ISSN 0147-6882, Scientific and Technical Information Processing, 2016, Vol. 43, No. 4, pp. 236–240. © Allerton Press, Inc., 2016.
Original Russian Text © N.A. Mazov, V.N. Gureev, D.V. Kosyakov, 2016, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 1: Organizatsiya i Metodika Informatsionnoi Raboty, 2016, No. 11, pp. 9–14.

On the Development of a Plagiarism Detection Model Based on Citation Analysis Using a Bibliographic Database

N. A. Mazov*, V. N. Gureev, and D. V. Kosyakov*****

Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia

**e-mail: MazovNA@ipgg.sbras.ru*

***e-mail: GureyevVN@ipgg.sbras.ru*

****e-mail: KosyakovDV@ipgg.sbras.ru*

Received July 15, 2016

Abstract—A relatively new method for the detection of text plagiarism is proposed based on a search for original sources with an identical or similar list of references. First of all, this is applied to the most difficult to detect forms of translated plagiarism and the plagiarism of ideas. This method successfully proved itself in test studies of groups of foreign authors and is a continuation of our studies. The peculiarity of the approach that is proposed by the authors is the use of multidisciplinary bibliographic databases rather than full-text databases, which are often unavailable due to the high price of subscriptions under Russian conditions. The advantage of their use is the access to the maximum possible number of article references, which significantly extends the base for the search for originals while analyzing suspicious scientific publications. The step-by-step algorithm for addressing a query to the Web of Science and Scopus databases and analysis of the obtained data can be recommended for implementation into systems of plagiarism detection as an additional component.

Keywords: plagiarism, plagiarism detection, citation analysis, bibliometric analysis, Scopus, Web of Science, scientific translation

DOI: 10.3103/S0147688216040092

INTRODUCTION

Detection of plagiarism in scientific texts, which is acquiring increasingly subtle forms in response to the modern technologies for searching for it, is becoming an international problem. As the experience with the detection of plagiarism by copying and pasting shows, the automated computer analysis of texts provides efficient results. In the recent studies with the application of the linguistic processing of full texts including morphological, syntactic, and semantic analysis, satisfactory results were obtained even for plagiarism with significant rephrasing [1, 2] whose detection was quite recently considered to be a difficult problem [3]. At the same time there are still forms of plagiarism that can hardly be automatically processed. First of all, this is plagiarism connected with translation of a scientific publication into another language, as well as plagiarism of ideas [3].

The detection of such secondary texts is now possible only by attracting the experts in some area of knowledge [4]. Such an approach has some significant disadvantages; the main one are the impossibility of its mass use, as well as the high price and time expenditures. Thus, analysis with respect to translated plagiarism or plagiarism of ideas is not used to check most scientific papers, is not performed when checking reports written in the frameworks of grants and governmental programs, and is not involved while checking Ph.D theses for originality. It should be noted that according to the notes of specialists, Russian editors, as a rule, do not use any system, including the publicly available [4] ones, to check scientific papers, while a relatively regular check is made only for thesis works [5]. Under such conditions dishonest scientists feel safe and often refuse independent and state-financed work in favor of less labor-intensive translation of another person's publications.

The difficulties of the detection of such a type of plagiarism also lead to aggravation of the problem of self-plagiarism, where authors often translate their own previously published Russian-language works into foreign languages. This problem is less discussed [6], although it contradicts publication-ethics standards no less [7]. As well, in the case of self-plagiarism judicial norms are violated related with copyrights, as authors do not give a reference to the journal that previously published the scientific material, which in most cases contradicts the author's contract provisions [6]. It should be mentioned that self-plagiarism

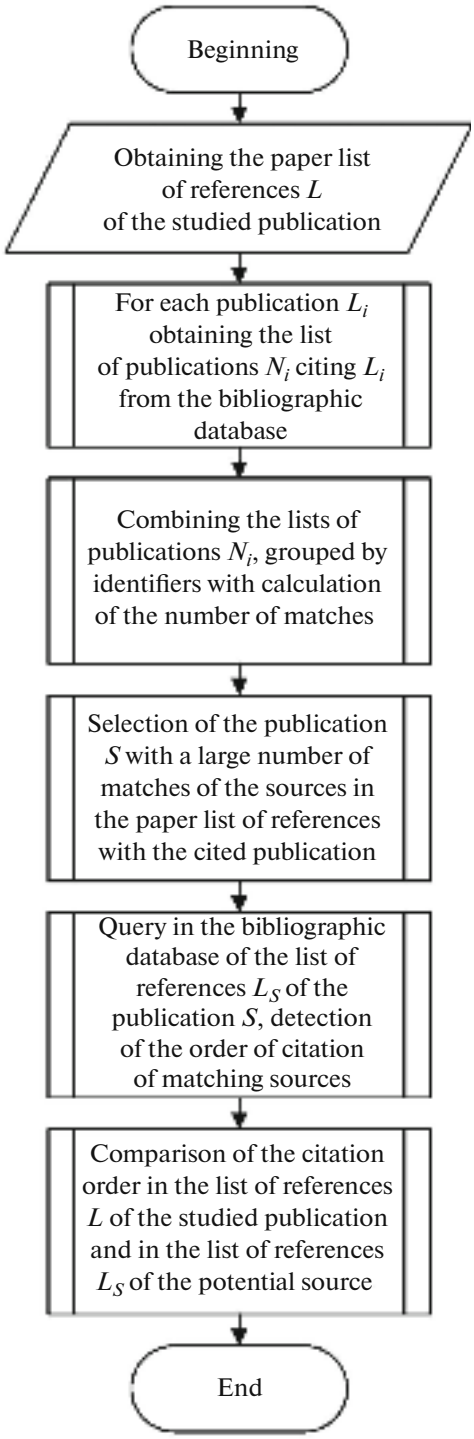

Fig. 1. A schematic block diagram of the algorithm for queries to the bibliographic database.

is widespread not only in Russia; the problem is of an international character [8].

It is notable that plagiarism and self-plagiarism are most widespread in countries with an insufficiently balanced system for the evaluation of the results of scientific activities. In particular, significant volumes of plagiarism and an increase of the translation services of previously published works occur in China [8], where the career growth of scientists and financing of their scientific projects mainly depend on their number of scientific publications. Recently, a similar situation has also been observed in both Russia and in the countries of the Commonwealth of Independent States. Specialists have already noted this fact [6, 9–11]. A separate problem is plagiarism in the works of students and postgraduates, which is often connected with insufficient information literacy, particularly with a lack of knowledge of the citation standards [5, 12].

PLAGIARISM-DETECTION METHODS

In this paper, which continues the research of the authors on this topic [13], proven technology for the analysis of suspicious publications is proposed based on a search for possible original sources in multitopic bibliographic databases (DBs) using the citation-analysis method. The basis for proving a high degree of similarity of publication contents is the identical or very similar lists of references of compared papers, especially in the cases where in the later text the sequence of the original source references is kept. Thus, only the lists of the cited literature and their sequence are analyzed, which allows one to avoid using the publication texts themselves and as a result to avoid the as-yet unsolved problem of the collation of texts in different languages (the case of translated plagiarism) or completely rewritten texts in one language (the plagiarism of ideas).

When collating the text of a suspicious publication with an original source four variants depending on the citation style are possible:

(1) The bibliographic list in both the original and secondary texts is arranged in the citation order;

(2) The bibliographic list both in the original and secondary texts is arranged in alphabetical order;

(3) The bibliographic list in the original text is arranged in the citation order, while in the secondary text it is in alphabetical order;

(4) The bibliographic list in the original text is arranged in the alphabetical order, while in the secondary text it is in the citation order.

The access to a significant number of full texts, that is, the original sources on whose basis a suspicious publication could be created, certainly permits one to perform the most accurate comparison of the list of references of a suspicious publication with the lists of references of all of the other publications in the database to which access is available. A group of foreign researchers under the supervision of B. Gipp first proposed this method and evaluated it based on some examples [14–17]. This group of researchers created a prototype of a program for the detection of translated

plagiarism (http://www.citeplag.org), which provides the ability to perform automatic comparison of two texts without using a full-text database; thus, it permits one to confirm or dispute that plagiarism occurred, but does not permit one to carry out a search of the original itself.

The authors have already noted [13] that the method proposed by their foreign colleagues is not entirely applicable to Russian realities, due to several factors, among which the main ones are as follows:

• in Russian organizations there is often no access to full-text databases, although it should be mentioned that recently the situation has changed, which is due to the increasing popularity of the open-access model. As well, recently in Russia the requirement that defenders of a thesis make the thesis text available to the public has been established by law;

• the absence of a common global full-text database and division of such databases by editors or topics;

• the impossibility of performing automatic simultaneous searches for several databases (simultaneous searching is possible only for certain databases that are available in an organization).

As a consequence, addressing the multidisciplinary databases is more logical, such as Web of Science, Scopus, or the Russian Science Citation Index (RSCI) with the ability to carry out searches on tens of millions of paper references (in the case of the Web of Science their number has already exceeded one billion). In the framework of the programs of the Russian Foundation for Basic Research and Ministry of Education and Science of the Russian Federation, access to these bibliographic databases is available for most national universities and scientific organizations. Their use, unlike that for full-text databases, significantly extends the fact base for analysis and provides automatic generation of a query to a database according to the sources cited in a suspicious text. The disadvantages of the use of the bibliographic databases include the impossibility of using the sequence of references in the original if they are given in alphabetical order; in this case access to the full text is required.

Since the time when the authors' first research demonstrated the possibility in principle to use this method, they have developed an algorithm for queries to bibliographic databases and an algorithm for further analysis of the results.

The proposed algorithm basically consists of the following steps (Fig. 1):

(1) for each cited source of the list of references of the suspicious publication a query is formed to the bibliographic database for the purpose of extracting the list of publications that also cited this source;

(2) the lists obtained for each cited source are combined and the number of matches is calculated. Thus, a list of publications is obtained that cite the same sources as the suspicious paper, specifying the number of matching sources. The list of publications ranked by the decrease of the number of matching sources is the subject of further analysis for plagiarism. Depending on the circumstances, this list can be reduced by separation using an absolute (for example, the number of matching sources is more than four) or relative (for example, the number of matching sources is more than 50% of the total list of references of the studied paper) boundary;

(3) if the list of references in the studied work is arranged in the citation order, an additional query to the bibliographic database is formed to obtain the list of references of the potential source of plagiarism. If this list is also arranged in the citation order, an analysis by matching the citation order is performed.

Any database that supports the ability to review the list of works cited the publication can be used as a bibliographic database. Two approaches are possible for extraction of the list of publications that cite the same source as the studied work. The first approach uses the ability to form a query by the lists of papers (Cited Reference Search) that are available in Scopus and Web of Science databases. The second approach assumes the search for the source in the database with further extraction of the list of publications cited in this source. This mode can also be used while working with the RSCI DB. The advantage of the first approach is the ability to find the lists of citing publications for the sources not indexed in the database; the second approach permits one to reduce errors while finding the source in database.

Scopus presents a program interface (API) that is sufficient for complete automation of the performance of the algorithm proposed by the authors. The query for the list of publications citing the specified source is performed by means of Scopus Search API. The query can be made using such metadata of the source as the authors, title, year of publication, name of the journal or collection, or number of the first page. Depending on the type of source and the availability of the metadata the authors use the first author, title, year of publication and number of the first page, which permits one to almost exclude false responses. In this case the query is as follows:

REF(REFAUTH(*ln*) AND REFTITLE(*t*) AND REFPAGEFIRST(*fp*) AND REFPUBYEAR IS *y*),

where *ln* is the surname of the first authors, *t* is the title, *fp* is the number of the first page, and *y* is the year of publication.

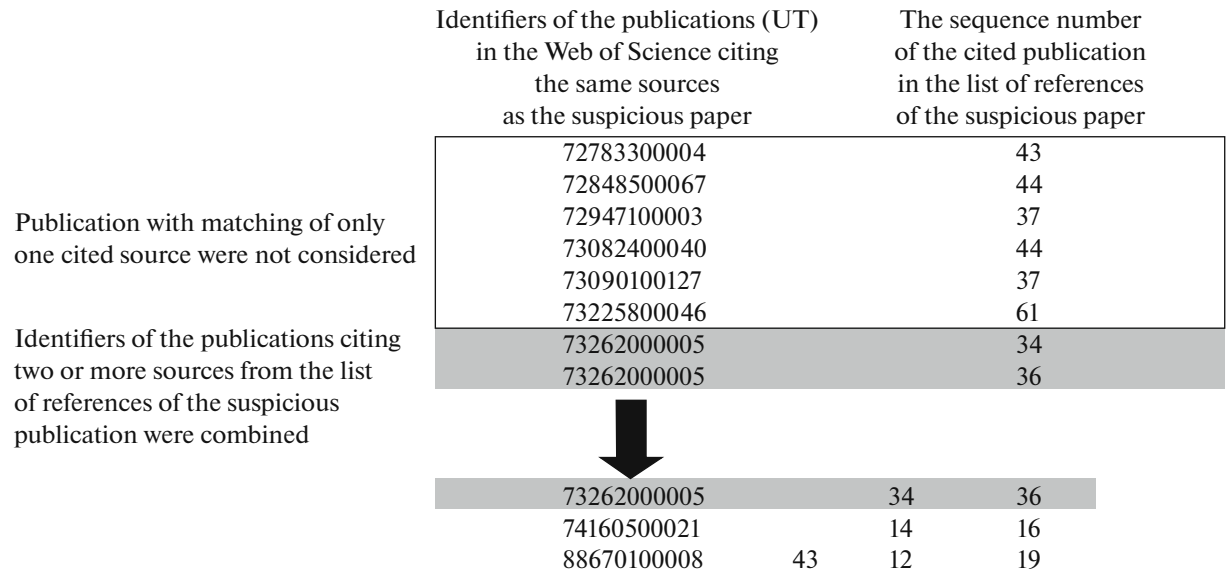

Fig. 2. The stages of processing of the paper lists of references where the same sources are cited as in a suspicious publication.

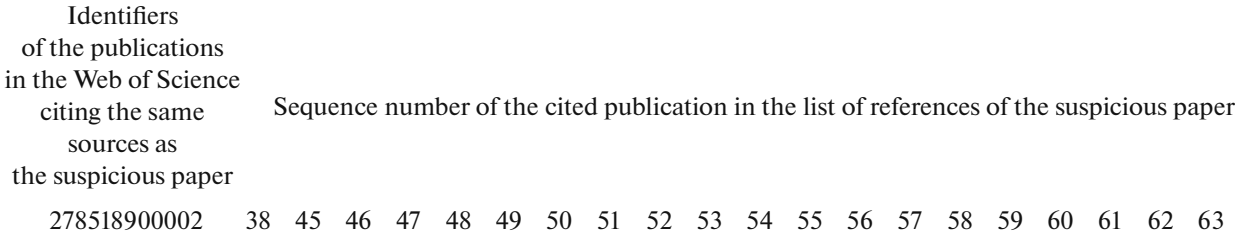

Fig. 3. The identifier of the publication in Web of Science with the largest number of sources (20) also present in the list of references of the analyzed suspicious publication.

The query of the list of references at the third step of the algorithm is made using Scopus Abstract Retrieval API via one of the identifiers of the potential source of plagiarism: Scopus ID, EID (Electronic Identifier), or DOI.

It should be noted that public access to the service of searching for potential sources of plagiarism by this algorithm using Scopus API will violate the policy of use of the program interfaces of the copyright holder and requires a special agreement.

Web of Science API does not provide possibility to solve this problem. In RSCI the API is not currently available. In this respect automation of the algorithm proposed by the authors while using these databases is limited in general to the generation of query lines, and in the Web of Science, to processing according to lists uploaded to the text format. A high degree of automation can be achieved by interaction with the web interfaces of these databases with simulation of the user's activity, but without a special agreement with the copyright holders this will violate the policy of use.

The results of searching for a possible original source for a suspicious publication using the Web of Science database are schematically demonstrated in Figs. 2 and 3.

A positive result of the search performed in Web of Science permitted the detection of a publication with 20 cited sources in the bibliographic database cited actually in the same sequence by the author of the suspicious publication is shown in Fig. 3.

A back query using the identifier in the Web of Science easily permits detection of the source that is necessary for further expert analysis and, if access to the full text is available, allows one to make a linguistic comparison of the two publications.

As is clear from Fig. 3, the analyzed suspicious article includes only one third of the references in common with the other source, while two thirds of it was apparently written by the authors. At the same time, the total sequence of the references in the suspicious part of the paper indicates that text fragments of the two publications with common references will be also similar with a high degree of probability.

CONCLUSIONS

The plagiarism-detection model based on collation of lists of references in the analyzed publications and their sequences was shown to be efficient, while even its partial automation can be efficiently applied to detect cases of text plagiarism. The algorithms included in the model can be directly applied in the computer programs for searching original texts and visualization of the results, which is performed by our group. The development and industrial use of such a system would permit, in the opinion of the authors, a significant reduction in the volumes of translated plagiarism and plagiarism of ideas, and, as a consequence, contribute to the growth of original national scientific results. While using the ability to create search queries based on the lists of references in the Russian Science Citation Index the fact base for research can be considerably extended to search for plagiarism in the national scientific literature.

The results of the research were discussed at the Detection of text plagiarism as a method of development of scientific activities section in the framework of the 23rd International conference Libraries and information resources in the modern world of science, culture, education, and business (June 4–12, 2016, Sudak).

ACKNOWLEDGMENTS

This research was financially supported by the Russian Foundation for Basic Research in the framework of scientific project no.16-07-00652\16.

REFERENCES

1. Osipov, G.S., Smirnov, I.V., Tikhomirov, I.A., Sochenkov, I.V., Zubarev, D.V., and Isakov, V.A., Technologies for semantic plagiarism detection in scientific texts, *Trudy 23-i Mezhdunarodnoi konferentsii "Biblioteki i informatsionnye resursy v sovremennom mire nauki, kul'tury, obrazovaniya i biznesa" (4–12 iyunya 2016 g., g. Sudak)* (Proc. 23rd Int. Conf. Libraries and Information Resources in the Modern World of Science, Culture, Education, and Business (June 4–12, 2016, Sudak)), Moscow, 2016.

2. Sochenkov, I., Zubarev, D., Tikhomirov, I., Smirnov, I., Shelmanov, A., Suvorov, R., and Osipov, G., Exactus Like: Plagiarism detection in scientific texts, *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016,* Padua, 2016, pp. 837–840.

3. Gipp, B., Meuschke, N., and Beel, J., Comparative evaluation of text- and citation-based plagiarism detection approaches using GuttenPlag, *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (June 13–17, 2011, Ottawa, Canada),* New York: ACM, 2011, pp. 255–258.

4. Chekhovich, Yu.V., About wrongful appropriations detection when conducting expert reviewing scientific articles, *Nauchn. Period., Probl. Resheniya,* 2013, vol. 4, no. 16, pp. 22–25.

5. Avdeeva, N.V., Nikulina, O.V., and Sologubov, A.M., Unscrupulous authors of dissertations vs "Antiplagiat.RSL"—who's the cutest?, *Nauchn. Period., Probl. Resheniya,* 2012, vol. 5, no. 11, pp. 11–16.

6. Kotlyarov, I.D., Plagiarism in scientific publications, *Nauchn. Period., Probl. Resheniya,* 2011, vol. 4, no. 4, pp. 6–12.

7. Scott-Lichter, D., *CSE's white paper on promoting integrity in scientific journal publications, in 2012 Update,* Wheat Ridge, CO: Council of Science Editors, 2012, 3rd ed.

8. Hvistendahl, M., China's publication bazaar, *Science,* 2013, vol. 342, no. 6162, pp. 1035–1039.

9. Kholodov, A.S., Citation indexes of scientific works, *Herald Russ. Acad. Sci.*, Vol. 85, No. 2, pp. 122–131.

10. Mazov, N.A. and Gureev, V.N., Publications at any cost?, *Vestn. Ross. Akad. Nauk,* 2015, vol. 85, no. 7, pp. 627–631.

11. Novikov, D.A., Compete by "Hirsch's"?, *Vyssh. Obraz. Ross.,* 2015, no. 2, pp. 5–13.

12. Abramova, N.Yu., The issue of plagiarism in research papers, *Nauchn. Period., Probl. Resheniya,* 2011, vol. 2, no. 2, pp. 25–28.

13. Gureev, V.N. and Mazov, N.A., Citation analysis as a basis for the development of an additional module in antiplagiarism systems, *Sci. Tech. Inf. Process.,* 2013, vol. 40, no. 4, pp. 264–267.

14. Gipp, B., Meuschke, N., and Breitinger, C., Citation-based plagiarism detection: Practicability on a large-scale scientific corpus, *J. Assoc. Inf. Sci. Technol.,* 2014, vol. 65, no. 8, pp. 1527–1540.

15. Gipp, B., Meuschke, N., Breitinger, C., Lipinski, M., and Nurnberger, A., Demonstration of citation pattern analysis for plagiarism detection, *36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2013* (July 28 – August 1, 2013, Dublin, Ireland), New York: ACM, 2013, pp. 1119–1120.

16. Meuschke, N., Gipp, B., and Breitinger, C., Citeplag: A citation-based plagiarism detection system prototype, *Proceedings of the 5th International Plagiarism Conference (July 17–18, 2012, Newcastle upon Tyne, United Kingdom),* Edinburgh: iParadigms Europe Ltd, 2012, pp. 1–10.

17. Gipp, B. and Meuschke, N., Citation pattern matching algorithms for citation-based plagiarism detection: Greedy citation tiling, citation chunking and longest common citation sequence, *Proceedings of the 11th ACM Symposium on Document Engineering (DocEng '11)* (September 19–22, 2011, Mountain View, USA), New York: ACM, 2011, pp. 1–10.

Translated by Yu. Bezlepkina