ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2019, Vol. 53, No. 6, pp. 329–342. © Allerton Press, Inc., 2019.
Russian Text © The Author(s), 2019, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2: Informatsionnye Protsessy i Sistemy, 2019, No. 12, pp. 25–38.

AUTOMATION OF TEXT PROCESSING

Classification of Scientific Texts Based on the Compression of Annotations to Publications

I. V. Selivanova[a], *, D. V. Kosyakov[a], **, and A. E. Guskov[a], ***

[a]*State Public Library of Scientific and Technical Information, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630102 Russia*

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

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

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

Received October 15, 2019

Abstract—This paper describes the possibility of establishing the semantic proximity of scientific texts by the method of their automatic classification based on the compression of annotations. The idea of the method is that the compression algorithms such as PPM (prediction by partial matching) compress terminologically similar texts much better than distant ones. If a kernel of publications (an analogue of a training set) is formed for each classified topic, then the best proportion of compression will indicate that the classified text belongs to the corresponding topic. Thirty thematic categories were determined; for each of them, annotations of approximately 500 publications were received in the Scopus database, out of which 100 annotations for the kernel and 20 annotations for testing were selected in different ways. It was found that building a kernel based on highly cited publications revealed an error level of up to 12 against 32% in the case of random sampling. The quality of classification is also affected by the initial number of categories: the fewer the categories that participate in the classification and the more terminological differences exist between them, the higher its quality is.

Keywords: classification of scientific texts, text compression, bibliographic databases, Scopus

DOI: 10.3103/S0005105519060062

INTRODUCTION

In recent decades, in connection with an increase in the number of scientific publications, the problem of text classification has become especially urgent. Under these conditions, both entire literary or poetic texts [1, 2] and short text messages, for example, SMS or tweets [3–5] are classified. However, a single classification method that would immediately have both high efficiency and low labor costs has not yet been found.

One of the areas where this task is of particular importance is the classification of scientific documents. Incorrectly classified publications make it difficult to find articles that are necessary for a scientist, which is a reason for a loss of relevant research in the field of science that is of interest to him.

The studies [6, 7] proposed a method for classifying scientific texts based on information compression. It showed more than 90% efficiency and low labor costs when classifying the English-language texts of the *arXiv.org* Scientific Publications Archive. One of the drawbacks of these works was the application of the method only to full-text documents. Full texts of articles often contain many unnecessary phrases because of their volume, which can lead to some errors in classification. Annotations of publications, on the contrary, must contain only key points used in the article [8, p. 35], which can facilitate the automatic classification of scientific papers. Moreover, many scientific bibliographic databases (BDBs) do not always make it possible to obtain the full text of an article and only its abstract is often available.

The aim of our study was to apply the method based on the compression algorithms to the classification of annotations of publications indexed in international databases. The initial results of this work were presented at the XVII "Distributed Information and Computing Resources" *DICR-2019* Russian Conference.

The most reputable BDBs for scientists are the *Web of Science* (*WoS*) and *Scopus*. They serve as a data source for a variety of scientometric studies, various university rankings, and are also used to evaluate the publication activity of researchers, organizations, and countries.

However, the classification of publications in these bibliographic databases has raised a large number of questions. One of its main drawbacks was that the topics of a publication were defined in both WoS and Scopus only at the level of the journal in which it was published [9, 10]. This had a particularly negative effect on articles from multidisciplinary journals. A publication that could have only one area was assigned to all the

same subject headings as this journal. This created "overloading" of scientific fields with publications that probably had no relation to them.

To improve the quality of classification, *Scopus* introduced the *Topic Prominence in Science* new classification system in 2017 [11], which was built in a similar manner to the technique described in [12]. The authors proposed a method for assigning publications certain categories, which consists of three stages: at the first stage, the connectedness of publications was determined using direct citation; at the second stage, a hierarchical clustering of publications was performed; the third stage was devoted to naming the resulting clusters by labels based on the terms of headings and annotations of publications included in this group.

An important role in improving the accuracy of the classification of scientific documents is played by the choice of a classification system. These systems can be divided into several types:

(1) Library classifiers (for example, UDC). The basis of this system is the decimal classification developed in 1876 by the American bibliographer M. Dewey. The central part of the UDC is basic tables that cover the entire assembly of knowledge and are constructed according to the hierarchical principle of dividing from general to particular using a digital decimal code [13]. The library classifiers also include the LBC, which is the national classification system of Russia. The principle of its formation is similar to the UDC.

(2) National classifiers, for example, the Fields of research (*FOR*) classification from *the Australian and New Zealand Standard Research Classification* (*ANZSRC*), which is a hierarchical classification with three levels: the first level covers extensive scientific areas, the second level covers related groups, and the third one includes narrower areas. *FOR* is used, for example, in the *Dimensions* platform [14]. Other examples of national classifiers are the Nomenclature of Scientific Specialties of Russia, which is used by the Higher Attestation Commission of the Ministry of Education and Science (VAK) [15], the All-Russian Classifier of Specialties in Education (ACSE), which is compared with the International Standard Classification of Education (ISCE) [16], and the State Rubricator of Scientific and Technical Information (SRSTI) [17].

(3) International classifiers, these include the *Field of science and technology* (*FOS*), which is a classification system that was published by the Organization for Economic Cooperation and Development in 2002 and amended in 2007 [18]. This classification is based on six scientific fields, which are divided into 42 areas. Another example of an international classifier is the nomenclature of *UNESCO*, that is, a system developed by UNESCO for the classification of scientific papers and dissertations [19].

(4) Classifiers in international BDBs. These classifiers, like the previous ones, are built on a hierarchical basis. In the *WoS*, the classifier consists of three levels. The basic level covers six fields of science, which are divided into 39 sublevels. The third sublevel of the *WoS* classifier includes 253 categories [20]. *Scopus* uses the *All Science Journals Classification* (*ASJC*). It includes publications distributed by four general scientific fields: biological sciences, physical sciences, medicine, as well as social and human sciences. These fields are divided into 27 large subject areas and more than 300 narrow categories [21]. Despite the wide coverage of topics, the *ASJC* has significant drawbacks. As an example, two different scientific areas include two related categories: *Language and Linguistics* (code, 1203; area, *Arts and Humanities*) and *Linguistics and Language* (code, 3310; area, Social Sciences). This problem was noted in 2016 in [22], where the authors proposed either to merge these categories or to indicate more clear differences between them.

The importance of choosing a classification system was noted in [23]. The examination of two Portuguese classification systems (*FCT* and *DeGóis platform*) revealed that the *Nursing* area of science was completely absent in the first system, and in the second it was not clearly defined. The study [24] showed two main problems of the *UNESCO* nomenclature: the first is related to the fact that smaller categories may be lost in this classification in large areas of science, as a result of which the classification will be insufficiently complete; the second is that this classification includes no areas of science that have appeared recently, which also significantly decreases the quality of the classification.

Many different methods are used to solve the problem of classifying text documents. One of the most commonly used algorithms is the method of k-nearest neighbors and its modifications, where the classified object belongs to the class that includes the objects of the training sample that are the closest to it. As an example, [25] was based on this method. Another algorithm is the Bayesian classification, which works based on the calculation of posterior class probabilities. This algorithm was applied in [26]. One representative of linear classifiers is the support vector method that consists in constructing a hyperplane, which separates the sample objects in the most optimal way. A comparison of the algorithms of the Bayesian classification and the support vector method for classifying article headings was given in [27].

Neural networks have recently been increasingly used to solve the classification problem. The study [28] proposed using recurrent convolutional neural networks to solve the problem of text classification. Its authors come to the conclusion that the use of neural networks in the classification of text documents will help to avoid the problem of sparseness of data, as well as to collect more contextual information about entities compared to traditional methods. Convolutional neural networks also showed high accuracy (83.98%) in the classification of patent documents [29].

On average, the accuracy of various text information classification algorithms varies from 70 to 86% [30, 31].

The main advantage of the classification of scientific publications is the commonality of terms, concepts, and phrases that are used in texts of the same area of science. Moreover, the more narrowly focused the scientific area is, the more specific the vocabulary of related articles is.

Scientific publications can be classified using the methods based on the following:

—Citation, which includes direct citation, co-citation, and bibliographic combination [32–34]. The most common source of data for such studies is the *WoS* BDB.

—Publication metadata (lists of co-authors, publication names, keywords, etc.). This method was used in [35] for clustering biomedical publications in the *MEDLINE* database;

—Combinations of citation and metadata. As an example, in [36], the authors used this approach to improve the quality of document classification in the ACM *Digital Library*;

—Annotations and full texts of publications. The classification of annotations of publications from the *Materials Science*, *Physics*, and *Chemistry* scientific fields of the *Scopus* BDB was performed by the authors of [37]. The study [38] considered checking the quality of classification of medical and biological publications from the *PubMed* database by comparing the results of applying the *k* nearest neighbors methods, the Bayesian classification, and support vectors to annotations. Full-text scientific documents were classified in [7, 39].

1. METHODS AND DATA OF CLASSIFICATION

1.1. Method

Let us consider the method based on the compression algorithms in more detail. Let there be n scientific fields X^1, ..., X_n, for each of which a set of texts that is characteristic of it is defined (the files which constitute the "kernel" of this area). There is also some test file y, whose subject needs to be determined, and an archiver φ, which can be used to compress any set of texts.

The work of the method is that the test file y begins to be sequentially compressed with each of n kernels using the archiver φ. As a result, the area of the test file y is determined by the kernel with which it has the best compression.

For the method to work, the following parameters were selected:

—The archiver is WinRAR with a maximum memory value of 128 MB; the compression algorithm is PPMd [6, 7];

—The kernel size is 100 files. This volume was recommended in [7], since the classification quality almost does not improve in case of a larger number of files in the kernel and the algorithm runs longer.

1.2. Data

The source of information for this study was the Scopus BDB. The data extraction process consisted of three stages.

(1) Extracting information about the name of the journal and of the publication, citation, and category through *Scival*, which is an analytical tool of the *Elsevier* Company based on *Scopus* data. The data for 30 categories were selected for the period 2009–2018 and are presented in Table 1.

(2) Generating annotation files by obtaining their texts via *the Scopus Abstract Retrieval API*

(3) Removing annotation files with missing text.

In total, 15000 files were downloaded (500 for each category).

1.3. Classification Process

To determine how efficiently the method based on the compression algorithms works on annotations of publications that are indexed in the *Scopus* BDB, the classification was made for the following:

(1) test files with one category,

(2) test files with several categories.

In the first case, the operation of the method was checked by classifying files with one category, and kernels that were optimal in composition were selected.

At the second stage, the classification was carried out for tests with several categories, for which the kernels were selected at the first stage. This case is useful in checking the quality of classification of publications from multidisciplinary publications.

Let us consider these two stages in more detail.

1.3.1. Classification of test files with one category

When classifying test files with one category three types of category definition errors were introduced:

• Type I errors are associated with an incorrect category definition within the scientific field, for example, instead of the *Aquatic Science* category, the *Plant Science* category from the same area *Agricultural and Biological Sciences* was identified.

• Type II errors include those test files, for which another area was determined in case of the same scientific field. As an example, instead of the necessary *Cell Biology* category from the *Biochemistry*, *Genetics* and *Molecular Biology* area, the *Pharmocology* category from the *Pharmacology*, *Toxicology and Pharmaceutics* area was defined. Meanwhile, the *Life Sciences* general scientific field was preserved.

• Type III errors are the most serious; this includes those test files for which an error in the classification occurred at the highest level. As an example, the *Social*

Table 1. The studied areas of science according to the classification levels

Field	Area	Category
Life sciences	Agricultural and biological sciences	Animal science, zoology
Life sciences	Agricultural and biological sciences	Aquatic science
Life sciences	Agricultural and biological sciences	Plant science
Social sciences	Arts and humanities	History
Social sciences	Arts and humanities	Literature, literary theory
Life sciences	Biochemistry, genetics and molecular biology	Cell biology
Life sciences	Biochemistry, genetics and molecular biology	Endocrinology
Social sciences	Business, management and accounting	Marketing
Physical sciences	Chemical engineering	Catalysis
Physical sciences	Chemistry	Inorganic chemistry
Physical sciences	Chemistry	Organic chemistry
Physical sciences	Computer science	Artificial intelligence
Physical sciences	Computer science	Computer vision and pattern recognition
Physical sciences	Computer science	Hardware and architecture
Physical sciences	Earth and planetary sciences	Geology
Physical sciences	Earth and planetary sciences	Oceanography
Physical sciences	Mathematics	Algebra and number theory
Physical sciences	Mathematics	Geometry and topology
Physical sciences	Mathematics	Logic
Physical sciences	Mathematics	Numerical analysis
Physical sciences	Mathematics	Statistics and probability
Health sciences	Medicine	Ophthalmology
Health sciences	Medicine	Surgery
Life sciences	Pharmacology, toxicology and pharmaceutics	Pharmocology
Physical sciences	Physics and astronomy	Astronomy and astrophysics
Physical sciences	Physics and astronomy	Condensed matter physics
Physical sciences	Physics and astronomy	Nuclear and high energy physics
Social sciences	Psychology	Social psychology
Social sciences	Social sciences	Library and information sciences
Social sciences	Social sciences	Sociology and political science

Sciences scientific field was determined instead of the *Physical Sciences* field.

The classification was carried out using two types of kernels:

(1) **Arbitrary kernels** were composed of annotations of publications selected at random.

(2) **Selected kernels** were composed of the 100 most cited publications in each area of science. This approach presumably increases the quality of the kernel due to the fact that publications from the same area that cite articles selected in the kernel are likely to inherit a characteristic vocabulary.

In both cases, the same set of arbitrary test files was used; 20 test files were selected for each of the 30 categories; a total of 600 tests were made.

The influence of the presence of stop words [40] and the names of publishers that are present in annotation texts (for example, © 2009 *The American Physical Society*) on the quality of the classification was also determined. The creation of kernels and preparation of test files were performed both with the original texts of annotations and with the removal of stop words (for example, *always*, *every*, *just*, etc.), replacing the uppercase letters with lowercase letters and removing all characters except numbers, letters, and punctuation marks: ., !, ?, :, , , -.

1.3.2. Classification of test files with several categories

The classification of test files with several categories was made using only kernels with the most highly cited publications.

Table 2. An example of calculating a normalized compression ratio

Test areas	Algebra and number theory, %	Geometry and topology, %	min, %	Algebra and number theory	Geometry and topology
Algebra and number theory, geometry and topology	26.70	26.80	26.70	OK	OK
Algebra and number theory, geometry and topology	33.25	32.85	32.85	Missed	OK

The selection of 20 tests was carried out arbitrarily from publications in which at least two categories coincided with the categories from Table 1. As in the previous cases, a total of 600 tests were selected.

Let us introduce the following groups of results:

—The test file was correctly identified to have at least 50% of these categories. As an example, the test was indicated to have four categories: *Algebra and Number Theory*, *Numerical Analysis*, *Geometry and Topology*, and *Discrete Mathematics and Combinatorics*. The categories we study include only the first three ones. Accordingly, to get into this group, the test file must be defined to have at least two of the first three categories. The correctly defined categories are those that have a normal compression ratio (the compression percentage minus the minimum compression percentage for all categories) that is less than the minimum compression percentage for all categories * 0.50% (the number of errors was almost unchanged at a higher threshold). As an example of such a calculation, we consider two test files with the *Algebra and Number Theory* and *Geometry and Topology* categories (Table 2). The minimum compression percentage for the first test was determined for the *Algebra and Number Theory* category. In this case, if the selected threshold value is not only the minimum compression percentage, but the minimum compression percentage for all categories * 0.50%, then the second indicated category *Geometry and Topology* will also be correctly defined with this test. In the second test file, only the *Geometry and Topology* category was defined.

—The test file was identified to have at least one of the indicated categories;

—All categories of the test file were defined incorrectly. This case will include those test files, which were defined to have some other category.

In this case, one test file can relate to only one of these groups.

1.4. The Influence of the Number of Categories on the Quality of Classification

In [30], the authors came to the conclusion that the number of categories affects the classification; if categories with similar terms are combined into one, then the quality of the classification will improve.

To assess the influence of the number of categories on the quality of the classification of annotations, we conducted an experiment with a sequential increase in the number of categories under consideration from 5 to 30 in increments of 5 (5, 10, 15, 20, 25, 30). For the test files, 20 publications were randomly selected from the first five categories (100 files in total).

2. DISCUSSION OF THE RESULTS

2.1. Classification of Test Files with One Category

2.1.1. Classification in case of arbitrary kernels

Table 3 gives the results of the classification of test files with one category in case of using arbitrary kernels by types of errors.

The proportion of erroneously defined test files was 32% of the total number (192 out of 600 tests).

The definition of an incorrect scientific field is most often due to the categories that are similar in terminology. As an example, *Aquatic Science* and *Oceanography* are such categories. However, the nature of the error sometimes cannot be determined, for example, in the case of a publication with the eid = 2-s2.0-67651018249 from the *Condensed Matter Physics* category, instead of which the *Marketing* category was defined. It was not possible to visually determine the reason from the text of the annotation (Fig. 1).

We performed the pairwise file compression from the kernels of these two categories and a test file with the eid = 2-s2.0-67651018249. For almost every one of the 100 files from the kernel of the *Marketing* category, a test with eid = 2-s2.0-67651018249 showed better compression. The average normalized compression coefficients for a test file with the *Condensed Matter Physics* category and the *Marketing* category are 9.76 and 9.13%, respectively.

The top ten files with the best compression of this test included three files from the *Condensed Matter Physics* category and seven files from the *Marketing* category (Table 4).

Figures 2 and 3 show the texts of two annotations of the *Marketing* category, with which the test file had the best pairwise compression. These files completely differ in terminology, both among themselves and with the test file under study. Thus, the results suggest that the error in determining the category of a test file with eid = 2-s2.0-67651018249 of the *Condensed Matter*

Table 3. The results of the classification of test files in case of using arbitrary kernels with a different number of citations

Category	Total number of tests	Number of errors	Errors		
			I type	II type	III type
Algebra and number theory	20	7	5	2	0
Animal science and zoology	20	1	0	1	0
Aquatic science	20	12	11	0	1
Artificial intelligence	20	8	6	2	0
Astronomy and astrophysics	20	1	1	0	0
Catalysis	20	5	0	5	0
Cell biology	20	1	0	0	1
Computer vision and pattern recognition	20	3	3	0	0
Condensed matter physics	20	11	2	4	5
Endocrinology	20	3	1	2	0
Geology	20	0	0	0	0
Geometry and topology	20	11	9	2	0
Hardware and architecture	20	0	0	0	0
History	20	8	1	7	0
Inorganic chemistry	20	13	4	1	8
Library and information sciences	20	12	2	8	2
Literature and literary theory	20	11	5	5	1
Logic	20	2	0	1	1
Marketing	20	2	0	2	0
Nuclear and high energy physics	20	4	4	0	0
Numerical analysis	20	0	0	0	0
Oceanography	20	11	0	1	10
Ophthalmology	20	9	7	0	2
Organic chemistry	20	3	0	2	1
Pharmocology	20	18	0	18	0
Plant science	20	20	13	7	0
Social psychology	20	1	0	1	0
Sociology and political science	20	8	0	7	1
Statistics and probability	20	6	0	2	4
Surgery	20	1	0	0	1
Total number	600	192	74	80	38
Share of the total number, %		32	12	13	6

Table 4. The top ten files with the best compression of the studied test with eid = 2-s2.0-67651018249 of the *Condensed Matter Physics* category

Incorrectly defined test from *Condensed Matter Physics*	File identifier	File category	Normalized compression percentage, %
2-s2.0-67651018249	2-s2.0-70350534620	Condensed Matter Physics	0.00
2-s2.0-67651018249	2-s2.0-80052140988	Marketing	1.17
2-s2.0-67651018249	2-s2.0-67149130202	Marketing	1.47
2-s2.0-67651018249	2-s2.0-79960889541	Marketing	2.70
2-s2.0-67651018249	2-s2.0-70449090433	Condensed Matter Physics	2.94
2-s2.0-67651018249	2-s2.0-70449127336	Condensed Matter Physics	3.16
2-s2.0-67651018249	2-s2.0-79959944133	Marketing	3.20
2-s2.0-67651018249	2-s2.0-67149101079	Marketing	3.49
2-s2.0-67651018249	2-s2.0-78650307261	Marketing	3.88
2-s2.0-67651018249	2-s2.0-78751585438	Marketing	3.90

Two-particle dispersion is of central importance to a wide range of natural and industrial applications. It has been an active area of research since Richardson's (1926) seminal paper. This review emphasizes recent results from experiments, high-end direct numerical simulations, and modern theoretical discussions. Our approach is complementary to Sawford's (2001), whose review focused primarily on stochastic models of pair dispersion. We begin by reviewing the theoretical foundations of relative dispersion, followed by experimental and numerical findings for the dissipation subrange and inertial subrange. We discuss the findings in the context of the relevant theory for each regime. We conclude by providing a critical analysis of our current understanding and by suggesting paths toward further progress that take full advantage of exciting developments in modern experimental methods and peta-scale supercomputing. Copyright © 2009 by Annual Reviews. All right reserved. All rights reserved.

Fig. 1. Annotation of the publication with eid = 2-s2.0-67651018249.

Franchisee selection is a major input for franchising success. In this article, we argue that franchisee selection criteria do not differ between social and commercial franchising. They may be even more relevant for obtaining social franchising success. We discuss criteria for franchisee selection and present details of our multiple case study research to support the argument. Our study finds that evolved social franchisors do adopt similar selection criteria as commercial franchisees. In addition, constraints faced with franchisee selection among commercial franchisors are reflected also among social franchisors. We contribute to franchising literature by extending commercial franchisee selection criteria to social franchisee selection. A major managerial implication of this research is that existing franchising professionals could easily assist new social franchisors in developing their social franchisees. Future research could be study criteria weights and methodology adopted for making final selection. A new research direction could involve studying if selection criteria would differ based on (a) social cause and (b) franchisee location. © Taylor & Francis Group, LLC.

Fig. 2. Annotation of the publication of the *Marketing* category with eid = 2-s2.0-80052140988.

Despite the popularity of online digital music and the broad application of digital music sampling, in the existing literature, there is a lack of substantial studies that examine online digital music sampling. This study uses a laboratory experiment to explore the determinants of the five effectiveness dimensions, i.e., evaluation, Willingness-to-Pay (WTP), perceived sampling usefulness, sampling cost and the likelihood of being a free rider, of online digital music sampling. Digital music samples with a higher quality and longer segments were found to increase the sampler's music evaluation and make the evaluation process more useful. Also, the sampler's music evaluation significantly determines his/her WTP. Higher music evaluations not only decrease the sampler's sampling cost during the sampling process, but also reduces the probability that the sampler will take the music sample as a substitute for the original music. This study also shows that the current practice of online digital music sampling is not ideal and music retailers could improve their music sampling strategies by providing digital music samples with longer segments and of higher quality. All of these findings have significant implications for music retailers to use digital music sampling strategies better. Copyright © 2009, Inderscience Publishers.

Fig. 3. Annotation of the publication of the *Marketing* category with eid = 2-s2.0-67149130202.

Table 5. Results of classifying the test files in case of using the kernels with the most highly cited publications

Category	Total number of tests	Number of errors	Errors		
			I type	II type	III type
Algebra and number theory	20	8	7	1	0
Animal science and zoology	20	7	6	1	0
Aquatic science	20	2	2	0	0
Artificial intelligence	20	5	2	2	1
Astronomy and astrophysics	20	0	0	0	0
Catalysis	20	4	0	4	0
Cell biology	20	4	1	3	0
Computer vision and pattern recognition	20	3	3	0	0
Condensed matter physics	20	2	0	2	0
Endocrinology	20	1	0	1	0
Geology	20	0	0	0	0
Geometry and topology	20	3	3	0	0
Hardware and architecture	20	0	0	0	0
History	20	4	2	1	1
Inorganic chemistry	20	3	0	3	0
Library and information sciences	20	3	1	1	1
Literature and literary theory	20	2	1	1	0
Logic	20	3	2	1	0
Marketing	20	1	0	1	0
Nuclear and high energy physics	20	3	3	0	0
Numerical analysis	20	0	0	0	0
Oceanography	20	4	0	0	4
Ophthalmology	20	0	0	0	0
Organic chemistry	20	1	0	0	1
Pharmocology	20	2	0	2	0
Plant science	20	2	1	1	0
Social psychology	20	2	0	2	0
Sociology and political science	20	2	1	0	1
Statistics and probability	20	1	0	1	0
Surgery	20	0	0	0	0
Total number	600	72	35	28	9
Share of the total number, %		12	6	5	2

Physics category may be related to the operation of the method.

2.1.2. Classification in case of selected kernels

Table 5 shows the results of the classification of test files with one category in case of using the kernels, which include the most highly cited publications.

The use of selected kernels improved the classification results by 20%. The number of type III errors decreased by three times. Such errors mainly occurred due to the fact that categories or files, in which similar terms are used, were in different scientific fields. As an example, a test from the *Sociology and Political Science* category was incorrectly defined to have the *Aquatic Science* category; however, the text of this annotation uses many terms that are used in the *Aquatic Science* category (Fig. 4).

Some incorrect definition of tests may be due to the incorrect initial classification. So, in the case of the *Library and Information Sciences* category, the test file was related to the *Artificial Intelligence* category. The annotation text contains a significant number of terms, which are specific to the *Computer Science* area (Fig. 5).

In some patterns, no definition of an incorrect category was found. As an example, a test file from the

This paper seeks to understand how the Brazilian Amazon, which many thought unsuitable for ***agricultural development****, has yielded to a dynamic cattle* ***economy*** *in only a few decades. It does so by embedding the Thunian model of* ***location*** *rents within the regime of* ***capital accumulation*** *that has driven the Brazilian* ***economy*** *since the mid-20th century. The paper addresses policies that have created location rents in Amazônia, the effect of these rents on* ***land managers****, and the* ***spatial implications*** *of their* ***behavior*** *on* ***forests****. Thus, the paper connects macro-****processes*** *and* ***structures*** *to* ***agents*** *on the* ***ground****, in providing a political* ***ecological explanation*** *relevant to* ***land*** *change science. The policy discussion focuses on* ***reductions*** *in transportation* ***cost****s, improvements in* ***animal health****, and monetary and* ***trade*** *reforms. To illustrate the impact of policy, the paper presents data on the* ***geography*** *of Amazonian herd* ***expansion****, on the growth of Amazonian exports, and on the profitability of the* ***region****'s cattle* ***economy****. It follows the empirical presentation with more abstract consideration of the spatial relations between cattle ranching and* ***soy farming****, and implications for* ***deforestation****. The paper concludes on a speculative note by considering the likelihood of* ***forest transition*** *in the* ***region****, given the transformation of Amazônia into a global* ***resource*** *frontier. © 2008 Elsevier Ltd. All rights reserved.*

Fig. 4. Annotation of the test with eid = 2-s2.0-70449527784 from the *Sociology and Political Science* category (the terms that often occur in the *Aquatic Science* category are **bolded**).

This chapter provides a tutorial overview of distributed optimization and game theory for decision-making in networked systems. We discuss properties of first-order methods for smooth and non-smooth convex optimization, and review mathematical decomposition techniques. A model of networked decision-making is introduced in which a communication structure is enforced that determines which nodes are allowed to coordinate with each other, and several recent techniques for solving such problems are reviewed. We then continue to study the impact of noncooperative games, in which no communication and coordination are enforced. Special attention is given to existence and uniqueness of Nash equilibria, as well as the efficiency loss in not coordinating nodes. Finally, we discuss methods for studying the dynamics of distributed optimization algorithms in continuous time. © 2010 Springer London.

Fig. 5. The text of the annotation of the publication with eid = 2-s2.0-77958562700 from the Library and Information Sciences category.

History category was incorrectly related to the *Condensed Matter Physics* category (Fig. 6).

It is worth noting that the incorrectly determined file from paragraph 2.1.1. with eid = 2-s2.0-67651018249 was correctly determined in case of selected kernels. Thus, the composition of the kernel has a great influence on the quality of the classification.

2.3.3. The influence of stop words and publisher names on the classification

To study the influence of the presence of publisher names in annotations on the quality of the classification, selected kernels were used (see Section 2.1.2).

Table 6 compares the quality of the classification of annotations with stop words and publisher names and without them. When removing the names of publishers, the number of errors increased by 3%. The number of errors in the *History*, *Geometry and Topology*, *Literature and Literary Theory*, and *Sociology and Political Science* categories almost doubled. Perhaps this is due to the fact that highly cited publications are most often printed by publishers whose names indicate important terms for the category. As an example, in one of the tests that was incorrectly defined after removing the publisher, the following line was found: © 2010 *English Literary Renaissance Inc. Published by Blackwell Publishing Ltd*.

When removing stop words from annotations where publisher names were present, the number of errors decreased to 11%. However, this decrease did not occur evenly for all categories: while the removal of stop words had a positive effect on the quality of the

The horse skeleton found in the autumn of 1958 at the fortress of Buhen in northern Sudan has become one of the most prominent, but also one of the most enigmatic equid remains from the second millennium BC: Firstly, because of its assumed early date of c. 1675 BC, deduced by W.B. Emery after analysing the stratigraphical data, This - according to our knowledge at the time - being several decades before the oldest known equid remains in Egypt. Secondly, because of wear on the lower left second premolar (LP2), which has led to the conclusion that it was most probably caused by bit-wear. Since the 1960s, both conclusions have been subject to criticism. The purpose of this study is to provide a review of the history of research and reception of the Buhen horse in its interdisciplinary context over the last fifty years with the result that only modern scientific techniques might be able to solve some of the outstanding questions. © 2009 Brill.

Fig. 6. A test file with eid = 2-s2.0-77951062083 from the *History* category.

Table 6. The influence of stop words and the presence of publisher names on the quality of the classification

Category	Total number of tests	Number of errors	Number of errors		
			with stop words, without publisher names	without stop words, with publisher names	without stop words and without publisher names
Algebra and number theory	20	8	8	8	7
Animal science and zoology	20	7	7	6	8
Aquatic science	20	2	2	1	2
Artificial intelligence	20	5	6	6	7
Astronomy and astrophysics	20	0	0	0	0
Catalysis	20	4	3	5	5
Cell biology	20	4	3	2	3
Computer vision and pattern recognition	20	3	3	1	3
Condensed matter physics	20	2	2	1	1
Endocrinology	20	1	2	1	2
Geology	20	0	0	0	0
Geometry and topology	20	3	6	2	6
Hardware and architecture	20	0	0	0	0
History	20	4	7	4	8
Inorganic chemistry	20	3	5	3	5
Library and information sciences	20	3	4	3	5
Literature and literary theory	20	2	4	4	4
Logic	20	3	3	3	4
Marketing	20	1	1	1	1
Nuclear and high energy physics	20	3	3	2	5
Numerical analysis	20	0	1	0	1
Oceanography	20	4	4	3	5
Ophthalmology	20	0	1	0	1
Organic chemistry	20	1	2	1	2
Pharmocology	20	2	3	2	2
Plant science	20	2	3	2	2
Social psychology	20	2	1	1	3
Sociology and political science	20	2	3	2	4
Statistics and probability	20	1	2	1	2
Surgery	20	0	0	0	0
Total number	600	72	89	65	98
Share of the total number, %		12	15	11	16

In California, the toxic algal species of primary concern are the dinoflagellate Alexandrium catenella and members of the pennate diatom genus Pseudo-nitzschia, both producers of potent neurotoxins that are capable of sickening and killing marine life and humans. During the summer of 2004 in Monterey Bay, we observed a change in the taxonomic structure of the phytoplankton community-the typically diatom-dominated community shifted to a red tide, dinoflagellate-dominated community. Here we use a 6-year time series (2000-2006) to show how the abundance of the dominant harmful algal bloom (HAB) species in the Bay up to that point, Pseudo-nitzschia, significantly declined during the dinoflagellate-dominated interval, while two genera of toxic dinoflagellates, Alexandrium and Dinophysis, became the predominant toxin producers. This change represents a shift from a genus of toxin producers that typically dominates the community during a toxic bloom, to HAB taxa that are generally only minor components of the community in a toxic event. This change in the local HAB species was also reflected in the toxins present in higher trophic levels. Despite the small contribution of A. catenella to the overall phytoplankton community, the increase in the presence of this species in Monterey Bay was associated with an increase in the presence of paralytic shellfish poisoning (PSP) toxins in sentinel shellfish and clupeoid fish. This report provides the first evidence that PSP toxins are present in California's pelagic food web, as PSP toxins were detected in both northern anchovies (Engraulis mordax) and Pacific sardines (Sardinops sagax). Another interesting observation from our data is the co-occurrence of DA and PSP toxins in both planktivorous fish and sentinel shellfish. We also provide evidence, based on the statewide biotoxin monitoring program, that this increase in the frequency and abundance of PSP events related to A. catenella occurred not just in Monterey Bay, but also in other coastal regions of California. Our results demonstrate that changes in the taxonomic structure of the phytoplankton community influences the nature of the algal toxins that move through local food webs and also emphasizes the importance of monitoring for the full suite of toxic algae, rather than just one genus or species. © 2008 Elsevier B.V.

Fig. 7. Annotation of the publication with eid = 2-s2.0-57649228732.

classification in the *Cell Biology* category, the number of errors in the *Literature and Literary Theory* category increased by two times.

When removing both stop words and publisher names, the number of errors increased to 16%. Similarly to the previous case, the removal of stop words and publisher names had a positive effect on a number of categories, while others began to be defined incorrectly. Thus, for example, in the *Geometry and Topology* category, the test with eid = 2-s2.0-84055189802 was defined correctly with removing stop words, and the test with eid = 2-s2.0-77956268008, which had been previously defined correctly, was now related to the *Numerical Analysis* category. Perhaps this is due to the fact that the length of the annotation decreases when removing stop words.

Thus, the absence of publisher names in annotation texts affects the quality of the classification negatively. One cannot draw an unambiguous conclusion about the influence of stop words.

2.2. Classification of Test Files with Several Categories

The results of classification of test files with several categories are given in Table 7.

Here, 23% (140 out of 600) of the test files were determined incorrectly. The scientific field was incorrectly determined in 6% (37 of 600) of the tests. It is

Table 7. Classification results for files with several categories

Test group	Number of tests	Share of 600 tests, %
No less than 50% of categories were defined	413	69
At least one category was defined	47	8
All categories were defined incorrectly	140	23

Current records management methodologies and practices suffer from an inadequate understanding of the 'human activity systems' where records managers operate as 'mediators' between a number of complex and interacting factors. Although the records management and archival literature recognizes that managing the active life of the records is fundamental to their survival as meaningful evidence of activities, the context where the records are made, captured, used, and selectively retained is not explored in depth. In particular, the various standards, models, and functional requirement lists, which occupy a vast portion of that literature, especially in relation to electronic records, do not seem to be capable of framing records-related 'problems' in ways that account for their dynamic and multiform nature. This paper introduces the idea that alternative, 'softer' approaches to the analysis of organizational functions, structures, agents, and artifacts may usefully complement the 'hard', engineering-like approaches typically drawn on by information and records specialists. Three interrelated theoretical and methodological frameworks-namely, Soft Systems Methodology, Adaptive Structuration Theory, and Genre Theory-are discussed, with the purpose of highlighting their contributions to our understanding of the records context. © 2010 Springer Science+Business Media B.V.

Fig. 8. Annotation of the publication with eid = 2-s2.0-79451471007.

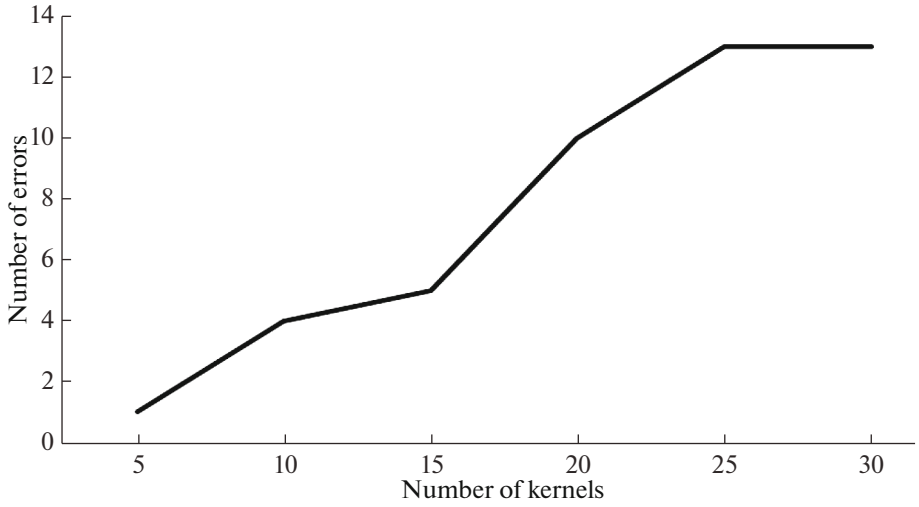

Fig. 9. The dependence of the classification quality on the number of kernels.

worth noting that in some cases this error occurred because of categories that were similar in terminology, but related to in different scientific fields. As an example, these are the *Aquatic Science* category from *Life Sciences* and the *Oceanography* category from *Physical Sciences*.

As an example, we give a test file with eid = 2-s2.0-57649228732, which was indicated to have two categories: Aquatic Science and Plant Science. The method defined the *Oceanography* category. The text of the annotation is shown in Fig. 7.

In other cases of error, its character failed to be determined. Thus, in the test file with eid = 2-s2.0-79451471007, the *Aquatic Science* category of the *Life Sciences* scientific field was defined instead of the *Library and Information Sciences* and *History* categories from the *Social Sciences* field (Fig. 8).

2.3. The Influence of the Number of Kernels on the Quality of the Classification

Figure 9 shows the dependence of the number of errors on the number of kernels that participated in the classification. The results show that the expansion of the number of categories leads to the increase in the number of errors.

While in case of five kernels only one test from the *Algebra and Number Theory category* was defined

incorrectly, the number of errors increased to 13 with 30 categories. Meanwhile, for both 5 and 30 kernels, the tests from the *Surgery* category were defined correctly.

Thus, the initial selection of the number and composition of categories affects the accuracy of the classification.

CONCLUSIONS

This study has shown that the method of classification of scientific texts based on the information compression algorithms shows different efficiencies depending on the conditions in which the classification is carried out. Using arbitrary annotations of publications indexed in the *Scopus* BDB, the number of classification errors in the kernel was 32%. The selection of kernels from annotations of highly cited publications made it possible to reduce their number to 12%. This is presumably due to the fact that such annotations use the vocabulary inherited by other publications, which is more characteristic for the area of science.

Removing stop words and publisher names in most cases affects the quality of the classification negatively. Perhaps this is due to fact that the name of publishers uses special terms, as well as the fact that length of the annotation is reduced.

In addition to the composition of kernels, the quality of the classification is affected by the initial number of categories: the fewer the categories that are involved in the classification and the greater the terminological difference between them, the higher the quality of this classification is.

However, the original classification of the articles was based on the *Scopus* journal classification, which also has errors, whose number is extremely difficult to measure. This introduces an objective error in our research.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Barakhnin, V.B., Kozhemyakina, O.Yu., Pastushkov, I.S., and Rychkova, E.V., Automated classification of Russian poetic texts by genres and styles, *Vestn. Novosib. Gos. Univ., Ser.: Lingvist. Mezhkul't. Kommun.,* 2017, vol. 15, no. 3, pp. 13–23.
2. Batura, T.V., Formal methods for determining authorship of texts, *Vestn. Novosib. Gos. Univ., Ser.: Inf. Tekhnol.,* 2012, vol. 10, no. 4, pp. 81–94.
3. Dos Santos, C.N. and Gatti, M., Deep convolutional neural networks for sentiment analysis of short texts, *COLING 2014—25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers,* 2014, pp. 69–78.
4. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M., Short text classification in twitter to improve information filtering, *SIGIR 2010 Proceedings—33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2010, pp. 841–842.
5. Kiritchenko, S., Zhu, X., and Mohammad, S.M., Sentiment analysis of short informal texts, *J. Artif. Intell. Res.,* 2014, vol. 50, pp. 723–762.
6. Ryabko, B.Y., Gus'kov, A.E., and Selivanova, I.V., Information-theoretic method for classification of texts, *Probl. Inf. Transm.,* 2017, vol. 53, no. 3, pp. 294–304. https://link.springer.com/article/10.1134/S0032946017030115.
7. Selivanova, I.V., Ryabko, B.Ya., and Guskov, A.E., Classification by compression: Application of information-theory methods for the identification of themes of scientific texts, *Autom. Doc. Math. Linguist.*, 2017, vol. 51, no. 3, pp. 120–126.
8. Hall, G.M., *How to Write a Paper,* John Wiley & Sons, Ltd., 2013.
9. Perianes-Rodriguez, A. and Ruiz-Castillo, J., A comparison of the Web of Science and publication-level classification systems of science, *J. Inf.,* 2017, vol. 11, no. 1, pp. 32–45.
10. Shu, F., Julien, C.A., Zhang, L., Qiu, J., Zhang, J., and Lariviere, V., Comparing journal and paper level classifications of science, *J. Inf.,* 2019, vol.13, no. 1, pp. 202–209.
11. Topic Prominence in Science is now available to SciVal users. http://elsevierscience.ru/news/428/topic-prominence-in-science-stali-dostupny-polzovatelyam-scival. Accessed October 14, 2019.
12. Waltman, L. and van Eck, N.J., A new methodology for constructing a publication-level classification system of science, *J. Am. Soc. Inf. Sci. Technol.,* 2012, vol. 63, no. 12, pp. 2378–2392.
13. UDC, LBC, ISBN as required elements of the publication's output. https://www.ipu.ru/structure/information-services/polygraphy/20804. Accessed October 14, 2019.
14. 1297.0—Australian and New Zealand Standard Research Classification (ANZSRC), 2008. https://www.abs.gov.au/Ausstats/abs.nsf/Latestproducts/1297.0Main%20Features32008?opendocument&tabname=Summary&prodno=1297.0&issue=2008. Accessed October 14, 2019.
15. Passports of scientific specialties. http://arhvak.minobrnauki.gov.ru/316. Accessed October 14, 2019.
16. OKSO, All-Russian Classifier of Education Specialties. https://classifikators.ru/okso. Accessed October 14, 2019.
17. GRNTI, The State Register of Scientific and Technical Activities 2019. http://grnti.ru/. Accessed October 14, 2019.
18. Revised field of science and technology (FOS) classification in the Frascati Manual. http://www.oecd.org/science/inno/38235147.pdf. Accessed October 14, 2019.
19. Proposed international standard nomenclature for fields of science and technology. https://unesdoc.unesco.org/ark:/48223/pf0000082946. Accessed October 14, 2019.
20. Parfenova, S.L., Dolgova, V.N., Bogatov, V.V., Khaltakshinova, N.V., and Korobatov, V.Ya., Methodologi-

cal approach to the formation of rubricators-adapters for the analysis of Web of Science and Scopus directions in the context of the priorities of the Strategy for Scientific and Technological Development of the Russian Federation, *Ekon. Nauki,* 2018, vol. 4, no. 2, pp. 143–153.

21. Scopus. Content Coverage Guide. http://elsevierscience.ru/files/Scopus_Content_Guide_Rus_2017.pdf. Accessed October 14, 2019.
22. Wang, Q. and Waltman, L., Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus, *J. Inf.,* 2016, vol.10, no. 2, pp. 347–364.
23. Mendes, A.C., Science classification, visibility of the different scientific domains and impact on scientific development Scopus, *Rev. Enferm. Ref.,* 2016, vol. 10, no. 4, pp. 143–149.
24. Martínez-Frías, J. and Hochberg, D., Classifying science and technology: Two problems with the UNESCO system, *Interdiscip. Sci. Rev.,* 2007, vol. 32, no. 4, pp. 315–319.
25. Tan, S., Neighbor-weighted K-nearest neighbor for unbalanced text corpus, *Expert Syst. Appl.,* 2005, vol. 28, no. 4, pp. 667–671.
26. Jiang, L., Li, C., Wanga, S., and Zhanga, L., Deep feature weighting for naive Bayes and its application to text classification, *Eng. Appl. Artif. Intell.,* 2016, vol. 52, pp. 26–39.
27. Wang, S. and Manning, C.D., Baselines and bigrams: Simple, good sentiment and topic classification, *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012—Proceedings of the Conference,* 2012, vol. 2, pp. 90–94.
28. Lai, S., Xu, L., Liu, K., and Zhao, J., Recurrent convolutional neural networks for text classification, *Twenty-Ninth AAAI Conference on Artificial Intelligence,* 2015, pp. 2267–2273.
29. Li, S., Hu, J., Cui, Y., and Hu, J., DeepPatent: Patent classification with convolutional neural networks and word embedding, *Scientometrics,* 2018, vol. 117, no. 2, pp. 721–744.
30. Li, Y.H. and Jain, A.K., Classification of text documents, *Comput. J.,* 1998, vol. 41, no. 8, pp. 537–546.
31. Xia, R., Zong, C., and Li, S., Ensemble of feature sets and classification algorithms for sentiment classification, *Inf. Sci.,* 2011, vol. 181, no. 6, pp. 1138–1152.
32. Šubelj, L., van Eck, N.J., and Waltman, L., Clustering scientific publications based on citation relations: A systematic comparison of different methods, *PLoS ONE,* 2016, vol. 11, no. 4, pp. 1–23.
33. Liu, X., Yu, S., Moreau, Y., Janssens, F., Moor, B.D., and Glanzel, W., Hybrid clustering by integrating text and citation based graphs in journal database analysis, *IEEE International Conference on Data Mining Workshops,* Miami, 2009, pp. 521–526.
34. Waltman, L., Boyack, K.W., Colavizza, G., and van Eck, N.J., A principled methodology for comparing relatedness measures for clustering publications, arxiv:1901.06815. https://arxiv.org/ftp/arxiv/papers/1901/1901.06815.pdf. Accessed October 14, 2019.
35. Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N., and Börner, K., Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches, *PLoS ONE,* 2011, vol. 6, no. 6, pp. 1–11.
36. Zhang, B., Chen, Y., Fan, W., Fox, E.A., Gonçalves, M.A., Cristo, M., and Calado, P., Intelligent GP fusion from multiple sources for text classification, *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management,* Bremen, 2005.
37. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., and Jain, A., Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature,* 2019, vol. 571, pp. 95–98.
38. Borrajo, L., Romero, R., Iglesias, E.L., and Redondo Marey, C.M., Improving imbalanced scientific text classification using sampling strategies and dictionaries, *J. Integr. Bioinf.,* 2011, vol. 8, no. 3, pp. 1–15.
39. Sinclair, G. and Webber, B., Classification from full text: A comparison of canonical sections of scientific papers, *Proc. of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications,* Geneva, 2004, pp. 66–69.
40. Riloff, E., Little words can make a big difference for text classification, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Seattle, WA, 1995, pp. 130–136.

Translated by L. Solovyova