of the contents of the given texts gave us a chance to assume that the peak metrics of discussion intensity can appear as a basis for detection of such sides of communicative practice which can be linked to aggression manifestation. There was a hypothesis that it is aggression that allows creating such structural characteristic as discussion density (see Fig. 3), which is required to be tested using a considerable array of cases of electronic communication. Because of the connection with the complexity of the solution of the given task by methods of sociology, the extension of research practice received reinforcement by the development of the automated analysis system of electronic communication texts. The architecture of this system is presented in the third part of the article. Currently, the system functions in the mode of accumulation of knowledge bases.

It was found that modern practices of electronic communication have a wide set of means allowing participants to express aggression. First of all, it includes capability to give a Like or a Dislike to support or condemn the comments of interlocutors. The analysis of forms of addresses (nicknames) in texts of communicative practices allowed us to proceed to the building of structural (sociometric) interaction models in each of the objects of our research. An example of a sociometric representation of a discussion is presented in Fig. 4.

The main problems of construction of such sociometric models are linked to the variability of the means to create electronic discussions. These could be sets of means to support electronic communication:

- visual means in the form of pictures, memes, smiley faces, video etc.;
- sound means: sounds, music;
- hypertext means in the form of codes, links (including links to user profiles), pop-up windows, functions such as Like and their variations, etc.;
- text means in the form of fonts, CapsLock, Leet, specific syntax, etc.

The necessity to consider the given sets in the qualitative analysis of the text complicates the process of formalizing in the analytical software. Thus, regarding the development process of algorithmic software means to detect aggression, it is offered to use quantitative criteria of intensity, density and reference to design preprocessing of web-resources for the decision-making on the further qualitative analysis of manifestations of aggression.

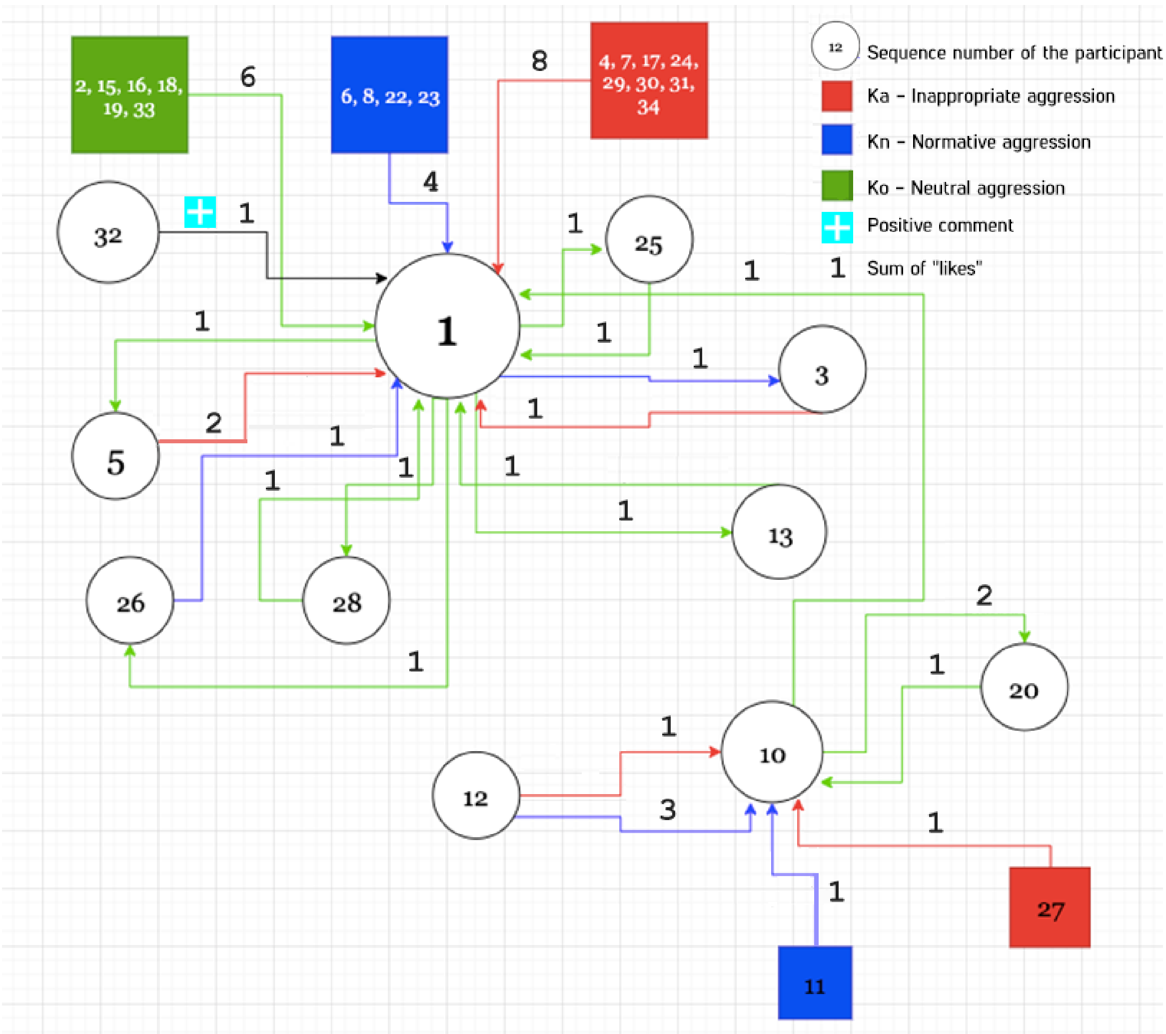

Fig. 4. Sociometric representation of aggression in the practice of electronic communication (example)

III. Process of Automation of Target Platform Crawling

Let us assume further that the structure of a target web-resource represents a strongly connected graph. Thus, having started data gathering on any page of a platform, it is possible to reach any other point. This is a serious enough assumption. On the one hand, it is obvious that the resource can contain hidden sections that do not have any links from the mainframe of the project. This fact could be considered as insignificant, because, as a rule, a web page without any links addressed to it is not attractive from the point of view of the search of destructive materials [11,12]. We examined the main graph traversal algorithms:

At the heart of the Depth-first search (DFS) algorithm, there is recursion. The algorithm represents the following sequence of operations:

1) *To mark the current point as processed.*
2) *To process the current point (in the case of a search robot, processing is simply copying).*
3) *For all points to which it is possible to get from the current one: if the point has not been yet processed, to process it recursively too.*

The given method is advisable to apply in order to traverse web-pages on a small Internet platform; however, to traverse big social networks, it is not applicable because of the following reasons:

- The given algorithm cannot be used in a mode of parallel data processing in view of the presence of recursive calls
- The usage of the given algorithm on great volumes of data leads to the overflow error. The crawler completely fills the stack of recursive calls in the process of passing “into the depth” through the links.

The breadth-first search (BFS) algorithm works similarly to Depth-first search; however, it traverses points of the graph in the course away from the home page. For this purpose, the algorithm uses “queue” data structure – in the queue, it is possible to add elements in the end and to take them away from the beginning.

1) *To add the first point in the queue and in the set of “seen” ones.*
2) *If the queue is not empty, we get the next point from the queue for processing.*
3) *To process the point.*
4) *For all edges which start at the processed point and do not belong to the “seen” ones:*
   a) *To add to the “seen” ones;*
   b) *To add in the queue.*
5) *To proceed to item 2*

In this case, the queue and set of the “seen” points use only simple interfaces (to add, take, check the occurrence) and can be easily directed for processing by a separate server connected to the client through these interfaces. This feature allows realizing multithreaded traverse of the graph. Thus, there is a possibility to start several simultaneous output agents using the same queue.

IV. Problem of the Data Parsing Automation

Absolutely all investigated platforms for electronic communication represent web-oriented systems [13,14]. “Parsers” are used for the automated gathering and structuring of the information from similar resources [15-20]. Ready solutions allow one to extract the information from a web-site after preliminary configuration. However, similar systems are directed on the extraction of the information from the readily available handbooks, databases, directories, Internet shops. They do not possess flexibility which the solutions developed under a specific platform can give. Thereof, for the complex distributed systems of monitoring, gathering and analysis of the information, parsers are developed individually, taking into account structural and technical features of the target resource.

The process of extraction of information from a separate web-page is possible to present in the form of the following procedures:

1) *Construction of an inquiry for information obtaining.*
2) *Execution of the inquiry and obtaining of the answer.*
3) *Processing of the answer, extraction and structuring of the necessary information.*
4) *Transmission of the received information for the subsequent processing.*

The investigated platforms for communication do not allow unauthorized users to extract information. For the purpose of protection against robots and spam, it is required to carry out authorization procedure in the system; after this, it is necessary to make an inquiry from headers and values of JavaScript-variables on the page. Besides, the name of these headers and variables frequently changes with each inquiry, and the JS-code of some resources is minimized and obfuscated.

During the system development, the following problems were revealed (Table 1).

TABLE I. THE MAIN PROBLEMS AND THE SOLUTION PATHS

Problem	***Way of the solution***
Dynamic modification of the structure of page code for the purpose of complicating of extraction of information from blocks.	Application of “screen scraping” methods assuming the extraction of information not from the page text but from its image.
Demonstration of CAPTCHA at authorization/registration/exceeding of inquiry number for a time unit.	Use of third-party services for CAPTCHA recognition
Blocking of clients according to the average volume of traffic in a time unit.	Increase of parsing process in time (setting the intervals of compulsory delays)
Analysis of client behavior and blocking/requirement to pass CAPTCHA to continue working in case of suspicious behavior.	Cloning of behavior of a real user: clicks, mouse movements, page scrolling.

V. Architecture of the Automated System

For the purpose of automation, a multithreaded system of crawling and parsing of materials of informational and communication platforms of the Internet was developed. The system architecture representing the aspects of multithreading and reservation is presented in Fig. 5.

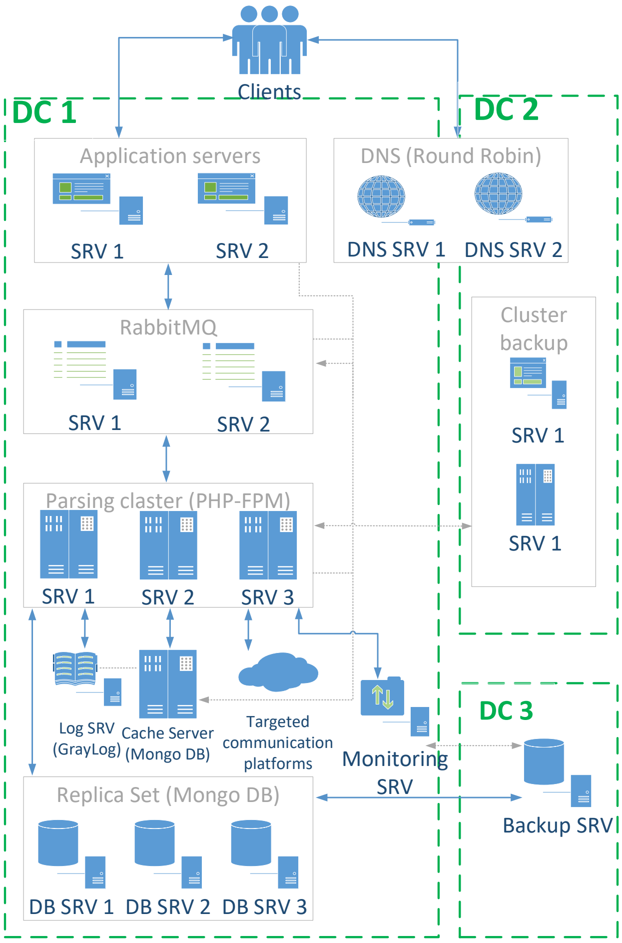

Fig. 5. The architecture of the developed system

Parsing servers gather the data from target resources (blog-sphere resources, social networks) and carry out analytical operations with great volumes of data, using Replica Set databases. The result of analytics is stored in Replica Set. One analysis can last from several seconds to about several weeks. To reduce this time, Cache Server is used which caches the inquiries to social networks. The work of all servers is logged in a log-server. The metrics of productivity and status of all servers of the system are collected in the monitor server. In the case of approximation to critical thresholds, the message is sent to the Telegram-channel of the system administrator.

All Replica Set data is continuously archived on Backup Server which is located on a remote data-center. If there is a failure in data-center 1 (DC 1) and within 15 minutes the system does not restore the work, backup service 2 (DC 2) where a simplified copy of the data-center 1 (DC 1) architecture is deployed.

On order to data-center 2 (DC 2) to be always ready to sudden reception of the load, the key data from data-center 1 (DC 1) are continuously synchronized with data-center 2 (DC 2) (every 10 minutes).

The total amount of Replica Set storage is 24 TB. It stores only recent analysis data (for the last 30 days). The older data are transferred to the back-up-server in data-center 3 (DC 3).

The operator (data analyst) uses a browser to log in into the application. The cluster from two front-end servers is used, which raises the convenience of the work and increases the general fault tolerance: in case of failure of one of the servers, the second one continues to service the inquiries of the operator. Server DNS 1 is responsible for the distribution of inquiries between servers.

To increase fault tolerance, the system provides 2 frontend-servers. If the inquiry concerns the already gathered data, the business logic server fulfils it using the data earlier stored in Replica Set. If the inquiry concerns the gathering of new data, the business logic server sends it to the parsing cluster for analytics. Since the parsing cluster is usually strongly loaded (load average is 79 %), the inquiries of business logic servers are put in the queue in the Queue Management block that regulates the load on the parsing server.

VI. Conclusion

Thus, at the current stage of research, the conditions for automatic search of structural characteristics and construction of sociometric models of electronic communication accompanied by the presence of aggression are defined. Their primary detection is the basis to carry out of the procedure of the linguistic analysis of the electronic communication text that promotes the reducing of the expenses on the management of aggression processes in the socio-cyber physical environment.

The examined in the article aspects of crawling and parsing of modern web-platforms, focused data gathering, automatic and automated extraction of comments with the application of various technologies allowed us to generate ways of the solution of the main defined problems and to consider them during the development of the system of automated data parsing from web-platforms.

The future investigation phases and developments will be directed on the perfection of the system architecture, the research regarding the optimization of traversal of the source DOM-structure, and also the analysis, experimental estimation of efficiency of system functioning, estimation of input parameters and search of the dependence of prototype productivity on data flow increase.

Acknowledgment

The reported study was partially funded by RFBR, project number 18-29-22104.

References

[1] B. Zhang, and M. Vos, “Social media monitoring: aims, methods, and challenges for international companies,” Corporate Communications: An International Journal, vol. 19(4), pp. 371-383. September 2014.

[2] F. Agrafioti, D. Hatzinakos, and A.K. Anderson, “ECG Pattern Analysis for Emotion Detection,” IEEE Trans. Affect. Comput. vol. 3, pp. 102-115, January-March 2012.

[3] I. Kulagina, A. Iskhakova, and R. Galin, “Modeling the practice of aggression in the socio-cyber-physical systems,” Tomsk State University Journal of Philosophy, Sociology and Political Science, vol. 52, pp.147-161, December 2019.

[4] R. Lawrence, P. Melville, C. Perlich, V. Sindhwani, S. Meliksetian, P.-Y. Hsueh, and Y. Liu, “Social media analytics,” OR-MS Today, vol. 37, issue. 1, pp. 26-30, February 2010.

[5] B. Bolloba's, and O. Riordan, “Mathematical results on scale-free random graphs,” Handbook of Graphs and Networks: From the Genome to the Internet, pp. 1-34, November 2002.

[6] E. Garin, and R. Meshcheryakov, “Method for determination of the social graph orientation by the analysis of the vertices valence in the connectivity component,” Bulletin of the South Ural State

University», series «Mathematics. Mechanics. Physics, vol. 9(4), pp. 5-12, 2017.

[7] J. Hu, Y. Liu, and R. Lawrence, “Graph-based transfer learning,” Proceedings of the 18th ACM conference on Information and knowledge management (CIKM ’09), pp. 939-946, November, 2009.

[8] L. Gritsenko, and T. Demidova, “The discrediting speech strategy in internet communication (on the example of the use of trolling),” Tomsk State University Journal of Philology, vol. 55. pp. 29-42, October 2018.

[9] I. Vasenina, and T. Kukhtevich, “Language aggression on the internet - call moral imperatives,” // Higher education in Russia, vol. 4, pp. 121-125, 2014.

[10] J. Donath, “Identity and Deception in the Virtual Community,” Peter Kollock and Marc A. Smith eds. Communities in Cyberspace. Routledge, pp. 11-12, 1999.

[11] A. Iskhakova, A. Iskhakov, and R. Meshcheryakov, “Research of the estimated emotional components for the content analysis,” Journal of Physics Conference Series, vol. 1203(1), article no. 012065, April 2019.

[12] A. Iskhakova, “Processing of Big Data Streams in Intelligent Electronic Data Analysis Systems,” Advances in Intelligent Systems Research, vol. 169, pp. 14-19, September 2019.

[13] D. Levonevskii, O. Shumskaya, A. Velichko, M. Uzdiaev, and D. Malov, “Methods for Deter-mination of Psychophysiological Condition of User Within Smart Environment Based on Complex Analysis of Heterogeneous Data,” Proceedings of 14th International Conference on Electromechanics and Robotics “Zavalishin's Readings”. Smart Innovation, Systems and Technologies, vol. 154, pp. 511-523, August 2019.

[14] A. Iskhakov, “Adaptive Authentication Technologies in Behavioral Analysis Solutions of Robotic Systems,” Advances in Intelligent Systems Research vol. 169, pp. 49-54, September 2019.

[15] A. Iskhakova, A. Iskhakov, R. Meshcheryakov, and E. Zharko, “Method of Verification of Robotic Group Agents in the Conditions of Communication Facility Suppression,” IFAC-PapersOnLine, vol. 52, issue 13, pp. 1397-1402, January 2019.

[16] R. Mesheryakov, A. Moiseev, A. Demin, V. Dorofeev, and V. Sorokin, “Using parallel computing in queueing network simulation,” Key Engineering Materials, vol. 685, pp. 943-947, 2019.

[17] P. Chahal, M. Singh, and S. Kumar, “Ranking of web documents using semantic similarity,” 2013 International Conference on Information Systems and Computer Networks, pp. 145-150, March 2013.

[18] S. Iskhakov, A. Shelupanov, and R. Mescheryakov, “Assessment of security systems complex networks security,” 2014 Dynamics of Systems, Mechanisms and Machines : Dynamics 2014 Proceedings, article no. 005656, 2014.

[19] S. Upadhyay, V. Pant, S. Bhasin and M. K. Pattanshetti, “Articulating the construction of a web scraper for massive data extraction,” 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-4, November 2017.

[20] T. Panum, R. Hansen, and J. Pedersen, “Kraaler: A User-Perspective Web Crawler,” 2019 Network Traffic Measurement and Analysis Conference (TMA), pp. 153-160, August 2019.