ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 166

7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)

Decision Support in Assessing the Quality of Students' Educational and Scientific Work Based on Automated Text Analysis of the Document

Natalia Minasova
Department of Computer Science
Ufa State Aviation Technical University
Ufa, Russia
e-mail: minasova@mail.ru

Sergey Tarkhov
Department of Computer Science
Ufa State Aviation Technical University
Ufa, Russia
e-mail: tarkhov@inbox.ru

Liaylia Tarkhova
Mechanics and Engineering Graphics
Bashkir State Agrarian University
Ufa, Russia
e-mail: tarkhova@inbox.ru

***Abstract*—The article describes the software product "Multifunctional Text Analyzer" (MTA), designed for informational decision support in assessing the quality of educational and scientific works. The main algorithm is based on the original method for semantic analysis of both the individual structural components of the educational and scientific texts, and the logical relationships at the substantive part of their sections. The main actions which expert performs during the process of automated analysis of the documents' substantive part are described. The software allows: to evaluate the references at the analyzed document, and quality of using references at the text of the document also; to search for information of interest to the expert contained in the document; to evaluate the quality of the document by analyzing the congruence of the texts to its sections.**

Keywords—training, assessment, semantic analysis, text mining

I. INTRODUCTION

Modern educational processes in educational institutions at various levels are inextricably linked with the processing, analysis and evaluation of a significant number of textual educational and scientific documents. As text documents we mean different types of students' works: essays, explanatory notes to projects, theses, articles, reports and etc. The analysis of the substantive part of such documents and their subsequent evaluation constitutes a considerable part of the working time of teachers and staff whose obligations include checking and processing documents. Traditional approaches to the analysis of the content of the educational and scientific work of students are largely subjective and require essential formalization in order to provide effective information support in making decisions about the work's quality. With using computer technologies, providing comprehensive information support for the processes of formalized automated analysis of text documents, it is possible to significantly reduce the time of an unproductive work performed during the analysis and subsequent evaluation of students' educational and scientific work, to improve the quality of analysis, and significantly reduce the number of errors that are inevitably arouse during the checking process.

The specificity of the textual information representation quite strongly depends on the scope of its application, as well as the semantic content of the document. Within this context, the task of analyzing is poorly formalizable and rather complicated [1, 2]. The task of analyzing and evaluating scientific documentation is a bit simpler, as its structure, composition and design are governed by certain requirements or standards. Thus, in most cases, both quantitative and qualitative criteria can be applied for evaluation. The more clearly formalized the document, the easier to develop and use a list of formal criteria, which will subsequently be evaluated when analyzing the information contained in the document.

Often, the amount of information presented in various documents is so large that comparing their semantics, even in accordance to the principle of comparing individual fragments, their frequency of mention, as well as other indicators that are amenable to a clearly formalized analysis, is a rather laborious task that requires a lot of time. The use of modern information technologies allows us to sufficiently simplify this task by applying the methods of multifunctional automated analysis of the text of a document.

Currently, two main approaches are used for analysis and evaluation of the content of documents: the traditional (qualitative) and the formalized (quantitative).

The traditional approach can be defined as a set of certain logical actions, the purpose of which in the general case is to disclose the contents of the document. Such an analysis depends on the point of view and experience of the analyst, the conditions and objectives of the research being conducted. Based on this, the main disadvantage of the traditional analysis is the dependence on the researcher, the subjectivity of the results of the evaluation of the document.

A formalized approach to the analysis of a document is based on the measurability of the characteristics of the information contained in it, for example, the frequency of use of certain terms, the number of coincidences of text fragments in two or more analyzed documents (plagiarism test) [3], etc.

Currently, various software tools and online services have been developed and are used to automate the process of documents analysis. They differ in goals and objectives. We can note programs and Internet services for semantic analysis and text checking for uniqueness, for example, Advego Plagiatus [4, 5]; internet services and programs for calculating the statistical characteristics of the texts, for example, TextAnalyze [6], tools for content analysis of texts, for example, BAAL [7]. Fast Dublicate File Finder [8] an effective tool which was designed to search and determine the degree of coincidence of the text of documents should be noted also.

One of the methods of formal quantitative evaluation of information contained in a document is content analysis, which is the subject of scientific research of both foreign and Russian scientists [9, 10, 11, 12].

The main interest of this article is the paper "Instrumental tool for assessing the quality of scientific documents" [13]. A distinctive feature of the research is authors' suggestion to use a combined approach which takes into account various categories of automatically calculated characteristics of the documents' quality, both existing bibliometric and scientometric characteristics, and types of characteristics based on semantic analysis of scientific texts. They also used technical documents, applied heuristic rules, as well as the methods for assessing the availability of direct text borrowings (plagiarism). Authors presented the experimental system which is aimed to improve the quality and accuracy of scientific documents.

This article discusses the software implementation of the multifunctional text analyzer Multifunctional Text Analyzer, as well as its underlying models and algorithms, an information support for the processes of analysis and evaluation of students' educational and scientific works (hereinafter "Documents"). In contrast to journalistic and artistic texts, such "Documents" contain clearly defined definitions and an unambiguous conceptual apparatus, as well as a fairly strict structure. They, usually, do not allow synonyms for basic concepts, which somewhat simplifies the process of semantic analysis of the text of the "Document".

II. The Program of Automated Text Analysis Document Multifunctional Text Analyzer

To solve the problem of decision-making, assessing the quality of educational scientific studies performed by students, we have developed the Multifunctional Text Analyzer (MTA) software [14]. It allows to automatically analyze the text of the "Document" and to determine the quantitative assessment of its relevant parameters (basic / basic characteristics). The main window of the Multifunctional Text Analyzer software is shown at Figure 1.

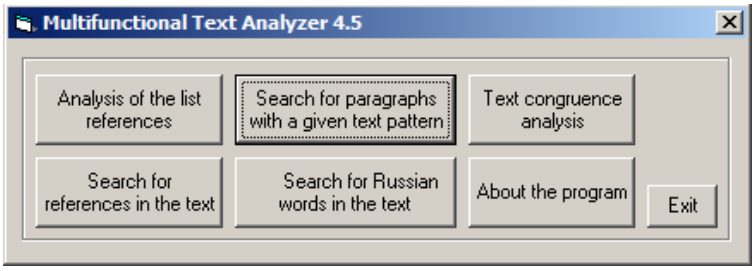

Fig. 1. Main window of the Multifunctional Text Analyzer software

Analysis of the list of references of the "Document" (over the text bibliographic references, designed in accordance with GOST 7.0.5.-2008), the implementation scheme of which is shown at Figure 2, it allows:

- to determine the number of papers for each year (the search is performed in a given date range; the following restrictions are set by default: the lower bound of the search is 1900, the upper bound of the search is the current year);
- to determine the number of conference papers;
- to search records in the list of references with a given text pattern.

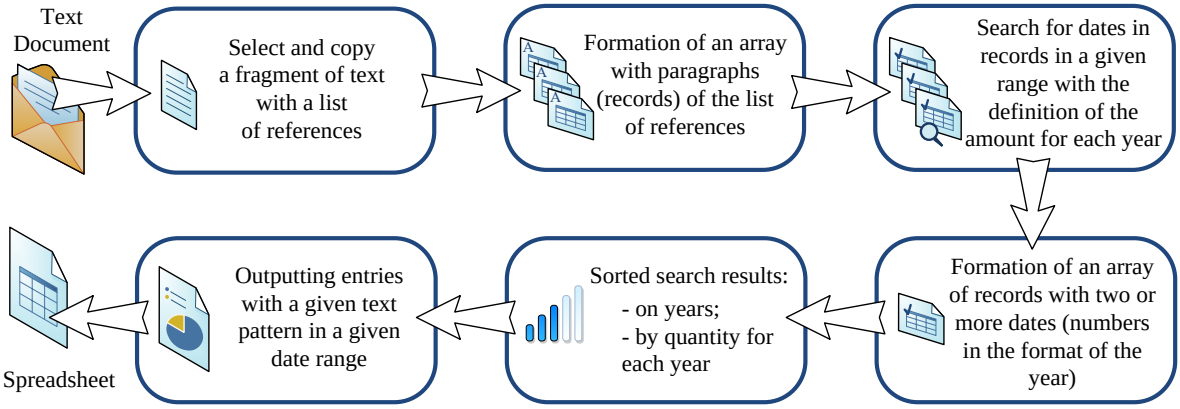

Fig. 2. The main stages of the analysis of the bibliography by year of publication

A feature of the implementation of the algorithm for the formal search of bibliographic records is that in one record there can be two or more different dates indicating the year, as the program informs the user. For example, the bibliographic reference "Tarkhov S.V. System of automated network and distance learning with multi-agent architecture // Information Technologies in Education (ITO-2003): collection of works of the XIV International Conference-Exhibition, Part III. Information computer technologies in the educational process, M., 2004. P. 288-291." contains two dates. In this case, the records can be corrected manually in the analysis software to get the correct result.

The search algorithm of the software allows not only to search the list of used literature for entries with a specified search pattern with or without case-sensitive characters in a

specific date range. The search algorithm also allows searching and displaying in the results window a list of bibliographic records with specified numbers, including searching by number range. For example, when using the search image "1.11-14.29" in the search process, we will see bibliographic records with the numbers 1, 11, 12, 13, 14, 29. Using the specified search mechanism in the text analysis process of the "Document" and by copying the bibliographic references given in brackets, we will get a compiled list of bibliographic records, allowing us to look at which works from the list of references used the author of the "Document" refers to. The results of the analysis can be saved in a text file as follows:

- date and time of the analysis;
- the first source from the list of references given in the "Document" and copied into the working window of the program for analysis;
- date range in which the analysis was conducted;
- the number of dates (in year format) found in the list of references (including several dates in one record);
- number of references to conference materials;
- number of records with two or more dates or numbers containing data in the format of the year;
- sources from the list of references with two or more dates or numbers containing data in the format of the year;
- a list of dates found in the list of references in the year format, indicating the number of records for each year and sorting the sources in the list of references by the year of publication;
- a list of dates found in the list of references in the year format, indicating the number of records for each year and sorting by the number of publications per year.

The search for references to the literature in the text of the "Document", designed in accordance with the requirements of GOST 7.0.5-2008 (references in the text in square brackets) and the analysis of the literature used in the literature list of the "Document" (the process is shown in Figure 3) allows :

- get the entire list of references in square brackets, available in the text of the "Document" and designed in accordance with the requirements of GOST 7.0.5-2008;
- determine the number of references in the text of the "Document" to each source listed in the list of references;
- identify the numbers of bibliographic references in the list of references, for which references are absent in the text of the "Document";
- calculate the values of the indicators: the total number of references to the literature, the number of unique references, the use of literature in the text of the document (as a percentage).

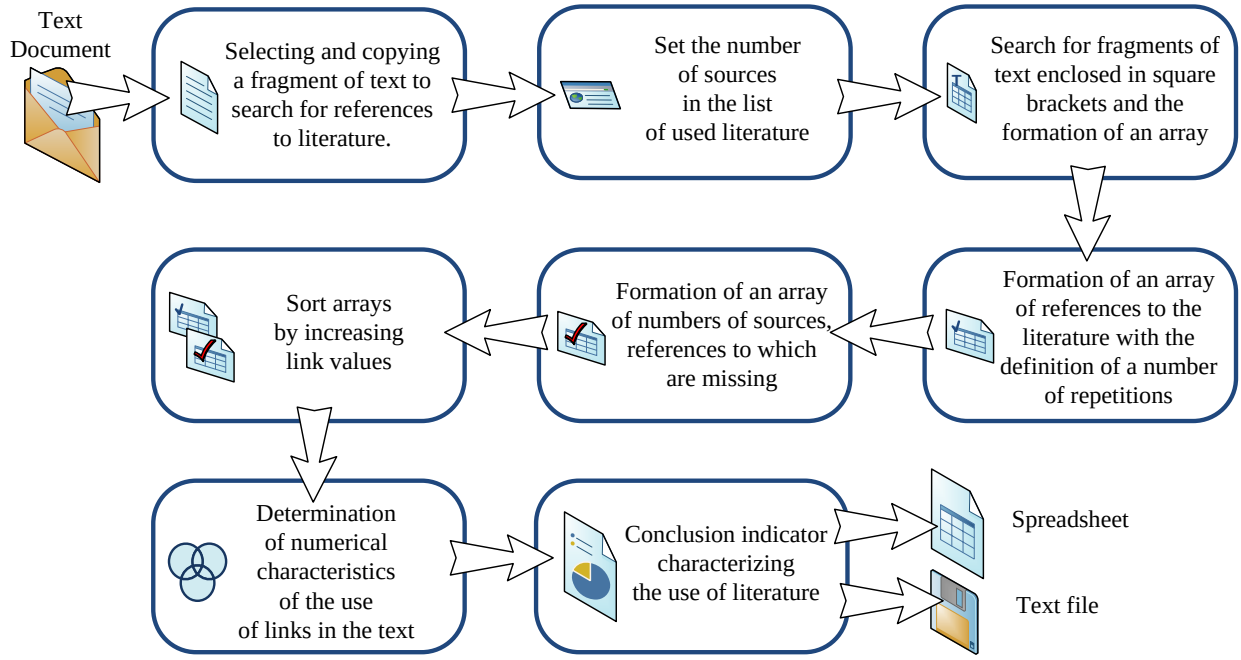

Fig. 3. The main stages of analysis avilability of references to literature in the "Document"

To analyze references to literature in the text of the "Document", the user must specify the number of sources in the literature list of the "Document". The results of the analysis can be saved in a text file in abbreviated or full form:

1. In abbreviated form are saved:

- date and time of the analysis;
- a fragment of the analyzed text with a length of 200 characters;
- the number of sources in the list of references;
- total number of links found;

- number of unique links;
- use of sources from the list of references in the text of the “Document” in percentage terms;
- numbers of sources from the list of references that were used in the analyzed text of the “Document”;
- numbers of sources from the list of references, references to which are absent in the analyzed text of the “Document”.

2. In the full form are saved:

- data stored in abbreviated form (see above);
- results of the search for links in the form in which they appear in the analyzed text (in square brackets);
- results of a search for links with the number of repeated links to the same source.

Searching for paragraphs with a given text pattern in text fragments T_i (text as a whole) from selected R sections of the “Document” with / without case-sensitive characters (Figure 4) for a given text pattern (search pattern) allows you to form text consisting of paragraphs and, if necessary, save it to a file or to the clipboard. The search image in the found text is highlighted with a symbol (group of symbols) specified by the program user.

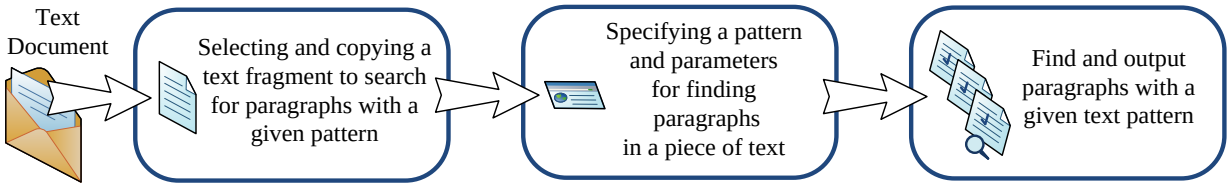

Fig. 4. The main stages of finding paragraphs with a given text pattern

Before searching for paragraphs with a given text pattern, after pasting text from the clipboard into the program's working window, the total number of paragraphs in the analyzed text is calculated. When specifying the "Register of Characters" key, the search will be performed in strict accordance with the character register of the search image, otherwise without case-sensitive. Selecting a search image in the found text with a symbol or group of symbols allows, when reading text formed from the found paragraphs, to visually find the fragment of interest to the user and to perform further informal evaluation of the content of the fragments of the “Document” being analyzed. Paragraph search results can be copied to the clipboard for insertion into the document with text analysis results. The search for Russian words (taking into account the limitation of their length) in fragments of the text (the text as a whole), from selected sections R of the document (Figure 5) allows:

- get the entire list of found Russian words and word forms, taking into account the specified limit of the minimum word length and indicating the number of each found word form;
- determine the values of indicators: the total number of words and word forms in the text, including numbers; the total number of Russian words, the number of Russian words, taking into account the length limit; the number of unique Russian words, taking into account the length limit.

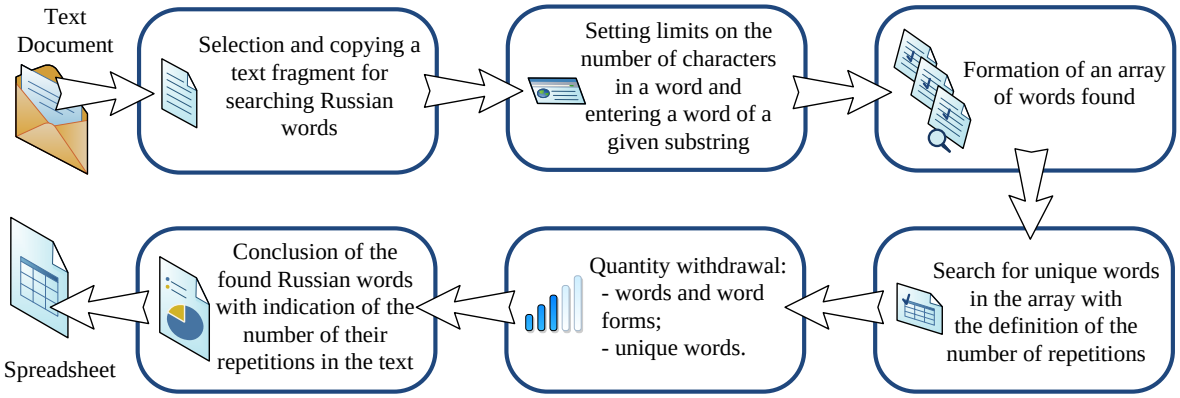

Fig. 5. The main stages search for Russian words in the text of “Document”

The analysis of the congruence of text fragments copied from the “Document” is based on the use of the algorithm for searching for the foundations of words (the part of the word representing its immutable part expressing its lexical meaning) found in the text. Stemming is used to find the basis of a word in a given source word (found word form). The Multifunctional Text Analyzer program allows you to compare up to four selected T_i text fragments, including keywords or the title of “Document” D (Figure 6). Analysis of the congruence of the text fragments of the document allows to determine:

- congruence of text fragments as the ratio of the number of matching word bases in the two compared fragments to the total number of word bases in one of them (as a percentage);
- a comprehensive indicator of congruence of text fragments (as a percentage) both with and without repeating words;
- values of indicators for each analyzed fragment of the text: the total number of words, taking into account the restriction given when searching for the length of

words; the number of unique stem words, the number of repetitions of each found word stem.

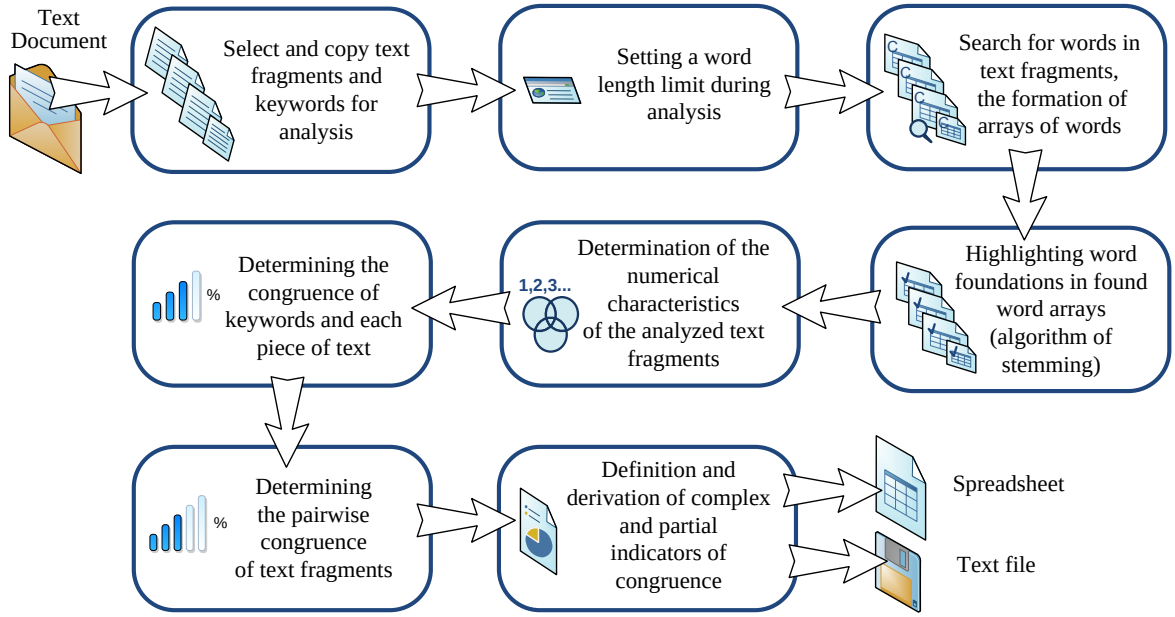

Fig. 6. The main stages analysis congruence of text fragments

The results of the analysis on the congruence of the fragments of the text of the "Document" can be saved in text files:

- in the file with detailed results of the analysis (date and time of analysis; information about the selected length of text fragments stored in the file by the number of characters; analyzed text fragments (up to four fragments) of a given length; information about the user's established minimum word length analysis of congruence of texts; for each of their three text fragments: the number of words of a given length, the number of unique word bases, the degree of similarity with keywords (in rotsentah); congruency to each pair of the analyzed text fragments as the ratio of the number of matching them bases words to the total bases of words in one of them (in percentage); complex index congruence texts (in percentage) with or without repetitions of words in texts.
- in the data analysis file, the spreadsheet contains records in the form of data lines with delimiters (date and time of analysis; information on the text analysis settings: limit on the minimum word length; the value of the key word repeat; for each text fragment being analyzed: the number words of a given length, the number of unique foundations of words, the degree of similarity with keywords; for each pair of analyzed text fragments: the number of matching foundations of words, the congruence of the analyzed texts fragments as the ratio of the number of words found in them to the total number of words found in one of them (as a percentage);
- a complex indicator of congruence of texts (as a percentage) with or without the repetition of words in texts.

During the development of the Multifunctional Text Analyzer software, considerable attention was paid to the improvement of evaluation mechanisms and the preparation of reporting documentation. A distinctive feature of the use of the software product Multifunctional Text Analyzer in the process of analysis and subsequent evaluation of the "Document" is the possibility of multi-factor text analysis. The main algorithm of the Multifunctional Text Analyzer software is based on the search and selection of words and word forms in natural and formal alphabets. So, to search for the basics of Russian words (Cyrillic alphabet), stemming is used - Porter's modified Stemmer algorithm, which is based on the use of a set of rules for the formation of words and word forms. In the process of analyzing the literature used in the "Document", the words formed on the basis of the artificial alphabet associated with the rules for the formation of bibliographic references according to GOST 7.0.5.2008 are highlighted. The mathematical model and algorithms developed in the process of creating the Multifunctional Text Analyzer software allow the user to quickly and at the same time effectively carry out a comprehensive analysis of the Documents.

III. Analysis and Evaluation of Development

The Multifunctional Text Analyzer software product is designed to support decision-making when evaluated by experts (teachers or staff) educational scientific works using the multifunctional automated text analysis of the "Document". Obviously, when evaluating "Documents" one should not completely rely on the results of automated analysis, as, incidentally, one should not rely only on the subjective assessment of an expert. To assess the quality of educational and technical educational work performed, an integrated approach should be used, which implies the use of both qualitative methods for evaluating work by an expert and quantitative methods for evaluating work using computer tools.

For a qualitative assessment of the "Document" by an expert without the use of automated analysis tools, it is advisable to use a mathematical model built on the basis of fuzzy logic methods. Such a model contains a membership

function and a ranking scale in which the actual values of the corresponding indicators and indicators are given a specific meaning associated with the assessment of work performed. An n-level linguistic scale can be adopted as a rank scale [15]. This approach allows us to bring the values of the qualitative assessment of the "Document" to quantitative indicators. If used to evaluate quantitative methods that are largely implemented by the automated text analysis tools of the Document, in particular the indicators obtained in the Multifunctional Text Analyzer software, the membership function will be identical to the actual measured value of a particular parameter on the selected measuring scale. The value of the final indicator for the entire set of parameters and indicators characterizing the estimated work in quantitative form can be calculated as a weighted sum of all measured or determined using the rank scale values of indicators.

The authors analyzed a significant amount of scientific works (explanatory notes to final qualifying works, competitive research projects, theses) using the Multifunctional Text Analyzer software. Practical use of the Multifunctional Text Analyzer software when analyzing the quality of scientific studies has shown its high efficiency. Thus, when evaluating scientific works, the expert reviewer had a real opportunity to analyze in detail the use of modern science and technology by the author based on the references in the text and references to publications from the references. The text of the analyzed work can be assessed as qualitative if the complex indicator of the congruence of fragments of the text of the work (name, purpose, objectives, novelty, conclusion (conclusions)) exceeds 70%.

IV. CONCLUSION

The practical implementation of the multifunctional text analysis method of students' educational and scientific works is considered. Authors have created the Multifunctional Text Analyzer software. Experts can use it as informational support system during the process of analysis and evaluation of students' educational and scientific works. The software allows:

- to reduce significantly the unproductive work of teachers and staff (experts), who should make the verification and evaluation of educational and scientific texts;
- to reduce significantly the effect of human factors on the results of a text evaluation;
- to determine the values of indicators, which evaluation without computer technology is impractical because of high labor costs (for example, using references in the text of the document), and in some cases it is practically impossible (analysis of text congruence).

REFERENCES

[1] D Korencic, S Ristov, J Snajder Document-based topic coherence measures for news media text / Expert systems with applications. 2018. V. 114. P. 357-373.

[2] J Misra Terminological inconsistency analysis of natural language requirements / Information and software technology. 2016. V. 74. P. 183-193.

[3] S.D. Pertile, V.P. Moreira, P. Rosso Comparing and Combining Content- and Citation-Based Approaches for Plagiarism Detection / Journal of the association for information science and technology. 2016. V. 67, release 10. P. 2511-2526.

[4] Online semantic text analysis, seo-analysis of a text. [in Russian] URL: https://advego.com/text/seo/ (дата обращения 24.02.2019)

[5] Software for the uniqueness of the text checking – Advego Plagiatus. [in Russian] URL: https://advego.com/plagiatus/ (дата обращения 24.02.2019)

[6] Text analyzer URL: https://www.textanalyzer.ru/ [in Russian] (дата обращения 24.02.2019)

[7] Content analysis of texts VAAL. [in Russian] URL: http://www.vaal.ru/ (дата обращения 24.02.2019)

[8] Fast Duplicate File Finder. URL: https://www.mindgems.com/products/Fast-Duplicate-File-Finder/Fast-Duplicate-File-Finder-About.htm (дата обращения 24.02.2019)

[9] R Jindal, Shweta A modified knowledge discovery process in the text documents / International journal of innovative computing information and control. 2018. V. 14, release 3. P. 817-832.

[10] K Zupanc, Z Bosnic Automated essay evaluation with semantic analysis / Knowledge-based systems. 2017. V. 120. P. 118-132.

[11] M.A Butakova., E.V. Klimanskaya, V.I. Yants A measure of information similarity for semistructured information analysis // Modern problems of science and education. 2013. №6, URL: www.science-education.ru/113-11307 [in Russian] (дата обращения: 20.05.2015).

[12] V.S. Simankov, D.M. Tolkachev Automatic assessment of semantic similarity of texts // Proceedingsof the XXXVII International Scientific and Practical Conference. №8 (33). Novosibirsk: «SibAK», 2014. 104 p. URL: http://sibac.info/15679 [in Russian] (дата обращения: 12.05.2015).

[13] Tools for assessing the quality of scientific documents / S.V. Gerasimov [et al.] // Proceedings of the Institute for System Programming of the Russian Academy of Sciences. M., 2013. V.24. P. 359-378.

[14] S.V. Tarkhov, N.S. Minasova, G.R. Kalimullina Certificate of state registration of computer programs No. 2015612998. Multifunctional text analyzer (MTA) / Russian Federal Service for Intellectual Property, Patents and Trademarks. M. 2015.

[15] S.V. Tarkhov, N.S. Minasova Support for decision making in the evaluation of educational scientific works on the basis of multifunctional automated analysis of text document // Proceedings of the 17th Workshop Computer Science and Information Technologies (CSIT'2015). Rome, Italy, September 22-26, 2015, vol.1. P. 186–190.