Irada Y. Alakbarova
Institute of Information Technology of ANAS, Baku, Azerbaijan
airada.09@gmail.com
DOI: 10.25045/jpit.v07.i2.09

THE PROBLEMS ASSOCIATED WITH BIG DATA IN WIKI-ENVIRONMENT AND THEIR SOLUTION WAYS

The article analyzes the problems emerged due to the collected and continuously growing large-scaled data in wiki-environment. The prospects of using wiki-metric research in decision-making for various purposes are identified. The necessary conditions for the storage and processing of large amounts of data in a wiki-environment are presented. It is proposed to jointly use some existing approaches to solve the problems associated with Big Data.

Keywords: *Wiki environment, Wiki technology, Wikipedia, Big Data, wiki-analytics, hromogram, data mining, Map-Reduce.*

Introduction

In the modern era, Big Data (BD) issues have become topical with the spread of the Internet and the expansion of the network [1]. Today, information and communication technology professionals are concerned with the storage, structuring and processing of information generated each second by the giant Internet projects, such as Facebook, Twitter and other social networks, search engines - Google, Yahoo, and Yandex, open encyclopedias - Wikipedia, WikiaMapia, Wikitravel and others, various forums, blogs and other systems. As the Internet users enter the contents of the above-mentioned projects, it is impossible to prevent the rapid increase of information in the virtual space.

Wiki technology-driven Wikipedia virtual encyclopedia and its affiliated projects (Wikibooks, Wikisource, Commons, Wikispecies etc.) make up the overall wiki-environment, and its giant database is being enriched each second with new articles, news, books, photos, audio and video files [2, 3]. The main objective of the study is to identify BD problems related to wiki-environment, and to show the possibility of resolving a number of issues based on the data collected in Wikipedia database (WB), and to provide the right solution for an efficient data processing.

Big Data problems of Wiki-environment

Wiki environment connects millions of web pages. According to the report of January 2016, only Wikipedia has more than 60 million registered users, about 40 million encyclopedic articles [4], 30 million photos, audio and video-file [5]. The total volume of only more than 4 million English articles in Wikipedia makes up 1.7 terabytes [6]. The total content collected in Wiki environment is about tens of terabytes, and this information is continuously growing. If we add of the tens of millions of users' personal and discussion pages, their surveys, generated templates, and bots to this list, it is not difficult to imagine how overwhelmingly big the information would be. The growing number of the users and their super activeness highlight the urgency of resolving these issues with the use of BD technology. Likewise in BD, the problems are especially occurring due to unstructured data and its excessive volume in the Wiki environment.

The main problems are:

- *Data volume*. The volume of the data in the database of Wikipedia projects is measured in bytes. The volume of data collected from millions of Wikipedia pages and files defines whether it is analyzed as BD or not. The specific conditions are required for the processing and storage of the large-scale data.
- *Data variety*. Wiki contents can be of different types. As texts, voice records, photo and video files can be uploaded to Wiki pages, and wiki technologies support almost all file formats, the data of different types and structure are collected here. They cannot be processed altogether and at the same time. First of all, the data is required to be in a format suitable for analysis.

- *Data velocity*. Wiki servers save all the data (requests, media-files, software codes, texts, etc.) received from the browser through the processing. It should be noted that each character included in Wikipedia is generated and stored in the database, and even though it is deleted from the web-pages, it remains in the database and can be restored by the user at any time. Undeleted information in the database of Wikipedia, which contains millions of information per second, means the excessive proliferation of data over time. The greater data processing speed, the better interactivity conditions are provided.
- *Data variability*. Wiki pages are open and can be edited directly from the browser at any moment. The reviewed page can be updated and submitted with completely different structure and design. All the old versions of the page (establishment history, comments, deleted and modified contents, referring to the authors) are stored in the database. These factors create certain problems in the data management during the analysis.
- *Complexity*. The data are included from different sources. There is a complex semantics between the Wiki pages. The links between pages may differ: inter-project, according to the language factor (inter-wikis), according to the categories, according to other wiki-projects, and finally, links to external projects (various websites on the Internet). These links can be canceled and modified at any time. Such complexity also complicates the Wiki-metric studies.

During the traditional data processing, the data is checked, set in a certain format, and uploaded on the system. This type of sequence is not promising for the storage and processing of data collected in the wiki environment. As described in a scientific literature, the process of big data processing includes solution of the issues as data stream visualization, Data Mining, BD-based description of the situation, advanced analysis hardware and software [2, 7–9]. Solution of such issues still remains relevant for wiki-analytic studies.

Wiki-analytics

Wiki-analytics is a science about the measurement, analysis and assessment of data collected in wiki-environment. The measurement, analysis and description of Wiki pages contents and information about the users visiting Wikipedia projects is carried out through Wiki-analytics. The main objective of the Wiki-analytic studies is to evaluate the content quality uploaded to the wiki-pages based on the data that defines the audience of the website, and to study the behavior of the Wikipedia users in making decisions on traffic monitoring and for different purposes. With the use of the Wiki-analytic studies, it is possible to explore information support of wiki pages, the accuracy of the submitted information, its relevance and coverage, the safety of relations between the wiki-pages and their logical structure, and it is possible to define the objectives of the users who are directly involved in the development of wiki pages and in the implementation of the information war, and to detect the secret social networks supporting specific ideological goal.

The following information can be attained through defining and evaluating the quality of the content posted on Wiki pages:

- cases of vandalism on Wiki-pages;
- compatibility of the title of the page with its context;
- semantics between the Wiki-pages;
- conflicts in Wiki-environment;
- detection of the social groups involved in the conflict.

The assessment of the Wiki-pages should also include the parameters, such as the scope of the context, the number of structural elements, the number of relations with other Internet resources, the number and date of the users' access to the wiki-pages, the volume of information at wiki-pages, quality and relevance of downloaded multimedia files to the subject, the number of grammatical and spelling errors, information reliability, availability, and the number and

dynamicity of lists of hyperlinks, presentation style of the information, comfort and reliability of the material presented in other languages and etc.

The following information can be obtained through the statistics of Wiki-page traffic:

- number of access to wiki-pages;
- keywords and phrases used for the search of the page;
- socio-demographic portrait of the users;
- time spent by users on Wikipedia projects and individual wiki-pages;
- relevance of Wiki-pages' data;
- number of users by categories: permanent and random.

From the analytical point of view, above-mentioned opportunities increase the accuracy of the evaluation of certain situations, and enable the development and implementation of possible scenarios in the course of the events. BD generated in Wiki environment, today, has turned Wikipedia into research field, and has created great opportunities for studies. The acquisition of hidden information from the large-scale data processing is the main objective of all scientific fields. The problems related to BD in Wiki environment create the need to seek new approaches to data processing. It should be noted that each wiki-page is an individual electronic document so that multiple parameters used to measure the quality and quantity of the documents are also used to measure the wiki pages [2, 3, 7].

The massive collection and processing of data in Wiki environment can also affect certain decisions made in public authorities, business and education. Comparative analysis methods, fuzzy sub-sets theory, clustering methods, and graph theory are dominating in the study of the establishment process of constantly updated collective knowledge in Wiki projects, and to what extent the Wikipedia articles cover all the fields, and to determine the influence and role of Wiki-content in the information conflicts. However, semantic analysis methods are widely used in Wiki-environment analysis, i.e., in the classification of articles by their certain signs and in defining the social network. Semantic analysis methods widely use the algorithms HITS (*Hyperlink-Induced Topic Search*), PageRank, Random Forest [8, 9, 10].

In Wikimetric Studies, first of all, the public databases, which are available for everyone, should be mainly referred. The databases indicators are as follows:

- size of databases, where wiki-content is stored (including volume of all wiki pages, discussion, categories and directing pages);
- number of encyclopedic articles of wiki-projects;
- number of words in each encyclopedic article;
- number of access to wiki pages;
- number of internal links of wiki-pages;
- number of external links of wiki-pages;
- volume of wiki pages in bytes;
- number of registered and anonymous users of wiki-projects;
- number of active users (users who have more than 5 editing per month);
- number of more active users (users who have more than 100 editing per month).

Wiki environment makes it possible to conduct more in-depth research in data mining and to obtain confidential information. Confidential information is used in solving problems and making decisions. Wiki-environment has great potential for Data mining, and these opportunities are as follows:

a) Wiki environment manages documents:

- availability of millions of wiki pages.
- availability of hundreds of different types of links (Dense link structure).

b) Wiki environment manages knowledge:

- continuously generated new content.
- dynamically changing content.

• content systematization.

c) Wiki environment is a giant social network.

In a connection with the above-mentioned Wikipedia possibilities, the methods, such as Wiki Mining, Text Mining and Link Mining, make it possible to analyze semantic relations between the pages in wiki-environment, to determine the scope and quality of encyclopedic articles, as well as to detect secret social networks [11, 12].

Existing approaches to the problem solution

Wikimetric studies use different methods (counter, analysis of journal files, access, observations, etc.). The problems related to BD on the Internet require the use of wiki-analytics as a powerful tool to measure Wikipedia resources and to study the users' behavior. IBM has worked out a particular algorithm to define the activities of Wikipedia users and the consistency between the events. This algorithm, which has been applied since 2007, is based on data visualization. The image presented as a hromogram and the colors in the hromogram depend on the colors of words of the texts. The method of presenting the text in colors is the main distinguishing feature of the hromogram [13]. For example, the hromogramprovided in Figure 1 is set out and based on the revisions made to the articles about the history of naval forces.

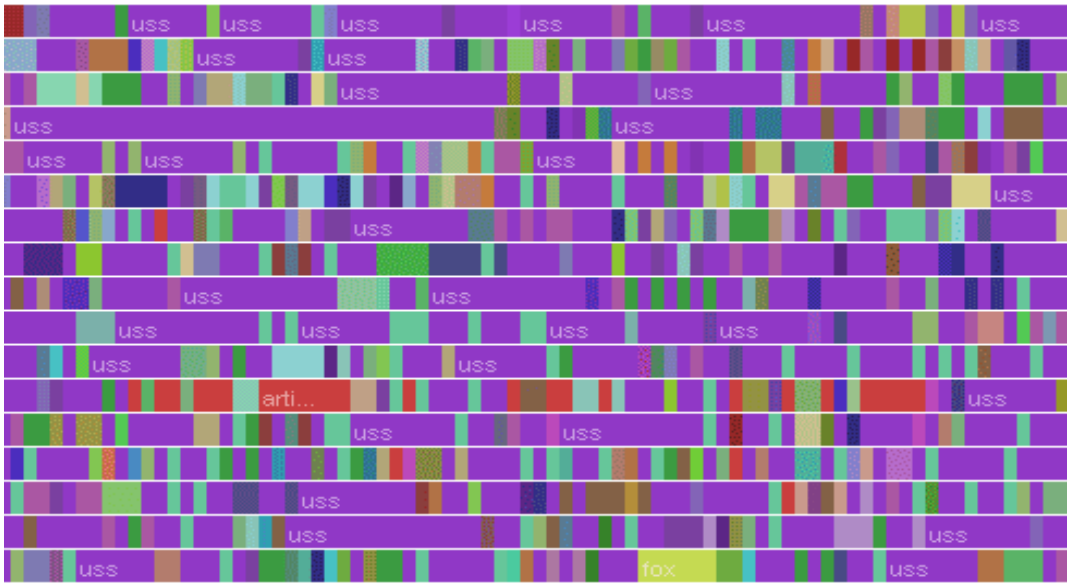

Figure 1.Hromogram providing the scope of users' interest in Wiki environment

The prevailing parts of the hromogram in purple indicate the articles containing the names of naval vessels. It means that the hromogram shows that, though the user accesses variety of topics, his interest in military ships is prevalent. With the use of hromogram, it is possible to analyze the topics, which the users are interested in, vandalism incidents, social networks in wiki-based, conflict situations and other interesting issues [12, 13].

It is also possible to define the nature of the users' activities with the use of hromogram. The system marks each type of activity with special color. Therefore, each user has own style in wiki-environment, and they are as follows:

- creating a new page;
- including a content on the available page;
- deleting or modifying the information of the pages;
- applying design and adding custom templates;
- controlling semantics between the pages;
- participating in discussions on specific topic.

SGI UV 2 (Silicon Graphics International, ultraviolet) is another system, which provides the data analysis in Wikipedia [13]. Developed by SGI company and the University of Illinois employee Kalev H. Leetaru, the system generates chronological order full-text content in Wikipedia by their date and space, and enables their search. The system defines the development stages of the world history by analyzing the data collected from the English Wikipedia and presents a visual description of the modern history. Using Wikipedia data, the system also shows how the world has been visualized in the last 2 centuries. It defines the content of wiki pages, and refined the semantic relations between the pages and establishes the great network based on these relations (Figure 2).

Figure 2. Visual historical map of the world for the year 1900 defined by SGI UV 2 system

SGI Chief Marketing Officer Franz Aman compares SGI UV 2 system to Google Earth and says: "We like to revise the overall situation while reducing the scale of the map on Google Earth, and applying this concept to the big data of Wikipedia, we have achieved this description" [14]. The system implements in a memory data-mining. Data mining algorithms, first of all, reveals the nature of the text, its content, and the links between the pages. The results are stored in metadata, and used for subsequent analysis.

Some approaches explain BD challenges with 3C parameters, rather than the traditional 3V (volume, variety, velocity) [15]: cardinality, continuity and complexity [16]. According to the authors, mathematical and statistical analysis, methods are easier to be set with 3C parameters to solve the problem. It should be noted that the data velocity in 3V conditions is not justified in most cases for Wikipedia. Thus, the data variety according to its content type and format in wiki-environment, their unstructured format, and the probability of being changed every second makes prompt data processing difficult.

Problem solution based on Big Data technologies

Analysis of some existing approaches to BD problem solution suggests that, when solving the problem, it is necessary to choose the best solution, thus the algorithm will automatically predict the development of events in wiki-environment in advance using the data collected during the previous experiences. Whilst working with the traditional data warehouse and solving the certain problems, such as defining the inconsistency among the data, the previously collected and aggregated data is used. When working with BD. such problem solution causes some difficulties. Before the data processing, the information shall be transferred to servers. BD in Wiki environment

also makes this approach impossible. Another problem is that the data in wiki-environment is distributed, structured, and set as per the users' wish, as they deal with content creation. For example, the data is collected in regional servers, and the structuring and collection of the data of different servers differ. In terms of data analysis, it would be better to distribute this data to the servers in accordance with time rather than location, and each server would be responsible for specified time interval. However, the capabilities of wiki servers do not provide this. Taking into all account, the following types of data in wiki-environment can be distinguished for the analysis of BD:

- by development date (old data/new modified data);
- by acquisition methods (data included from computers/mobile phones);
- by type (text/video/audio/picture).

Working with BD requires very prompt analysis. For example, the issues such as detecting secret social networks in wiki-environment, preventing conflict situations, monitoring wiki-pages and users, identifying the rapid spread of information and its source and so on, should be resolved as soon as possible. However, as information is too much, there is a need for prompt data mining algorithms. These algorithms should also be capable to process the texts in the wiki-pages, discussions of the users, as well as the photos, video and audio-files included by them. Therefore, the following conditions must be provided to work with BD collected in the wiki-environment:

1. The methods used for data analysis in Wiki-environment shall be determined;
2. Intellectual system capable to analyze large-scale data in a short time shall be provided.

It should be noted that the massifs set out during the big data analysis are not always randomly obtained. Sometimes, statistical data are of great importance. The main problem here is that statistical analysis of large, complex and heterogeneous data may lead to the selection of wrong and inappropriate variables and algorithms. Thus, the risks of errors related to coverage, choice, measurement and obtained responses should be taken into account while working with large-scale data [17].

Along with the requirements not only of the specific software, but also the hardware used for Big data analysis, the introduction of approaches together with special computer modeling in this field should also be taken into account. Agent modeling may be cited as an example. Agent modeling studies behavior of distributed data, and subsequently, behavior of the entire system is determined by this method [18]. At the same time, the hidden structure of data is possible to be explored using the multidimensional data analysis methods as factor and cluster analysis [19]. The use of Hadoop Distributed File System (HDFS) [20] model, Hive database and ZooKeeper techniques [21], Cloud technologies [22] in Big Data clustering can achieve more effective results.

HDFS provides data storage and organization in special clusters. ZooKeeper is a coordination service high availability and used in building distributed applications. Hive is a distributed data warehouse. Hive works with HiveQL (Query Language) and manages the data collected on HDFS.

A brief analysis of approaches to BD problem solutions shows that the processing of distributed analysis of data may ease its prompt analysis. Efficient big data processing is based on the distributed file systems technology. The data in distributed systems is stored and indexed not in a single file system, but in the data warehouse on a server (Figure 3). In this regard, parallel and distributed database management systems (DBMS) should be used in Big Data processing, and the data should be processed in the nodes in which they are located. In this case, the work quality of the parallel DBMS depends on the accuracy of data distribution. Data distribution for database may minimize the number of information exchange between the nodes during the implementation of requests.

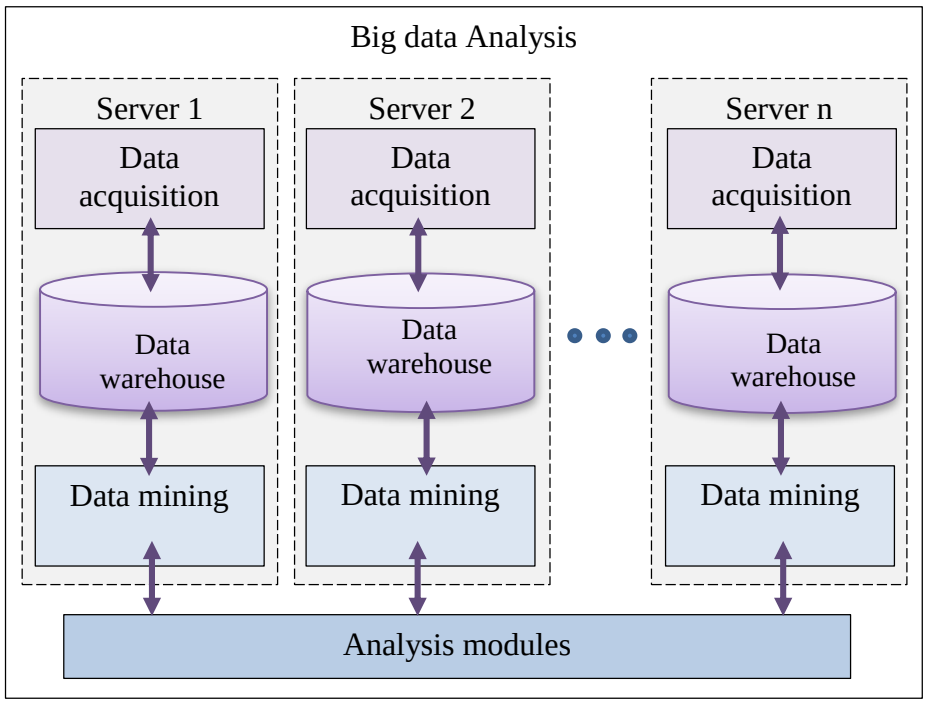

Figure 3. Distributed system used in the big data analysis

It should be noted that if all the data in the traditional warehouses is provided from a logical block, which is responsible for its conversion, cleaning, inspection and loading, the use of single logic blocks in BD processing is not efficient. The tools used in the traditional warehouses include the search for multidimensional analysis (OLAP), regression, classification, clustering and patterns. Today, the systems as SAP HANA, Greenplum Chorus, Aster Data Cluster apply these methods for the analysis of BD. For example, the most popular technologies to work with BD, are Map-Reduce model offered by Google Inc. in 2004 and Hadoop open-source software offered by Apache Software Foundation. These systems do not require to collect the data in tables and to follow a strict hierarchy. Before the data analysis, it is enough to distinguish the data species. Map-Reduce model provides the efficiency of the accurate data distribution. Map-reduce is a distributed data calculation model and used for parallel calculations over the large-scale data. The Map designed for initial processing of input data shows the individual nodes in which the distributed data is located, and stores the information about the results and the algorithms used in the processing. The results are summed up at the end of the operation (Reduced). For example, the algorithm calculates the intermediate sums at each nodes of parallel distributed file system and sums up found values in the subsequent phase in order to the final sum.

Splitting up the graph into sub-graphs with certain characteristics for the effective data distribution is also noteworthy. The main condition is the minimum number of nodes connecting these sub-graphs. This is "NP-completeness" problem. It means that for the problem solution, the problem should be split up into NP (non-deterministic polynomial) problems. If a solution algorithm for any of these problems is found, then this algorithm is possible to be used for the solution of other NP problems. The use of data mining methods and hybrid algorithms in the problem solution allows achieving more effective results. BD processing cannot be limited to the specific technology providing only distributed data processing. In the course of processing, the parameters typical for BD should also be considered. For example, network interactivity, intensity, data volume, and so on.

As in other areas, in Wiki-environment, BD problems also raise new questions, which are necessarily to be explained. For example, verification and validation of the approaches and models related to BD are of great importance. Verification and validation is used for software quality control

to identify the system failures. The verification provides a comparison of the approaches proposed for the solution of any issue with the previous ones and checks the software requirements, the standard norms, and the correlation between user documentation. Validation checks the compliance of the proposed approach with the requirements of the users or customers. Even though these requirements are not often officially documented, they are provided in the requirements' descriptions. It should be noted that, when the data is too large, measurement of errors is also important. The numbers of errors occurred in the course of big data processing and the measurement are interrelated.

Conclusion

The study revealed that data collection in wiki-environment creates great challenges related to the data processing and storage, and requires new approaches to the problem solution. The working principles of the existing systems used in data visualization and analysis in Wiki environment allow to conclude that the data accumulated in Wikipedia during the recent years has the features of Big Data: the volume is great – which requires special conditions for the efficient structuring and storage, and the variety – which cannot be collected and processed in a database through the traditional methods. Furthermore, there is a complex semantics between the wiki-pages that also complicates its processing.

The examined approaches to the BD related problem solutions suggest that, parallel and distributed data processing enables to achieve more efficient results. On the other hand, data diversity and variability in wiki-environment requires the use of new technologies in computer graphics for its analysis and description.

References

1. Aliguliyev R.M., Hajirahimova M.S. "Big Data" phenomenon: Challenges and Opportunities // Information Technologies Problems, 2014, No 2, pp.3-16.
2. Alakbarova I.Y. Some Approaches to the Development of Information Influence and Hidden Communications Detection Systems in Wiki-Environment / Proceedings of the 4th International Conference "Problems of Cybernetics and Informatics" (PCİ), Baku, Sept. 12–14, 2012, vol.I, pp.119–120.
3. Alguliev R.M., Aliguliyev R.M., Alakbarova I.Y. Wikimetric research: current status and prospects // Telecommunications 2014, No 5, pp.15–31.
4. https://meta.wikimedia.org/wiki/List_of_Wikipedias
5. https://commons.wikimedia.org/wiki/Main_Page
6. https://www.openhub.net/p/mediawiki
7. Alakbarova I.Y. Analysis of some information warfare technologies in Wiki-environment//Information Society Problems, 2011, No 2, pp.18-28.
8. Arazy O., Stroulia E.A. Utility for estimating the relative contributions of wiki authors / Proceedings of the Third International ICWSM Conference, San Jose, California, May 17–20, 2009, pp.171–174.
9. Halfaker A., Keyes O., Taraborelli D. Making Peripheral Participation Legitimate: Reader engagement experiments in Wikipedia / Proceedings of the 2013 conference on Computer supported cooperative work, ACM, NY. USA, 2013, pp.849–860.
10. Aliguliyev R.M., Alakbarova I.Y. Wikimetric research: state, problems and prospects, Express-information, Baku, "Information Technology" Publishing House, 2015, p. 87.
11. Iba T., Nemoto K., Peters B., Gloor P.A. Analyzing the creative editing behavior of Wikipedia editors: through dynamic social network analysis // Procedia – Social and Behavioral Sciences, 2010, vol.2, no.4, pp.6441–6456.
12. Müller C., Meuthrath B., Baumgra A. Analyzing wiki-based networks to improve knowledge processes in organizations // Journal of Universal Computer Science, 2008,vol.14, no.4, pp. 526–545.

13. Wattenberg M., Viégas F.B., Hollenbach K. Visualizing Activity on Wikipedia with Chromograms / Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction, 10 Sept. 2007, Berlin, vol.4663, pp.272–287.
14. http://www.sgi.com/go/wikipedia/
15. Alguliyev R., Imamverdiyev Y. Big Data: Big Promises for Information Security / Proceedings of the 8th International Conference on Application of Information and Communication Technologies (AICT), 15–17 Oct. 2014, IEEE, Astana, pp.1–4.
16. Hilbert M. Big Data for Development: From Information – to Knowledge Societies, 2013. http://ssrn.com/abstract=2205145
17. Couper M. Is the sky falling? New technology, changing media, and the future of surveys // Survey Research Methods, 2013, vol.7, no.3, pp.145–156.
18. Suthaharan S. Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning // ACM Sigmetrics Performance Evaluation Review archive, 2014, vol.41, no.4, pp.70–73.
19. Alguliev R.M., Aliguliyev R.M., Alekperova I.Ya. Cluster approach to the efficient use of multimedia resources in information warfare in Wikimedia // Automatic Control and Computer Sciences, 2014, vol.48, no.2, pp.97–108.
20. Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Zhang N, Antony A., Liu H., Murthy R. Hive – A Petabyte Scale Data Warehouse Using Hadoop / Proceedings of the 26th International Conference on Data Engineering (ICDE), 2010 IEEE, 1–6 March, 2010, Long Beach, pp.996–1005.
21. Hunt P., Konar M., Junqueira F.P., Reed B. ZooKeeper: wait-free coordination for internet-scale systems / Proceedings of the 2010 USENIX conference on USENIX annual technical conference. Berkeley, CA, USA: USENIX Association, 2010, pp.11–12.
22. Carlin S., Curran K. Cloud Computing Technologies // International Journal of Cloud Computing and Services Science (IJ-CLOSER), 2012, vol.1, pp.59–65