An Algorithm for Finding Contradictions in Multiformat Data using Apache Spark[1]

A. A. Vorobyev, Associate Professor, awa@mail.ru, **S. M. Makeev,** PhD, maksm57@yandex.ru, Russian Federation Security Guard Service Federal Academy, Oryol, 302015, Russian Federation

Corresponding author:

Makeev Sergey M., PhD, Employer, Russian Federation Security Guard Service Federal Academy, Oryol, 302015, Russian Federation
E-mail: maksm57@yandex.ru

Received on May 20, 2021
Accepted on March 17, 2022

The quality of managerial decision-making is significantly influenced by the inconsistency and heterogeneity of information obtained from various sources with the inability to unambiguously determine their reliability, for example, social networks, electronic media, opinion polls, as well as the types of representations used, for example, texts, graphs or tables. The purpose of the work was to conduct theoretical and experimental studies that ensure the choice of methods and their implementation in the algorithm for processing multiformat data to solve the problem of inconsistency and heterogeneity of information. To achieve this goal, the following tasks were solved: comparative analysis of the possibilities of methods for finding contradictions in heterogeneous information: latent-semantic analysis, neural networks and others; development of an algorithm for intelligent processing of big data using the Apache Spark module; evaluation of the algorithm's performance for obtaining a qualitative result within a given time interval. As a result of the research, in the framework of solving the problem of finding contradictions for processing media publications, it is proposed to consistently use latent semantic analysis to select articles on a given topic, and then the method of determining the tonality of articles, and for processing the results of sociological surveys, the method of calculating the integral indicator for the question selected from the questionnaire. Based on the selected methods, a multi-step algorithm was developed and then implemented in Python using the Apache Spark platform in the form of a software product registered in the Register of Computer Programs. Based on the results of the experiments conducted in the work, it was concluded that the use of the Apache Spark module with the developed algorithm makes it possible to ensure an effective search for contradictions in information with the fulfillment of the requirements for efficiency.

Keywords: *heterogeneous information, contradiction, latent semantic analysis, text tonality, search algorithm, Apache Spark*

For citation:

Vorobyev A. A., Makeev S. M. An Algorithm for Finding Contradictions in Multiformat Data using Apache Spark, *Programmnaya Ingeneria,* 2022, vol. 13, no. 5, pp. 239—246.

DOI: 10.17587/prin.13.239-246

УДК 004.9 DOI: 10.17587/prin.13.239-246

А. А. Воробьев, канд. техн. наук, доц., awa@mail.ru, **С. М. Макеев,** канд. техн. наук, сотр., maksm57@yandex.ru, Академия Федеральной службы охраны Российской Федерации, Орел

Алгоритм поиска противоречий в разноформатных данных с использованием Apache Spark

На качество принятия управленческих решений существенно оказывают влияние противоречивость и разнородность информации, получаемой из различных источников, с невозможностью однозначного определения их достоверности, например, социальные сети, электронные СМИ, социологические опросы, а также применяемых видов представлений, например, текстов, графиков или таблиц. Цель работы, результаты которой представлены в статье, — прове-

[1] The article is based on the materials of the report at the Seventh International Conference "Actual problems of Systems and Software Engineering" APSSE 2021.

дение теоретических и экспериментальных исследований, обеспечивающих выбор методов и их реализации в алгоритме обработки разноформатных данных для решения проблемы противоречивости и разнородности информации. В результате исследований в рамках решения проблемы поиска противоречий для обработки публикаций СМИ предложено последовательно использовать латентно-семантический анализ для отбора статей по заданной тематике, а затем метод определения тональности статей, а для обработки результатов социологических опросов — метод расчета интегрального показателя по выбранному из анкеты вопросу.

Ключевые слова: *разнородная информация, латентно-семантический анализ, тональность текста, алгоритм поиска, Apache Spark*

Introduction

Currently, when processing big data, many mathematical methods are used to extract useful information obtained from different data sources that are heterogeneous and diverse in structure. It provides the ability to improve the efficiency of decision support processes [1]. The paper proposes a solution to the problem of inconsistency (when some sources contradict others, and it is impossible to unambiguously determine a reliable source) and heterogeneity of information (when information from different sources: electronic media, sociological survey, etc. comes in a different form: texts, graphs and tables etc.) by choosing methods and their implementation in the algorithm for processing multiformat data. To achieve this goal, theoretical and experimental studies were carried out, including a comparative analysis of the capabilities of the selected methods, the development of an algorithm for intelligent processing of big data using the Apache Spark module, an assessment of the algorithm's performance to obtain a high-quality result in the required time.

To solve the problem considered above, an analysis of the literature was carried out [1—4], as a result of which the following methods were chosen:

- neural networks;
- latent semantic analysis.

Neural networks are a method that belongs to the class of machine learning methods, the characteristic feature of which is not direct obtaining a solution, but based on training on several examples of similar problems. During the application of the method, previously unknown and practically useful knowledge for decision support is discovered in the data. The method is capable of working with various data formats, including textual information of various structures. Neural networks need not only be programmed, but also trained based on the solution of several similar problems. Learning is one of the main advantages of neural networks over traditional algorithms [5]. Training neural networks consists in finding the coefficients of connections between the nodes of the network. During training, a neural network is able to identify complex dependencies between input and output. It means that in case of successful training, the network will be able to return the correct result based on data that was absent in the training set.

Neural networks have a wide range of applications. They can be used to build algorithms for pattern recognition, classification, clustering, approximation, forecasting, solving optimization problems and analyzing data for hidden relationships [5, 6]. To solve problems related to the processing of textual information, neural networks are used for the following [7]:

- automatic annotation;
- extraction of key concepts;
- text navigation;
- search for associations;
- determination of the tonality of the text;
- automatic selection of sets of semantically similar documents among a fixed set.

The main advantage of neural networks is their fault tolerance, that is, if a part of the neural network is damaged or deleted, only the quality of the result obtained is reduced, but not completely lost. Also, if a part of the network is disrupted by training, the neural network can be restored to its original state.

Despite the broad capabilities of neural networks, they also have disadvantages [8]:

— to build a model of neural networks, it is required to perform a multi-stage adjustment of internal elements and connections between them;

— there are problems arising in the preparation of a training sample associated with the difficulty of finding a sufficient number of training examples;

— it often takes a lot of time to complete the training procedure, which does not allow the use of neural networks in real-time systems;

— the impossibility of predicting the result of the trained neural network, since the mechanism for obtaining the output is opaque for the user.

Latent semantic analysis (hereinafter referred to as LSA) is a method of processing information in natural language, which determines the relationship between documents and concepts encountered in them, and also allows you to determine the subject of texts and classify them. This type of analysis is intended to work with any format of text information, which is relevant in the processing of multi-format data. The method is used to extract context-sensitive values of lexical units using statistical processing of large volumes of texts. LSA can be compared to a simple kind of neural network. In this case, the neural network will consist of three layers: the first layer contains a set of words (terms), the second — a set of documents under study, and the third, hidden layer, is a set of nodes with different weights connecting the first and the second layer [9, 10].

LSA uses a "term-document" matrix describing the dataset used to train the system as initial information. The elements of this matrix usually contain weights that take into account the frequency of use of each term in each document and the participation of the term in all documents.

The main advantage of the LSA method is the identification of hidden dependencies within a set of documents. Also, the method has a linear mechanism for obtaining the result, based on the singular value decomposition of the "term-document" matrix.

Among the disadvantages are the following:

— when using the LSA method, it is assumed that the distribution of terms and documents is close to normal, although most often there is a Poisson distribution [9];

— it is necessary to use a qualitatively compiled thesaurus on a specific topic, for the possibility of interpreting the natural language of a PC.

Comparative analysis of the capabilities of neural networks and latent semantic analysis for the most appropriate analysis method in the information presented in table 1.

Table 1

Comparative analysis of opportunities method of latent semantic analysis and neural networks

Capabilities		Neural networks	The method of latent semantic analysis
Ability to work with various formats of text information		+	+
The presence of a training stage, before obtaining a true result		+	–
Transparency of obtaining results		–	+
Operational data processing regardless of the amount of information		–	–
Ease of interpretation of the results		+	+
Ability to work with any amount of information		–	+
Ease of software implementation		–	+
Result:	"+"	3	5
	"–"	4	2

As a result of the comparison, the following conclusion can be drawn: to solve the problem posed in the work, the LSA method is chosen the most suitable for use, since it has the ability to work with any text formats and has a convenient mechanism for obtaining results.

Analysis of mathematical methods chosen to solve the problem of finding contradictions in heterogeneous information

The results of sociological research by RPORC [11] and Internet sites of mass media in the region [12] on the topic "The effectiveness of management of the country's regions in the socio-economic sphere" were considered as data sources in the work. To search for contradictions in the information obtained from the sources under consideration, it is necessary to analyze the data format (text, tabular, graphic) and, depending on this, use the appropriate methods and methods of information processing.

The analysis of the results of the sociological study (fig. 1) showed that the calculation of indicators of the effectiveness of management of the country's regions in the socio-economic sphere: the index of assessments of the economic situation, the index of self-assessments of the material situation, the index of assessments of the general vector of the country's development, the index of life satisfaction is carried out by subtracting the percentage of responses in two gradations (positive, average) and the percentage of answers in negative gradation, since the answers to the questions that form the indicators are presented in an ordinal scale with three gradations (positive, medium and negative).

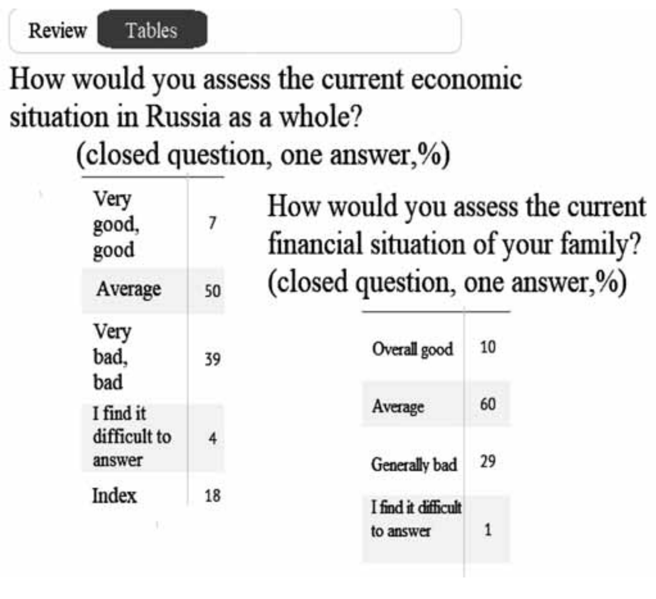

Fig. 1. Fragment of the results of sociological research

Since the document with the results of the sociological research has a clearly structured format in the form of answer tables, it is possible to compile a matrix of the form $\mathbf{W} = \mathbf{C} \times \mathbf{V}$:

$$\mathbf{W} = \begin{array}{cc} & \begin{array}{ccc} v_1 & v_2 & v_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \end{array} & \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix} \end{array},$$

where $\mathbf{C} = \{c_i\} = \{c_1, c_2, c_3, c_4\}$ — research category (lines): c_1 — "Economic situation"; c_2 — "Financial situation"; c_3 — "Country development"; c_4 — "Life satisfaction"; $\mathbf{V} = \{v_j\} = \{v_1, v_2, v_3\}$ — answer options (columns): v_1 — "Overall good"; v_2 — "Average"; v_3 — "Bad"; the option "difficult to answer" is not considered when calculating the indicator.

To calculate the generalized indicator, it is necessary to find for each category the difference between positive and negative assessments of answer options in percentage points:

$$\begin{cases} c_1 = a_{11} + a_{12} - a_{13} \\ c_2 = a_{21} + a_{22} - a_{23} \\ c_3 = a_{31} + a_{32} - a_{33} \\ c_4 = a_{41} + a_{42} - a_{43} \end{cases},$$

where c_{1-4} — research categories; a_{ij} — element of the matrix **W**, which determines the value of the number of respondents who answered a specific version of the question of the questionnaire (indicated in percentage points).

To process electronic media data (text information), it is necessary to use the method of latent-semantic analysis, the algorithm of which consists of the following stages (fig. 2) are below.

1. Pre-processing of documents:
- exclusion of stops;
- words and punctuation marks;
- exclusion of words that occur once in the text;
- highlighting in terms the stem of the word (stemming).

2. Extracting the necessary information:
- definition of terms weights;
- construction of the "term-text" matrix, showing the belonging of a particular term to a specific text.

3. Transformation of the received information: singular value decomposition of the matrix obtained in the previous step. As a result, we get the decomposition of the original matrix into three other matrices.

4. Interpretation of results: selection from a set of articles those that fit a given condition, on a given topic.

The algorithm of latent semantic analysis can be formalized and presented in the following form.

Input data:
- $D = \{d_i\},\ d = \overline{1, i}$ — a set of various media publications (texts) from several sources;
- $S = \{(s_1^+, s_2^+, \ldots, s_m^+), (s_1^-, s_2^-, \ldots, s_m^-)\},\ s^{(+,-)} = \overline{1, m}$ — set of terms.

Mass media publications are a collection of text articles of various structures. A set of terms are two given dictionaries, one with positively colored words on a given topic, the other with negatively colored words extracted from the analyzed articles, where particular indicators are reflected, and selected as a generalized indicator, depending on the index of interest in the management efficiency of the country's regions in socio-economic sphere, obtained by expert advice.

It is necessary to construct $\mathbf{A} = s \times d$ a "term-text" matrix, where the elements of this matrix are weights that take into account the frequency of using a particular term in each text from the set, as well as the presence of the term in all articles. To determine the weighting factors, the TF-IDF method is used — a statistical measure that is used to assess the importance of a term in a specific text from the analyzed set of articles. This measure consists of two components TF — a measure of the frequency of occurrence of words in the document:

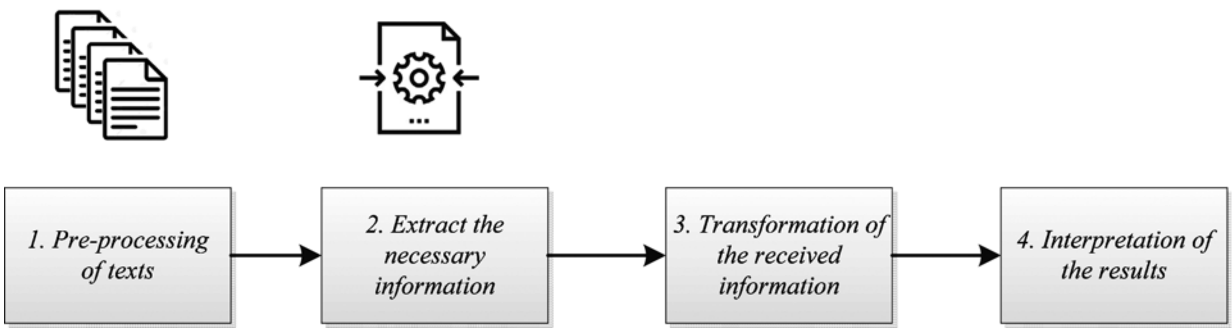

Fig. 2. Stages of the latent semantic analysis algorithm

$$TF(s, d) = \frac{n_s}{\sum_m n_m},$$

where n_s — the number of occurrences of the word s in a specific text d; $\sum_m n_m$ — the total number of words in a specific text.

$$IDF(s, D) = \log \frac{|D|}{|\{d_i \in D \mid s \in d_i\}|}, \text{ при } n_s \neq 0,$$

where $|D|$ — total number of texts in a set of articles; $|\{d_i \in D \mid s \in d_i\}|$ — the number of texts from the set D, in which the word s occurs.

The choice of the base of the logarithm does not matter, as changing the base of the logarithm will result in to change the weight of each word by a constant value.

$$TF\text{-}IDF(s, d, D) = TF(s, d) \times IDF(s, D).$$

When calculating this measure, a large weight will be given to those words that have a high frequency within a specific text (article), and with a low frequency of occurrence in other documents. After calculating the weights of terms, matrix **A** will contain rows — indexed words, columns — articles. Each cell of the matrix will indicate the number of occurrences of each word in the corresponding text.

Further, a singular value decomposition is applied to the constructed matrix **A**, which consists in decomposing the original matrix into three others: two orthogonal and one diagonal:

$$\mathbf{A} = \mathbf{U} \times \mathbf{M} \times \mathbf{K}^{\mathrm{T}},$$

where **U**, $\mathbf{K}^{\mathrm{T}}$ — orthogonal matrices, whereby $\mathbf{K}^{\mathrm{T}}$ — is transposed; **M** — is a diagonal matrix.

This decomposition is necessary to highlight the key components of the original matrix, while removing the information "noise", that is, those columns and rows, the values of which make the least contribution to the total product of three matrices [13]. Thus, the singular value decomposition makes it possible to isolate the key components of the original matrix.

After carrying out the latent semantic analysis, we get a set of articles on a given topic: $D^* = \{d_i^*\}$, $d^* = \overline{1, i}$.

Then it is necessary to prepare the selected articles to compare them with the results of opinion polls and search for contradictions. When analyzing media publications, each of the articles is considered as the opinion of the correspondent, that is, one article — one opinion. This ratio also corresponds to the results of sociological surveys, namely, one answer to a question corresponds to the opinion of one respondent. Thus, the analysis of media publications and the results of opinion polls is analyzed in two gradations: "positive" and "negative".

To refer to one of the gradations of the analyzed articles, tokenization is performed — breaking the text into words. This procedure is necessary so that each of the articles can be compared with a set of particular indicators reflecting the positive and negative characteristics of the situation in the region, thereby determining the tone of each of the articles:

$$q_i = \sum \left(s^+ \in d_i^*\right) - \sum \left(s^- \in d_i^*\right),$$

where $\sum (s^+ \in d_i^*)$ — the sum of positively colored words found in the i-th document; $\sum \left(s^- \in d_i^*\right)$ — the sum of negatively colored words found in the i-th document; q_i — assessment of the color of the i-th article.

If $q_i > 0$, then the article has a positive tone, if $q_i \leqslant 0$ — negative tone.

After analyzing articles and opinion polls, the results are compared for the selected indicators (GI_p — generalized indicator used to analyze media publications, GI_{ss} — generalized indicator used to analyze the results of sociological surveys):

- if $GI_p = GI_{ss}$, then no contradictions were found in the documents under study;
- if $GI_p \neq GI_{ss}$, then there is a contradiction in the documents under study.

Thus, to solve the problem of finding inconsistencies in information, the following actions are performed:

1) media publications are selected on a specific topic that corresponds to one of the question category from the questionnaire, that is, either "Economic situation", or "Financial situation", or "Country development", or "Life satisfaction", the tonality of these articles is determined;

2) based on the results of sociological surveys, indicators of the effectiveness of managing the country's regions in the social and economic sphere are calculated for a specific category;

3) the results of the first and second stages are compared.

General system of equations for solving the problem of finding contradictions the information is as follows:

$$\begin{cases} q_i = \sum \left(s^+ \in d_i^*\right) - \sum \left(s^- \in d_i^*\right) \\ c_i = a_{i1} + a_{i2} - a_{i3}. \end{cases}$$

Thus, a survey of methods showed the following:

- to process media publications within the framework of the problem of finding contradictions, it is necessary at the first stage to select articles on a given topic to apply latent semantic analysis; at the second stage, apply the method for determining the sentiment of articles;

* to process the results of sociological surveys within the framework of the problem of finding contradictions, it is necessary to use the method of calculating the integral indicator for the question selected from the questionnaire.

Taking into account the selected methods for solving the problem, an algorithm for searching for contradictions in information was developed [14].

Algorithm for intelligent processing of big data using the Apache Spark module

The developed algorithm has a multi-step structure, that is, the result is obtained after passing through three stages. At the first stage *A1*, the media publications are processed according to the mathematical model described above and using the method of latent semantic analysis (fig. 3).

At this stage, the data diversity and the thematic difference of the processed articles are taken into account. At the second stage *A2*, the results of sociological surveys are processed, namely, the integral indicator "satisfaction with the socio-economic situation of the region" is found. The block diagram of the algorithm is shown in fig. 4.

The third stage is the presentation of the results obtained at stages *A1*, *A2* and their interpretation by the decision-maker (fig. 5).

Results of an experimental study to assess the impact of the algorithm on the quality of the result

The quality of the result obtained by the algorithm means the search for such a number of contradictions in the information, which will correspond to the number of contradictions found by the expert.

As part of the experiment, the expert was provided for verification with test arrays of articles and documents with the results of sociological surveys of various sizes, where the presence of a contradiction was recorded in each of them. When checking these documents, the time of their checking for contradictions was measured. The measurement results are presented in table 2.

Table 2

Results of checking information by an expert for contradictions

Number of analyzed documents (volume, Kb)	Checking time, min	Found a contradiction/ not found (+/–)
10 (88)	12	+
30 (188)	35	+
50 (306)	52	+
80 (548)	90	–
150 (649)	130	–

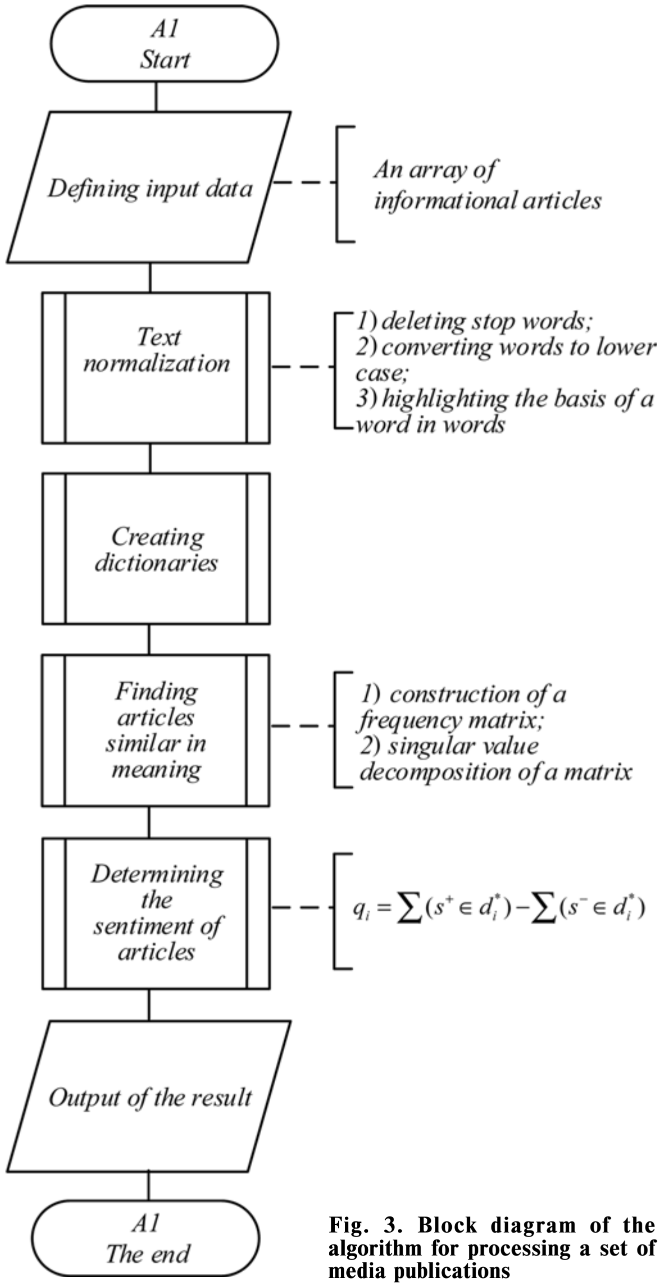

Fig. 3. Block diagram of the algorithm for processing a set of media publications

As a result, the expert verification of information showed that with an increase in the amount of information, the verification time increased, and the effectiveness of the search for contradictions decreased. The measurement results are shown in fig. 6, where horizontally reflects the number of documents being checked, and vertically — the time of verification.

Similar measurements carried out using the Apache Spark module showed the following experimental results (table 3).

As a result of the analysis of table 3, it showed that when using the module Apache Spark, test times are shortened and productivity is increased. The final presentation of the results of experiments in the form of graphs of the execution time of searching for inconsistencies in

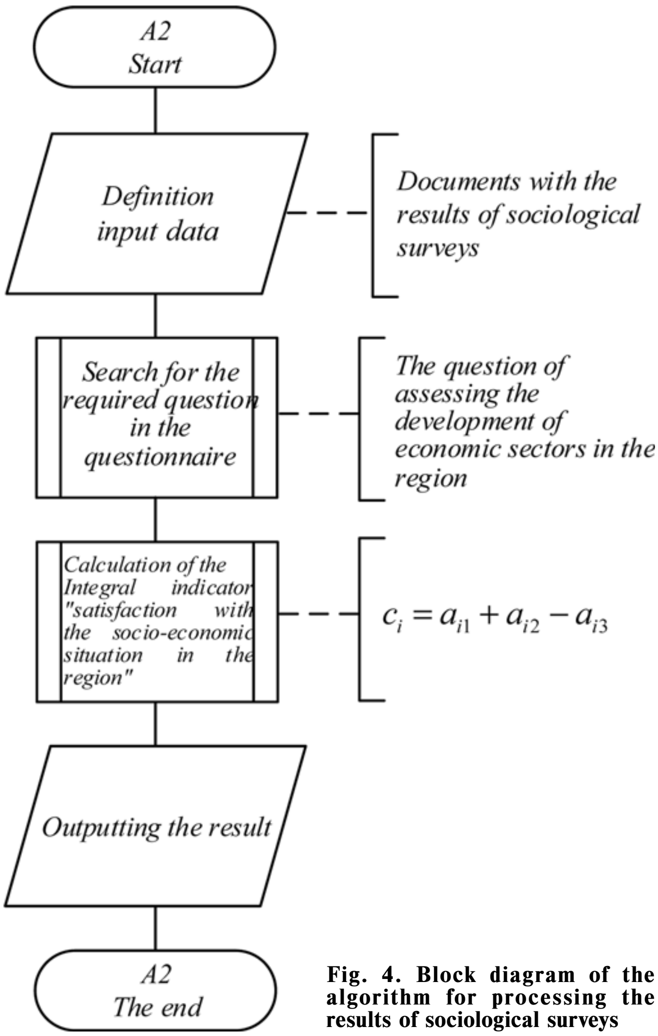

Fig. 4. Block diagram of the algorithm for processing the results of sociological surveys

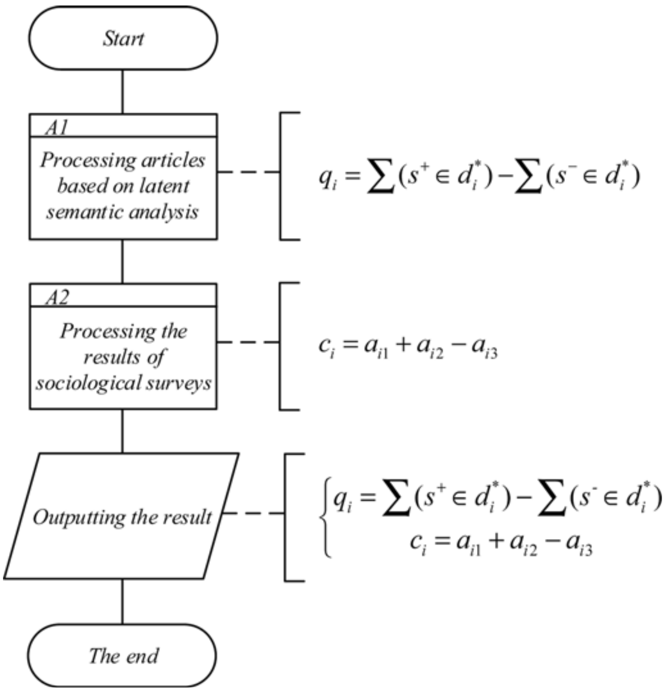

Fig. 5. Block diagram of the algorithm for processing multiformat data (results of stages *A1* and *A2*)

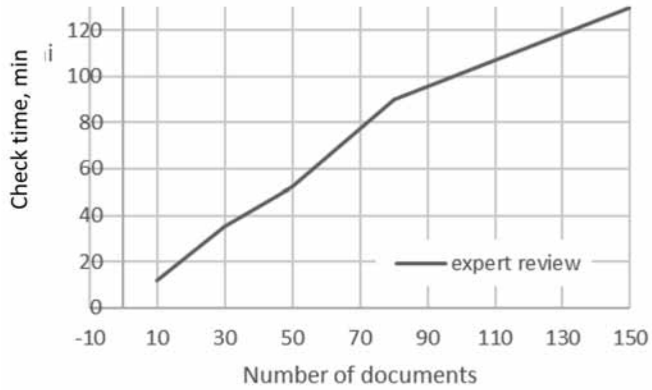

Fig. 6. The graph of the dependence of the time of checking documents on their number (volume)

Table 3

Results of checking information for inconsistencies using the module Apache Spark

Number of analyzed documents (volume, Kb)	Checking time, s/min	Found a contradiction/ not found (+/−)
10 (88)	0.2/0.0033	+
30 (188)	0.3/0.005	+
50 (306)	0.54/0.009	+
80 (548)	0.83/0.014	+
150 (649)	0.9/0.015	+

information of a fixed amount by an expert and using the Apache Spark module is shown in fig. 7.

Based on the results of the experiment, the following conclusions can be drawn:

— when using the Apache Spark module to search for inconsistencies in information, the time for checking information is reduced;

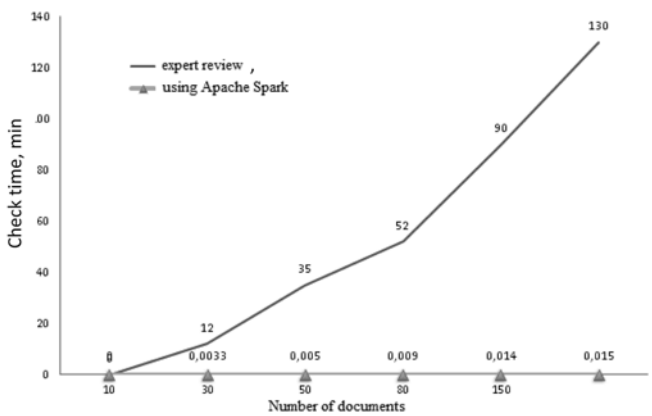

Fig. 7. Graph of comparison of the results of the execution time of the search for contradictions by the expert and using the module Apache Spark

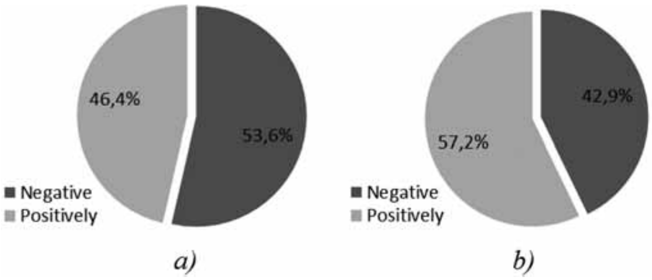

Fig. 8. The result of the algorithm implemented in the software module Apache Spark:
a — analysis of the results of opinion polls; *b* — analysis of publications

— when using the Apache Spark module to find inconsistencies, performance is improved.

The proposed algorithm was implemented in a computer program [15]. Based on the results of processing a sociological survey and media articles using the developed program, diagrams are displayed, an example of which is shown in fig. 8. As a result of analyzing the diagrams, a decision-maker can determine the presence or absence of contradictions in the information obtained from two different data sources.

Conclusions

The paper presents an algorithm for intelligent processing of big data using the Apache Spark module, which made it possible to solve the problem of inconsistency and heterogeneity of information by choosing methods (latent semantic analysis, a method for determining the sentiment of articles and a method for calculating an integral indicator for a question selected from the questionnaire).

The algorithm was implemented in a computer program registered [15].

Experimental results showed that using the Apache Spark module to find inconsistencies in information reduces test run time and improves performance.

It is advisable to conduct further research in the direction of improving the developed methods and algorithms and increasing the reliability of the data.

References

1. **Kachalov D. L., Mishustin A. V., Farhadov M. P.** Modern Methods of Big data Analysis in Large-Scale Systems, *Institute of Control Sciences of the Russian Academy of Sciences named after V. A. Trapeznikov*, 2017, no. 11, pp. 65—71 (in Russian).
2. **Ivanov P. D., Lopukhovsky A. G.** Big Data Technologies and Different Methods of Their Presenting, *Inzhenernyj zhurnal: nauka i innovacii*, 2014, vol. 33, no. 9, article 2 (in Russian).
3. **Cielen D., Meysman A, Ali M.** *Basics of Data Science and Big Data. Python and Data Science*, St. Petersburg, Piter, 2017, 336 p. (in Russian).
4. **Maggeramov Z. T., Abdullaev V. G., Maggeramov A. Z.** Big Data: problems, methods of analysis, algorithms, *Radiojelectronika i informatika*, 2017, no. 3, pp. 42—52 (in Russian).
5. **Haykin S.** *Neural networks: a complete course,* Moscow, Williams, 2006, 1104 p. (in Russian).
6. **Callan R.** *Basic concepts of neural networks*, Moscow, Williams, 2001, 287 p. (in Russian).
7. **Belous R. O., Chernyatina Yu. A.** The use of neural networks in the tasks of processing textual data. *Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki*, 2008, no. 46, pp. 28—33 (in Russian).
8. **Disadvantages** of neural network, available at: https://neuronus.com/theory/nn/240-osnovnye-nedostatki-ispolzovaniya-iskusstvennykh-nejronnykh-setej-i-puti-ikh-resheniya.html
9. **Landauer T., Foltz P., Laham D.** An introduction to Latent Semantic Analysis, *Discourse Processes*, 1998, vol. 25, no. 2-3, pp. 259—284.
10. **Bondarchuk D. V.** The use of latent-semantic analysis in the problems of text classification by the emotional color, *Byulleten rezultatov nauchnykh issledovaniy*, 2012, vol. 3, no. 2, pp. 146—152 (in Russian).
11. **Analytical** review "Social well-being of Russians: monitoring", available at: https://wciom.ru/analytical-reviews/analiticheskii-obzor/soczialnoe-samochuvstvie-rossiyan-monitoring-6
12. **Media** rating, available at: https://www.mlg.ru/ratings/media/regiona
13. **Fedyushkin N. A., Fedosin S. A., Savinov I. A.** Latent semantic analysis of the text. *Actual problems of technical sciences in Russia and abroad. Collection of scientific papers on the results of the international scientific and practical conference*, 2018, no. 5, pp. 15—17 (in Russian).
14. **Makeev S. M., Vorobiev A. A., Grushevaya E. V.** Investigation of the possibilities of using the Apache Spark module for intelligent processing of heterogeneous data, *Izvestiya Tul'skogo Gosudarstvennogo universiteta. Tekhnicheskie nauki*, 2019, no. 3, pp. 263—269 (in Russian).
15. **Vorobiev A. A., Makeev S. M., Grushevaya E. V., Mysin O. D., Shnibaev V. V.** Software complex for collecting, storing and intelligent processing of big data. Certificate of registration of the computer program RU 2019615471, 04.26.2019. Application no. 2019614183 dated 15.04.2019.