Advances in Intelligent Systems Research, volume 169
VIth International Workshop 'Critical Infrastructures: Contingency Management, Intelligent, Agent-Based, Cloud Computing and Cyber Security' (IWCI 2019)

Processing of Big Data Streams in Intelligent Electronic Data Analysis Systems

Anastasia Iskhakova
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russia
shumskaya.ao@gmail.com

***Abstract*—This article is devoted to a problem of design of a system of collecting and stram data processing in the virtual environment. Information in the Internet is presented in the form of diverse electronic material which for a person is today also a source of information about the outside world, both entertainment, and way of self-expression, and in connection with all above-mentioned - the tool for achievement of different purposes. The important question of assessment and the analysis of the data, extending in global information network and available to each person is brought up in the paper. The author listed distinctive features of the studied materials regarding creation of an analytical system unlike systems with lower volume of data. The components of technology of collecting and processing of larger data are offered and requirements to design and realization of final systems are made, in particular examples of their structurally functional scheme. Criteria of effectiveness are made. Researches on creation of the intelligent detector of the materials having negative impact on users, based on the offered technology are conducted today.**

Keywords—data processing, big data, the Internet, content, electronic data, data analysis, artificial intelligence, efficiency, data mining, network user, database, replication, database shard, remote access, cyberspace

I. INTRODUCTION

The data analysis is one of the most significant directions of researches today, and development of intellectual methods of digital information processing only promote this process. Creation of systems based on artificial intelligence, including self-training, is the most popular way of the solution of difficult tasks in data processing in the last decade.

It is possible to carry to tasks of the data analysis [1]:

- data minimization for reduction of non-informative arrays and storage, more optimum in terms of volume;
- receiving or allocation of new data, summary data, expansion and specification of the available knowledge;
- assessment of system functioning efficiency;
- scientific and technical forecasting;
- decision-making in various conditions.

The listed tasks are especially relevant for information systems of objects of critical information infrastructure as methods of forecasting and decision-making are capable to solve the most important problems of ensuring their information security. It is supposed that the control system adequately displays regularities of functioning and the direction of development of the operated object. Proceeding from the developed functions of management, the tree of specific objectives of management is under construction. The methods and procedures used for the solution of these tasks are analyzed, potential opportunities of improvement and modernization of the applied algorithms are defined. The information flows accepted on an object providing realization of management tasks are exposed to obligatory inspection [2-4]. In particular, creation of new methods of identification of additional knowledge about information which is transferred in the system will allow to improve the protection system due to expeditious decision-making on a possible incident.

II. THE RESEARCH PROBLEM

The author offers to consider technology of processing of big data flows in the intellectual systems of the electronic data analysis with the purpose of increase in accuracy of identification of materials having negative impact on users. This task belongs to the direction of development of methods of a thematic-focused crawling of web-space, in particular improvement of procedures of parallel processing of big data streams. Generally, the problem of crawling can be considered as a multi-purpose problem of search with restrictions where a big variety of criterion functions and the shortage of the corresponding knowledge of the search place does the task very difficult [5].

The problem of this research can be decomposed onto two components:

- research of methods of a thematic crawling of web-space;
- formation of set of approaches and technologies for ensuring stream data processing in the virtual environment with a possibility of loading of a system with difficult models and algorithms of decision-making.

The greatest practical interest in introduction of similar methodical and program complexes bears a

segment of the systems anyway connected with interaction with the person: for example, with an operator or a client (both local, and removed). In such cases the probability of putting threat due to placement of illegitimate inquiry increases, and need of quick reaction of the system also increases. The proposed solution will allow to reach higher level of efficiency in decision-making at the analysis of transmitted data and identification of potentially dangerous inquiries.

Thus the research problem can be defined as formation of the set of approaches and technologies for ensuring stream data processing in the virtual environment with a possibility of loading of a system with complex models and algorithms of decision-making. At the same time the difference of approaches from available is treated by need of interaction with third-party resources which it is impossible to operate, control their state, working capacity, and also the heterogeneity and ambiguity of contents of the studied web resources and their structures.

III. Requirements To The Subsystem Of Thematic-Focused Searching In Network (Crawling)

The stage of information content processing is preceded by not less important process of collecting of materials in the Internet. Within the real research the model operation of social processes in the virtual environment and creation of the formalized representation considering features of text and multimedia Internet-content is carried out.

Dynamics of growth and scale of web-space create a number of problems, significantly complicating effectiveness of application of the existing methods and algorithms of crawling for the purpose of identification of diverse content of the studied subject. Even indexes of such search engines as Google and Yandex cover an insignificant part of the complete volume of content in web-space. Considering this fact, it is apparent that for the multi-agent thematic-focused crawler, as one of the planned to receiving instruments of automation in the developed technology, it is required to provide an optimum functional of ranging and focusing on particular information resources.

According to the existing researches [6-9], was made the decision to use the following categories of systems of collection of information from web space:

1. The Focused Web Crawler – that is a crawler which task is in loading connected from with each other page on a concrete subject. Still such type of crawlers is called Topic Crawlers. The principle of its work is based that passing every time to the new document, this crawler checks as far as it is relevant to the designated subject, transition is carried out only on the documents corresponding to a subject. Its advantages consist that it is economic and does not demand considerable computing resources.

2. The Incremental Crawler – a traditional type of a crawler, which periodically updates the documents, collected in the storage. Besides replacement of old versions of documents on new, it can update a rank of documents, sorting them on less important and more important.

3. The Distributed Crawler is a type of the crawler which is based on the distributed calculations. As a rule it includes several computing nodes, one of which is appointed the master, and others in affiliated knots. This type of crawlers uses an algorithm Page Rank for improvement of relevance of search. The advantage of this type of crawlers is their reliability and fault tolerance.

4. The Parallel Crawler is the crawler that consists of several crawler processes, each of which works on the chosen set of data.

5. The Cross-platform Crawler is the type of crawlers that has to be established and be adjusted equally by machines with various operating systems.

The research assumes improvement of the existing methods of intellectual processing of diverse data and within development of this technology the subsystem of a krauling has to carry out primary filtration of not substantial references and parts of the imported documents. On the basis of it the following requirements to methods and systems for data collection were formulated:

1) The crawling subsystem has to use lists of keywords and phrases for search, and also for exceptions of materials;

2) The crawling subsystem has to fulfill the requirement for installation of restrictions for the number of URL received in search result and an opportunity to cut off these addresses;

3) The crawling subsystem should be cross-platform, so there has to be an equal opportunity it to configure and to configure on computing nodes with different operating systems;

4) Authorization modules with transfer of all accompanying parameters of a form and cookie values [10] are necessary, because one of the main sources of the studied content are the social networks;

5) The scalability and adaptability of parameters of productivity of processing for different volumes of the studied content are necessary properties of a subsystem. In that case if it is more data for collecting and the analysis than estimated, the crawler should give an opportunity to increase its productivity easily by providing bigger number of streams for work or adding of additional computing nodes;

6) The crawling subsystem should be integrated with the database for storage of collected information and the full text index allowing to retrieve the data quickly for the subsequent analysis;

7) It is necessary to use the crawling subsystem for data collection and vertical search as in the specified task it is necessary to take information on concrete subject domain, but not a narrow set of the facts;

8) In case of application of ready decisions on automation of process of the crawling the system has

to be supported and continuous updated. The outdate and closed projects can possess a number of the critical problems which are strongly complicating the solution of an objective.

IV. The Concept of Processing Of The Big Data Streams In Information Networks

The automate analysis of data on the Internet means the following stages:

- collecting;
- preprocessing (preparation);
- storages;
- processings;
- the direct analysis – data processing;
- adoption of analytical decisions, maintaining result for the subsequent actions and responses.

Due to the real specifics of dissemination of information on the Internet the analysis of electronic content attracts some features of processing which impose restrictions practically on each step of processing.

A. Large volumes of data

If to speak about processing of the electronic data distributed in the virtual environment, then their volume increases promptly and is expressed in exabytes (billions of gigabytes). At the same time, content of one resource is not so big and averages several megabyte. The stream analysis of electronic content of a set of virtual resources (for example, a Russian-speaking segment of the Internet – Runet) requires the solution of the certain tasks connected with the large volume of these data. Container technologies, methods of algorithms optimization due to fragmentation of the processed streams and other technologies are for this purpose applied [11].

A series of approaches, tools and methods of processing of the structured and unstructured large volumes of considerable variety of data for obtaining the results perceived by person carry to approaches to big data processing. As the defining characteristics for big data note: volume, speed, variety (simultaneous processing of various types the structured and semi-structured data), variability and value of data.

B. Heterogeneity and combination of materials

The most widespread types of virtual content are text, graphics images, video files, sound files and a diverse combination of the listed forms. Forming of optimum technology to stream data processing requires forming of the most universal approaches including recognition and categorization.

Addition of functions of recognition on the basis of a rubrication, identification of primary distinctive signs can help for unloading of further processes and more successful categorization at primary stages. It is important to add them for the improvement of quality of recognition and a categorization in addition to usual methods of recognition of content realized by means of various programming languages, for example such as Phython, JScript, Java, PHP, Perl, etc.

C. Reliability

Necessary high reliability of functioning of the database and management system database. Processing of the mentioned materials is connected with classification and use of the developed models which are stored in the special database in the set format. In connection with the large volume of data, requests to base can exceed feasible constraints and cause failures and violations of work. For solution and prevention of similar situations it is necessary to apply special technologies, for example replication, sharding [12], etc.

Replication according to the master-master (multi-master) principle is the most preferable for data processing in the Internet (Fig. 1). Scaling of systems allows to balance load on reading and writing. Balancing is made due to narrower specialization of machines in a cluster: for example, a part of nodes are assigned responsible for data-refresh while other nodes are responsible for the data choice.

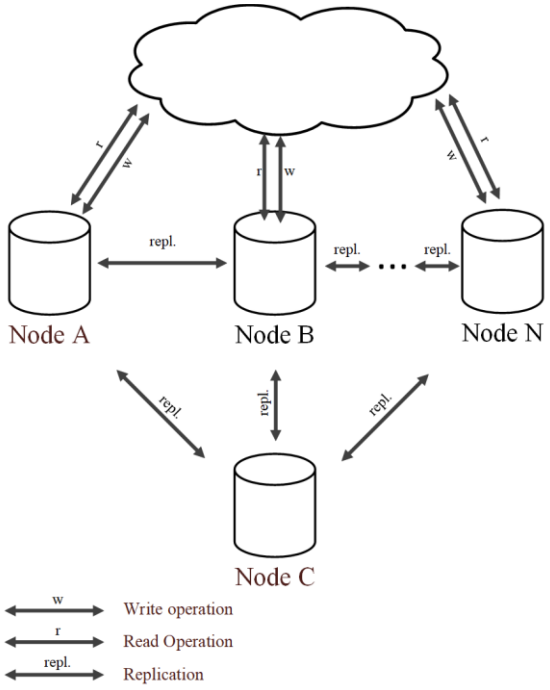

Fig. 1. The scheme of replication of the database according to the master-master principle with lack of a uniform point of failure.

The lack of a uniform point of failure allows to provide availability, to have a time reserve for resumption of work of a node and by that to create effective reliability of functioning of the database in the conditions of need of continuous processing.

The listed applied approaches have a number of considerable pluses today:

- They have the considerable practical level of development and application, are widespread thanks to the fact that data processing and

intellectual systems of search and data analysis are one of the leading spheres of information technology development now;

- Their functionality allows to combine them flexibly and functionally for the solution of difficult tasks;
- Support of the approaches realized today (relevant) is made at the high level and additionally is amplified thanks to large social community of developers and analysts of data in the uniform global environment Internet.

However, despite the high level of development of these methods, as their shortcomings it is possible to allocate lack of universal approach, complexity in system design in each case – for the solution of a specific objective and need of high skill level for carrying out these works. In this regard for the solution of the problem of cyberphysical system creation for the analysis of stream data of the Internet it is offered to create own approach based on methods of optimization of work with big data and an author's algorithm of data preprocessing for the solution of the specific objective.

V. Offered Approach

The author offers the scheme of processing of data streams during the collecting and the analysis of data from web resources. This scheme is based on the MapReduce model – the recognized decision for realization of parallel calculations over very big data sets (several petabyte) [13]. The main advantage of the model taken as a basis is the fact that the model allows to make operations of preliminary processing and convolution distributed. MapReduce has architecture of "Share-nothing". Work of MapReduce consists of two steps: Map and Reduce. On a Map-step there is a preliminary processing of input data. Then on a Reduce-step the working node receives from them answers, and on their basis forms the end result. One of the most important advantages of this construction is the fact that it allows to work reliably at platforms with low indicators of reliability. MapReduce maintains the von Neumann's principle: realization of reliable systems from unreliable elements.

The difference from classical model in the offered option (Fig.2) is the fact that it is offered to bind replication by the master-master (multi-master) method for increase in level of data accessability at quantity of flows more than several thousands. At the same time it is offered origin algorithmic solutions at the stages Map and Reduce.

A. Preprocessing stage

At the Map-stage the algorithm of preprocessing includes the analysis of web structures and a categorization of data according to a solvable task: distribution of data on the streams depending on a form of their representation and the relevant further requirement to their processing. For example, forming of a stream for processing of text these certain categories (longwise, to structure, a heading) and also information from the database for decision-making for the selected categories. Thus to flows 1, 2, 3, …, *n* will correspond data for which we accept a uniform way of processing and decision-making.

The algorithm of preprocessing (Map-function) in this case it is possible to present in the form of the sequences of steps. Input data is a set of web resources and, optionally, additional information on the uploaded data. The output data – the created strems 1, 2, 3, …, *n* (as it is designated on the scheme).

Step 1. Data uploading from each web service to the temporary storage.

Step 2. Data division on the streams depending on the category: the text, graphics images, media files, the mixed inseparable objects, etc., and also depending on properties in the called categories if it is necessary.

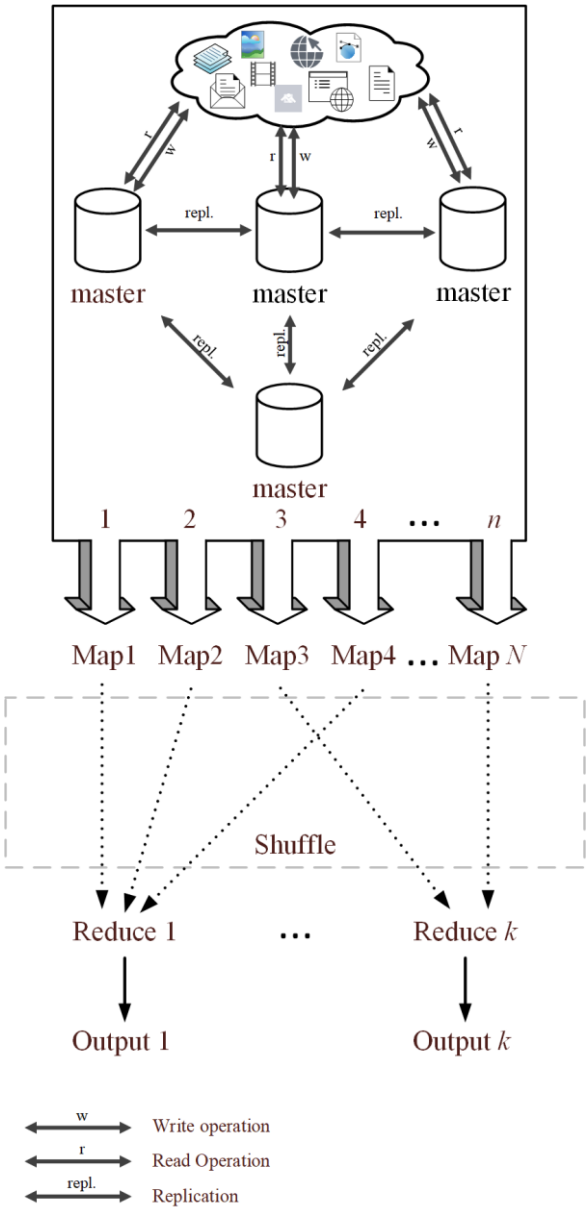

Fig. 2. T he offered scheme of distribution of data flows when collecting, processing and the analysis of information in the virtual environment (web resources).

Step 3. Addition of the streams with information from the database for processing of each one: 1, 2, 3, …, *n*.

Step 4. Forming of separate group of processing which corresponds a certain set of the pretrained models for each stream.

Step 5. Transfer of the list of streams 1, 2, 3, …, *n* on the following processing stage (output).

B. Processing stage

At the Shuffle-stage working nodes redistribute data on the basis of keys for the streams created by the Map-function. As a result all data belonging to one key lie on one working node.

At the Reduce-function execution (which is also called "convolution") working nodes process each group of results on a sequence of keys. The main node receives intermediate answers from working nodes and transfers them to free nodes for performance of the following step. The result which turned out after passing of all necessary steps is a solution of a task which was initially formulated. At a stage of implementation of the Reduce-function in this specific case the decision-making model format on the basis of the allocated streams and data stored in special storages with replication support is offered.

C. Check of results

Earlier created database of a key word and phrases on scope of detection of harmful impact on the person through virtual space (1729 elements) was used for carrying out an experiment for the purpose of approbation of the offered approach. Twelve predeterminated blog-platforms in which materials of target contents are often published participated in procedures of search. Results of the made experiment are presented in the Table 1.

TABLE I. AVERAGE VALUES OF PRODUCTIVITY AT INFORMATION PROCESSING

The used web crawler	**The spent time**		
	Loading and processing of 10,000 pages (min)	***Loading and processing of 50,000 pages (min)***	***Loading and processing of 100,000 pages (min)***
OpenWeb Spider	14.9	38.8	49.0
Apache Nutch	9.2	26.7	38.1
Scrapy	10.5	32.5	43.2

The used crawlers together with the offered original complemented data processing model showed satisfactory results of work. It should be noted that on each subsequent iteration of the analysis of web pages the model of a system of detecting is characterized by increase in productivity. The main delays of time are caused by increase in turns of data which arrive on processing, and the systems of protection against DDoS-ttacks used in the studied resources. In the provided table time for deployment and adaptation of a configuration of crawlers for each of resources is not considered. The configuration is understood as installation of external modules (web space round algorithms, parsers, methods of updating of the index, etc.). On average, control time under each resource takes no more than 20 minutes.

VI. CONCLUSION

Approach for multistream data processing of Internet resources with providing the high level of reliability and calculation capacity was offered in the article. In addition, the requirements for a solution of such tasks and the main types of crawlers which are singled out in literature today are considered.

The offered approach is a model of subprocesses of data processing using set of computing tools, such as model of distributed computing MapReduce and replication of the database on the principle of master-master (multi-master). During the check of the approach at detection of harmful impact on the person through virtual space the data of efficiency meeting requirements of an objective were obtained.

The further area of work assumes development of methodical providing, adding of subsystems of the intellectual analysis at the level of a categorization and data processing and also development of the direction of identification of Internet content in the set distinguishers in a general view.

ACKNOWLEDGMENT

The reported study was partially funded by RFBR according to the research project № 18-29-22104.

REFERENCES

[1] T. Matsumoto, W. Sunayama, Y. Hatanaka, K. Ogohara, "Data Analysis Support by Combining Data Mining and Text Mining," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, 2017, pp. 313-318.

[2] A.O. Iskhakova, R.V. Meshcheryakov, "Automatic search of the malicious messages in the Internet of things systems on the example of an intelligent detection of the unnatural agents requests," Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, Russia, 2017, pp. 85-89.

[3] B. Usmonov, O. Evsutin, A. Iskhakov, A. Shelupanov, A. Iskhakova, R. Meshcheryakov, "The cybersecurity in development of IoT embedded technologies," 2017 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, 2017, pp. 1-4.

[4] A. Iskhakova, A. Iskhakov, R. Meshcheryakov, S. Timchenko, "Analysis of the vulnerabilities of the embedded information systems of IoT-devices through the honeypot network implementation," Proceedings of the 4th International Scientific Research Conference on Information Technologies in Science, Management, Social Sphere and Medicine, vol. 72, pp. 363-367.

[5] G. Pant, P. Srinivasan, F. Menczer, "Crawling the Web", Web Dynamics, eds. M. Levene, Poulovassilis, Springer, 2004, pp. 153–178.

[6] S. Bai, S. Hussain, S. Khoja, "A framework for focused linked data crawler using context graphs," 2015 International Conference on Information and Communication Technologies (ICICT), Karachi, 2015, pp. 1-6.

[7] A.V. Patil, V.M. Patil, "Search Engine Optimization Technique Importance," 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN), Lonavala, 2018, pp. 151-154.

[8] T.V. Udapure, R.D. Kale, R.C. Dharmik, "Study of Web Crawler and its Different Types," Journal of Computer Engineering, vol. 16, issue 1, 2014, pp. 1-5.

[9] Public search engines using Nutch [Electronic source], URL: http://wiki.apache.org/nutch/ (access date: 05 March 2018).

[10] A.Y. Iskhakov, R.V. Meshcheryakov, "The information system for enterprise visitors' verification, identification and authentication utilizing modern identification features," Journal of Physics: Conference Series, vol. 803 (1), 2017, 012056.

[11] K. Ye, Y. Ji, "Performance Tuning and Modeling for Big Data Applications in Docker Containers," 2017 International Conference on Networking, Architecture, and Storage (NAS), Shenzhen, 2017, pp. 1-6.

[12] C. Roy, M. Pandey, S. SwarupRautaray, "A Proposal for Optimization of Data Node by Horizontal Scaling of Name Node Using Big Data Tools," 2018 3rd International Conference for Convergence in Technology (I2CT), Pune, 2018, pp. 1-6.

[13] W. Qing, Y. Yue, Y. Yi, W. Liang, "A method of pre-sentence text based on Map/Reduce storage and indexing classification," 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, 2014, pp. 195-199.