The Reference Analysis as a Quality Characteristic of a Scientific Article

Yulia Hlavcheva[1], and Olga Kanishcheva[1]

[1] National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
(glavjul, kanichshevaolga)@gmail.com

Abstract. Nowadays the qualitative characteristics of a scientific document are becoming more and more relevant because a large number of conferences, seminars, and journals generate a huge amount of scientific papers. A scientific paper consists of such elements as title, information about the authors, abstract, keywords, body and a list of references. References are one of the important factors that affect paper quality. In this paper, the authors analyze the qualitative characteristics of the cited sources and highlight the formal features that characterize the quality of the references. Ukrainian scientific papers were used as data for the experiments. In this paper, the authors have developed an application that allows the reviewer to analyze the list of paper references and an approach to the analysis of the bibliography, which allows identifying those sources that may not be relevant to the research topic. This approach allows determining the artificial increase of irrelevant references.

Keywords. Article Quality, Citation Analysis, References, Academic Plagiarism, Academic Integrity.

1 Introduction

The scientific activity assessment is based on the analysis of information flows, which are presented as documents. The quality of these documents affects the evaluation quality, therefore the task of developing the indicator and methods for assessing the quality of scientific documents is very timely and relevant.

This paper discusses how quality is interpreted and how it is measured. Research quality is a multidimensional concept, where plausibility/soundness, originality, scientific value, and societal value commonly are perceived as key characteristics [1].

According to the final report of the European project "European Educational Research Quality Indicators", a separate project area is the development and testing of internal and external quality indicators [2]. Internal quality indicators can be identified from the text and external quality indicators are metadata, bibliometric and/or

webometric information. The list of indicators are rigour; originality; significance (for other researchers, policy, and practice); integrity (considerations of authenticity, honesty and ethical requirements in the conduct of research); style (including clarity, communicability, eloquence, and elegance).

In our opinion, the rigor, originality, and integrity indicators depend on the quality of the scientific sources studied to a certain extent. Therefore, quotes and a reference list also can be considered as indicators of document quality. The authors of [3] confirm the citation influence not only on other paper characteristics but also has an influence on the whole document.

2 Background

A scientific paper consists of such elements as title, information about the authors, abstract, keywords, body and list of references. We focus on the study of the scientific reference list.

Citation is an essential component of any scientific work and one of the important means of scientific communication. In scientific publications, citation can be considered in various aspects. Citation is used to solve problems in many directions. Some of them are presented in Table 1.

Table 1. Research directions of citation.

Direction	Publications
Assessment of scientific results for scientific groups, departments and institutions (bibliometric methods)	[4], [5], [6], [7]
Academic rankings (local and global) and distribution of research funding	[8], [9]
Document quality	[10], [11], [12]
Determine of Manipulation of scientometric indicators	[13], [14], [15]
Identification of intellectual plagiarism (citation-based plagiarism detection, definition of an idea)	[16], [17], [18], [19], [20], [21]

All listed directions are important. Bibliometric methods are used widely nowadays. The indicators which are determined on the basis of citation effects the result of academic ratings and the distribution of funding. This contributes to the emergence case of academic fraud and manipulation of indicators in the academic environment. Therefore, ensuring a qualitative assessment of research is relevant.

In practice, the responsibility for the article quality lies with the scientific reviewer, who is an expert in the researched field. He makes decisions based on an analysis of all the substantive and formal properties of a publication, including analyzing a reference list. The authors independently determine the appropriateness and rationality for using a quote.

The authors of this paper aim to investigate and describe the qualitative characteristics of the reference list in already published articles and, based on the mentioned

characteristics to propose a software to the reviewer (expert). The proposed software will be recommendatory and will help to reviewer quickly examine the paper and pay attention to certain formal features of the reference list and may help to indicate the unreasonable use of sources.

3 Research of Ukrainian Papers in Web of Science Core Collection and Scopus

The reviewer's task is to perform a comprehensive evaluation of scientific work. Peer review can be divided into two stages: i) analysis and evaluation of formal features; ii) scientific review of the publication content. Formal features include the following characteristics: total number of citations; time ranges for quotes; number (percentage) of unique source names; quality of sources (indicators and types); a percentage of self-citation; inconsistency of certain citations with the subject of publication; non-specific sections with excessive citations.

By the example of already published articles that were scientifically reviewed and included in the Web of Science Core Collection (WoS CC) and Scopus, we analyze and characterize some of them of formal features.

We used the 2018 publications on a theme of own scientific work (indexation in WoS CC as of 06.2019) for the analysis of paper formal properties. It is the 83 publications, they were selected from publications included in WoS CC (topic "computational linguistics", 2018 year, categories "computer science", "artificial intelligence", "language linguistics", "information science library science", "computational linguistics").

It should be noted that the average number of sources for publications in various thematic areas is different. The average number of sources in the list for the indicated topic is 39, with the exception of 6 review ones (more than 100 citations).

We determined the percentage of sources included in WoS CC, the year of publication, and the quantity and quality of unique sources for the 20 publications from the list (83 publications).

The citation number in the bibliographic lists of 20 publications is 1087; 506 citations (47%) of them are included in WoS CC. The reliability of publications data included in scientometric databases is not in doubt. The average percentage of links to external sources in the references list is 53%. Possible reasons are certain scientific sources are not included in scientometric databases or scientists have limited access to foreign publications.

The relevance and modernity of the study are evidenced by the use of a significant number of publications in recent years. Review papers are exceptions since a thorough study of the topic is necessary for a qualitative examination of the topic.

The citation structure by years of publication is shown in Fig. 1. Authors often use publications of recent years (2010-2018 – 57%). But older publications (1950-1969) may be considered in review papers.

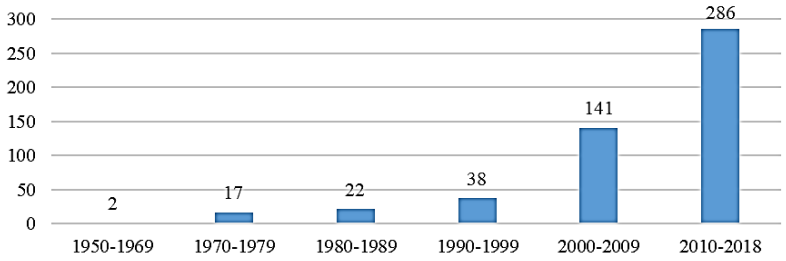

Fig. 1. Citations by publication year.

To ensure the completeness of the study, it is necessary to analyze materials from different sources. The more diverse the list is the better. Therefore, we investigate links from our collection and select unique names of sources for each paper. The total number of unique source names is 245. Fig. 2 shows the percentage of unique sources for 20 publications. The average percentage of unique sources is 48%.

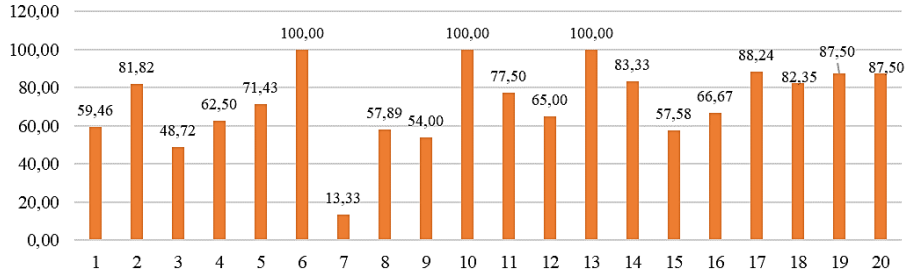

Fig.2. Percentage of unique sources for 20 publications.

The quality of the reference source is also important. The reviewer, who is an expert on the topic, has information about heavyweight journals in the field and can identify their names on the list. Table 2 presents scientometric indicators of journals which often used in reference lists (TOP-10). More citations (9 out of 20 documents) were identified in the “Computational Linguistics” journal. All journals are well-known and respected.

An important characteristic of the references is the percentage of self-citation. Authoritative publishers recommend authors to limit the amount of self-citation to 30% of the total number of sources in the citation list. It is believed that this is enough to demonstrate the previous and related works of the authors. In practice, self-citation can be different (from 0% to 100%).

The paper [15] describes the self-citation analysis for a data set of 7 million scientists in the world. The result, the median self-citation rate is 15.5%. Scientists from the United Kingdom, United States, Turkey have self-citation rates below the median; Japan, China have on the median level. Ukrainian scientists belong to a group of scientists with self-citation rates up to 40%.

In our paper, we investigated the effect of self-citation and obtained a similar to [15] result for Ukraine. We analyzed 100 author profiles of Ukrainian scientists (Scopus, Computer Sciences area) and defined the part of self-citation in the total citation and its influence on the author's h-index: average self-citation is 35%; MAX % self-citation is 96%; MAX % growth of the h-index is 80%; h-index unchanged in 12 profiles.

Table 2. TOP 10 unique sources from citation lists and their indicators.

Name of journal	Number of documents	Impact Factor	Impact Factor (5 years)	Quartile
Computational Linguistics	9	1.319	2.202	Q3, Q3, Q2
Journal of the American Society for Information Science and Technology	7	2.452	2.762	Q1, Q1
Science	6	41.058	40.627	Q1
Communications of the ACM	5	3.063	5.29	Q1, Q1, Q1
Journal of Machine Learning Research	5	2.281	5.805	Q2, Q2
ACM Transactions on Information Systems	4	1.767	2.203	Q2
Biometrics	4	1.524	1.962	Q3, Q3, Q2
Journal of the Association for Information Science and Technology	4	2.835	2.931	Q2, Q1
Plos One	4	2.766	3.352	Q1
Artificial Intelligence	3	3.034	4.156	Q1

The authors' profiles of Ukrainian scientists distribution by influence self-citation degree on the h-index is shown in Fig. 3. The index increased from 1% to 20% for 56 profiles; for 24 profiles – 21%-30%; for 6 profiles – 31%-40%; for 6 profiles – 41%-50%. The h-index in-creased by more than 51% for 8 profiles.

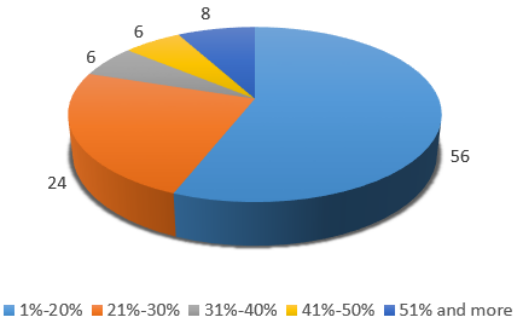

Fig. 3. Authors' profiles distribution by influence self-citation degree on the h-index.

It is determined that self-citation affects the scientometric indices of the authors. It is very difficult to determine the authors' abuse based on the count of self-citation. The authors determine the expediency and justification for the use of quotes, so the problem of self-citation is on an ethical plane. The self-citation percentage of the total number of citations for 100 author profiles is presented in Figure 4.

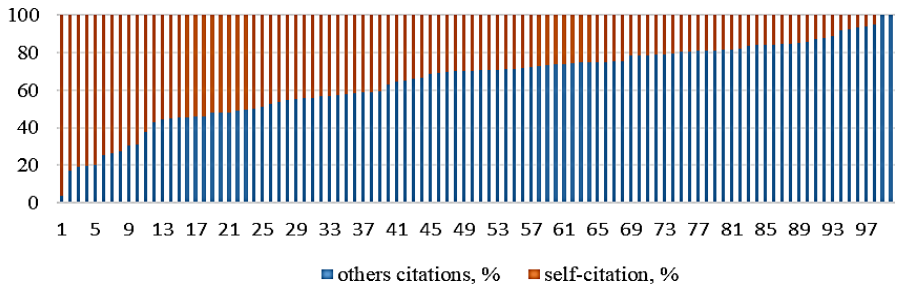

Fig. 4. The self-citation percentage of the citation total number of citations for 100 author profiles.

The use of information tools to automate scientific activity accelerates the scientific process. The reference list, formed in the required format, allows the use of software for data analysis.

We use VOSviewer for quick visualization and analysis of information about authors and the subject of links. It's a software tool for creating visualized scientific landscapes based on textual data. For experiments, we employed the data of Web of Science CC. Through VOSviewer we can quickly analyze such indicators as i) excessive self-citation – a network of author citing; ii) relevance of the quotation topic to the paper topic – keyword analysis from the title and annotation. Fig. 5 shows the author citation network from the reference list for paper1 with using VOSviewer.

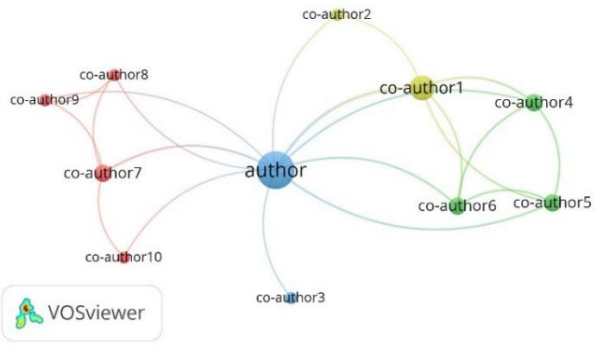

Fig. 5. VOSviewer – author analyze with using citations.

The published author is associated with all links and he is present in all 9 citations. In addition to self-citation, unscrupulous authors may cite papers “on order”. Formally, such sources differ in topic and are not related to a specific study. VOSviewer visualizes the semantic relationships between words (title, annotations, keywords). In Fig. 6 we presented the publication titles and annotations from the paper2 citations.

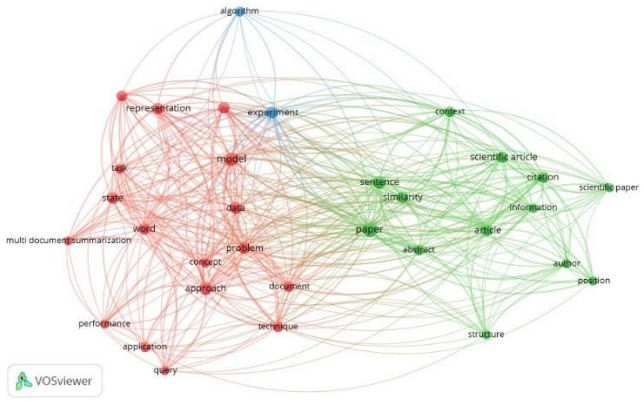

Fig. 6. VOSviewer – publication name analyze and annotations from citations.

All words are closely related and demonstrate a certain interdisciplinary interaction (Fig. 6). Thus, we examined the following features and identified them for our data sets. Our results are presented in Table 3.

Table 3. Data sets features.

		Min value of citations	Max value of citations	Average value of citations
1	Sources on the list (for paper)	5	89	39
2	Sources on the list (for a review paper)	128	390	196
3	Year of publication: • 2019 (2 quotes) • 2018 (15 citations) • 2017 (25 citations) • 2016 (45 citations) 2015 (39 citations)	1950	2019	-
4	Unique sources	13%	100%	72%
5	Self-quoting author profile,% of total citations (except for profiling without self-quoting)	5%	96%	36%
6	Increase in the author’s h-index due to self-citation (except for profiling without self-quoting)	5%	80%	25%

The described characteristics are not clearly formalized, and therefore the appropriateness of using sources is confirmed by reviewing the content of the publication. Thus, our task is to develop the special software in order to distinguish and present the characteristics of the list for the expert review.

4 Experiments and Application for Reviewer

The reviewer's task is to perform a comprehensive evaluation of scientific work. In this section, we show 1) a developed application that allows the reviewer to analyze the paper reference list; 2) an approach to the analysis of the bibliography, which allows identifying those sources that may not be relevant to the research topic and, accordingly, artificially increase the performance of other authors.

4.1 Software for Reviewer

In the application development for analyzing the reference list, we tried to take into account not only our research but and the recommendations of conferences and journal's editorial boards. These main functions we presented below.

Analysis of the publication year. This function helps with the issue of publication relevance, how relevant they are at the time of writing. Our application has a threshold field in which the user can enter the year and the program calculates the number of publications before and after this year. This function allows you to quickly understand whether the author has analyzed the latest research in this area or not (Fig. 7).

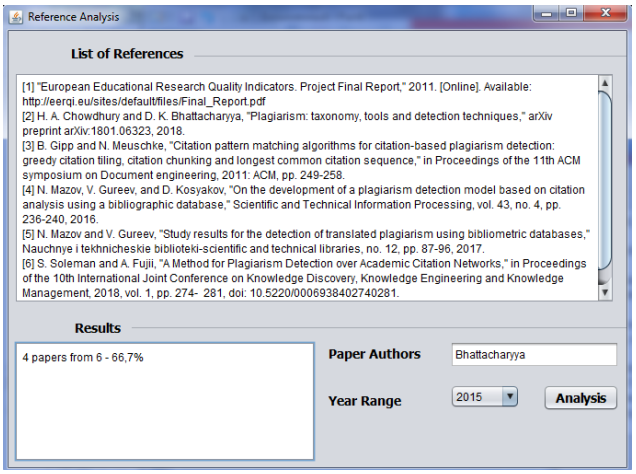

Fig. 7. The example analysis of the publication year.

As an example, in the field "Results" Fig. 7 a reviewer can see how many publications belong to the 2015 year and higher and a percentage value for these papers towards the total number.

Analysis of self-citation. The user needs to enter the authors of the papers in the "Authors" field and the program gives him the papers of these authors from the reference list and calculates the percentage of the total number (Fig. 8).

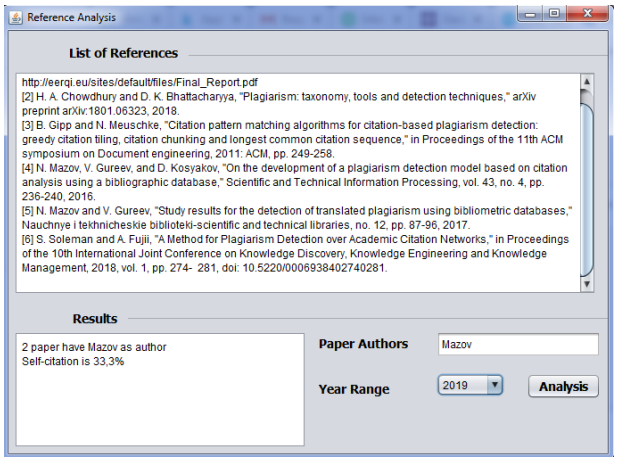

Fig. 8. The self-citation analysis example.

As an example, in the field "Results" Fig. 8 a reviewer can see that an author with the surname "Mazov" has 2 papers and a percentage value for self-citation is 33,3%.

4.2 Identification of Irrelevant Sources in Reference List

Analyzing the difficulties which reviewers face, we found such a problem as an artificial increase in the citation for a publication. This is realized by citing a source irrelevant to the main topic.

In order to identify such publications, we propose an approach that uses the methods of computational linguistics and determines the proximity between the sources of references, and can also take into account paper keywords if it is necessary. Consider the following example, we have the next reference list (as an example, we take the reference list from our paper), which consists of 21 sources [1-21] from this paper. Define these papers as P_1, … P_{21}. We artificially added the paper to this list that is not relevant to this topic. This is the following source:

Lefèvre T, Gouagna L-C, Dabiré KR, Elguero E, Fontenille D, Renaud F, et al. (2010) Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes. PLoS ONE 5(3): e9546. https://doi.org/10.1371/journal.pone.0009546

Denote this source as P_{22}. In order to determine the less similar source to the paper topic, we compare the title of each paper with the title of our work "*The References Analysis as a Quality Characteristic of a Scientific Article*". The comparison we implement with similarity measure from Spacy library and word embedding models. In the Spacy library, a full sentence word-embedding calculates as an average over all words in the sentence. Before processing, we deleted stopwords in each sentence. As a result, we got Fig. 9.

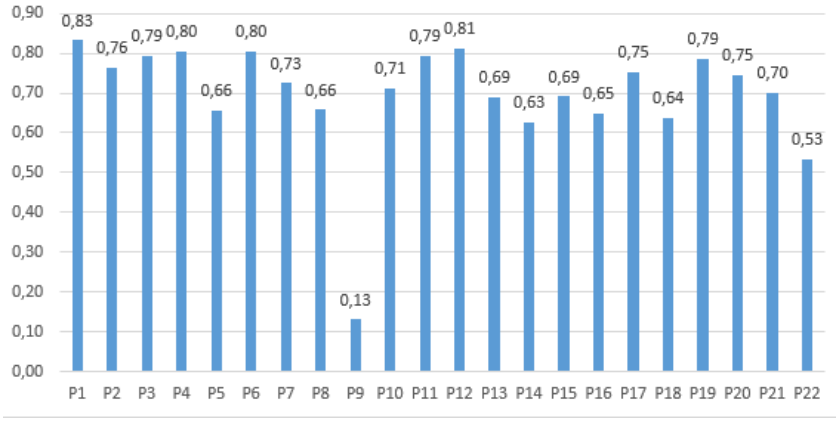

Fig. 9. Results of semantic comparison of the paper title with the article titles from the reference list.

We received the minimum value of 0,13 for P_9. The title of this publication was obtained by transliteration from Russian. Therefore, such proximity coefficient was obtained. However, for the publication with number P_{22}, we received low value and this confirmed our hypothesis. Because this is our "artificial" publication. Other elements from our reference list have very close values from 0,6 to 0,8. It should be noted that this approach does not provide a 100% guarantee for identifying such irrelevant sources, but can help identify candidates for such references.

5 Conclusions and Recommendations

In this work, we have analyzed the qualitative characteristics of a scientific document and focused attention on the paper reference list as an object for research. For research, we selected articles from journals included in the Web of Science Core Collection and data profiles of Ukrainian scientists from Scopus. For analysis, we used the capabilities of Web of Science Core Collection, Scopus, VOSviewer, MS Excel.

According to our research results, we recommend the reviewer, first of all, pay attention to the formalized properties of the citation list. The reference list is researched in this publication and results demonstrate that the quality of the paper content could e defined through the citation list analysis. Due to the allocation and formalization of the citation list properties, it became possible to create a special software tool for reviewers.

We developed an application that allows the reviewer to analyze the reference list of paper and proposed the approach to the analysis of the bibliography, which allows identifying those sources that may not be relevant to the research topic. Our experiments showed that the proposed approach is worked well enough and our next step will be to experiment on the big data sets.

References

1. Aksnes, Dag W., Liv Langfeldt, and Paul Wouters: Citations, citation indicators, and research quality: An overview of basic concepts and theories. Sage Open 9(1), 1-17 (2019).
2. European Educational Research Quality Indicators. Project Final Report, http://eerqi.eu/sites/default/files/Final_Report.pdf, last accessed 2020/02/01.
3. Tahamtan, Iman, Askar Safipour Afshar, and Khadijeh Ahamdzadeh: Factors affecting number of citations: a comprehensive review of the literature. Scientometrics 107(3), 1195-1225 (2016).
4. Biagioli M.: Quality to Impact, Text to Metadata: Publication and Evaluation in the Age of Metrics. KNOW: A Journal on the Formation of Knowledge 2(2), 249-275 (2018).
5. Hyland K.: Self-citation and self-reference: Credibility and promotion in academic publication. Journal of the American Society for Information Science and technology 54(3), 251-259 (2003).
6. Moed, H. F.: Citation analysis in research evaluation. Springer, Dordrecht, The Netherlands (2005).
7. Cabezas-Clavijo, A., Robinson-Garcia, N., Escabias, M., & Jimenez-Contreras, E.: Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? PLoS ONE 8(6), (2013). – doi:10.1371/journal.pone.0068258.
8. Piro, F. N., Sivertsen, G.: How can differences in international university rankings be explained? Scientometrics 109, 2263-2278 (2016).
9. Akoev, M., Markusova, V., Moskaleva, O., Pislyakov, V.: Rukovodstvo po Naukometrii: Indikatori Razvitiia Naukii Tehnologii [Handbook for Scientometrics: Indicators of science and technology development]. Ural Federal Univ, Ekaterinburg (2014).
10. Shibayama, S., Wang, J.: Measuring originality in science. Scientometrics 122, 409–427 (2020). – doi:10.1007/s11192-019-03263-0.
11. Tahamtan I., Afshar A. S., Ahamdzadeh K.: Factors affecting number of citations: a comprehensive review of the literature. Scientometrics 107 (3), 1195-1225 (2016).
12. Krapež K.: The (Un) Originality of Scientific Papers–An Analysis of Professional Quality Standards. In: Management, Knowledge, and Learning International Conference. Make Learn, Zadar (2013).
13. Baas J., Fennell C.: When peer reviewers go rogue-Estimated prevalence of citation manipulation by reviewers based on the citation patterns of 69,000 reviewers. In: ISSI 2019. Rome, Italy (2019).
14. Ioannidis, John PA, Richard Klavans, and Kevin W. Boyack.: Thousands of scientists publish a paper every five days. Nature 561(7722), pp. 167-170 (2018).
15. Van Noorden, Richard, and Dalmeet Singh Chawla: Hundreds of extreme self-citing scientists revealed in new database. Nature 572, pp. 578-579 (2019). – doi: 10.1038/d41586-019-02479-7.
16. Gañan D.: Plagiarism Detection. In: Baneres D., Rodríguez M., Guerrero-Roldán A.: Engineering Data-Driven Adaptive Trust-based e-Assessment Systems. Lecture Notes on Data Engineering and Communications Technologies, vol. 34. Springer, Cham (2020).
17. Foltýnek T., Meuschke N., Gipp B.: Academic plagiarism detection: a systematic literature review. ACM Computing Surveys (CSUR) 52(6), 1-42 (2019).
18. Gipp B: Citation-based plagiarism detection. In: Citation-based plagiarism detection, pp. 57–88. Springer Vieweg, Wiesbaden (2014).
19. Mazov N., Gureev V., Kosyakov D.: On the development of a plagiarism detection model based on citation analysis using a bibliographic database. Scientific and Technical Information Processing 43(4), 236-240 (2016).

20. Mazov N., Gureev V.: Study results for the detection of translated plagiarism using bibliometric databases. Nauchnye i tekhnicheskie biblioteki-scientific and technical libraries 12, 87-96 (2017).
21. Soleman S., Fujii A.: A Method for Plagiarism Detection over Academic Citation Networks. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 1, pp. 274-281 (2018). – doi: 10.5220/0006938402740281.