Author’s Identification within the Subject Area in the Semantic Library

Olga Ataeva[1][0000-0003-0367-5575], Vladimir Serebryakov[2][0000-0003-1423-621X],
Natalia Tuchkova[3][0000-0001-6518-5817]

[1,2,3]Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, 11933, Moscow, Russia
[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract. The peculiarities of the task of authors identifying and determining author's contribution to publications in digital bibliographic codes are considered. The features of the problem of insufficient identification are manifested in the repetition of information, doubling, the presence of authors with completely coincidental names, self-quotation, autoplagiate and plagiarism itself. It is proposed to use publication information that has already been accumulated in the digital library in the form of related object area data and a variety of target thesaurus data, as the author and user of the library. This information contains links whereby keyword contexts, multiple co-authors, and term associations in dictionaries and thesauruses can be used to identify authorship. It is important that an array of scientific publications is considered, since they have an established traditional structure, which allows comparing fixed text elements (annotations, keywords, classifier codes, etc.). Thus, even if the names in the publications are fully matched, the question of authorship can be raised if the publications in the digital library correspond to different subject areas. Resolution of such contradictions is accomplished by evaluating a plurality of links of all elements of secondary publication information. The result of the comparison could be the addition of the author to a specific area, i.e. the extension of the addressee's thesaurus and the author's personal thesaurus, or the appearance of full namesakes in the library, but from different areas of knowledge. It has been shown that modern data analysis tools allow you to evaluate the author's contribution to publication, despite the fact that of course, only the scientific community can evaluate the real contribution to scientific research.

Keywords: Comparison of Scientific Texts, Semantic Search, Thesaurus for the Ontology of Knowledge, Information Query using the Thesaurus, Methods of Author’s Identification, Thesaurus of Addressee, Secondary Information, Individual Frequency Dictionary, LibMeta.

1 Introduction

The problems of determining who deserves to be the author of a scientific article and what is his contribution to the collective publication, if there is no reliable information in the digital collection, are resolved in various ways. Basically, a comparison of related articles and a survey of registered authors is performed, as in ResearchGate. Almost all digital resources known today face the issues related to the identification of authors in bibliographic systems. When the information is updated, a "controversial" author, a full namesake, an "old" author with a different transcription in the spelling of the surname, etc. may appear. Everyone knows the difficulties of their own identification even in such authoritative databases as WoS and Scopus, when, despite all the filters set, we get as a result of a search a list of a "mixture" of their own and other people's works, which is reflected, for example, in publication [1]. Quite often it is necessary to manually generate the necessary list, despite the mechanism of automatic formation of the author's index that exists in these systems (as well as in many others). The only exceptions are publications and editions in which the ORCID of the author is initially required. ISTINA system (IstinaResearcherID, IRID), elibrary (author's SPIN code), Scopus (Scopus Author ID), Web of Science ResearcherID, Google Scholar Citation ID have also introduced their own identifiers. The more indices the author indicates when registering in these systems and in articles when transferring to publishers, the more accurately he is identified, naturally. Some publishers make it mandatory to reference the indexes of the authors of the respective databases with which these publishers cooperate. The fact that the authors' identifiers accompany publications suggests that other methods, despite the accepted identification rules, are not reliable enough.

There are a number of requirements for articles and authors in certain specific subject areas, and they were approved, for example, for authorship in medical research, but later became generally accepted. An author is someone who participates in the development of an idea, collection and analysis of data, writing a work, and making relevant and ideologically justified changes to the text. Nevertheless, these tools are not enough to determine the author's contribution to collective research, as indicated, for example, in [2]. Moreover, in the digital age, there are options in some scientific communities: peer review of authors' contributions to research; granting publishers the right to express opinions about authorship based on the accumulated information. This weakens the previously accepted traditional norms [3].

The level of reliability, transparency and documentation of data about authors has changed. Thus, the problem of authorship is posed wider, and is not limited to secondary information when indexing in databases. This problem includes the human factor, that is, interviewing experts, editors and co-authors. In general, there has been a tendency towards an increase in the number of coauthors over the past 30 years [4]. For domestic scientists, this leads to known problems in reporting to foundations and ministries.

This paper discusses the options for using the data that are available in the arsenal of modern information technologies of semantic libraries for indexing publications, authors and their contribution to collective work.

2 On the means of identification of authors

2.1 The data to author’s identify

The structure of a scientific publication is a feature of scientific articles that is quite well-established for many domestic and international journals. The strictness that authors are encouraged to adhere to in accordance with the instructions from the publishers is dictated to some extent by the process of digitizing publications for their subsequent indexing in bibliographic databases. In the 70s of the last century, a family of standards for machine-readable cataloging (MARC) [5] appeared with the further development of the ISO 2709 standard (GOST 7.14-84 (ST SEV 4269-83) SIBID and GOST 7.14-98 SIBID). These standards were originally proposed by the US Library of Congress as formats for interlibrary bibliographic data exchange, and were later adapted for national libraries, and began to be used in one form or another in all English-language library systems. Naturally, standard bibliographic record fields for machine-readable cataloging have become components and fixed positions in the structure of scientific articles.

Thus, a list of required fields of secondary information about the document "scientific article" was formed: author, affiliation of authors, title, keywords, classifiers (MSC, UDC and / or specialized), output data (publisher, pages, year). In the future, an abstract, a list of cited literature and identifiers such as ORCID, etc. were added. All these fields are used for indexing publications and can be used as search fields when forming a request and identifying authors.

The difficulty arises if this information is not enough, or it is not in full in the database, or if the user has not it. Refinement is carried out through expert knowledge or through semantic links, which can be implemented in the form of hints from the database.

The body of the publication, as a rule, is not searchable, even if the publication is in the public domain, but is available to publishers for preliminary lexical, syntagmatic, paradigmatic, and semantic processing when placed in bibliographic databases.

2.2 Dataset for thesaurus of addressee

The concept of “addressee in the information environment”, formulated for the convenience of identifying users and authors from databases, implies a person - a participant in the information process, search and exchange of information. The term “addressee (individual) thesaurus” (TA) was introduced into computer science by Yu.A. Schrader [6] to represent the author's subject domain (SD) based on the author's conceptual background. The term is also associated with the representation of "knowledge" in the information system as "structured information" [7]. For a more detailed acquaintance with the use of thesauri in search processes and knowledge extraction, you can refer to [8]. In the future, the importance of this representation, as a basis for describing the ontology of the addressee (OA) in modern databases, was manifested [9].

The composition of the data (information) of the addressee's thesaurus depends on the conceptual reserve of the individual. For a semantic library, you can focus on the following data set: the frequency vocabulary of the individual; variants of combinations of terms; contexts of frequency terms; special designations and formulas; lists of cited literature; lists of citing authors; list of publications with cross-references. If the information system contains enough data and publications on a certain subject area, then on the basis of the set of data about the addressee's thesaurus and metric analysis, it is possible to build a dictionary-thesaurus of the author's subject domain. Further, comparing subject thesauri, it is possible to more accurately identify their authors, as well as establish the belonging of the text to a certain author and his contribution to research.

2.3 Text comparison tools to author's identify

Methods of text comparison for attribution are considered, such as frequency algorithms [10], contextual comparison [11], thematic clustering and deep text analysis algorithms associated with machine learning methods [12], [13].

Using this set of methods, it is possible to form an information processing technology for newly received data in an information bibliographic system.

The first stage of preprocessing publications for each author includes:

- frequency processing of texts to obtain a list of terms with their weight (frequency of use);
- compiling a list of co-authors;
- forming the set of contexts for terms.

As a result, the following data (parameters) of the author are accumulated: list (dictionary) of terms, rank (weight) of terms, word forms of terms, relative frequency of terms (in relation to other terms), absolute frequency of terms, concordance dictionary (dictionary with contexts), Fig. 1. At this stage, it is also possible to select a list of unique terms, designations, formulas and other features of the text, typical for some authors and subject areas.

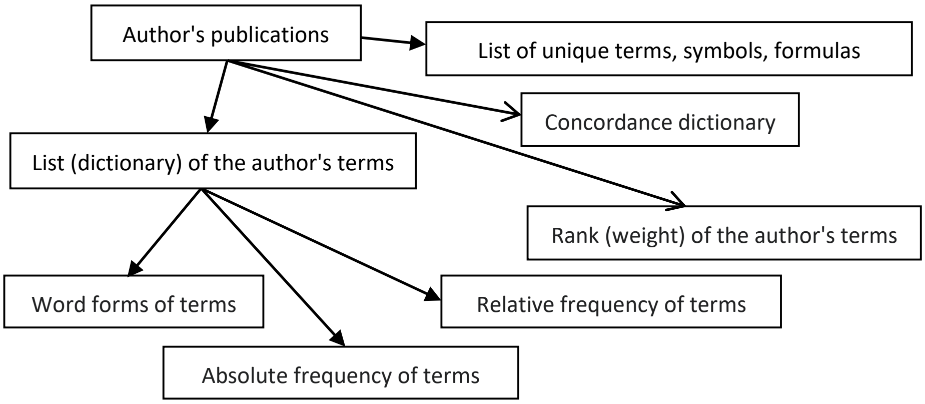
Author's publications
List of unique terms, symbols, formulas
Concordance dictionary
List (dictionary) of the author's terms
Rank (weight) of the author's terms
Word forms of terms
Relative frequency of terms
Absolute frequency of terms

Fig. 1. Author's publications preprocessing scheme.

The second stage consists in the procedure for comparing authors according to the available (accumulated) parameters. The intersection of the sets of terms, contexts and unique terms, designations, etc.

After comparing and identifying a set of publications belonging to a specific author, an author's index and an index of cited publications are compiled. At the same time, it is possible to vary the severity of the belonging of "controversial" publications to one or another author, taking into account the degree of coincidence of the revealed parameters (in%, for example).

At this point, the preprocessing of the newly received data about the author can be completed. The whole set of related information obtained can be attributed to the addressee's thesaurus.

Note 1. If a series of publications of one group of authors is supposed to be loaded into the system, then at the preliminary stage of processing it is possible to compose a *thesaurus of co-authors*.

Note 2. If a single work has been received, then preprocessing (according to the scheme in Fig. 1) is used to be included in the existing author index, or in the absence of matches and questionable properties of the publication (variants of surnames and other secondary documents), it is stored in the *confirmation status*, but participates in further subject semantic processing. Confirmation can be done automatically if the system accumulates additional information about the author or upon request to the author.

For further semantic processing of publications, it is necessary to use dictionaries (thesauri) of professional terms from SD (for example, mathematical SD). Publications must be indexed in accordance with the subject and thematic focus, determining the belonging of the terms of publications to dictionaries (thesauri) of subject areas. Thus, to fix the *links* of the thesaurus of the addressee (author) with SD. These links are further additional features for the *subject identification of the author*. As a result, publications that are semantically linked in ontologies, after preprocessing, will have a number of author identification features.

3 Examples on LibMeta datasets

Using the example of a number of works in the fields of higher mathematics, we can consider options for identifying authors of publications with similar sets of secondary documents.

For text processing, a free library for high-performance full-text search Apache Lucene, implemented in the Java language, is used.

3.1 Establishment of authorship

To highlight the meaningful expressions of the document, the calculation of the *tf-idf* measure was used for the terms of the document extracted from the index, taking into account the morphology [13]. At the first stage, only nouns and terms were considered that were identified as proper nouns.

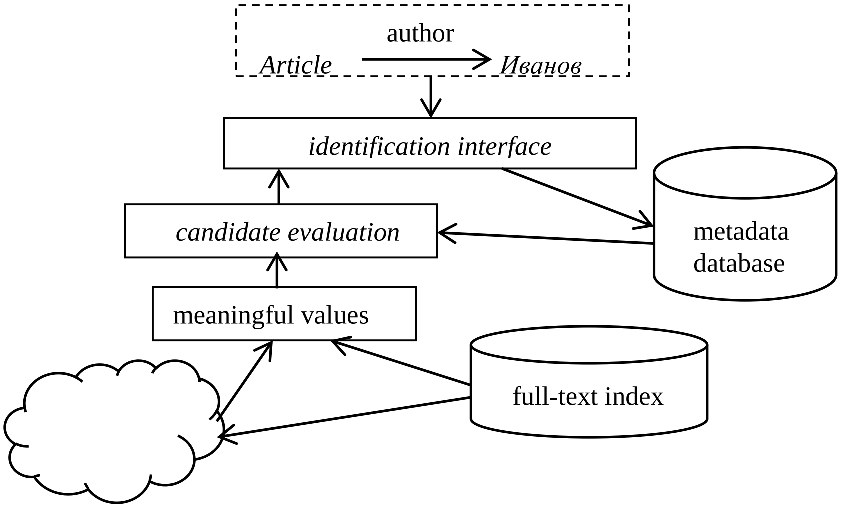

Fig. 2. General scheme for working with terms and authors.

Further, terms for which the *tf-idf* measure was less than the threshold were excluded. Composing *combinations of two and three words* was performed based on the use of the context of the selected words, and rules that take into account morphology. The context is understood as *N* words in the text before the word for which the vector is constructed, and *N* words after this word. To highlight the context, a shallow neural network model word2vec [14–16] is used, in the "skip-grams" mode. In Fig. 2 shows the general scheme of work.

As an example, further in Fig. 3 shows the stage of formation of the thesauri of subject areas of individual authors (Russian), on the basis of which one can reason about their (authors') identity.

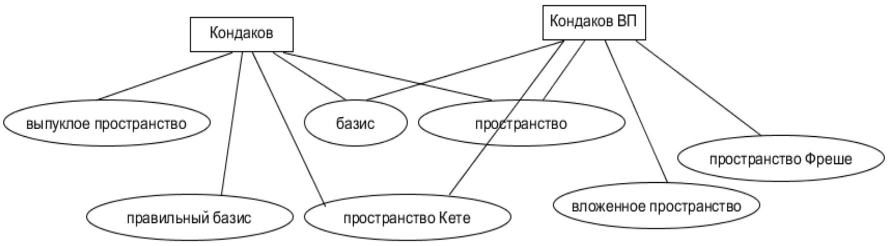

Fig. 3. Fragment of the authors comparison scheme.

The example (Fig. 3) shows that the works of the authors were obtained with an incomplete set of secondary information. Application of the described algorithm makes it possible to identify terms, relationships and intersections of subsets of terms, taking into account their contexts.

Further, we use additionally the connections of terms from the encyclopedia, classifiers UDC, MSC and other works from the field of analytical spaces, such as shown in Fig. 4.

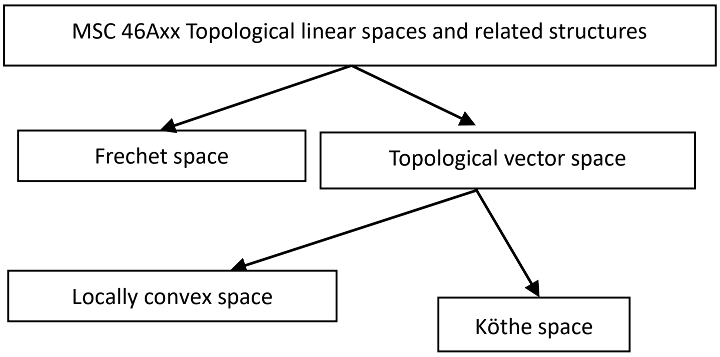

Fig.4. Links of the identified terms of the authors.

Further, we use additionally the connections of terms from the encyclopedia, classifiers UDC, MSC and other works from the field of analytical spaces, such as shown in Fig. 4.

About 5000 authors of publications were processed. Separately, work is being done to process formulas and include them in the author's thesaurus. A formula comparison algorithm based on a vector model is used. The algorithm is conventionally divided into two parts: the initial selection of candidate formulas and their subsequent ordering by similarity. A description of this algorithm is beyond the scope of this article.

3.2 On author's contribution

To take into account the author's contribution to the publication, it is required to investigate the history of the author's work and his affiliation to scientific schools, as well as the author's research in subject areas. This is of particular importance as co-authorship has become commercial in nature and paid publications, “senior authorship” and citation have become possible [17].

The set of "historical" data about the author and publication is formed on the basis of the addressee's thesaurus as follows. The history of publications is collected and stored: co-authors, cross-references, keywords, internal system indicators of publications belonging to subject areas (LibMeta).

Almost all modern bibliographic collections collect and display the listed data. The development of information technology makes it possible to use various methods of analysis to establish the authorship of publications. It should be noted that for works of fiction, such as "pocket" detectives, such an examination has been carried out for a long time, since this process was originally built on a commercial basis and it is necessary to take into account the contribution of each participant. The scientific community needs to avoid this approach, as it inevitably leads to downplaying the value of research work.

The criteria by which the publications are distributed are selected, these are the *history of the issue, novelty*, the number of publications on related topics, many co-authors, expert opinion expressed in the process of discussions and peer review.

Publication history. The structure of a scientific article assumes the presence of an introduction, which *lists previous research*. Analyzing this text, it is possible to compile lists of researchers and corresponding bibliographic references on the selected topic, example Fig. 5.

Next, select the intersections within these sets and identify the "main" authors and their co-authors. For co-authors to *identify frequency characteristics* and belonging to the SD. Thus, to obtain a "map" of publications on the topic, where there will be areas of intersection of authors' collectives, where {k1, k2,…, kN} of *authors intersect* (k1> k2>…> kN). Authors from citation lists are also included in these areas. Individual authors (A, B, C, ...) can belong to many “invitees” to participate in publications, and then their role is evaluated by experts from the scientific community. These can be authors of publications without co-authors working in a given subject area, and then, naturally, their contribution to the work is not disputed.

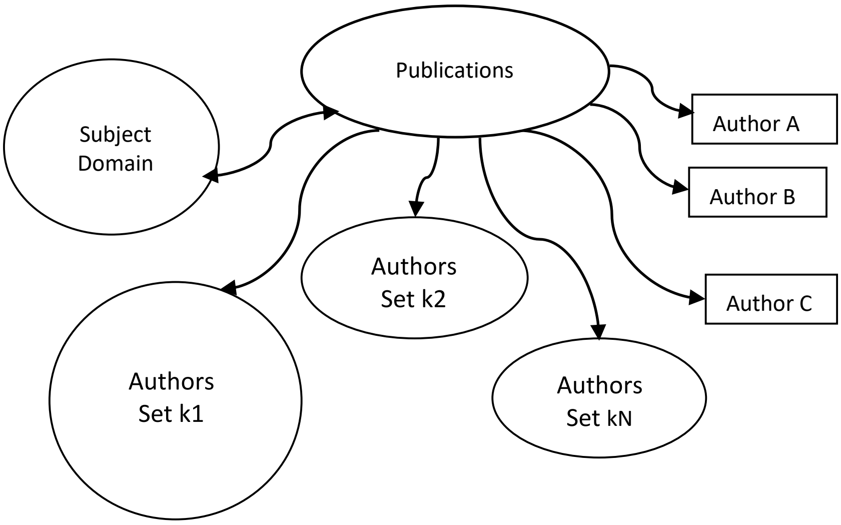

Fig. 5. A diagram of the links between sets of authors through sets of publications.

The set of authors {k1}, which is larger than others, can apply for many leading scientists, leaders of scientific schools and research projects (grants, etc.).

Keyword scoring. The intersection of keywords in the authors' thesauri indicates the closeness of the studies.

Keyword novelty score. The analysis of the collectives that make up the sets {k1, k2, ..., kN} allows us to identify “new” members of the group of authors, for a certain period of time, “new” keywords for the same period of time. Since, thanks to the TA, it is possible to find out to which author the “new” keywords belong, it can be concluded that thanks to whom the “new” contribution to publications and to the SD.

Metric for evaluating author's participation. Based on the data of the TA, the LibMeta system has introduced a metric for assessing the author's participation in publications in mathematical SD.

The following sets of metrics are calculated for the ODE domain thesaurus:

- core of key concepts of the SD (*Concept Kernel*) – $\{CK=K_1 \cap K_2 \cap K_3\}$, where K_1 set from ODE domain thesaurus, K_2 set from special function dictionary and K_3 set from mathematical encyclopedia;

$$|KK|=|K_1|+|K_2|+|K_3|,\ |K_1|=184,\ |K_2|=151,\ |K_3|=6263,\ |KK|=6\,598;$$

- core of information object keywords for different types of resources of the domain resources (*Keyword Kernel*) – {KK}, |KK|=6810;

- core of authors' collectives by years (*Kernel of Copyright Teams*) – {KCT}.

Consider, for example, 2015, for publications dealing with *Bernoulli* ODE[1]

We get:

$|KK_{2015}|=754$, KCT_{2015}={*'Лазарев', 'Неустроева', 'Шишкина', 'Бочкарев', 'Лекомцев', 'Сенин', 'Янковский', ‘Кольцун'*};

- core bibliographic references (*Bibliographic Reference Kernel*) – {BRK} for these authors are represented by 34 references, |BRK|=34.

Next, the intersection of data from the author's TA is estimated:

- keywords {KWA}, $|KWA_{Лазарев}|=14$, $|KWA_{Янковский}|=79$;

- co-authors {CA} $|CA_{2015}|=163$;

- bibliographic lists {RL} $|RL_{Лазарев}|=3$, $|RL_{Янковский}|=16$, with the sets $\{KK_{2015}\}$, $\{KCT_{2015}\}$, $\{BRK_{2015}\}$ to the general characteristics of the SD:

$$\{KWA\}\cap\{KK\},\ \{CA\}\cap KCT\},\ \{RL\}\cap\{BRK\}.$$

Based on these sets, estimates of the author's contribution to the SD are introduced $KWA_{Лазарев}/KK_{2015}=14/754$, “average” author's contribution to the SD this year $CA_{2015}/KCT_{2015}=163/8$, “average” author's contribution $|RL_{Лазарев}|/|BRK|=3/34$, $|RL_{Лазарев}|/|BRK|=16/34$.

These estimates show the author's contribution to the SD and to specific research (publications) "over time". We emphasize that these estimates do not reflect the picture of the real world, but they are valid for characterizing the set of objects that are loaded into the system.

In reality, it can be difficult to draw a line between authorship claims, and sometimes it is a matter of controversy among academic schools. There are cases when an idea and its implementation in research belongs to different people who may or may not know about each other's works. This raises issues of plagiarism and priorities in science. An example of this is the history of disagreements between Newton and Lebniz on the contribution of each to the development of mathematical analysis [18].

[1] http://libmeta.ru/concept/showRelatedValues/404?attribute=119

Authors who have the highest percentage of "overlaps" with the SD of ontology can be considered “key” researchers in the SD.

Note 4: Our study does not provide any assessment of the rationale for the authors' studies and the quality of scientific papers.

Note 5: All assessments are made only on the basis of publications, secondary information or full texts (if available) and author's methods of tracking links in the semantic library.

Note 6: The real contribution of the author to publication and research can only be assessed by the scientific community. In a digital library, you can only set the number of links according to the selected characteristics and on the basis of the data array that is already in the library. This gives a picture of the contribution of the publication and the rating of the author in the scale of the available data, rather than the quality of the publication and the knowledge of the author in general.

Note 7: The LibMeta library has a technology for creating a subject author's thesaurus and on its basis you can get an idea of the addressee's thesaurus as a participant in the exchange of information in the information environment. This technology allows us to consider the meaning and contribution of the author's publications in relation to various subject areas that make up the intersection of sets within the author's subject thesaurus.

4 Conclusions and outlook

The technology of preliminary processing of publications for further placement in the digital library is proposed. Using the data of the addressee's thesaurus allows accumulating structured information about authors and publications, which helps to identify authors at the preliminary stage and evaluate their contribution to research.

An ideal scheme for assessing the role of the author and attribution is presented, and of course, there are controversial factors in it, but it can be used as a first approximation if the authorship of the article is in doubt due to the inaccuracy of secondary data in the digital library. However, it is in digital libraries that you can take into account, if not all, but many of the attributes of authorship, which is shown in the examples of mathematical articles in LibMeta.

The use of a personal environment for scientific research on the basis of individual bibliographic collections and the results collected by the author in the process of research allows us to consider the problems of identification and determination of the author's contribution as part of the functioning of the semantic library.

This work was carried out with partial support from the Russian Foundation for Basic Research, project #20-07-00324 and within the framework of the topic ‘Mathematical methods for data analysis and forecasting’.

References

1. Krämer, T., Momeni, F., Mayr, P.: Coverage of Author Identifiers in Web of Science and Scopus. arXiv preprint arXiv:1703.01319 (2017).
2. Clement, T.P.: Authorship Matrix: A Rational Approach to Quantify Individual Contributions and Responsibilities in Multi-Author Scientific Articles. Sci. Eng. Ethics, 20, 345–361 (2014) https://doi.org/10.1007/s11948-013-9454-3.
3. Frische, S.: It is time for full disclosure of author contributions. Nature, 489 (2012). http://www.nature.com/news/it-is-time-for-full-disclosure-of-author-contributions-1.11475.3.
4. Cozzarelli, N.R.: Responsible authorship of papers in PNAS. Proceedings of the National Academy of Sciences of the United States of America, 101, 10495 (2004).
5. http://www.loc.gov/marc/marcdocz.html, last accessed 2020/11/25.
6. Shrejder, Yu.A.: Tezaurusy v informatike i teoreticheskoj semantike. Nauchno-tekhnicheskaya informaciya. Ser. 2 (3), 21–24 (1971).
7. Gavrilova, T.A., Horoshevskij, V.F.: Bazy znanij intellektual'nyh sistem. Piter, Saint-Petersburg (2000).
8. Lukashevich, N.V.: Tezaurusy v zadachah informacionnogo poiska. Izd-vo MGU, Moscow (2011).
9. Muromskij, A.A., Tuchkova, N.P.: Ob ontologii adresata v matematicheskoj predmet-noj oblasti. Elektronnye biblioteki, 21 (6), 506-533 (2018).
10. Borisov, LA, Orlov, Yu.N., Osminin, K.P.: Identification of the author of the text by the distribution of the frequencies of letter combinations. Keldysh Institute preprints M.V. Keldysh, 27 (2013). URL: http://library.keldysh.ru/preprint.asp?id=2013-27.
11. http://neon.niederlandistik.fu-berlin.de/textstat/.
12. Mohsen, A.M., El-Makky, N.M., Ghanem, N.: Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 898–903, https://doi.org/10.1109/ICMLA.2016.0161.
13. Manning, K.D., Raghavan, P., Shyutce, H.: Vvedenie v informacionnyj poisk. Dialektika, Moscow (2011).
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR (2013).
15. Mikolov, T., Yih, W.T., Zweig, C.: Linguistic Regularities in Continuous Space Word Representations. Proceedings of NAACL HLT (2013).
16. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. International Conference on Machine Learning, pp. 1188–1196 (2014).
17. Strange K. Authorship: Why not just toss a coin? American Journal of Physiology-Cell Physiology, 295 (3), 567–575 (2008). https://doi.org/10.1152/ajpcell.00208.2008.
18. Meli, D.B.: Equivalence and Priority: Newton versus Leibniz: Including Leibniz's Unpublished Manuscripts on the Principia. Clarendon Press (1993).