The Use of Thematic Analysis Methods in Scientometric Systems

Alexander Kozitsin [0000-0002-8065-9061], Sergey Afonin [0000-0003-3058-9269]
and Dmitry Shachnev[0000-0002-5940-9180]

Research Institute of Mechanics, Lomonosov Moscow State University
alexanderkz@mail.ru

Abstract. In many modern scientometric systems and citation systems, various mechanisms of thematic search and thematic filtering of information are presented. In most cases, a full-text approach is used for thematic analysis of articles and journals, which has a number of limitations. The use of algorithms based on graph analysis, both independently and in conjunction with full-text algorithms, eliminates these limitations and improves the completeness and accuracy of thematic search. The algorithm developed by the authors and presented in this work uses the co-authorship graph 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 information analysis 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 in terms of studying the materials available in them, and in terms of publishing the user's own articles. In the future, the developed algorithm can be adapted to search for similar 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 the citation between journals. The results of the algorithm for determining thematic proximity between journals, collections, conferences and scientific projects can also be used to build rules for data access control models based on domain ontologies.

Keywords: thematic classification, bibliographic data, co-authorship graph, information systems.

1 Introduction

The use of modern methods of thematic analysis for analytical processing of large amounts of information is currently used in almost all spheres of human activity, including scientometrics. The results of thematic analysis of scientific information can be used to clarify scientometric indicators, make management decisions, search for information, and determine the rules for access to information.

The calculation of scientometric indicators is used to assess the significance of articles (citation), the credibility of journals (impact factor, h5-index, h5-median), the impact on the scientific community of individual authors (Hirsch index and g-index), assessing the activities of organizations in general (i-index) [1]. However, many authors note that the characteristics of the distribution of the absolute numerical values of scientometric indicators have a significant dependence on the analyzed thematic area [2]. For example, the values of the citation index of articles for the last 2 years have a different median for physics and mathematics, since mathematical articles are cited for a longer period, but are slower to "gain" the number of citations. A similar inconsistency is observed in the journals as a whole. For example, the best Russian journals according to the Russian Science Citation Index (RSCI), presented on the statistics page elibrary.ru/titles_compare.asp, for 2018 as of 20.04.2020 have the following citation indicators for different rubrics: physics 9200; biology 4600; mathematics 3600; mechanics 1500; computer science 1100. In this regard, it is incorrect to compare the absolute values of scientometric indicators of articles, journals or authors from different thematic areas. In such cases it is necessary to use the normalized average citation count [3] or other similar indicators that take into account the thematic area of the research. The construction of such normalized indicators requires thematic classification of large volumes of scientific articles and journals.

When performing management activities, the use of the results of thematic analysis makes it possible to assess the state of various areas of research, to compare them with the world level, to identify new thematic areas to determine the policy of allocating material resources for stimulating scientific activity. At the same time, it is necessary to assess not only the current values of indicators, but also their dynamics over time, as well as world indicators. For example, a decrease in indicators for a certain research topic while the same indicators in the world are increasing may mean an outflow of scientific personnel in organizations from this research area or equipment obsolescence.

Another important area of application of thematic analysis is the creation of effective mechanisms for conducting information search. Search objects can be publications, journals, persons, organizations and other objects. Based on the thematic classification and clustering, the following demanded tasks can be solved: searching for published materials on a given topic, finding the most authoritative experts in a certain subject area, determining a list of journals for publication and assessing their significance, determining new thematic directions in any area, and search for research teams.

Determination of thematic links between objects of the information system [4] can also be used to automatically build ontologies and define data access rules in the attribute-based logical access control models (ABAC) [5], which have now largely supplanted the old access control models: role-based (RBAC),mandatory (MAC) and discretionary (DAC).

Many large scientometric and citation systems have tools for thematic data analysis.

2 Use of thematic analysis in modern systems

The capabilities of thematic analysis of various scientometric systems differ in the type of information being processed, types of classifiers, information sources, and the set of classification and clustering methods used.

The Web of Science (WoS) project uses keyword indices and thematic classifiers to conduct thematic search. Indexing by keywords is performed using the Author Keywords, which the authors specify manually when adding the article. Also, there is indexation of keywords and terms (KeyWords Plus), that are automatically extracted from the names of articles cited in the work. Keyword indexing allows search and additional filtering using user-defined terms. For indexing by thematic classifiers, two main classifiers are used: the one-level Web of Science Categories classifier for journals (containing 250 categories) and the two-level Research Area classifier for articles (150 science areas). Apart of that, an additional Essential Science Indicators classifier that consists of 22 categories is used. The project implements the Manuscript Matcher service, which is able to build recommendations on the selection of a journal based on the text of the manuscript proposed for publication [6].

The Google Scholar project uses a two-level classifier with 8 first-level elements and 400 second-level elements. Subdivision by topic can be used at scholar.google.com/citations to filter journals when displaying their indicators (h5-index and h5-median), which allows building more objective rankings for each of the thematic areas specified in the classifier. Thematic filtering is only available for English language journals. For Russian-language journals, as well as for journals published in other languages, thematic classification is not supported.

To correct the data, in accordance with its basic methodology, Google actively uses user interaction to collect information "from bottom to top", allowing authors to create their own pages with a list of articles, a photo, descriptions of interests (Google Scholar Citations). Adding items to the profiles can be done automatically, automated (selected options are shown to the user), or manually with specifying full bibliographic data.

The Scopus project uses the All Science Journal Classification Codes (ASJC) two-level classifier, containing 4 first-level records and about 350 second-level records to classify journals. On the page www.scival.com one can see the distribution of journals by area and the relative normalized characteristics for the selected sections of the classifier, the change in the number of publications by topic over time and other indicators. The data is available with a paid subscription.

The RSCI project uses the three-level State Rubricator of NTI of Russia (GRNTI), containing about 8 thousand rubrics (elibrary.ru/rubrics.asp). Thematic classification can be used to search for journals and articles, as well as to filter the results of the selection of journals when issuing their scientometric parameters.

The Open Academic Graph (OAG) Project, an extended version of the Microsoft Academic Graph (MAG), contains 170 million articles with citation references. The project is not a scientometric system or a citation system, but the project data can be used to test the algorithms of scientometric systems. The data can be freely downloaded from the project website www.aminer.org/open-academic-graph.

In addition to the above classifiers of commercial systems, there are a number of generally accepted classifiers that are not associated with any particular citation system. At the world level, the most famous is the three-level classifier OECD Fields of Science, containing more than two hundred rubrics, which was planned to be used, among other things, in the "Map of Russian Science" project. In many Russian journals, for the thematic classification of articles by the authors themselves, a more detailed Universal Decimal Classification (UDC) is used, containing more than 150 thousand rubrics. Also, for the thematic classification of various scientific materials, the VINITI Rubricator is used, containing more than 53 thousand rubrics, and a number of other thematic classifiers: Classifier of the Russian Scientific Foundation (RSF) [7]; Classifier of the Russian Foundation for Basic Research (RFBR) [8]; International Patent Classification (IPC) [9], All-Russian Classification of Standards (RCS) [10], Mathematics Subject Classification (MSC) [11]; Journal of Economic Literature Classification (JEL) [12] and others (scs.viniti.ru/MapService/treeList.aspx). In the presence of such a variety of classifiers, it is natural that various projects for their matching appear. For example, the project on comparing the Scopus and OECD classifiers [13], as well as the VINITI project [14].

The projects enumerated above aimed to develop systems of counting citation indicators of scientific publications and to classify them by field of science. The next development step was the emergence on their basis of systems for assessing the scientific activities of organizations as a whole.

The Spanish project SCImago Journal & Country Rank of the University of Granada (or "Atlas of Science") evaluates aggregated data on scientific activities in Spain, Portugal and South America based on Scopus data. The project website www.scimagojr.com provides indicators not only for scientific journals, but also for countries as a whole. The SJR index, developed by the authors of the project, is an alternative to the impact factor.

The Faculty Scholarly Productivity Index (FSPI) project evaluates the metric indicators of USA universities based on Scopus data. In addition to the number of publications and citation rates, this project uses data on the received awards and prizes, as well as on the volume of federal research funding to calculate the rank of the university. More than 350 universities are ranked based on the aggregated data.

The Times Higher Education (THE) project aims to assess universities around the world [15]. The World University Rankings index developed within the framework of the project is built on the basis of WoS citation data, which make up 32.5% of the rating [16]. In addition, the subjective assessments of experts, the amount of funding for the research carried out, the attraction of foreign students and teachers, as well as the introduction of the university's developments into industry are taken into account.

The QS World University Rankings project [17] assesses in terms of research and teaching performance, student-teachers ratio, average citation index per faculty member, reputation with employers, and the number of international students and teachers.

The Academic Ranking of World Universities (ARWU), often referred to as the "Shanghai Ranking" (www.shanghairanking.com), takes into account the receipt of Nobel Prizes by university alumni, the number of published articles in the "Nature" and "Science" journals and citation rates.

It should be noted that making such comparisons without taking into account the language of instruction, as well as evaluating journals without taking into account their thematic area, does not give quite accurate results [18]. For example, a comparison of universities all over the world in terms of citation only in English-language journals indisputably shows only the fact that the percentage of teachers and students who are fluent in English at universities in the USA, England and Canada is significantly higher than in Russia or other non-English-speaking countries. Likewise, a comparison of the proportion of foreign students and teachers at universities with English and Japanese as the language of instruction shows not so much the level of education in the institution as the number of foreigners who are fluent in this language.

For scientometric systems that have a goal to obtain objective and balanced scientific assessments of the quality of products, taking field of study into account in the analysis of scientometric data, including language, subject area, and other similar characteristics, it is a necessary requirement when constructing objective scientometric indicators.

3 Thematic analysis using textual information and classifiers

In the process of developing and maintaining the ISTINA scientometric system, special attention has always been paid to the development of methods for intelligent information analysis, including methods of thematic analysis. The volume of data processed by this system is significantly inferior to the world citation systems, since it covers only 28 organizations, 900 thousand publications, 70 thousand monographs and 13 thousand patents. However, the number of types of data used is much higher. In addition to publications and patents, the system contains complete information about data on scientific projects (R&D, grants), conference talks, dissertations and diplomas, on the participation of employees in the activities of various councils and editorial boards, prizes and awards they receive, courses taught and other data [19]. In addition, the information in the system is double checked. The basic principle of the system is the movement of information "from bottom to top". At the first stage, the user, as the most interested person, registers all their works in the system, which are displayed on their personal page. At the second stage, the responsible employees of the departments confirm the accuracy of the data. A similar method of collecting information using the creation of personal pages is currently used in the Google Scholar Citations project by Google corporation, which is the leader in the text processing market. But due to objective reasons, it is impossible to organize the second stage of verification in this system.

One of the simplest ways to conduct thematic analysis is to use classifiers with manual comparison of objects and thematic classes, including the use of thematic classification of journal cards. This approach is used in Scopus, WoS and RSCI.

When the ISTINA scientometric system was created, methods of analysis using the categorization of journals on various static rubricators were implemented to analyze employees' and organizations' activity in different thematic areas. When using an interactive interface on the organization's statistics page [20], data on the distribution of the number of articles, citations in WoS, the number of authors and other aggregated characteristics by Scopus and GRNTI rubrics are presented. Data can be provided both for

individual departments and for the organization as a whole with the ability to filter by year of publication, by overcoming the threshold value of the analyzed indicator and by belonging to a group of journals: journals from Scopus, journals from WoS Top 25%, journals from the Higher Attestation Commission list, collections of articles and others. It is also possible to separately specify the metric for filtering by the threshold value and the metric to be displayed on the chart. For example, filter by the number of articles, and display the number of links to an article.

Analysis is possible both at the level of the organization as a whole, and at the level of each department separately. It should be noted that the choice of the level of aggregation is especially useful given the ambiguity of determining the department for each individual publication. In large scientific organizations, a large number of articles are published in co-authorship by employees of different departments. With the traditional method of counting, when aggregated data on individual departments are counted independently and then added together, these articles are counted several times, resulting in distortion of the totals. Using the ability to aggregate source data both at the organizational level and at the departmental level makes such estimates more accurate and objective.

This approach provides the user with the opportunity to assess the degree of publication activity of employees in various thematic areas. However, it does not allow analyzing information with a sufficient degree of detail. The rubricator is static and additional detailing within one rubric is not possible.

The second possible approach is to determine the subject and search for information by keywords, abstracts or full-text articles. Keywords can be specified by the authors of a work when it is registered in the system or calculated during the indexing process from the abstract, full-text articles, or a list of cited literature, for example, Author Keywords and KeyWords Plus in WoS. This approach makes it possible to better concretize the search topic, which is necessary for tasks such as determining new thematic areas or searching for information on a specific information need of the user. It should be noted that the use of such thematic analysis is not limited to information search. For example, in [21], it is proposed to use thematic analysis to assess the quality of the journal. The main hypothesis is that in "good" scientific journals, articles should be devoted to a fixed set of topics, and these topics should change over time. Thus, after training the thematic analysis algorithm on a training set of articles from all analyzed journals, it is possible to carry out a thematic and temporary classification of articles from these journals. The quality of the journal will be proportional to the classification accuracy with which the articles contained in it have been correctly classified by journal affiliation and publication time interval.

The main difficulty in using keywords for thematic analysis is the limited set of keywords. When describing articles, authors usually specify less than 10 keywords. For example, the average number of keywords that authors specify when registering articles in IAS ISTINA is 3.8. An additional obstacle is the subjectivity of choice. At the first stage, the authors extract from the article the basic concepts that, according to their assessment, are significant at the moment. At the second stage, for each concept they specify only one of keywords describing it, excluding possible synonyms. Thus, articles with similar topics may have a non-overlapping set of keywords, and the accuracy of

determining their thematic similarity is significantly reduced. At the same time, a similar approach to finding articles that are close in subject is implemented in some citation systems. For example, one can test the quality of the selection of articles when searching for keywords in Russian-language journals on the search page of the RSCI project [22].

The WoS project provides users with the Manuscript Matcher service for selecting a journal for publication according to the text of the article. The service requires prior registration of the user to operate. After submitting the title of the article to be published and the abstract, the service determines the keywords and searches for a match with the keywords of the journals. The result is shown as a list of journals with a description of each journal, as well as a list of common keywords with a measure of similarity to the uploaded article. The service can be useful for authors who use highly specialized terms, for example, in chemistry, biology or astronomy. For more general topics, comparing terms is not very accurate. For example, for the article "Determining the thematic proximity of scientific journals and conferences using Big Data technologies" the best journals in the results of the search are "Scientometrics" and "Journal of the association for information science and technology", however, the top 5 list also contains "Journal of medical systems" and "Journal of digital imaging" which are matched by "create software tools" and "full-text information" keywords.

The RSCI project offers users a service for searching for similar articles. The user can select one of the articles already indexed in the system and request a search for similar articles by topic. But the results of such a search are even less accurate than the results of keyword search and the results of the Manuscript Matcher service. For example, for the article "Architecture, methods and means of the basic component of the ISTINA system of scientific information management" 14 thousand related articles are determined, and in the top 10 list there is not a single article that would be related to the system considered in the article or any analogue, and only one article deals with the issues of scientometrics. The top 3 results of the search by thematic similarity are: "Information technology of software architecture structural synthesis of information system", "Analysis of the asp.net development information system", "General overview of agris (agricultural research information system)".

One of the possible ways to improve the completeness of keyword search and to resolve issues of homonymy is to expand the set of keywords based on building relationships between keywords [23], as well as using translations of terms. The ISTINA project uses Wikipedia materials to automate the translation process, as well as free services from Abbyy. Keyword search is used as the first stage of thematic analysis in the developed algorithms for finding experts and selecting journals, which are currently being tested on the data of the ISTINA system. The results of ongoing research in this direction cannot yet be used for the implementation of industrial software, however, it can already be argued that using only thematic analysis based on keywords, abstracts and texts does not allow obtaining a satisfactory classification result. In this regard, the developed algorithms use a combination of full-text analysis methods and analysis methods of graph theory, which use explicit or implicit connections between classified objects.

4   First Section

The use of links between objects (or graph of objects) allows one to supplement or refine the analysis data in case of lack of information. Objects in the graph can be of the same type, for example, articles and links between articles, or they can be of different types, for example, employees and their projects. The goal of graph analysis can be to expand the search scope or clarify the significance of objects in an existing search scope.

One example of supplementing data in a graph with objects of different types is the problem of finding experts on a given topic [24]. To search for experts, the objects in the authorship graph most related to experts (articles, monographs, projects, reports, etc.) are determined, the degree of edge is determined, the keywords of the objects are extracted, an information portrait of the user is built on the basis of an expanded set of keywords and weights of the graph edges, proximity to the original search query is estimated. To test the algorithm, the data of the ISTINA scientometric system were used. The use of such algorithms in citation counting systems is difficult, since the graph of object links in them contains only two types of vertices: authors and publications. In full-fledged scientometric systems, the user's information portrait is composed of a larger number of object types, which improves the quality of the results.

An example of solving the problem of data refinement on the basis of links between similar objects is the algorithm for determining authorship of articles [25], which is implemented in the ISTINA system. It is assumed that the authors' groups have a certain stability, and the probability of publication by two authors of several joint articles is much higher than writing an article in which one of the authors is replaced by a full namesake. In accordance with this hypothesis, to resolve the ambiguity in determining the authors of the article among all possible namesakes, a co-authorship graph is constructed and the most connected component is selected.

Using the co-authorship graph, it is also possible to solve the problem of determining the thematic proximity of journals without using data of full-text analysis. The main hypothesis in the implementation of this method is the assumption that a significant part of the authors publish articles in their subject area, and, therefore, several journals in which the same set of authors are published are similar in topics. Based on this hypothesis, the thematic proximity of two journals is calculated as the weighted sum of authors who have publications in both journals. This takes into account not only the number of publications made by the author, but also the position of the author in the bibliographic metadata of the article. The link weight of an article is distributed among all authors, but the first authors carry more weight than the rest. A formal description of the algorithm is given in [26]. The main difference between this algorithm and similar algorithms that use full-text analysis or keyword analysis is insensitivity to the language of journals and, as a result, the ability to search for links between journals in different languages. In addition, the algorithm does not require lengthy training on large arrays of texts, while showing a fairly high accuracy of 78%.

The further development of the algorithm described in [26] was the work on automating the expansion of the search area for journals in the co-authorship graph. The

main premise is the assumption that the proximity relation is transitive for highly specialized journals. If two highly specialized journals are close on the subject to the third, then they are close to each other. At the same time, the generalization of this rule for all journals, including broad-scope ones, is incorrect. For example, the presence of common authors in any two journals with the journal "Bulletin of the Russian Academy of Sciences" with general thematic does not mean mutual thematic similarity of the original two journals. In this regard, it is necessary to use mathematical models with normalizing the weights of the edges in the graph of journals' connections [26]. In the course of the studies, it was shown that the best result is achieved by normalizing the weights of the edges using the total sum of the edges outgoing from each vertex. After normalization, the proximity matrix between journals is calculated based on the comparison of paths in a graph of length 3. This approach can significantly increase the completeness of thematic search. The final result is constructed by combining two lists: the closest journals in the original thematic proximity matrix and the closest journals in the extended thematic proximity matrix. Combining these lists before showing to the user can increase the completeness of search, not much reducing its accuracy.

The software implementation of the algorithm is used in the ISTINA system to provide users with a convenient interface for thematic search of journals. To perform a search, the user must select one journal known to them on a given topic, finding it by name.

After that, follow the "Related journals" link. For convenience, in the row of each journal in the list presented, an assessment of its thematic similarity with the original journal, various citation characteristics and the number of articles from this journal loaded into the scientometric system are indicated.

The user can go to the page of a journal, or continue moving through the graph of thematic links of journals using the links in the "Similar journals" column.

It should be noted that this algorithm can search for thematic links not only between journals, but also between other groups of objects with authors. In this example, the algorithm also searches for conferences similar in topic to the given journal.

Another important practical task, which can be solved using the description of relationships between objects in a scientometric system, is the task of determining the authority of experts when searching for them by thematic description. For directed graphs, the classical algorithm for assessing the authority of vertices in a graph is PageRank, which was used by Google to rank search results. The algorithm is based on the assumption that an incoming edge in the graph confirms the authority of the vertex. Moreover, the significance of this confirmation is higher when the authority of the outgoing vertex is higher. In scientometric systems, the algorithm can be effectively used to analyze the citation graph. To analyze the undirected graph of co-authorship and other similar graphs in scientometric systems, it is possible to use a number of other characteristics: degree of connectivity (the number of edges for each vertex); the degree of proximity (average shortest distance to other vertices of the graph); the degree of mediation (the number of shortest paths between all pairs of vertices passing through a given vertex); the degree of influence (the degree of connectivity in which the contribution of each edge depends on the degree of influence of the neighboring vertex, for

example, PageRank); cross-clique centrality (the number of cliques that a vertex belongs to) and others. Preliminary experiments carried out on the data of ISTINA system show that this approach can be quite effective for use in ranking the results of experts search, automatic detection of stable research teams and other similar tasks.

5   Conclusion

The use of thematic analysis algorithms for solving a number of information processing problems in scientometric systems allows us to create convenient services for searching and processing information. The combination of full-text and graph analysis methods allows to increase the accuracy and completeness of the presented results. Currently, such services are not widely used in scientific citation systems. Scientific research in this area, carried out using the data of the ISTINA project, can provide new mechanisms for searching and processing scientometric information.

References

1. Akoev, M.A., Markusova, V.A., Moskaleva, O.V., Pisliakov, V.V.: Rukovodstvo po naukometrii: indikatory razvitiia nauki i tekhnologii. Izdatelstvo Uralskogo universiteta, Ekaterinburg (2014).
2. Orlov, A.I.: Naukometriia i upravlenie nauchnoi deiatelnostiu. Upravlenie bolshimi sistemami. Spetsialnyi vypusk 44: Naukometriia i ekspertiza v upravlenii naukoi, 538–568 (2013).
3. Brichkovskii, V.V.: Naukometricheskii analiz v informatsionnom obespechenii innovatsionnoi deiatelnosti. V mire nauki, 8(174), 64–67 (2017).
4. Afonin, S.A., Kozitsyn, A.S., Shachnev, D.A.: Software Mechanisms for Scientometrical Data Aggregation Based on Ontological Representation of the Relational Database Structure. Programmnaia inzheneriia, 7(9), 408–413 (2016).
5. Afonin, S.: Ontology models for access control systems. Proc. of the 3rd International Conference Russian-Pacific Conference on Computer Technology and Applications (RPC), pp. 1–6 (2018).
6. WoS journal recommendation service, http://mjl.clarivate.com/home, last accessed 2020/10/10.
7. RSF Classifier, http://www.rscf.ru/node, last accessed 2020/10/10.
8. RFBR Classifier, http://www.rfbr.ru/rffi/ru/contest_documents, last accessed 2020/10/10.
9. IPC Classifier, http://www.fips.ru, last accessed 2020/10/10.
10. RCS Classifier, http://classinform.ru/oks.html, last accessed 2020/10/10.
11. MSC Classifier, http://www.ams.org/msc/, last accessed 2020/10/10.
12. JEL Classifier, http://www.aeaweb.org/journal/jel_class_system.html, last accessed 2020/10/10.
13. Scopus and OECD classifiers matching project, http://report03.metrics.ekt.gr/en/ appendixIII, last accessed 2020/10/10.
14. VINITI classifiers matching project, http://scs.viniti.ru/MapService/mapform.aspx, last accessed 2020/10/10.
15. Times Higher Education, http://www.timeshighereducation.com, last accessed 2020/10/10.

16. World University Rankings, http://gtmarket.ru/ratings/the-world-university-rankings/info, last accessed 2020/10/10.
17. QS World University Rankings, http://www.topuniversities.com, last accessed 2020/10/10.
18. Kincharova, A.V.: Metodologiia mirovykh reitingov universitetov: analiz i kritika. Universitetskoe upravlenie: praktika i analiz, (2) 70–80 (2014).
19. ISTINA project data, http://istina.msu.ru/statistics/activity/, last accessed 2020/10/10.
20. Organization statistics in ISTINA, http://istina.msu.ru/statistics/organization/214524/ dynamic, last accessed 2020/10/10.
21. Krasnov, F.V.: Sravnitelnyi analiz kollektsii nauchnykh zhurnalov. Trudy SPIIRAN, 18, 767–793 (2019).
22. Keywords search in RSCI, https://www.elibrary.ru/querybox.asp, last accessed 2020/10/10.
23. Afonin, S.A., Lunev, K.V.: Topic Analysis in Collection of Keyword Tuples. Programmnaia inzheneriia, (2), 29–39 (2015).
24. Vasenin, Valery, Lunev, Kirill, Afonin, Sergey, Shachnev, Dmitry: Methods for intelligent data analysis based on keywords and implicit relations: The case of "ISTINA" data analysis system. In Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings, pp. 151–155, United States (2019).
25. Kozitsyn, A.S., Afonin, S.A.: The Resolution of Ambiguities in the Identification of Authors of the Publication with the Use of Co-Authors' Graphs in Large Collections of Bibliographic Data. Programmnaia inzheneriia, 8(12), 556–562 (2017).
26. Kozitsyn, A.S., Afonin, S.A.: Discovering hidden dependencies between objects based on the analysis of large arrays of bibliographic data. Proc. of the International Conference Actual Problems of Systems and Software Engineering (APSSE 2019), IEEE Conference Proceedings, pp. 320–328, Moscow (2019).