Methodology for the Development of Dictionaries for Automated Classification System

Alexey Artamonov[*], Dmitry Kshnyakov[†], Valeriya Danilova[‡], Andrey Cherkasskiy[§], Ilya Galin[**]

National Research Nuclear University MEPhI, Moscow, Russian Federation.
aaartamonov@mephi.ru, kshnyakov.do@yandex.ru, danilova_lera94@mail.ru, aicherkasskij@mephi.ru, ilia.galin@gmail.com

Abstract
The paper describes a relevant task for research area specialists to define categories for the incoming information and scientific materials by using automated systems (software) and the methodology of developing of dictionaries for such systems. The methodology is based on studying of existing classification codes and developing dictionaries, which contain the most relevant and frequent keywords.

Keywords: Information classification, classification code, scientific category, keyword dictionary

1 Introduction

The issue of classification of the incoming data stream invariably arises while working with information and analytical data. This task is relevant for any type of information systems, both bibliographic and multimedia. The article presents a method for solving the problem of classification of the scientific and technical information flow. The urgency of the current task is also related to a significant increase in the amount of information that generates the Big Data problem.

Historically, the solution of classification problems was solved by the development of classification codes, so in the Russian Federation a GASNTI (now GRNTI) classification code was developed in 1984, in Europe the most used code is OECD Fields of Science. In the Russian Federation one of the mandatory condition for the submission of scientific papers is assigning a GRNTI category.

[*] Development of methodology concept
[†] Development of dictionaries
[‡] Conduction of the experiments
[§] Conduction of the experiments
[**] Analysis of the working documents

However, the processes of globalization have led to a blurring of the boundaries between categories in classification codes and to the emergence of interdisciplinary studies and categories. Also, new sources of scientific and technical information appeared, such as specialized web-sites, which do not display GRNTI category. As well the issue of working with a collection of documents in multiple national languages is not still resolved.

The article discusses a method for developing dictionaries for automatic thematic categorization of incoming data materials. In this case, the thematic areas in the broad sense are discussed, because of information specification for a specific user will be carried out later in accordance with his individual preference.

2 Methodology

The primary task in the development of an automated classification system the flow of information materials is the definition of a classification code on the basis of which a proper categorization will occur. The main types of information materials with which the system will operate are a news message from a website, a scientific publication, a patent, etc.

Various world classification systems of information were reviewed. In all of them it can be highlighted an essential feature associated with the complex structure and oversaturation of the number of research areas. So, for example, OECD Fields of Science contains 200 categories, Web of Science Subject Categories - 255, GRNTI - 868. The developed system should allow to dynamically change the description of each thematic area without significant expenses. The solution of this problem was developed on the basis of the Web of Science categories. This choice is validated by the presence of most of the research areas in this database and the fact that they are common.

After determining the classification code, it was necessary to describe each of the areas. The main way of describing the research area is to compose a thematic dictionary.

One of the approved ways of composing the dictionary is the dialogue mode with a research area specialist. In such a way dictionaries were developed in the MIAS on the thematic areas "Plasma Physics" and "Laser Industrial Technologies", which have approved themselves in evaluation on the search for thematic information on the Internet.

However, in this case, the process of interviewing research area specialists in 235 areas is too time-consuming, and, probably, due to narrow-thematic focus of specialists appears an issue of large number of errors. Therefore, it was necessary to develop a unified method for constructing such dictionaries. The solution of this problem became possible due to the use of frequency analysis of words to the corresponding articles of the Web of Science.

In the pilot project, the collection of information materials was received, namely scientific articles on several topics. In this case, author's keywords for each article were analyzed. In the thematic area “Materials Science, Coatings & Films”, a sample of 50 billion keywords was received. After the removal of the duplicate keywords, 95.6 thousand terms were received (Figure 1).

Keyword	Count
THIN-FILMS	4951
THIN FILMS	1793
CHEMICAL-VAPOR-DEPOSITION	1628
COATINGS	3330
ELECTRODEPOSITION	1254
FILMS	4203
PHOTOLUMINESCENCE	1139
SOLAR-CELLS	1046
XPS	1133
MICROSTRUCTURE	3321
DEPOSITION	3463
MAGNETRON SPUTTERING	668
OPTICAL-PROPERTIES	1659
SILICON	1658
FABRICATION	1940
SPUTTERING	707
TRIBOLOGICAL PROPERTIES	547
NANOSTRUCTURES	987
ATOMIC LAYER DEPOSITION	566
CORROSION RESISTANCE	576
PULSED-LASER DEPOSITION	554
NANOPARTICLES	2610
CORROSION-RESISTANCE	566
ELECTROCHEMICAL ELECTRODES	244
TIO2	931
ALLOYS	1068

Figure 1. Dictionary with keywords

The first 500 most frequent keywords were picked from the dictionary and then ranked from 500 to 1 in accordance with frequency.

Thus, dictionaries were compiled on 10 thematic areas on which the system was tested. As a result of the work, there was developed a system that allows to extract the relevant keywords from the incoming information materials and to rank them in accordance with the category.

3 Results

Based on the received dictionaries, the experiments were conducted on the informational materials classification. It was taken four articles from three different research areas, which were represented by dictionaries. For each of the articles, a distribution was made for the entered categories. The results were exported to the .xlsx document.

For example, the article "Self-assembled 20-nm Cu-64-micelles enhance accumulation in rat glioblastoma" was examined, in Web of Science this article referred to following categories: Chemistry, Multidisciplinary; Pharmacology & Pharmacy.

After processing the article following data was received, which are presented in Figure 2.

The system defined that categories “Chemistry, Multidisciplinary” and “Pharmacology & Pharmacy” are first and second in the list of the most relevant categories to the article. System effectiveness approved.

Web of Science Category	%%
Chemistry, Multidisciplinary	4.4
Pharmacology & Pharmacy	3.8
Material Science, Biomaterials	3.6
Radiology, Nuclear Medicine	3.5
Biophysics	3.5
Medicine, Research & Experimental	3.4
Biochemical Research Methods	3.1
Chemistry, Analytical	2.9
Engineering, Biomedical	2.9
Others	68.9

Figure 2. Category distribution for the first article

The second article "GaN-based flip-chip LEDs with highly reflective ITO/DBR p-type and via hole-based n-type contacts for enhanced current spreading and light extraction" was examined, in Web of Science this article referred to following categories: Optics; Physics, Multidisciplinary.

After processing the article following data was received, which are presented in Figure 3.

The system defined that categories “Physics, Multidisciplinary” and “Optics” are first and third in the list of the most relevant categories to the article. System effectiveness approved.

Web of Science Category	%%
Physics, Multidisciplinary	7.0
Materials Science, Coatings & Films	6.6
Optics	5.5
Physics, Condensed Matter	5.5
Materials Science, Multidisciplinary	5.4
Physics, Fluids & Plasmas	5.3
Chemistry, Physical	3.9
Chemistry, Multidisciplinary	3.9
Instruments & Instrumentation	3.6
Others	53.3

Figure 3. Category distribution for the second article

The third article "Study of Co, Ru/SBA-15 type materials for FischereTropsch synthesis in fixed bed tubular reactor: I. Effect of the high Ru content on the catalytic activity" was examined, in Web of Science this article referred to following categories: Chemistry, Physical; Electrochemistry; Energy & Fuels.

After processing the article following data was received, which are presented in Figure 4.

The system defined that categories “Chemistry, Physical”, “Electrochemistry” and “Energy & Fuels” are first, second and fifth in the list of the most relevant categories to the article. System effectiveness approved.

Web of Science Category	%%
Chemistry, Physical	6.0
Electrochemistry	5.7
Engineering, Chemical	5.5
Chemistry, Multidisciplinary	5.3
Energy & Fuels	5.0
Chemistry, Applied	4.6
Physics, Applied	4.6
Chemistry, Inorganic & Nuclear	4.4
Materials Science, Coatings & Films	4.2
Others	54.7

Figure 4. Category distribution for the third article

The fourth article "Mutational profiling of brain metastasis from breast cancer: matched pair analysis of targeted sequencing between brain metastasis and primary breast cancer" was examined, in Web of Science this article referred to following categories: Oncology; Cell Biology

After processing the article following data was received, which are presented in Figure 5.

The system defined that categories “Cell Biology” and “Oncology” are third and eights in the list of the most relevant categories to the article. System effectiveness partially approved.

Web of Science Category	%%
Medical Laboratory Technology	5.7
Medicine, Research & Experimental	5.5
Cell Biology	5.5
Genetics & Heredity	5.3
Biophysics	5.2
Biology	4.9
Biochemistry & Molecular Biology	4.4
Oncology	4.3
Anatomy & Morphology	4.2
Others	59.4

Figure 5. Category distribution for the fourth article

Furthermore, all articles have been read, and their content analysis was performed. Based on comparative data the experiment was in general successful, however, it is necessary to scale the system to all 235 categories and to build the experiment on a larger sample.

Moreover, the analysis of the words, identified in the text, showed that not all the terms from the dictionary were selected which related to the algorithms and words. It is necessary to check the keywords from the dictionary at a further stage in the dialog mode and lead them to a form available to work with regular expressions. For example: reaction -> react, accelerated radiotherapy -> accelerat radiotherap.

In addition, while filtering dictionaries it was excluded the following types of terms:

- The names of scientists, universities, research centers, laboratories;
- Designation of dates, years, centuries, numbers and numerals;
- Abbreviations that cannot be uniquely determined;
- The terms "not similar" to the keywords / that cannot be labeled as a keyword – for example, "100th anniversary symposium";
- Terms that cannot strictly define the subject or category – for example, "research", "data", "test";
- Latin expression as "ad hoc", "ab initio" "de jure".

Then the experiment was repeated. Based on the experiment it was found that 3 of the 4 articles was defined correctly for the category, and it was made on the basis of 500 words.

4 Conclusion

Based on the data obtained as a result of the experiments, it can be concluded that it is advisable to use this method in order to determine the thematic direction of informational materials. In addition to that, it is necessary to mention some difficulties associated with the need to obtain a representative array of documents for each thematic area, which is impossible due to closed sources and the need to process all keywords in an interactive manner and bring them to a convenient view. It is also worth mentioned that the system of automated definition of the incoming material category is developed as a tool for the specialist in the thematic area and, moreover, the system helps to significantly reduce time input for working with informational materials.

5 Acknowledgements

This work was supported by the Competitiveness Program of NRNU “MEPhI”.

References

Anastasia Ananieva, Boris Onykiy, Alexey Artamonov, KristinaIonkina, Ilya Galin, Dmitry Kshnyakov (2016) Thematic Thesauruses in Agent Technologies for Scientific and Technical Information Search. *Procedia Computer Science, 88*, Pages 493-498

Ananieva, A. G., Artamonov, A. A., Galin, I. U., Tretyakov, E. S., & Kshnyakov, D. O. (2015, November). Algoritmizatiom of search operations in multiagent information-analytical systems. *Journal of Theoretical and Applied Information Technology, 81(1)*, pp. 11-17.

Kemp, D. A. (1974). Relevance, pertinence and information system development, Information *Storage and Retrieval, 10(2)*, 37-47.