ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2017, Vol. 51, No. 3, pp. 120–126. © Allerton Press, Inc., 2017.
Original Russian Text © I.V. Selivanova, B.Ya. Ryabko, A.E. Guskov, 2017, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2: Informatsionnye Protsessy i Sistemy, 2017, No. 6, pp. 8–15.

INFORMATION ANALYSIS

Classification by Compression: Application of Information-Theory Methods for the Identification of Themes of Scientific Texts

I. V. Selivanova[a], *, B. Ya. Ryabko[b, c], **, and A. E. Guskov[b, c], ***

[a] *The State Public Scientific Technological Library, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 123298 Russia*

[b] *Novosibirsk State University, Novosibirsk, 630090 Russia*

[c] *Institute of Computational Technologies, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia*

**e-mail: selivanova@ict.sbras.ru*

***e-mail: boris@ryabko.net*

****e-mail: guskov@spsl.nsc.ru*

Received February 3, 2017

Abstract—A method for automatic classification of scientific texts based on data compression is proposed. The method is implemented and investigated based on the data from an archive of scientific texts (arXiv.org) and in the CyberLeninka scientific electronic library (CyberLeninka.ru). Experiments showed that the method correctly identified the themes of scientific texts with a probability of 75–95%; its accuracy depends on the quality of the original data.

Keywords: classification, thematic classification of texts, information theory, text compression, arXiv.org, CyberLeninka

DOI: 10.3103/S0005105517030116

INTRODUCTION

The problem of automatic (without human participation) classification of scientific texts (texts, books, etc.) is of great practical interest, since the recent decades have been characterized by an avalanche-like increase in the number of scientific texts. Every year so many new scientific texts appear in many fields of science that specialists are not able to read them all. Under these conditions, information support for research becomes particularly important and requires a preliminary classification of new scientific texts in order to identify those that are of interest to a particular scientist. The existing scientific text-classification methods are based on the work of experts; they are expensive and hardly applicable. Therefore, the development of automatic methods is important and attracts the attention of linguists, as well as experts on artificial intelligence and in such relatively new disciplines as *machine learning, data mining,* and *Big Data.*

Despite numerous studies, the problem of constructing effective scientific-text classification methods is still far from being solved. Let us consider how researchers approach this problem in more detail. One of the most popular methods for the classification of scientific texts is the construction of word occurrence vectors [1, 2] and its modifications. As an example, in [3], it was suggested to consider an N-gram as a component of the word occurrence vector rather than an individual word.

This technology makes it possible to use different methods. In some cases, a measure of proximity between vectors is calculated; classification based on graph theory is then applied [4]. The naive Bayesian classification [5], decision trees [6], neural networks [7], k-nearest neighbors [8], etc. are often used.

This paper describes our method for automatic scientific text classification based on information-theory methods. The basic idea is quite simple and natural: in texts that relate to one field of science many common concepts, terms, and patterns are used; the narrower the scientific field is, the "closer" the vocabulary of texts that relate to it is. This obvious observation is widely used in many text classification methods (for example, identification of keywords and phrases, etc.). However, unlike other methods, we propose to evaluate the degree of lexical proximity using text compression by so-called *archivers* or *data comressors.* Previously, a similar approach was used in solving problems of classification and determining the authorship of texts [9–17]. However, this method was not used to classify scientific texts.

METHOD DESCRIPTION

Let there be n scientific fields: $X_1, X_2, \ldots, X_n$. For each of these fields, the *core,* a set of texts typical of this cat-

egory (hereinafter, training sample), is defined. The core of the first category we denote by $x_1^1, x_2^1, \ldots, x_{m_1}^1$, the second, by $x_1^2, x_2^2, \ldots, x_{m_2}^2$, and the nth, by $x_1^n, x_2^n, \ldots, x_{m_n}^n$. Let there be a scientific text y that belongs to one of these fields. The proposed method based on the training samples should determine which field the text y belongs to.

In order to solve this problem, we will use the archiver φ for the compression of any set of texts. By $K(u_1, \ldots, u_m)$ we denote the length of texts $u_1, \ldots, u_m$ represented as computer files compressed using the method φ. We now consider a case where some text file u is compressed together with $u_1, \ldots, u_m$ and the length of compressed texts $K(u_1, \ldots, u_m, u)$ is calculated. Note that the order of the texts $u_1, \ldots, u_m, u$ compressed by the archiver is important, i.e., u must be compressed after $u_1, \ldots, u_m$. It is clear that $K(u_1, \ldots, u_m)$ will be greater than $K(u_1, \ldots, u_m)$, since in the first case the number of compressed files is larger, u is added. We denote the difference between the length of the encoded files $u_1, \ldots, u_m, u$ and $u_1, \ldots, u_m$ by $K(u/u_1, \ldots, u_m)$:

$$K(u/u_1, \ldots, u_m) = K(u_1, \ldots, u_m, u) - K(u_1, \ldots, u_m). \quad (1)$$

Here, the file u is encoded after $u_1, \ldots, u_m$ and when it is compressed, the archiver uses information about the frequencies of letters and words as well as other regularities and features of text files $u_1, \ldots, u_m$. Since $K(u/u_1, \ldots, u_m)$ depends on the files $u_1, \ldots, u_m$, it will be smaller, the more the file u is similar to $u_1, \ldots, u_m$. The idea of the method is quite simple: to assign a text file u, whose category is unknown, to the group of texts with which it is compressed best. More formally, the proposed method looks as follows: for texts from each scientific field X_i, $i = 1,\ldots n$ we calculate $K\left(y \middle/ x_1^i, x_2^i, \ldots, x_{m_i}^i\right)$ and select such j that $K\left(y \middle/ x_1^j, x_2^j, \ldots, x_{m_j}^j\right)$ is at the minimum. We then assume that the text y belongs to the field X_j. Formally,

$$K\left(y \middle/ x_1^j, x_2^j, \ldots, x_{m_j}^j\right) = \min_{i=1,\ldots,n} K\left(y \middle/ x_1^i, x_2^i, \ldots, x_{m_i}^i\right). \quad (2)$$

THE DATA AND DESIGN OF THE EXPERIMENT

General Description of Data

The proposed method was implemented and investigated based on the data presented in two repositories: the English archive of scientific texts (arXiv.org) and the CyberLeninka Scientific Electronic Library (CyberLeninka.ru).

The arXiv.org archive contains an archive of scientific texts on physics, mathematics, biology, statistics, computer science, and financial mathematics. For each of these sections, fields of science are identified (for example, physics is divided into astrophysics, accelerator physics, atomic physics, and nuclear physics). When the text is published on a website, the author indicates one or more fields to which their work belongs. Thus, all archived texts are classified (i.e., referred to fields) by the authors themselves, which makes it possible to verify the quality of our method. In addition, the texts posted on arXiv.org are publicly available, presented in standardized form, and almost all in one language (English), which simplifies their processing. The first field of science indicated by the author will be called the main one; the rest are secondary. For the experiment, we selected 30 fields of science related to physics, mathematics, biology, or computer science. It should be noted that one field of science can refer to several sciences: for example, mathematical physics refers to both physics and mathematics and information theory refers to mathematics and computer sciences.

Classification of papers in CyberLeninka is based on the Code of State Categories Scientific and Technical Information (GRNTI). We selected 20 out of more than 80 categories: astronomy, automatics, biology, ecology, economics, philosophy, physics, geophysics, geography, chemistry, history, cybernetics, literature, mathematics, medicine, mechanics, politics, psychology, sociology, and jurisprudence). These categories were selected because they contained more than 300 files in Russian, which made it possible to conduct uniform experiments with the obtained data.

Data Processing

Full texts of the papers were received in the PDF format. Text layers were extracted from each of them using the PDF2Text Pilot software (http://www.colorpilot.com/pdf2text.html). Punctuation marks, numbers, and symbols that appeared after the conversion of formulas and stop words were removed. Russian texts from CyberLeninka were stemmed (the process of bringing the words to their initial form) using Yandex mystem (https://tech.yandex.ru/mystem/?ncrnd=4660). Files of less than 10 kb were deleted.

Compression Dynamics Test

In order to test the method, let us define the compression dynamics test (CDT). Take the core of one of the fields of science and break it into groups of files in the core: 0, 1, 2, 3, etc., where each previous group becomes a subgroup of the next one.

Next, we take the test file and compress it sequentially with each of the selected groups. In the end, we consider the dependence between the number of files in the core and the compression ratio (the smallest compression ratio corresponds to the best compression). In the case of the correct operation of the method, it is expected to see a decreasing curve, i.e.,

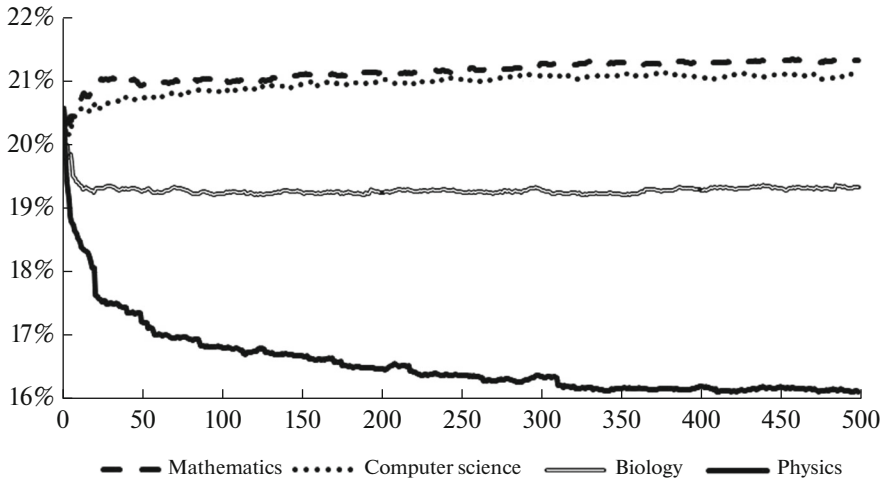

Fig. 1. The dependence of the compression degree of the test from Physics on the number of files in the core and other sciences.

the fewer the files are in the core, the greater the compression ratio is. However, when the dictionary is filled with the required number of terms, the curve decrease rate should slow down.

Archiver Selection

We selected the archiver that is the most suitable for our purposes. To do this, a series of experiments was carried out on the WinRAR (http://www.win-rar.ru), 7z (http://www.7-zip.org), and PeaZIP (http://www.peazip.org) archivers using the compression dynamics test. Different algorithms (PPMd, PPMd and LZMA, Deflate and BWT, respectively) were used in these archivers. All archivers were tested for different values of parameters. The WinRAR archiver with a maximum memory of 128 MB (memory is an archiver parameter) was selected based on experimental results; later it was used for all of the experiments.

Investigation of Method Properties

Let us describe an experiment that shows that the proposed method makes it possible to correctly determine the science to which a text belongs. We took the cores from four sciences presented in arXiv.org (physics, biology, mathematics, and computer science) and a test file from physics. We then conducted a compression-dynamics test.

Figure 1 shows that the curve corresponding to the compression of the test file from physics with the core from the same science travels below the others, i.e., the test shrinks best with the science it corresponds to.

Let us now show that the proposed method works for narrower disciplines. We took a test file from the subject area astro-ph.GA (Fig. 2) and carried out a test for other categories within Physics. We obtained the result that the curve that corresponds to the test of compression dynamics of the astro-ph.GA test file and the astro-ph.GA core goes below the rest. It can also be seen that just above it there is a curve that corresponds to another field of astrophysics, i.e., the method correctly identifies the field even from two terminologically close disciplines.

Training Sample Size

The text compression degree depends on the size of the core, which is shown in Figs. 1 and 2. Next, we define the number of texts in training samples, i.e., the value of the parameter m_i in $x_1^i, x_2^i, \ldots, x_{m_i}^i$ for X_1, X_2, ..., X_n (see the method description). The dependence of the number of errors on the number of files in the core is shown in Fig. 3. In total, the test was carried out in 450 runs, that is, on 15 test files of each category that is not included in the core.

It can be seen that the classification quality depends significantly on the number of files in the core. Moreover, if the training sample includes several dozen files, then the error is large, while if the core consists of 300–500 files, further increase of the sample size does not affect the quality greatly. Therefore, we recommend selecting a core with at least 100 texts; if computing resources are sufficient it is possible to increase its size to 1000 texts.

Taking these results into account, we will use the core of 100 files to classify the texts.

Training Sample Optimization

We investigated the dependence of the classification accuracy on the number of files in the core. A number of experiments showed that the number of correctly and incorrectly identified test files varies depending on what files are in the core. The optimal

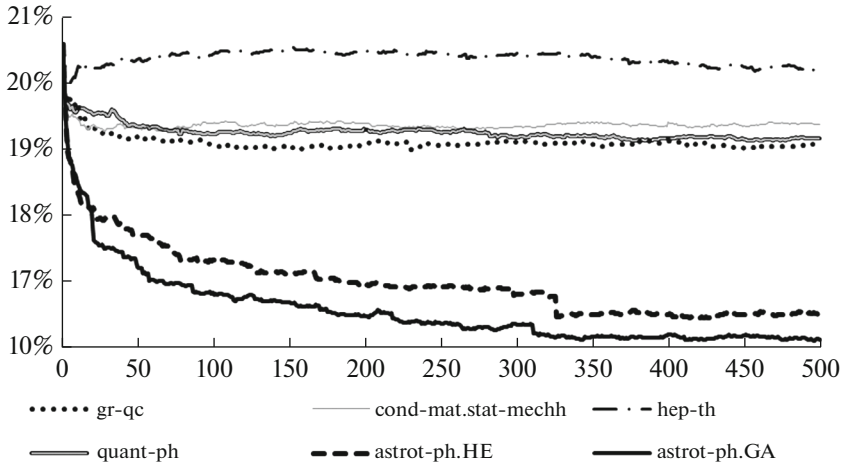

Fig. 2. The dependence of the compression degree of the test from the subject area astro-ph.GA on the number of files in the core and other categories within Physics.

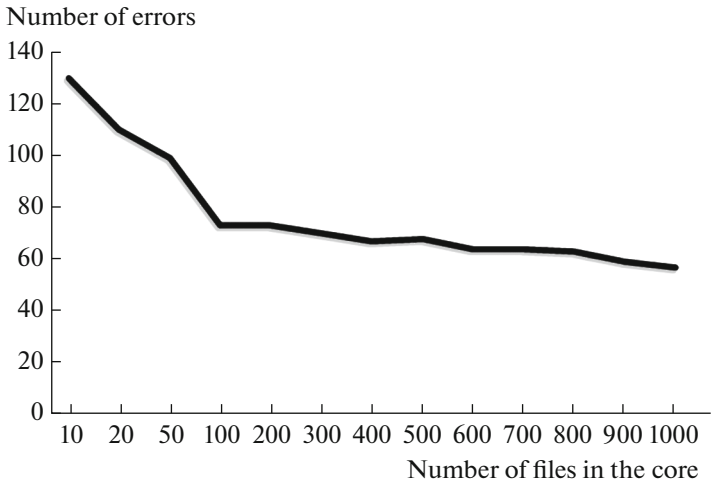

Fig. 3. The dependence of the number of errors on the number of files in the core.

core is the one that most fully shows the content of subject areas.

Next, we consider how the composition of the data in the training sample affects the probability of errors and describe the core method that makes it possible to reduce the number of incorrectly identified test files.

It is intuitively clear that the informativity of the files in the core is incomplete; therefore, in the case of random selection both the most typical and atypical texts can enter the core for the given field.

In order to reduce the probability of errors during classification, the following algorithm was developed for the core generation from typical files.

Let the files A, B, ..., E belong to one category. Take the file A and start compressing it sequentially with the files B, ..., E. We will indicate the compression ratio at the intersection of rows and columns. We obtain a compression matrix of files of one category, in which the rows show how the file A is compressed with each of the other files and the columns indicate how other files are compressed with the file A.

Since the lower the compression ratio is the better the result are, we calculate the average compression ratio for each column and sort the resulting values in ascending order. Next, we remove the file with the smallest average ratio (this file will be definitely included in the core) from the rows and columns and then repeat the procedure. As a result, the core will include 100 files that will be at the very beginning of these ratings and with which other files of this category are compressed best. In other words, the core will contain texts that carry the largest amount of information

Table 1. The compression matrix of test files of the cond-mat.stat-mech category with other categories

Category	Test file		
	test1 (cond-mat.stat-mech)	test2 (cond-mat.stat-mech, math-ph)	test3 (cond-mat.stat-mech)
cond-mat.stat-mech	**0%**	**0%**	**0%**
CS.IT	1.48	2.46	1.60
CS.LO	2.73	3.63	2.63
gr-qc	0.44	1.66	**0.01**
math-ph	0.48	**0.02**	1.60
Nucl-ex	2.00	2.83	1.56
physics.acc-ph	2.75	3.30	1.71
physics.atom-ph	1.43	2.46	0.98
physics.optics	1.59	2.50	1.03
physics.soc-ph	1.06	1.80	0.74
q-bio.BM	1.18	1.68	0.38
quant-ph	0.76	1.37	0.79

for other files of this category. As the tests showed, this core selection makes it possible to significantly improve the classification results.

It should be noted that in some cases (for example, if the initial classification of data is doubtful) the training sample should be generated manually.

RESULTS AND DISCUSSION

The results of this scientific text classification method were obtained in the form of a compression matrix (Table 1) of test files with files of each of the categories normalized by the compression ratio, i.e., the minimum compression ratio was subtracted from each matrix row. As a result, a zero compression ratio indicated to which category the test belongs. In the case of test2, where the main category is cond-mat.stat-mech and the secondary is math-ph, it can be seen that the method identified both categories correctly (0% compression corresponds to cond-mat.stat-mech and 0.02%, to math-ph).

Such a representation is convenient if the paper is interdisciplinary. In Table 1, this case is represented by test3 (the compression ratio with categories cond-mat.stat-mech and gr-qc differs only by 0.01%).

Let us consider the results obtained for CyberLeninka and arXiv.org separately. In the experiments with the texts from the CyberLeninka archive, 400 test files were randomly selected, at 20 files for each category. As a result, in the case of random selection of the core, 47% of the tests were identified incorrectly. With a more detailed study of texts from CyberLeninka, it was found that initially in some categories there were texts related both to close categories and to completely different fields of science. As an example, there were papers from the History, Sociology, etc categories in the Mathematics category. Thus, initially a large number of errors can be explained by the fact that in the case of random selection a large number of texts from other fields of science entered the core of the categories.

For the cores selected by the method described in *Training Sample Optimization,* the number of errors decreased by more than 1.5 times (Table 2). In two categories (History and Literature), all tests were identi-

Table 2. The results of the classification of scientific texts from the CyberLeninka archive

	Random cores	Selected cores
Automation	7	7
Astronomy	4	1
Biology	12	13
Economy	17	5
Ecology	16	6
Philosophy	8	4
Physics	8	9
Geophysics	7	4
Geography	8	10
History	2	0
Cybernetics	13	14
Literature	1	0
Mathematics	13	10
Medicine	7	2
Mechanics	15	5
Chemistry	7	2
Policy	5	8
Psychology	8	1
Sociology	17	9
Jurisprudence	12	2
Total	187 (47%)	114 (28%)

Table 3. The results of the classification of scientific texts from arXiv.org*

Science	Scientific field	Random cores	Selected cores			
		number of errors	number of errors	type 1	type 2	type 3
Physics	astro-ph.CO	2	2	2	0	0
Physics	astro-ph.GA	3	3	1	2	0
Physics	astro-ph.HE	2	2	1	1	0
Physics	cond-mat.dis-nn	2	4	3	1	0
Physics	cond-mat.stat-mech	3	1	0	1	0
Computer science	cs.AI	3	6	1	4	1
Computer science	cs.CR	0	2	0	1	1
Computer science	cs.IT	2	2	1	0	1
Computer science	cs.LO	2	0	0	0	0
Computer science	cs.SE	2	0	0	0	0
Physics	gr-qc	2	0	0	0	0
Physics	hep-ex	1	4	0	4	0
Physics	hep-th	1	1	1	0	0
Mathematics	math.AG	0	0	0	0	0
Mathematics	math.CO	1	1	0	1	0
Mathematics	math.DG	0	2	0	2	0
Mathematics	math.FA	0	0	0	0	0
Mathematics	math.GR	0	0	0	0	0
Mathematics	math.PR	1	2	1	1	0
Mathematics	math.ST	0	0	0	0	0
Mathematics	math-ph	18	4	1	3	0
Physics	nucl-ex	4	0	0	0	0
Physics	nucl-th	1	2	2	0	0
Physics	physics.acc-ph	0	1	0	1	0
Physics	physics.atom-ph	3	1	0	1	0
Physics	physics.ins-det	7	3	1	2	0
Physics	physics.optics	0	1	1	0	0
Physics	physics.soc-ph	0	0	0	0	0
Biology	q-bio.BM	2	2	0	0	2
Physics	quant-ph	1	1	0	1	0
	Total	63	47	16	26	5
	Share, %	11	8	3	4	1

* Types of errors. Secondary category is identified instead of the main one (Type 1). Another scientific category is identified (Type 2). Another science is identified (Type 3).

fied completely correctly. In other cases, a similar category was most often identified: for example, Mathematics tests were assigned to Automation or Cybernetics, Geophysical tests to Geography, and History tests to Politics or Jurisprudence.

A different situation was observed with the classification of data from arXiv.org. A total of 600 test files (20 from each category) were selected for the experiments. However, in contrast to CyberLeninka, even for random cores, the number of incorrectly identified tests was only 11% (Table 3).

Moreover, errors in the classification of texts from arXiv.org can be divided into the following types:

(1) A secondary category is identified instead of the main one.

(2) Another scientific category is identified.

(3) Another science is identified.

Thus, the number of errors decreased by approximately 3%. The secondary category was identified instead of the main category in 3% of the cases and another category, in 4%. Another science was identi-

fied only in 1% of the cases, and, for example, math.PR is identified in the scientific field cs.CR, but another mathematical category, which is not on the considered list, is assigned as the secondary category for this test.

CONCLUSIONS

In this paper, we proposed a method for automatic classification of scientific texts. Its accuracy was investigated based on two repositories: the archive of scientific texts arXiv.org and the CyberLeninka scientific electronic library.

As was expected, the accuracy of the method depends on the quality of the original classification used for the initial training. In the case of texts from arXiv.org, where the original classification is carried out by the authors themselves and has a low percentage of errors, our method works quite accurately and shows more than 90% efficiency.

For texts from CyberLeninka, it was found that initially some papers were classified incorrectly: for example, papers on history, sociology, etc. were found in the Mathematics category. This complicates the generation of training samples and decreases the efficiency of the method by 2–5 times.

Our investigation showed that for qualitative data, the percentage of errors does not exceed 5–8% and the method correctly identifies not only the general scientific field but also a narrower scientific category. The advantage of the method consists in its ability to estimate the proximity of a text to other subject fields.

The proposed method can be used to classify large volumes of scientific texts for the introduction of a new system of classifiers or the verification of an existing one, as well as in problems of detecting texts that are similar in content.

REFERENCES

1. Baghel, R. and Dhir, R., A frequent concepts based document clustering algorithm, *Int. J. Comput. Appl.,* 2010, vol. 4, no. 5, pp. 6 – 12.
2. Beil, F., Ester, M., and Xu, X., Frequent term-based text clustering, *Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD '2002),* Edmonton, Alberta, 2002, pp. 436–442.
3. Miao, Y., Keselj, V., and Milios, E., Document clustering using character n-grams: A comparative evaluation with term-based and word-based clustering, *CIKM '05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management,* New York, 2005, pp. 357–358.
4. Schaeffer, S.E., Graph clustering, *Comput. Sci. Rev.,* 2007, vol. 1, no. 1, pp. 27–64.
5. Kim, S., Han, K., Rim, H., and Myaeng, S.H., Some effective techniques for naïve Bayes text classification, *IEEE Trans. Knowl. Data Eng.,* 2006, vol. 18, no. 11, pp. 1457–1466.
6. Shevelev, O.G. and Petrakov, A.V., Classification of texts with decision trees and neural networks of direct propagation, *Vestn. Tomsk. Gos. Univ.,* 2006, vol. 290, pp. 300–307.
7. Wang, Z., He, Y., and Jiang, M., A comparison among three neural networks for text classification, *Proceedings of the IEEE 8th International Conference on Signal Processing,* 2006, no. 3, pp. 1883–1886.
8. Matyasko, A.A. and Khaustov, V.A., Classification of documents in vector space. Comparison of the Roccio methods and the k-nearest neighbor method, *Informatsionnye tekhnologii i sistemy 2012 (ITS 2012): Materialy mezhdunarodnoi nauchnoi konferentsii (g. Minsk, Belarus', 24 oktyabrya 2012 g.)* (Information Technologies and Systems 2012 (ITS 2012): Proceeding of the International Conference, BSUIR, Minsk, October 24, 2012), Minsk, 2012, pp. 140–141.
9. Li, M. and Vitanyi, P.M.B., *An Introduction to Kolmogorov Complexity and Its Applications,* New York: Springer-Verlag, 1997, 2nd ed., p. 637.
10. Cilibrasi, R. and Vitanyi, P.M.B., Clustering by compression, *IEEE Trans. Inf. Theory,* 2005, vol. 51, no. 4, pp. 1523–1545.
11. Cilibrasi, R., Vitanyi, P.M.B., and de Wolf, R., Algorithmic clustering of music based on string compression, *Comp. Music J.,* 2004, vol. 28, no. 4, pp. 49–67.
12. Li, M., Chen, X., Li, X., Ma, B., and Vitanyi, P.M.B., The similarity metric, *IEEE Trans. Inf. Theory,* 2004, vol. 50, no. 12, pp. 3250–3264.
13. Kukushkina, O.V., Polikarpov, A.A., and Khmelev, D.V., Determination of the authorship of the text using alphabetic and grammatical information, *Probl. Peredachi Inf.,* 2001, vol. 37, no. 2, pp. 96–109.
14. Khmelev, D.V., A complex approach to the problem of determining the authorship of the text, *Trudy i materialy Mezhdunarodnogo kongressa Russkii yazyk: Istoricheskie sud'by i sovremennost' (13–16 marta 2001 goda)* (Proc. Int. Congress The Russian Language: Historical Fates and the Present (March 13–16, 2001), Moscow: MGU, 2001, pp. 426–427.
15. Malyutov, M.B., *Authorship attribution of texts: A review*, *Springer Lect. Notes Comput. Sci.,* 2007, vol. 4123, pp. 362–380.
16. Malyutov, M.B., Wickramasinghe, C.I., and Li, S., *Conditional Complexity of Compression for Authorship Attribution. SFB 649 Discussion Paper No. 57,* Berlin: Humboldt University, 2007, p. 38.
17. Ryabko, B., Astola, J., and Malyutov, M., *Compression-Based Methods of Statistical Analysis and Prediction of Time Series,* Springer, 2016.

Translated by O. Pismenov