Combined Method of Text Classification

Alexander Osochkin, Vladimir Fomin, Alexsander Flegontov
The Herzen State Pedagogical University of Russia
St.Peterburg, Russia
osa585848@bk.ru , v_v_fomin@mail.ru , flegontoff@yandex.ru

Abstract— **A subject of this article is a method of classification of texts in Russian, which integrates the algorithms of frequential, morphological and intellectual data mining. The article proposes a procedure for the classification of texts with the use of frequential indices and regression trees. The results of experiments are presented showing an application of the method for recognition of individual writing styles and functional styles of articles and other publications.**

I. Introduction

There is a steady growth and accumulation of textual, semi structured information [5] in the environment of information and communication technologies and systems. The capacity of data storages (libraries, databanks, repositories, etc.) constantly increases. The need for efficient extraction of valuable knowledge from text arrays creates complexities and stimulates an emergence of new methods of information processing, such as:

- an intellectual analysis of texts (text-mining) [15], including application of resource-intensive statistical algorithms and algorithms of data mining [10],[17]:
- semantic search;
- use of network and Internet technologies, etc.

An ever-increasing need for additional capacity for data storage combined with the use of time consuming data processing algorithms (mainly due to their complexity) as consequence the cost escalation associated with the improvement of computer hardware.

The development of text-mining methods concentrates on the extraction of useful knowledge from information arrays[3], taking into account the specifics of natural language processing[4] including [20]: classification of text categorization, information extraction, referencing, data search and etc.

Text-mining methods are used in different software [21] and information technologies both as separate applications, library modules and as a part of the data mining toolkits, business analytic systems, corporate governance, etc [2]. One of the key tasks of text mining is classification of texts in natural language (NL)[14].

The growth of data arrays and associated demand for efficient extraction of valuable information from them stimulate an emergence of new methods of information processing and its complexity. It includes [17] application of intensive statistical algorithms, semantic search algorithms, neural and Internet technologies, etc. Due to the increasing volume of information arrays to be processed and ever-growing need for more complex and profound methods of analysis to be involved, the demand for more computing resources steadily increases too.

We have created a special combined method of classification for the purpose of solving the above problems. The combined method includes the step of extracting from the text numeric set of indicators, which allows to apply the classification techniques of Data mining, thereby expanding the range of possible techniques used for classification.

A main feature of the combined method of classification is the presence of a collection of algorithms which let to choose the most efficient algorithms for classification of the analyzed corpora of texts. This paper is focused on classification of corpora of text by author's style and functional style of the text. The method of decision trees was selected as the classification algorithm. Use of an algorithm of trees of decisions allows classifying corpora of text by functional styles with higher accuracy and allows to describe those rules of classification by means of "if then".

II. Background

One of the emerging trends in the struggle with increasing difficulty and cost of text processing technology is a return to classical methods of frequency-morphological analysis [22]. Text search methods and algorithms [23] focus on attempting to use a small, minimal arsenal of theoretical-linguistic methods combined with formal methods of statistical processing of simplified word forms. This tendency is traced in dynamics of scientific publication activity in classification of texts.

In our research we will base on results of the research Ben W. Medlock [1] from 1960 for 2008. The results are completed with own research (figure 1) for the period from 2008 to 2016, based on scientific Internet search engine "Base" [6].

If you look at Fig.1, the left column indicates a total number of related publications regardless of language, while the right column contains the papers published in English only. As a result, an average share of English publications between 2008 and 2016 in relation to the total number of published papers for this period is 76%. This suggests that this field of scientific activity is being actively developed by English-speaking authors. Hence, the data classification is considered to be done in English.

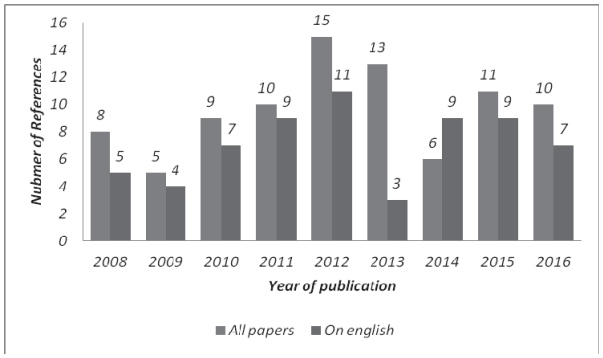

Fig. 1 Dynamics of publishing activities in the field of text-mining and NL processing between 2008 and 2016

According to result of the research was observed a decline of interest in this problem, and as a result, a decrease in number of papers in this field since 2008. In modern works, the question of practical application of technologies of classification of data and their efficiency is even more often discussed.

III. TEXT DATA CLASSIFICATION

While classifying text data based on statistical methods, an important step is the process of derivation [18] of the set of statistical indicators (see Fig.2).

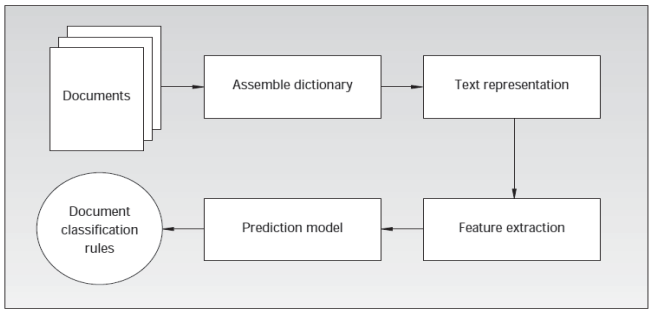

Fig. 2 A generalized classification scheme

Getting a set of indicators based on data in NL can be done by applying various procedures of analysis such as:

- Frequency analysis.
- Morphological analysis.
- Frequency-morphological analysis.

This stage of getting a set of initial indicators is the most resource consuming and important factor affecting the efficiency of the whole classification[19],[27].

Therefore, any future development in field of text mining and NL processing can benefit not only from applying highly accurate methods of classification, but also by employing the techniques that would reduce the use of computing resources for data classification.

Reduction of resource consumption for classification of data in NL can be achieved through the use of different classification methods, including those based on probability theory and well-known Bayes theorem [22],[28], as well as simplified statistical measurements [18] (frequency of occurrences of the word in the text). These two approaches demonstrate an issue of balance between resource consumption and classification accuracy.

It can be argued that there is a high potential in applying the methods based on «trees of decisions» (regression trees). The regression tree methods of text data classification are not used [17] as often as other methods. This is because there are other methods of text classification, especially the ones based on probability theory, which in some instances produced more reliable classification results.

However, there are other studies which experimentally prove the opposite [18], i.e. the comparative effectiveness of the decision tree method for text data classification. The decision tree method is based on machine learning and, as such, it requires a training sample. A significant important feature of decision trees is their ability to reduce the number of characteristics, leaving only the significant ones. Another important advantage of decision trees is a logical mechanism of interpretation and explanation of results. An application of decision tree methods allows to achieve a highly accurate classification of text data by using a minimum set of characteristics, therefore reducing the resource consumption of classification.

Consider the hypothesis that some text data classification tasks can be successfully resolved with a small consumption of computational resources through manipulation of a minimal set of frequency characteristics, which are derived based on a formalized NL set and some classification algorithms. The solution to this problem can be narrowed down to the application of various algorithms of morphological, frequency analysis to obtain a minimum data set, and the use of efficient classification method.

Reduction of resource consumption is possible through the use of frequency-morphological analysis, which allows to obtain a minimal set of theoretically-linguistic characteristics and statistically simplified word forms extracted from text data. This approach allows to reduce an algorithmic and computational workload associated with a process of text data evaluation and classification, but creates some limitations for interpretation of a semantic context of a particular data set.

The resulting set of theoretically-linguistic characteristics (derived with the use of frequency-morphological analysis) can be divided into:

- morphological word forms and their characteristics;
- Syntax features;

Syntax is a form or structure of the expressions, sentences and grammar units. Historically, a syntactic unit is more than just a word and represents a phrase or a sentence. An extraction of syntax from a text allows to parse the text and build syntax trees, which are often duplicated because sentences are based on common syntactic rules of a natural language.

The syntactic trees derived in a number of experiments [24],[25] have shown that texts in general contain no more than 3-10 different types of syntactic trees. Their variability can be a subject to a multitude of factors [7],[8],[11], such as an application of various speech styles for text writing (official, colloquial, etc.); an author's unique stylistic approach which may result in a non-conventional use of grammatical rules or formation of sentences etc. Thus, the sequential order of parts of speech allows to derive a number of different characteristics, which may display not always obvious, but nevertheless unique or specific features of a particular text cluster.

A number of sequential combinations of speech parts for each NL is different due to a varying number of parts of speech in them. However, a majority of languages on average have 10 speech parts. Thus, a quantity of various chains of parts of speech can be estimated as 10^{10}. The majority of speech parts have individual characteristics, for example, for nouns: nominative, animate, not animate, etc. Taking into account those individual characteristics further increases the need for computing resources.

IV. Combined Method of Text Classification

A combined method of text classification is a set of the algorithms, approaches, operations and tools applied to a given text data to derive its classification. The combined method is based on the idea of minimization of operations through the use of modules and replaceable components, including a possibility to extend the functions through third-party programs and libraries. A text-mining system based on the combined method consists of three blocks (Fig. 3):

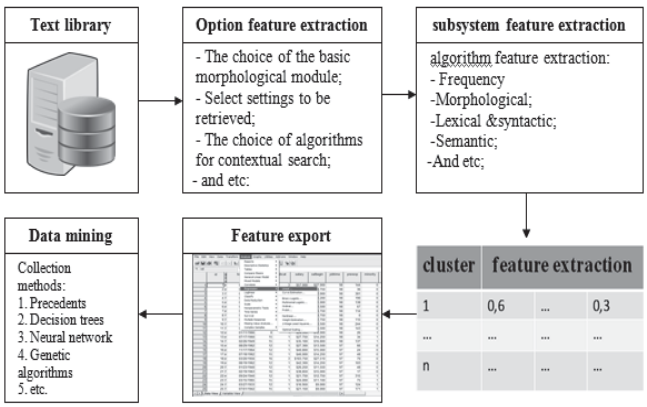

Fig. 3 Subsystems of the combined method

- ***Text library*** enabling to search, extract, catalogue and store text data. Should have a complete user interface: navigation, addition, removal, viewing, formatting, etc;

- ***Subsystem for extraction of text specific parameters*** including the frequency, morphological and syntax analysis of the text, maintenance of parameters database, etc.;

- ***Data mining subsystem*** allowing a supervised classification, clustering, identification of texts, accumulation of statistical data, generation of algorithmically compliant training samples. Should contain a number of algorithmical kits to allow the user to choose the most effective way of classification;

The key feature of the combined text-mining method is a possibility to choose the most effective algorithm of data mining (from a collection of algorithms) for classification of a given text data. The collection of algorithms isn't static; it grows in sync with relevant scientific developments and can be easily expanded to include new algorithms.

V. Frequency Analysis in the Combined Method

In the combined method of classification of the text, the most difficult analysis stages are the frequency and morphological analysis. In the implementation of frequency analysis the task was to obtain accurate frequency data, but also to optimize the process of frequency analysis, reducing the time needed for string search.

Generally the scheme of the frequency analysis in the combined method of classification of the text is submitted in the Fig. 4.

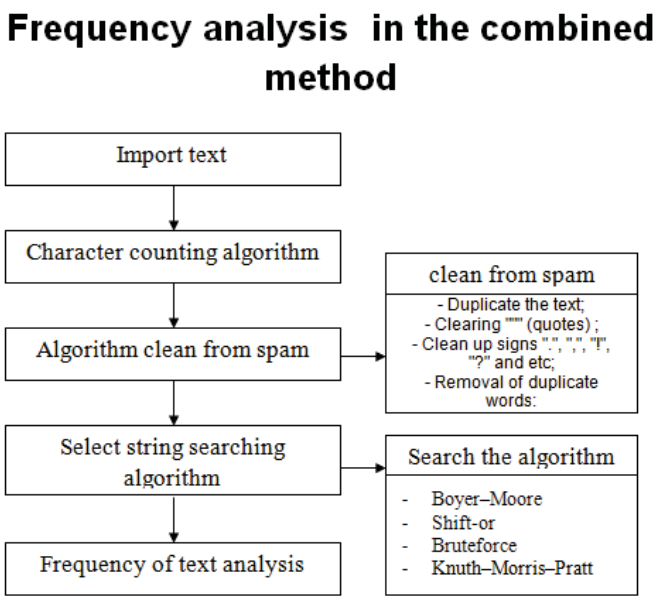

Fig. 4 Diagram of frequency analysis in the combined method

Initialization of an algorithm of the frequency analysis begins with the procedure of reading of the text. Basic frequency analysis is the algorithm of counting all symbols, determine the contents of foreign words in the text and the necessity of the use of morphological modules for other languages. The following stage is cleaning of the text of spam, meaning brackets, quotes, punctuation marks, conjoint symbol with the text, etc. This procedure allows you to clear the words from punctuation marks, thereby improving accuracy in the search of word forms in morphological analysis. In order not to destroy the original text, the algorithm creates a duplicate, which is filtered from spam.

After cleaning the text, the algorithm starts the frequency analysis itself, using an algorithm of contextual search of words.

To maximize processing speed, there is an additional subroutine enabling a contextual search through the use of two different algorithms: a Boeyr-Moore algorithm or brute force algorithm.[9]. The selection of algorithm is not random; they are selected on the basis of results of experiments measuring the speed of searching. The algorithm reads the text, if its volume is more than 800 characters, the algorithm automatically uses a Boeyr-Moore search algorithm, otherwise - brute force.

Thus, the algorithm of the frequency analysis prepares the analyzed text for the further morphological analysis.

VI. Realization of morphological analysis

A key element of the combined method of text categorization is the use of morphological analysis, the structure of which is shown in Fig. 5.

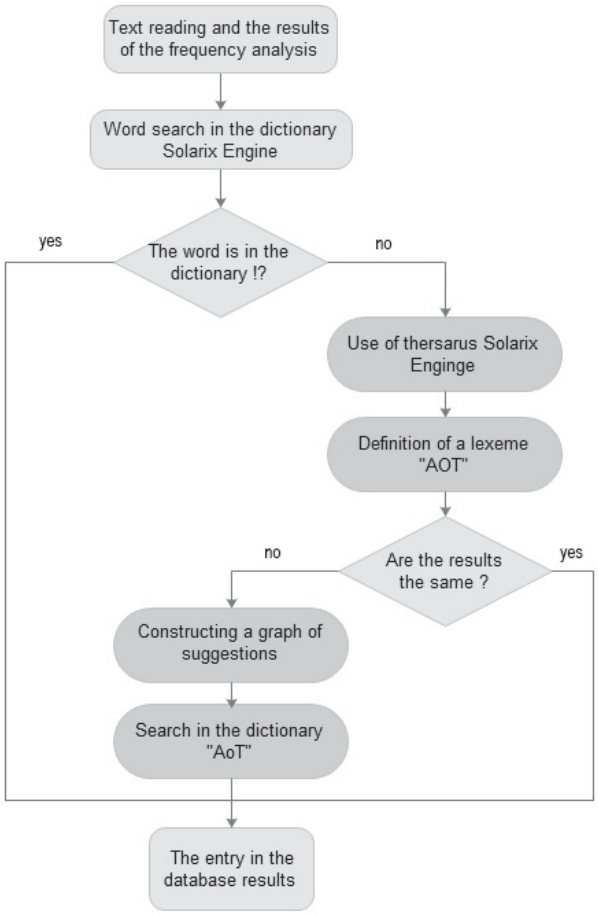

Fig. 5 Hybrid morphological algorithm in combined method

Morphological analysis is carried out by the special hybrid algorithm through the use of two built-in modules - automatic text processing or «AoT» and «Solarix Engine». Morphological module «AoT» [12] is an automatic text processing module, which contains Russian morphological dictionary with about 161 thousand words with different forms. It also includes syntactic, semantic text analysis. «Solarix Engine» [13] is the module of morphological analysis, which includes a dictionary of 1 800 000 words and 218 000 articles of thesaurus, containing information about the possible subordination relations and associative relations between words for machine learning.

A double check of parts of speech is carried out in the hybrid algorithm that allows you get a more accurate result. In cases where the results of the two modules conflict with each other, a proposal graph is built, in which the correlation between words and parts of speech is identified.

Thus, the use of the morphological analysis of two morphological modules of different types is one of the features of the hybrid algorithm. There was an experiment to evaluate the accuracy of morphological analysis. This experiment included five simple texts of no longer than 200 words, in which it was known in advance what part of speech is each word in the text. The accuracy of the analysis is determined by the formula 1:

$$Qx = \frac{\sum X_i}{\sum Y_i} \quad (1)$$

Where Xi is the number of referred words to parts of speech i by the morphological module, and Yi is the number of words, which do belong to parts of speech i.

TABLE I. Modules accuracy assessment

Text	«AoT»	Solarix Engine	Solarix Engine (with thesaurus)	«AoT» + Solarix Engine (with thesaurus)
№1	76%	76%	82%	86%
№2	73%	81%	89%	96%
№3	71%	74%	79%	93%
№4	77%	77%	86%	91%
№5	79%	68%	82%	88%
Average result	**75%**	**75%**	**84%**	**91%**

As it can be seen from the table, an average overall accuracy of the two modules is the same. It is worth noting that the «AoT» module was almost always wrong in determining gerunds, and the module «Solarix Engine» was wrong in the analysis of participles and adjectives. However, the analysis of the entire sentence as a whole (with thesaurus) allowed to increase the accuracy by 9%.

At the same time, the use of «AoT» with leksimization (bringing words into their basic form) in relation to 16% of words not identified by Solarix Engine, allowed to improve the accuracy by 7%.

VII. SET OF FEATURES

After getting the number of numeral indicators as a result of frequency and morphological analysis, the algorithm of combined method calculates relative indicators based on a set of formulas. A set of formulas contains of 76 indicators, 55 of which a user can disable or enable in settings to reduce the time required for calculation of coefficients. Table II shows some main formulas to calculate the major indicators.

TABLE II. RELATIVE INDICATORS IN A SET OF FORMULAS

№	Formula	Description
1	$D_i = \frac{k_i}{\sum_{j=1}^{n} k_j}$	Share of parts of speech (verbs, nouns, adjectives, adverbs, particles, conjunctions, etc.); a coefficient displaying a share of ***i*** "part of speech" in the total number of other "parts of speech" in the text; where «n» – quantity of "parts of speech"; «k_j» – the number of the words in the test belonging to ***j*** «part of speech».
2	$L_i = \frac{C}{K}$	An average length or averaged rate. Displays the relation of ***i*** number of information units (symbols, words, sentences) to the number of syntactically more complex structures in the text (words, sentences, paragraphs); Where «C» is the total number of units of information, «K» is the total number of words or sentences, paragraphs and etc. These indicators include: the average number of characters per word, words in a sentence, sentences in a paragraph and etc.
3	$C_{сим} = \frac{c}{K_{сим}}$	Characters share or coefficient reflecting an average number of vowels, consonants, numerals and other characters per total number characters. *c* – the number of vowels or consonants or numerals, etc., $K_{сим}$ - the total number of characters in the text
4	$N_{над} = \frac{n_{над}}{N}$	Nouns in a case or a coefficient displaying number of nouns in various cases in relation to a total number of nouns. $n_{пад}$ – noun in case , N – total number of nouns.

A possibility of calculation of a big number of coefficients increases a possibility that one of the indicators reflects a unique feature of both the object and class. Once the calculation of coefficients is finished, they are saved in Excel database, which is exported into intellectual toolkit of data analysis.

The obtained results can be exported into different data mining tools for data analysis. Three different data mining tools with multiple methods of data classification were chosen for this particular study: IBM SPSS Statistic 23, Rapid miner Studio Free 7.2, SciKit-Learn 0.18rc2.

A set of indicators obtained from the analysis of 100 original texts in Russian written in 4 functional styles is used as an input for the data mining tool with classification algorithm.

For the purpose of machine learning, a set of indicators contains a "conditional cluster" which can have 1 of 4 values, i.e. Literary, News, Scientific and Official. Also it should be noted that key assumptions and approaches for this study were as follows:

- Quantity of finite clusters equals 4, test set is 40%;
- The Euclidean distance is used for measurement of distances in a method of the closest neighbor,
- In a decisions tree method the number of iterations equals 100, quantity of the relations between the “father” and “son” tree is min 2;

The results of classification of text by different sets of methods and data mining tools are given below:

TABLE III. TEST CLASSIFICATIONS

Data mining tools	Set of methods	accuracy
Rapid miner	k-nearest neighbors algorithm	68%
Rapid miner	Decision tree	63%
Rapid miner	multi-level decision tree	95%
Rapid miner	Random decision tree	95%
SciKit-Learn	Decision tree	88%
SciKit-Learn	Linear SVM	88%
SciKit-Learn	Bayes method	87%
SciKit-Learn	NearestCentroid	50%
SciKit-Learn	k-nearest neighbors algorithm	76%
IBM SPSS	k-nearest neighbors algorithm	73%
IBM SPSS	Decision tree (method exhaustive CHAID)	98%
IBM SPSS	Decision tree (method. CHAID)	94%
IBM SPSS	Decision tree (method. CRT)	83%

For further experiments SPSS has been chosen as the main algorithm of the data mining analysis (decision tree method exhaustive CHAID). As can be seen from table №2, this algorithm provided the most accurate classification of corpora of texts, the accuracy of the classification was 98%, which is 3% higher than the accuracy of the decision trees method in the «Rapid miner» data mining tool.

Similar algorithms are widely used in data mining and possess a number of preferable benefits, including interpretability of results and the effective instrument of decrease in the set of features.

The result of the use of algorithms of "trees of decisions" is the set of rules «If <logic function> Then <class>», which easily can be verified, interpreted and not exact to computing resources.

VIII. EXPERIMENTS

In this research is made a comparative experiment to show the efficiency of the combined method of text classification. In the experiment used the same corpus for classification, similar to what was presented in table №3. The methods used in the experiment are based on the frequency and morphological analysis of the text.

Average method is the method in which the *Functional* text style is defined by the average values. As a standard we have taken the received by M. A. Zilbergleyt and A. S. Malyukevich [29] data (see the table №.4), but at the same time slightly changed them. The cluster, to which relates text, determined by approximation of the average value to the table indices. The weight of an indicator of each part of speech is identical.

TABLE IV. THE AVERAGE FREQUENCY OF THE PARTS OF SPEECH

Part of speech	Functional style			
	Literary	News	Scientific	Official
	XCp	**XCp**	**XCp**	**XCp**
Noun	0,243	0,335	0,396	0,497
Adjective	0,063	0,107	0,13	0,184
Pronoun	0,126	0,075	0,047	0,029
Numeral	0,006	0,007	0,005	0,009
Adverb	0,065	0,049	0,029	0,008
verb	0,162	0,12	0,09	0,048
Participle	0,045	0,066	0,091	0,091
prepositions	0,055	0,058	0,061	0,046
conjunctions	0,037	0,038	0,033	0,009
particles	0,158	0,13	0,09	0,071
interjections.	0,003	0	0,001	0
etc	0,006	0,007	0,008	0,002

Punctuation method is the method offered in paper of K.S. Tumanova [30]. Initially the purpose of this work was to carry out classification of text corpus by a gender-specific and age characteristics. The results of the research have shown that classification is considerably influenced by two factors: auxiliary parts of speech (conjunctions and particles) and punctuation (exclamation, question marks, etc.).

To implement this method, we added in the combined method of classification the accounting system of all punctuation and special characters.

For classification we used a corpus of text consisting from the 4 clusters (Literary, News, Scientific, Official).The results of experimentation on the classification of the corpora of text by different methods are presented in Table V.

A detailed study of the results of the classification rules and decision trees showed that punctuation is an important element in the classification of texts according to the styles, therefore, two methods, combined and Punctuation showed a higher accuracy.

TABLE V. BENCHMARK

method of text classification			
	Combined	Average	Punctuation
Literary	100%	80%	90%
News	70%	95%	100%
Scientific	100%	87%	95%
Official	100%	65%	70%
Overall accuracy	93%	81%	89%

A number of experiments was conducted on the possibility of identifying the unique author styles. Creating a set of such basic, classic styles will allow identification of a new composition. It will be possible to determine the degree of individuality of the manner of narration or conformity to the classic styles.

As initial basis the classics of Russian literature have been taken. We sought to determine how individual their styles.

Very good results were shown by experiments with the prose writers (Table VI).

TABLE VI. RESULTS OF CLASSIFICATION OF FICTION

	Clusters	**Test set 20%**	**Test set 50%**
Experiment №1	**Fiction (6 cluster)**		
	D.A Granin	75,00%	66,70%
	F.M. Dostoyevskiy	100,00%	94,10%
	A.I. Kuprin	100,00%	88,20%
	L.N. Tolstoy	100,00%	78,90%
	A.P. Chehov	100,00%	100,00%
	M.A. Sholokhov	87,50%	100,00%
	Overall accuracy	91,40%	86,70%
	Corpora of texts	36	90
Experiment № 2	**Fiction (5 cluster)**		
	F.M. Dostoyevskiy	75,00%	73,70%
	A.I. Kuprin	88,90%	83,30%
	L.N. Tolstoy	100,00%	82,60%
	A.P. Chehov	100,00%	100,00%
	M.A. Sholokhov	100,00%	100,00%
	Overall accuracy	91,20%	85,90%
	Corpora of texts	30	75
Experiment № 3	**Fiction (4 cluster)**		
	F.M. Dostoyevskiy	100,00%	73,70%
	A.I. Kuprin	100,00%	88,90%
	L.N. Tolstoy	100,00%	92,30%
	A.P. Chehov	100,00%	93,80%
	Overall accuracy	100,00%	93,30%
	Corpora of texts	24	60

Based on the results of the experiments No. 1-3 we can say that we found the individual style of the narration.

As the results of the use of decision trees, set of rules "If <logic function> THEN <class>", clearly reflect the individuality of style of the narration. These styles are rather individual, but have also common features. It explains reduction of accuracy of classification in case of increase in quantity of clusters.

Objects from the corpora of text are various fiction works of the author, which are written at different periods of his work in different genre and have different actors, etc.

In case of such distinctions, classification by keywords (for example, the TF-IDF method) will make smaller success, than combined methods of texts classification. Therefore the obtained data in case of classification of the fiction have the scientific value.

We have several times repeated experiments with random chosen fiction and received approximately the same results. Accuracy of classification changed in the range ± 3% of the values presented in the table № 6.

The experiments with poets showed a problem with classification of the authors poetic styles. An example of one of the experiments presented in Table № 7

TABLE VII. RESULTS OF CLASSIFICATION OF POETRY

	Poetry	accuracy
Experiment №4	A.A.Block	44,00%
	A.S. Pushkin	50,00%
	M.V. Lomonosov	16,70%
	M.Y. Lermontov	76,50%
	N.A. Nekrasov	66,70%
	F.I. Tyutchev	56,30%
	Overall accuracy	**54,50%**
	Corpora of texts	120
Experiment № 5	**Poetry**	**accuracy**
	A.A.Block	34,80%
	A.S. Pushkin	54,50%
	M.V. Lomonosov	60,00%
	M.Y. Lermontov	86,70%
	N.A. Nekrasov	53,80%
	Overall accuracy	**54,90%**
	Corpora of texts	100

The results of first experiments showed the low reliability of classification and, as a consequence, the difficulties of identifying the individuality of poetic style, even the classics of literature.

As can be seen from the results of just reducing the cluster size is not much increased accuracy. We conducted another experiment with the decreasing number of clusters. We have conducted experiments with decreasing number of clusters. In experiment №6 we removed the author who got the lowest accuracy in experiment № 5. In experiment №7-8, we removed authors with a very similar style of Pushkin and Lermontov.

TABLE VIII. RESULTS OF CLASSIFICATION OF POETRY WITHOUT SIMILAR STYLE

	Poetry	accuracy
Experiment № 6	A.S. Pushkin	47,40%
	M.V. Lomonosov	71,40%
	M.Y. Lermontov	60,00%
	N.A. Nekrasov	72,20%
	Overall accuracy	60,90%
	Corpora of texts	80
Experiment № 7	**Poetry**	**accuracy**
	A.A.Block	82,10%
	A.S. Pushkin	96,40%
	M.V. Lomonosov	82,10%
	N.A. Nekrasov	95,80%
	Overall accuracy	88,30%
	Corpora of texts	80
Experiment № 8	**Poetry**	**accuracy**
	A.A.Block	69,80%
	A.S. Pushkin	96,60%
	M.V. Lomonosov	86,70%
	N.A. Nekrasov	78,20%
	Overall accuracy	83,00%
	Corpora of texts	80

Based on the results received in experiments № 6-8 we can say that classification of the author's poetic styles is very difficult. Complexity is caused by strong similarity of some authors, and also specifics of a poetic form.

According to the results of the numerous experiments can be concluded that in finding the individuality of the author style poems showed the worst result. Bad results of classification of poems are the consequence of the similarity of the authors' styles, which wrote in the same epoch and the special structure of the poem.

IX. CONCLUSION

Through the use of integrated technology of frequential, morphological and data mining analyses for text analysis, the study allowed to reveal some common patterns (or regularities) which can be used for classification of weakly-formalized information. The achieved reliable results of classification of texts written in different functional styles suggest the possibility to use the proposed technology as one of the effective instruments for the analysis of natural language information.

The proposed procedure for text analysis is limited by full processing of texts only in Russian, which makes it narrowly specialized. However, the use of frequency and morphological analysis combined with the flexibility of the set of formulas, short execution time and the possibility to export the results of analysis into data mining tools allow to employ an impressive number of classification methods. Thus, the prospects of expanding of the suggested technology into analysis of other European languages are quite promising.

REFERENCES

[1] B.W. Medlock, *Investigating classification for natural language processing tasks*. Cambridge: University Cambridge 2008, pp.138.

[2] F.Thabtah, P.Cowling and Y.Peng , " MMAC: A New Multi-Class, Multi-Label Associative Classification Approach", *In Proceedings of the 4th IEEE International Conference on Data Mining* (ICDM'04), Brighton, UK, 2004, pp. 217–224.

[3] S.Bird, E.Klein and E. Loper, *Natural Language Processing with Python*. Sebastopol: O'Reilly Media, 2015, pp. 479.

[4] M.Z.Islam, A.Rahm, R, Mehler, "Text Readability Classification of Textbooks of a Low-Resource Language" *26th Pacific Asia Conference on Language, Information, and Computation*, 2012.

[5] D.Spergel, *Organizing information: Principles of data base a. retrieval systems*. Orlando: Acad. press, vol. 14, 1985, pp. 450.

[6] Internet papers base «Base» Web: https://www.base-search.net

[7] M.F.Caropreso, S.Matwin, F.Sebastiani "learner-independent evaluation of the usefulness of statistical phrases for automated text categorization" Published in book *Text databases and document management: Theory and practice*. Virginia: Virginia Commonwealth University 2001, pp. 78-102.

[8] S.Ibrahim., K.R. Chandran, "Compact Weighted Class Association Rule Mining using Information Gain", *International Journal of Data Mining & Knowledge Management Process* (IJDKP), vol.1, no.6, 2011. pp. 1–13.

[9] C.M.Bishop, "Pattern Recognition and Machine Learning" New York: Springer, 2006, pp. 729.

[10] S.Weiss, I.M. Kaufmann "Predictive Data Mining: A Practical Guide", San Francisco, 1998, pp. 228.

[11] J.Read, B.Pfahringer, G.Holmes, "Multi-Label Classification using Ensembles of Pruned Sets", *8th IEEE International Conference on Data Mining*, 2008, pp. 995 -1000.

[12] Official website of the program automatic text processing «AoT», Charter Russian morphological dictionary, Web: http://aot.ru/index.html

[13] Official website of the program Solarix Engine, charter "Computer Russian grammar" Web: http://www.solarix.ru/index-ru.shtml

[14] D.Manning, H.Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 2000, pp. 680.

[15] K.P Murphy, *Machine Learning A Probabilistic Perspective*. Cambridge: MIT press, 2012, pp. 997.

[16] C. Mengen, "Short Text Classification Improved by Learning Multi-Granularity Topics", *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2012, pp. 1176 – 1781.

[17] S. Dumais, H. Chen. "Hierarchical Classification of Web Content", *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 256-263.

[18] M.S. Weiss, "Maximizing Text-Mining Performance", M.S. Weiss and etc. - *Intelligent information retrieval*, 1999, Vol.14, pp. 63-69.

[19] N. Archakova, "Extraction of Low-frequent Terms from Domainspecific Texts by Cluster Semantic Analysis", *proceeding of the ismw-fruct 2016 conference*, pp. 86- 89.

[20] A. Kao, S.Poteet, *Text Mining and Natural Language Processing.* New York: Springer, 2006, pp. 277

[21] A.Srivastava, M.Sahami, "Text mining Classification, Clustering, and Application", *Minneapolis: Department of Computer Science and Engineering*, pp. 278.

[22] O.Durandin, "Using Probability Distribution over Classes in Automatically Obtained Training Corpora" O.Durandin N.Hilal, D.Strebkov, N. Zolotykh, *proceeding of the ismw-fruct 2016 conference*, p.90-93

[23] C.Aggarwal, *Data Mining: The Textbook.* New York: Springer, 2015,– pp. 734.

[24] O.Hellwig, "Improving the Morphological Analysis of Classical Sanskrit", Düsseldorf University, Web: http://aclweb.org/anthology/W/W16/W16-3715.pdf

[25] L.Xing, "The Multi-Tree Cubing Algorithm for Computing Iceberg Cubes", L.Xing, J.Howard, k.Hamilton, Karimi, Liqiang Geng, *Journal of Intelligent Information Systems (JIIS)*, Vol. 33 ,2009, pp. 179-208.

[26] Y.Adaskina, "STOPKA: Unbalanced Corpora Classification by Bootstrapping", Y.Adaskina, A.Popov, P.Rebrova. – *proceeding of the ainl-ismw fruct 2015 conference*, pp. 141 – 143.

[27] D.Lewis, M.Kaufmann, "Feature Selection and Feature Extraction for Text Categorization", *Proc. Speech and Natural language Workshop*, San Francisco, 1992, pp. 212.

[28] H.W.Press *Numerical Recipes in C.* H.W. Press, S.A. Teukolsky, W.T.Vetterling and B.P.Flannery, Cambridge: Cambridge University Press, 1992, pp.1018.

[29] M.A.Zilbergleit, A.S.Maliukevich "Assessing the possibility of using morphological analysis of text for bold style", *Proceedings BSTU*, vol. 9, 2012. pp. 93-98.

[30] K.C.Tumanova "algorithm of russian text classification by author's gender and age", st.Peterburg State University Mathematics and Mechanics Faculty, 2011 (Graduate work).