

Determination of Thematic Proximity of Scientific Journals and Conferences using Big Data Technologies

A.S. Kozitsin ^[0000-0002-8065-9061], S.A. Afonin ^[0000-0003-3058-9269] and D.A. Shachnev ^[0000-0002-5940-9180]

¹ Research Institute of Mechanics, Moscow State University Lomonosov
alexanderkz@mail.ru, serg@msu.ru, mitya57@gmail.com

Abstract. The number of journals published in the world is very large. In this regard, a software toolkit is needed that will allow thematic links of journals to be analyzed. The algorithm developed by the authors and presented in this work uses a graph of co-authorship to analyze the thematic proximity of journals. The algorithm is insensitive to the language of the journal and selects similar journals in different languages, which is difficult to implement for algorithms based on the analysis of full-text information. The algorithm was tested in the scientometric system IAS «ISTINA». In the interface developed for these purposes, the user can select one journal that is close to them by subject, and the system will automatically generate a selection of journals that may be of interest to the user both from the point of view of studying the materials available in them and from the point of view of publishing their own articles. In the future, the developed algorithm can be adapted to search for related conferences, collections of publications and scientific projects. The presence of such a tool will increase the publication activity of young employees, increase the citation of articles and citation between journals. The results of the algorithm for determining the thematic proximity between journals, collections, conferences and scientific projects can also be used to build rules in models for differentiating access to data based on domain ontologies.

Keywords: Thematic Classification, Bibliographic Data, Co-authorship Graph, Information Systems..

1 Introduction

The number of currently published scientific journals is very large. For example, in the information and analytical system (IAS) «ISTINA» [1] more than 70 thousand scientific journals and another 200 thousand different collections of scientific publications and conference materials are registered. In this regard, young scientists, graduate students and students need services that will automatically select the journals that are most relevant for their scientific interests. To solve this problem, the accumulated experience of the entire scientific community can be used.

Copyright © 2020 for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

There are several possible ways to solve the problem of determining the thematic proximity of journals. The first method is based on the use of thematic analysis of full-text descriptions of journals, texts of articles published in journals, their annotations and keywords. Based on the results of a full-text thematic analysis of such texts, it is possible to construct an assessment of the semantic proximity of the interests of users and publications in the journal. To be able to conduct such an analysis, the user must describe the area of his scientific interests with the help of keywords or upload the full texts of his articles to the system. In addition, it is necessary to have fairly accurately described thematic profiles of all journals or the full texts of articles published in these journals. Obtaining sufficiently complete full-text data is a difficult task, since in many journals open publication of the full texts of articles is not permitted.

Using only keywords for thematic analysis can give too general results. This is due to the fact that in many cases the keywords of the article do not characterize its topic, but the relationship of the article with one of the priority areas for the development of science, technology and technology in the Russian Federation. For example, the keyword “Nanotechnology” is found in articles of completely different subjects: “Development and production of new nanostructured diamond-like carbon coatings of tribological purpose”; “The development of new medical nanotechnology for the defeat of cancer cells in pediatric acute lymphoblastic leukemia”; “The use of radionuclides and sources of ionizing radiation in nanochemistry, nuclear medicine and for the study of processes occurring in the environment”; “Development and creation of supersensitive field and charge nanostructures for reading and sensor devices of nanoelectronics.”

In this regard, the use of full-text thematic analysis to solve the above problem in scientometric systems is difficult to implement.

An alternative method for assessing the proximity of journals in thematic areas is to analyze a column of co-authorship of articles published in these journals. When implementing this method, it is assumed that most authors publish their articles in thematically related journals. As a result, similar authors often publish similar authors. In contrast to the methods of full-text subject analysis, the approach based on the use of graphs of co-authorship does not require the availability of full-text information about articles, and uses only bibliographic data of articles published in journals. Such data can be obtained from scientometric systems (for example, IAS «ISTINA»), or citation systems (for example, WoS).

2 Algorithm for Assessing the Thematic Proximity of Scientific Journals

Formally, the problem of assessing the proximity of journals can be formulated as follows. It is necessary to construct a graph whose vertices are the journals, and the weights of the edges correspond to their thematic proximity.

The developed algorithm at the first step for each pair of journals calculates all pairs of articles published in these journals by one author. If only one pair of articles

corresponds to a pair of journals, then such pairs are considered unrelated. If a pair of journals corresponds to several pairs of articles, then journals are considered to be connected by an edge with a certain weight.

In the framework of this work, several methods for determining the weight of an edge were considered. The simplest method is to determine the weight of the edge equal to the number of unique authors among the corresponding pairs of articles. The main disadvantage of this method is the inability to take into account the importance of the authors for each article. In many cases, articles are written by only one author, whose last name is put first in her bibliographic description. The remaining co-authors may be involved in the work on the article slightly, and their main area of scientific activity may not coincide with its subject.

To test the hypothesis about the importance of the order of authors during a thematic analysis, an assessment was made of the proportion of articles in which the order of authors is determined by the lexicographic order, and not by significance in the work on the article. From the scientometric system of Moscow State University, all articles in journals for 2014–2017 with the number of authors from 2 to 7 were selected for analysis.

For each of the indicated number of authors, the percentage of articles L was calculated for which the correct set of authors is determined by the lexicographic order. The calculation results for a different number of authors are shown in Table 1.

Table 1. The percentage of articles with lexicographical order.

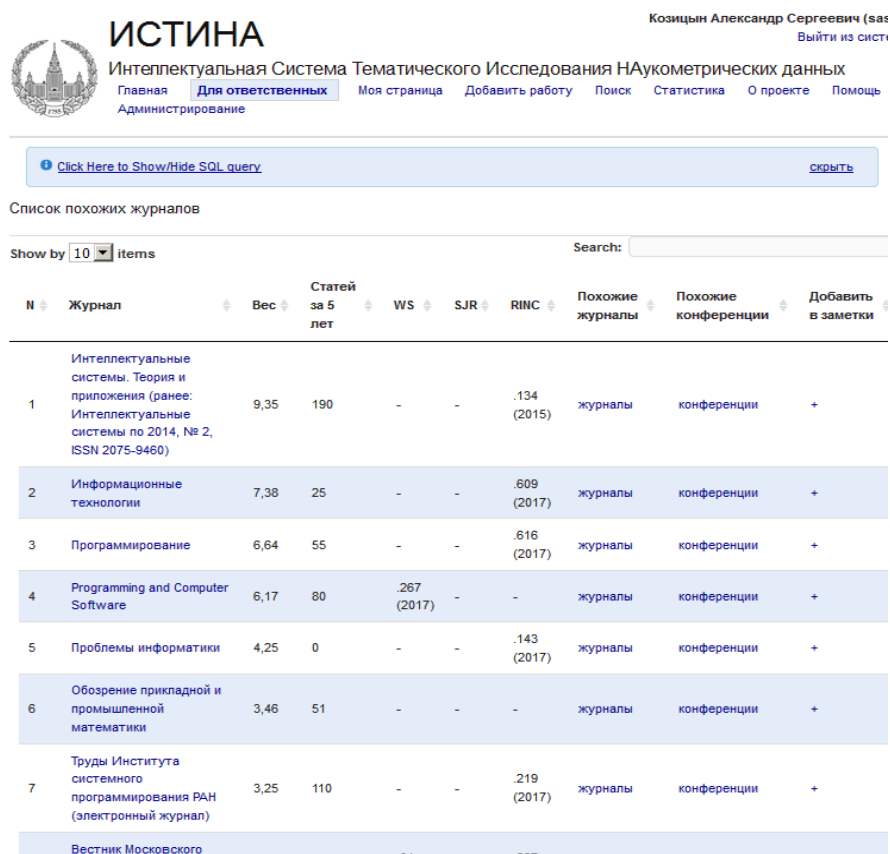
Number of authors	L
2	24%
3	16%
4	9%
5	6%
6	6%
7	3%

From the data given in the table, we can conclude that in most cases the main author is the author, who is indicated first in the bibliographic description. To account for this fact, a formula for calculating the weight of the edges was developed taking into account the position of the author in the bibliographic description of the article. The weight of the author for each article is defined as $1/2 + 1 / (2K)$ for the first author and $1 / (2K)$ for the remaining contributors, where K is the number of co-authors in the article. The degree of communication for a given author for two journals is determined from at least the maximums of his weights for the subsets of articles in each journal. The final weight of the connection edge between the two journals can be calculated as the sum of the degrees of their connection for all authors.

3 Software Implementation and Test Results First Section

When choosing a language for the software implementation of the algorithm, such features of the algorithm as the large amount of processed data, the need for quick access to the data stored in the DBMS, the small requirements for the amount of memory to create temporary data structures, and the absence of the need to conduct a dialogue with the user were taken into account.

Given these requirements, PL / SQL was chosen for implementation. The thematic proximity between the journals is calculated at specified time intervals and stored in the DBMS tables. Subsequent paragraphs, however, are indented.



Козицын Александр Сергеевич (sa) [Выйти из сист](#)

Интеллектуальная Система Тематического Исследования НАукометрических данных

Главная [Для ответственных](#) [Моя страница](#) [Добавить работу](#) [Поиск](#) [Статистика](#) [О проекте](#) [Помощь](#)
Администрирование

[Click Here to Show/Hide SQL query](#) [скрыть](#)

Список похожих журналов

Show by items Search:

№	Журнал	Вес	Статей за 5 лет	WS	SJR	RINC	Похожие журналы	Похожие конференции	Добавить в закладки
1	Интеллектуальные системы. Теория и приложения (ранее: Интеллектуальные системы по 2014, № 2, ISSN 2075-9460)	9,35	190	-	-	.134 (2015)	журналы	конференции	+
2	Информационные технологии	7,38	25	-	-	.609 (2017)	журналы	конференции	+
3	Программирование	6,64	55	-	-	.616 (2017)	журналы	конференции	+
4	Programming and Computer Software	6,17	80	.267 (2017)	-	-	журналы	конференции	+
5	Проблемы информатики	4,25	0	-	-	.143 (2017)	журналы	конференции	+
6	Обозрение прикладной и промышленной математики	3,46	51	-	-	-	журналы	конференции	+
7	Труды Института системного программирования РАН (электронный журнал)	3,25	110	-	-	.219 (2017)	журналы	конференции	+
	Вестник Московского			.01		.267			

Fig. 1. Search for related journals.

In the interface developed for these purposes [2], the user can select one journal that is close to the subject, and the system will automatically generate a selection of journals that may be of interest to the user both from the point of view of studying the materials contained in them, so and in terms of publishing their own articles (Fig. 1).

The web interface is implemented using the open DataTables library [3]. A link has been added to the information card of each journal to go to a table with a list of thematically similar journals. This table shows the names of related journals and measures of similarity. In addition, in order to be able to quickly assess the authority of each journal from the list, the table shows the number of publications in this journal over 5 years (registered in the ISTAINA system), as well as data from the Web of Science and the Russian Science Citation Index. In order to conveniently navigate the graph of proximity of journals, the developed interface also implements the ability to follow links to a list of similar journals directly from each element of the list. By means of the DataTables library for a quick search by journal names, a mechanism for quick filtering by part of the journal name is implemented.

For the convenience of the user, it is possible to add the selected journal to notes, which can later be viewed, edited, and also used for subsequent searches. Additionally, it is possible to select conferences similar in topic.

Testing of the developed software implementation of the algorithm was carried out according to the following methodology. From the obtained results, 200 pairs of log links were randomly selected. Experts carried out a manual assessment of the coincidence of the topics of the journals with setting points (2 - accurate; 1 - not entirely accurate; 0 - error). The total score was divided by twice the number of analyzed bonds. The accuracy rating for this technique was 78%.

As an example of algorithm errors, one can cite, for example, a list of journals that are defined as being similar in theme to the publication of Proceedings of the Higher School of the USSR Ministry of Internal Affairs: Philosophical Sciences; "Logical research"; "Proceedings of MSTU" MAMI ""; "Logical and philosophical research"; "Bulletin of Moscow University. Series 7: Philosophy. " Such errors may arise as a result of too broad a subject area of articles accepted in the journal

4 Conclusion

The algorithm described in this paper allows us to automatically evaluate the degree of thematic proximity of scientific journals based on bibliographic descriptions of articles and without using full-text versions of articles. It should be noted that the algorithm is insensitive to the language of the journal and selects similar journals in different languages, which is difficult to implement for algorithms based on the analysis of full-text information.

In the future, the developed algorithm can be adapted to search for related conferences, collections of publications and scientific projects. The presence of such a tool will increase the publication activity of young employees, increase the citation of articles and citation between journals.

The results of the algorithm for determining the thematic proximity between journals, collections, conferences and scientific projects can also be used to build rules in models of differentiating access to data based on domain ontologies [4].

This work was supported by the Russian Foundation for Basic Research, project 18-07-01055. Subsequent paragraphs, however, are indented.

References

1. Sadovnichiy, V., Vasenin, V.. Intelektualnaya sistema tematiceskogo issledovaniya naukometricheskikh dannyh: predposylki sozdaniya I metodologija razrabotki. Part 1. Programnaja inzhenerija, 9(2), pp. 51–58 (2018).
2. IAS ISTINA. <https://istina.msu.ru>, last accessed 2019/11/10.
3. Library datatables. <https://datatables.net/>, last accessed 2019/11/10.
4. Afonin, S.: Ontology models for access control systems. In: 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), pp. 1–6, Vladivostok (2018), <https://doi.org/10.1109/RPC.2018.8482178>.