Using Topic Modeling for Automation Search to Reviewer

Yuliia Hlavcheva[1], Olga Kanishcheva[1], Maryna Vovk[1], Maksym Glavchev[1].

[1] *National Technical University "KhPI", 2 Kyrpychova str., Kharkiv, 61002, Ukraine*

Abstract

Maintaining the scientific research quality and publications is a prerequisite for ensuring the science development. The solution of this problem is provided by work in various directions, such as popularization of ethics and research integrity; ensuring transparency of the peer-review process, use of intellectual text processing methods. The reviewer plays an important role in the paper evaluation during the process of expert scientific evaluation. This paper is presented the use of topic modeling and natural language processing methods to find reviewers who are experts in this paper's field. The proposed information-topic model not only helps to find potential reviewers for articles, but also helps to resolve conflicts of interest, based on the use of metadata. The proposed method can be used for any industry, for any language, with little adaptation of pre-processing methods of textual information. The presented method works well, quickly and really people who have been found by our software, fit the paper subject.

Keywords

Topic modeling, review process, natural language processing, article quality, article metrics.

1. Introduction

A significant number of issues related to the activities of a scientist and an academic institution depend on the quality of their evaluation. For example, in Ukraine, the serial number of the university in one of the international rankings is taken into account in the distribution of funding (state budget expenditures between higher education institutions) [1]; expert judgement based on scientometric and other indicators used in the international system of examination, which are used to assess the quality of scientific and scientific-technical activities [2]; the number of articles in WoS and/or Scopus is taken into account when including the publication in the list of scientific professional publications of Ukraine [3], when approving the license conditions for educational activities high institutions [4], etc.

The practice shows rapid growth in the number of academic products, they are academic articles, conference proceedings, books, reports, presentations. Evaluating such a large number of scientific documents is a complex multifaceted task.

Also, the quality of scientific publications is a crucial issue not only for researchers and academic institutions but also for organizations that fund research, shape science policy, engage in employment, organize scientific awards based on the evaluation of academic research.

The definition and implementation of scientific article quality indicators are described in many publications [5, 6, 7]. The author [7] investigates and determines the weights of each of the four basic quality indicator groups: Citation metrics and Engagement metrics [8]; Scientific collaboration metrics; Educational metrics. According to the study, the most important indicators from the group are Citation metrics – 0.45. Other groups received the following weights: Engagement metrics – 0.27; Scientific collaboration metrics – 0.18; Educational metrics – 0.09.

Studies [5, 6] also describe the use of quality indicators, which are calculated after the article has been scientifically reviewed and included in the scientometric databases of Web of Science or Scopus.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: yuliia.hlavcheva@khpi.edu.ua (Y. Hlavcheva); kanichshevaolga@gmail.com (O. Kanishcheva); Maryna.Vovk@khpi.edu.ua (M.Vovk); maksym.glavchev@khpi.edu.ua (M. Glavchev)
ORCID: 0000-0001-7991-5411 (Y. Hlavcheva); 0000-0002-9035-1765 (O. Kanishcheva); 0000-0003-4119-5441 (M.Vovk); 0000-0001-9670-9118 (M. Glavchev)

But, even after the publication of a scientific paper, as world practice shows, it can be retracted. According to the recommendations of the Committee on Publication Ethics, editors should consider retracting the publication in the following cases [9]:

• there is evidence that the conclusions are unreliable (errors in calculations, falsifications, manipulation);

• plagiarism was exposed;

• the document contains material or data without permission for use;

• copyright has been infringed;

• unethical research is described; the document was published based on a compromised peer review process.

That is why the authors of this paper pay attention to determining the article quality at the stage of scientific peer review and ensuring the objectivity of the peer-review process. To do this, measures should be taken in the following areas:

1. ensuring the transparency of the scientific review process by creating instructions and tools;
2. selectivity in choosing the reviewer.

To ensure the first direction we can identify the main factors that potentially indicate the quality of the article [10]: the novelty of the study; the potential impact of results and expected contribution to the sphere of activity; absence of errors and reliability; the validity of conclusions based on data; article structure; quality and validity of literature review; compliance with the article style recommendations. The importance of determining the quality and validity of the literature review and an auxiliary tool for the reviewer, the authors of this article described in [11]. To ensure the transparency of the scientific review process, publishers create and publish open access to extended review instructions [12, 13].

To guarantee the appropriate reviewer authors of the article propose an approach to the selection of a reviewer. A reviewer is an expert in the scientific area to which the peer-reviewed document belongs and has not biased against the authors.

2. Related Works

The urgency of the task of choosing an unbiased professional expert to conduct a scientific examination of the article is confirmed by the difficulties of choosing a reviewer, which takes place in a rapidly growing volume of information and development of interdisciplinary science [14].

A study conducted by the Swiss National Science Foundation found that when grant applicants are selected, grant authors are four times more likely to receive a positive response than a negative one [15].

Each editorial board decides on this issue in its own way. To facilitate this task, the web service Publons (Clarivate) [16] can be used, which contains information on more than 3 million researchers from around the world. Publons was launched in 2012 as a platform to recognize reviewers for their important work. The selection of potential reviewers is done by selecting applicants for a particular topic from the list of standard topics of Web of Science (Clarivate) and further analysis of their scientometric indicators and the list of publications [17]. In total, the Web of Science includes approximately 250 subject areas in the natural, social and human sciences, making it difficult to find reviewers for publications on narrow or interdisciplinary topics.

An alternative approach to finding a reviewer may be to create and use your own database of scholars. However, self-formation and maintenance of the relevance of the reviewer database take a long time and cannot be effective.

The search for a reviewer should be carried out very carefully and responsibly so as not to fall for unscrupulous scientists. There are known cases where reviewers use Publons to gain recognition for superficial or poor expert judgment [18].

Reviewing remains an important element in scientific publishing. Therefore, innovative technologies are being introduced to improve it. Artificial intelligence and specialized programs, which create auxiliary tools for the reviewer [10], are used at different stages of review: search for academic plagiarism, identification of incorrect statistical results, grammatical tests, language quality assessment, etc.

Springer has implemented a tool for finding reviewers "Springer's reviewer finder", which offers to select an expert based on the metadata of publications [19]. The correspondence algorithm returns a list of researchers who have a publishing profile similar to the profile of the manuscript author(s) [20].

Frontiers (Switzerland) has developed and presented to external editors the Artificial Intelligence Review Assistant (AIRA) software. AIRA helps editors, reviewers and authors assess the quality of manuscripts: language quality assessment, number integrity, plagiarism detection, and potential conflicts of interest. The software finds whether authors, editors, or reviewers of articles have co-authored in the past. It also notes articles on controversial topics that require increased attention from the editor [21].

The authors of this article aim to explore the use of modern means of choosing a reviewer for the scientific examination of manuscripts. The authors propose a method of optimizing and simplifying the process of finding potential reviewers by the editorial boards of scientific publishers.

3. Data Description

The ways to optimize/simplify the process of finding a reviewer are offered. The reviewer is an expert in the qualitative evaluation of scientific information. He is a specialist who has high educational, scientific and professional qualifications, has the experience, has scientific publications in authoritative scientific journals around the world.

That is why the authors consider global multidisciplinary academic resources as a source of data for the reviewer, which contains information about the authors and their publications, as well as the ability to export data for automated processing. Such resources include publications from various scientific journals and publishers, not limited to specific geography. In this study, the authors analyzed the following resources: Scopus (http://scopus.com), Web of Science Core Collection (http://webofscience.com), Dimensions (https://app.dimensions.ai/discover/publication).

Data from the Scopus bibliographic and abstract database were used to form the information-topic model (ITM). Scientific publications included in this database meet strict selection requirements and are considered authoritative scientific publications. Thus, authors published in these publications can be considered potential reviewers and involved in scientific reviews.

This paper uses data from the Computing science area (Scopus). In total, Scopus contains 7.5 million documents in the subject area of Computing Science. To ensure the use of exclusively relevant published data in experimental research, only articles from journals published in 2020-2021 were selected. Publication data is also limited to articles authored (one of the authors) by scientists from Ukraine. The proposed method can be easily adapted to other data (topic areas), regardless of the language of the data.

3.1. Data Statistics

For automated processing and formation of information-topic model from Scopus exported information about 2550 publications (Computing science, journal articles, 2020-2021 publications, authors from Ukraine). Statistics of publications are presented in Table 1.

Table 1
Statistic of publications

Indicator	Value
Total articles	2550
Year of publication 2020/2021	1341 /1209 papers
Authors with 4 or more publications	160
Maximum number of articles per author	25
Geography of the authors	81 country
Number of original sources (journals)	157
Subject Area to which the articles belong	24

It should be noted that articles may be interdisciplinary. It means that the same paper simultaneously relates to two areas. One field is Computing science, and the other field is determined by the scope. This is evidenced by 24 Subject Area, which includes 2550 publications in Computing science. The TOP-10 Subject Area is presented in Table 2.

Table 2
ТОП-10 Subject Area

Subject Area	Number of publications
Engineering	1671
Mathematics	1207
Business, Management and Accounting	1172
Energy	908
Decision Sciences	335
Social Sciences	269
Physics and Astronomy	185
Chemical Engineering	155
Chemistry	150
Materials Science	146

The interdisciplinary nature of articles submitted to the publication makes it difficult to solve the problem of finding a reviewer because he must be an expert in two disciplines (areas).

The method described in this publication takes into account these features and shows good results for interdisciplinary articles.

3.2. Metadata Description

As mentioned earlier, the choice of the source was influenced by the technical possibilities of exporting data for automated processing.

Scopus provides export of 5 metadata groups of publications in CSV text format, intended for presentation of tabular data.

Citation information:

- Author(s),
- Author(s) ID,
- Document title,
- Year,
- EID,
- Source title,
- Volume, issue, pages,
- Citation count,
- Source & document type,
- Publication Stage,
- DOI,
- Open Access.

Bibliographical information:

- Affiliations,
- Serial identifiers (e.g. ISSN),
- PubMed ID,
- Publisher,
- Editor(s),
- Language of original document,
- Correspondence address,
- Abbreviated source title.

Abstract & keywords:

- Abstract,
- Author keywords,
- Index keywords.

Funding details:

- Number,
- Acronym,
- Sponsor,
- Funding text.

Other information:

- Tradenames & manufacturers,
- Accession numbers & chemicals,
- Conference information,
- Include references.

To form an information-topic model, 6 types of metadata were used, which identify the author (authors) of the publication and reflect the topic area of the article:

1. Authors – name and surname of the author(s).
2. Author(s) ID – identification number of the author(s) in the Scopus database.
3. Title – the title of the article, which clearly reflects its content.
4. Abstract – is a brief summary of a research article.
5. Author Keywords – these are keywords chosen by the author(s) which, in their opinion, best reflect the contents of their document [22].
6. Index Keywords – these are keywords chosen by content suppliers and are standardized based on publically available vocabularies. Unlike Author keywords, the Indexed keywords take into account synonyms, various spellings, and plurals [22].

Metadata statistics are presented in Table 3.

Table 3
Metadata statistics

	Title	Abstract	Author Keywords	Index Keywords
Total tokens	37361	524962	27943	20066
Tokens with the length of 4 or more (meaningful words)	25401	316239	25504	19034
The average length of metadata field (meaningful words)	10	124	10	8
Number of original tokens	7902	28776	7004	4508
Percentage of original tokens to their total number	22%	6%	25%	23%

4. Method Description

Publishers' websites present a significant amount of material that describes the algorithm of actions and the review process, but not the search for a potential reviewer. In our opinion, in general, the process of finding a reviewer can be described by such an algorithm (Figure 1).

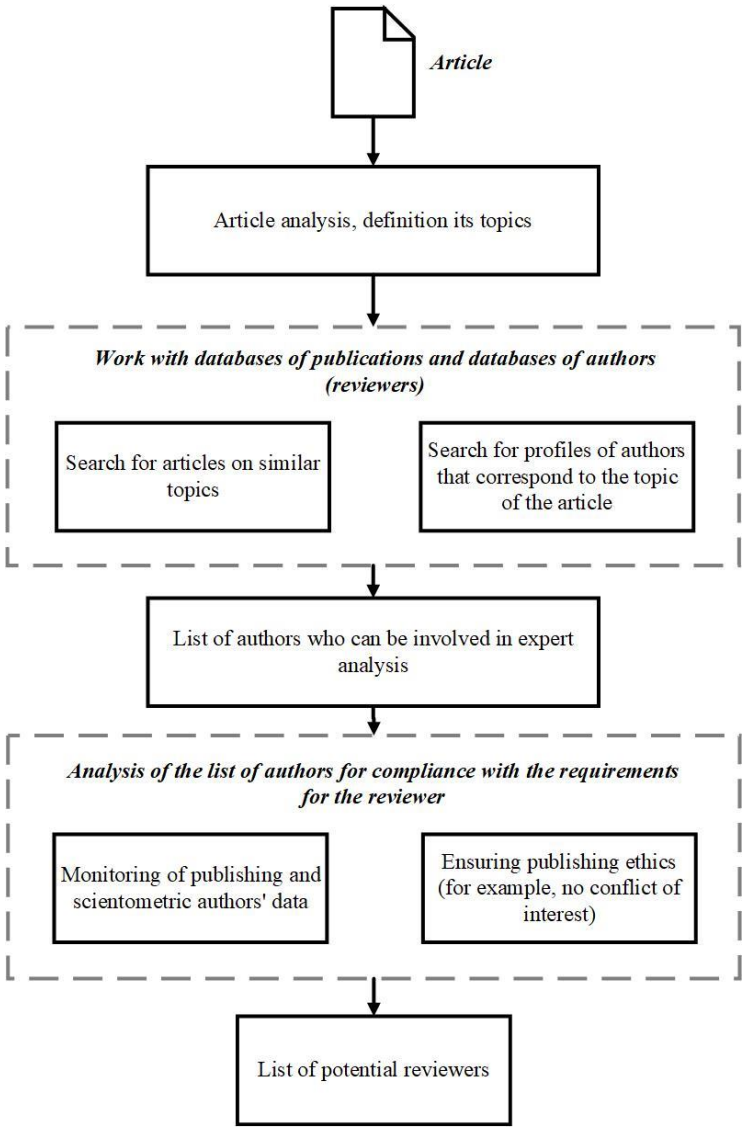

Figure 1: General reviewer search algorithm

The method described in this article optimizes the most resource-intensive stage "Working with databases of publications and databases of authors (reviewers)", the result of which is a list of authors who could be involved in the expert analysis of the article. Also, the presented method partially simplifies the stage "Analysis of the list of authors for compliance with the requirements for the reviewer" in terms of monitoring the publication of applicants. We believe that this monitoring is included in the method presented in the article. The general algorithm of the method is presented on Figure 2.

Due to the use of methods of natural language processing and topic modeling, the presented method differs significantly from the general algorithm for finding a reviewer. The new method does not require additional subject area analysis of the article and carefully analyzes the publishing activities of applicants. The topic branch is chosen based on the received information-topic model. A list of reviewers is offered on the basis of BMI publication information.

5. Experiments and Results

Data from the Scopus database, namely metadata from Compute Science articles, were used as experimental data, 2,550 articles were taken in 2020-2021. Fields such as *"Title"*, *"Abstract"*, *"Author Keywords"*, *"Index Keywords"* were used as basic data. There is enough data to use the Latent Dirichlet Allocation (LDA) method as a basic topic modeling algorithm.

Latent Dirichlet Allocation is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words [22, 23]. The LDA provides information on topic collections of any text in general, individual documents, and the relationships between them.

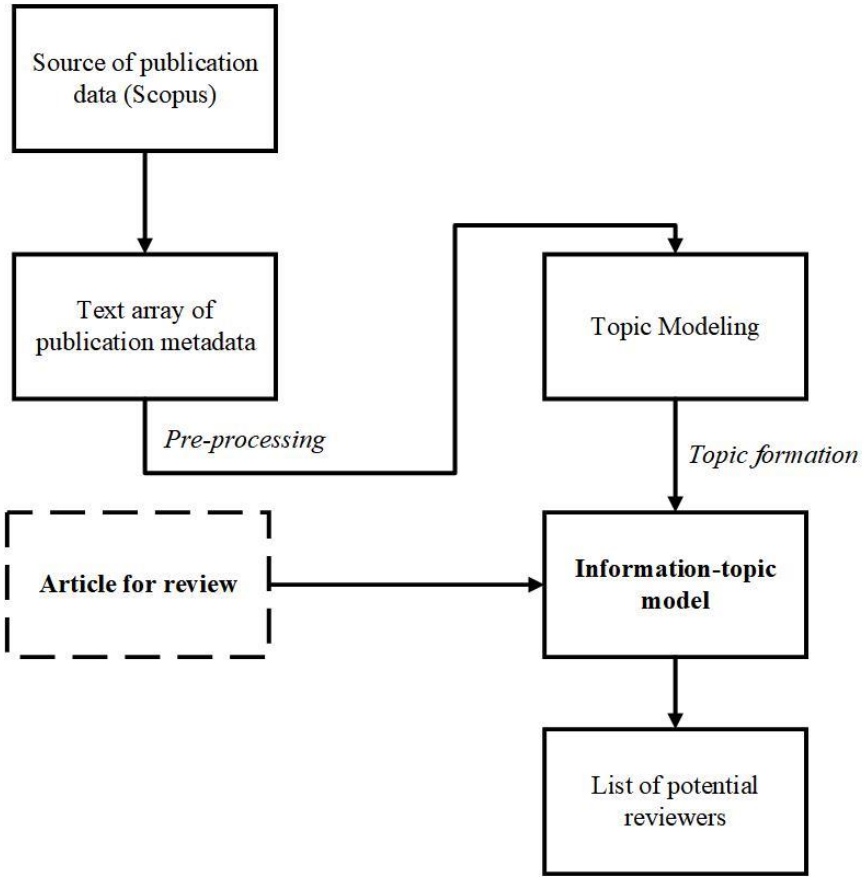

Figure 2: The general algorithm of the method of potential applicants for the reviewer role

All data passed the pre-processing stage, namely the removal of stop words, removal of punctuation, numbers, etc. Morphological analysis was also used to use only parts of speech such as nouns, verbs, and adjectives ('JJ', 'NN', 'NNP', 'NNS', 'VB', 'VBP').

As a result of experiments with the value of coherence, it was found that the best division into topics is for 34 classes. The formed information-topic model contains 34 topics. According to the statistics of publication data, publications are divided into 24 Subject Areas (Table 1). Thus, the LDA distribution forms a larger number of classes by 10 topics and is more detailed. Visualized representation of topics is presented in Figure 3.

For experiments on the reviewer selection, we took articles from the journal "Cybernetics and Systems Analysis". Some of these articles like,

1. A.Y. Bomba, M.V. Boichura, Identifying the Structure of Soil Massifs by Numerical Quasiconformal Mapping Methods, Cybern Syst Anal 57(6) (2021) 927-937.
2. V.V. Semenov, S.V. Denisov, A.V. Kravets, Adaptive Two-Stage Bregman Method for Variational Inequalities, Cybern Syst Anal, 57(6) (2021) 959-967.
3. V.I. Norkin, Stochastic Generalized Gradient Methods for Training Nonconvex Nonsmooth Neural Networks, Cybern Syst Anal, 57(5) (2021) 714-729.
4. L.O. Hnativ, Discrete Cosine-Sine Type VII Transform and Fast Integer Transforms for Intra Prediction of Images and Video Coding, Cybern Syst Anal 57(5) (2021) 827-835.
5. P.S. Knopov, E.J. Kasitskaya, Large Deviations of Empirical Estimates in the Stochastic Programming Problem for the Homogeneous Random Field with a Discrete Parameter, Cybern Syst Anal 57(5) (2021) 704-713.

As a result of the experiment with topic modeling, the following results were obtained.

Article 1 belongs to class 1 with a probability of 0.8500712. The probability for 11 topic is very low.
[(1, 0.8500712), (11, 0.14165811)]

Article 2 belongs to class 29 with a probability of 0.8471544. Other classes have a low probability value.
[(29, 0.8471544), (10, 0.066285245), (34, 0.050520007), (1, 0.024828823)]

Article 3 belongs to class 18 with a probability of 0.8551518. Other classes have a low probability value.
[(18, 0.8551518), (14, 0.03998026), (4, 0.039681997), (7, 0.033085056), (26, 0.021731576)]

Article 4 belongs to class 22 with a probability of 0.9812607. Other classes have a low probability value.
[(22, 0.9812607), (28, 0.010214581)]

Article 5 belongs to class 30 with a probability of 0.9861704. Other classes have a low probability value.
[(30, 0.9861704)]

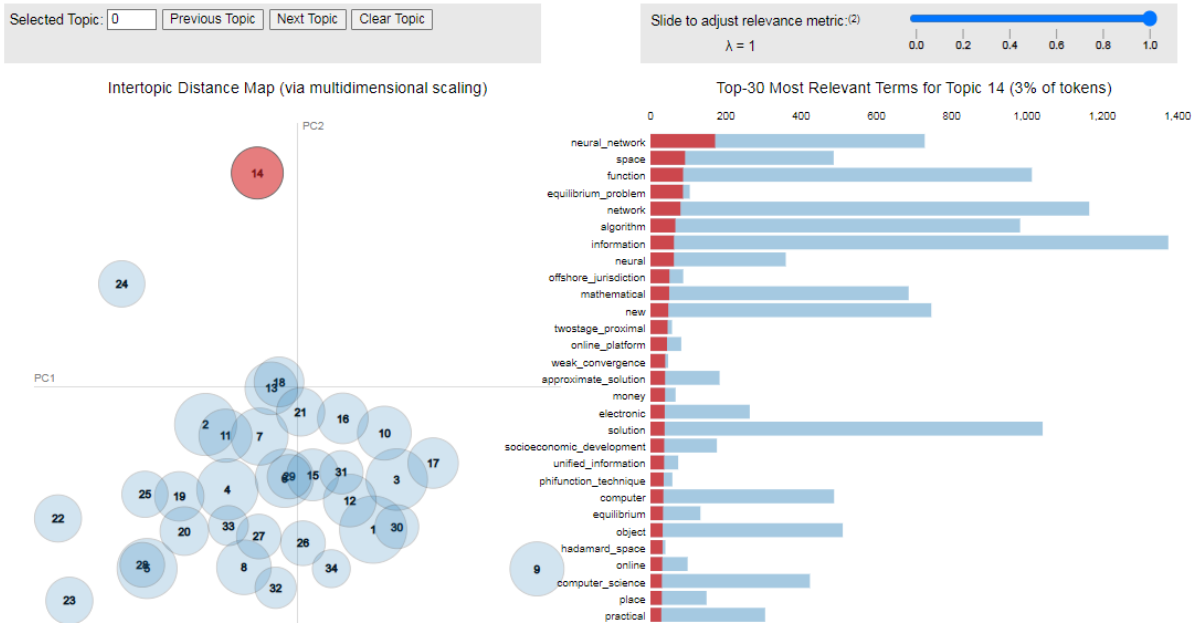

Figure 3: Visualization of topic modeling based on metadata in the field of Compute Science

Class 30 is described by the following words "*network*", "*artificial_intelligence*", "*mobile_robot*", "*mobile_robot*", "*measure*", "*code*", "*information*", "*service*", "*program*", "*geometric*" and others. Thus, we can say that the model correctly classified ***Article 5*** to 30 class.

Thus, we can conclude that the model works correctly and is suitable for optimizing/simplifying the process of finding a reviewer.

The next step was to find the names of potential reviewers. And for our data, the method found the following names of potential reviewers:

Article 1 Shevchenko I., Shastalo V., Kozak Y., Abramov Y., Basmanov O., Teslia I., Khlevna I., Yehorchenkov O., Latysheva T., Grigor O., Tryus Y., Prokopenko T., Polishchuk O.

Article 2 Kalyna V., Koshulko V., Ilinska O., Tverdokhliebova N., Tolstousova O., Bliznjuk O., Gavrish T., Stankevych S., Zabrodina I., Zhulinska O., Vedel Y.I., Denisov S.V., Semenov V.V.

Article 3 Alyokhina S., Maksymov M.V., Romashov Y., Osokina N., Kostetska K., Herasymchuk O., Tkachenko H., Podpriatov H., Pusik L., Falendysh N., Bobel I., Belinska K.

Article 4 Hnativ L.O., Luts V.K., Zhdaniuk V., Volovyk O., Kostin D., Lisovin S.

Article 5 Slyusar V., Protsenko M., Chernukha A., Kovalov P., Borodych P., Shevchenko S., Chernikov O., Vazhynskyi S., Bogatov O., Khrustalev K., Knopov P.S., Kasitskaya E.J.

There are some examples when the model with a small probability determines which topic the article belongs to. Examples of such cases are given below:

N.Y. Kuznetsov, I.N. Kuznetsov, Fast Simulation of the Customer Blocking Probability in Queueing Networks with Multicast Access, Cybern Syst Anal 57(4) (2021) 530-541.
[(27, 0.33189335), (30, 0.2498508), (2, 0.12633349), (22, 0.12138765), (15, 0.09545717), (11, 0.023521349), (29, 0.022272551), (25, 0.01759534)]

V.M. Bulavatsky, Mathematical Models with Local M-Derivative and Boundary-Value Problems of Geomigration Dynamics, Cybern Syst Anal 57(4) (2021) 563-577.
[(22, 0.64569837), (9, 0.18071947), (11, 0.16654922)]

V.M. Bulavatsky, Closed Solutions of Some Boundary-Value Problems of Filtration-Consolidation Dynamics within the Fractured-Fractal Approach, Cybern Syst Anal 57(3) (2021) 383-395.
[(11, 0.2851506), (12, 0.17952716), (1, 0.17065176), (19, 0.16245417), (22, 0.11737535), (25, 0.058207653), (32, 0.020618271)]

L.S. Stoikova, Exact Estimates of the Probability of a Non-Negative Unimodal Random Value Hitting Special Intervals under Incomplete Information, Cybern Syst Anal 57(2) (2021) 264-267.
[(12, 0.25641593), (25, 0.2526573), (15, 0.19586651), (8, 0.089407556), (11, 0.0873035), (0, 0.06457444), (10, 0.04478603)]

A.I. Ivaneshkin, A New Approach to Operating with Undirected Trees, Cybern Syst Anal 57(1) (2021) 124-132.
[(31, 0.2491094), (2, 0.22874023), (25, 0.21450539), (34, 0.09106176), (10, 0.060724407), (17, 0.05571771), (23, 0.051634632), (33, 0.024108535), (4, 0.01810252)]

For such publications, we can use the proposed reviewers, but with a more detailed analysis of the proposed candidates.

6. Discussion and Conclusion

The process of finding reviewers for a journal article or conference is always a rather complicated process. It is not always possible to understand by the scientific interests of the reviewer (by keywords), whether he will be able to review the paper qualitatively, whether it is in the area of his scientific interests.

The approach we proposed using the methods of topic modeling and natural language processing allowed us to obtain not only research topics, but also take into account possible conflicts of interest, thanks to information from metadata. The topic modeling model was built on the basis of Scopus data and tested on journal articles. The results showed that our approach works quickly and efficiently. The authors proposed by our method are relevant to the research topic.

In addition, reference metadata that reflects the content of the article includes "References". However, references can contain both positive and negative links. These metadata also contain many words (tokens) that cause white noise. The use of "References" metadata needs further analysis and is planned for future research. In the future, we plan to test and evaluate the performance of our method for a real journal with existing data about reviewers.

7. References

[1] Pro rozpodil vydatkiv derzhavnoho biudzhetu mizh zakladamy vyshchoi osvity na osnovi pokaznykiv yikh osvitnoi, naukovoi ta mizhnarodnoi diialnosti: Resolution of the Cabinet of Ministers of Ukraine dated 24.12.2019 No. 1146, 2019. URL: https://zakon.rada.gov.ua/laws/show/1146-2019-%D0%BF#Text.

[2] Pro naukovu i naukovo-tekhnichnu diialnist: Law of Ukraine, 2022. URL: https://zakon.rada.gov.ua/laws/show/848-19#Text.

[3] Pro zatverdzhennia Poriadku formuvannia Pereliku naukovykh fakhovykh vydan Ukrainy: Order of the Ministry of Education and Science of Ukraine dated January 15, 2018 № 32, 2018, 2021. URL: https://zakon.rada.gov.ua/laws/show/z0148-18#Text.

[4] Pro zatverdzhennia Litsenziinykh umov provadzhennia osvitnoi diialnosti: Resolution of the Cabinet of Ministers of Ukraine No. 1187, 2021. URL: https://zakon.rada.gov.ua/go/1187-2015-%D0%BF.

[5] R. Ghazavi, B. Taheri, H. Ashrafi-rizi, Article Quality Indicator: Proposing a New Indicator for Measuring Article Quality in Scopus and Web of Science, J. Sci. Res., 8(1) (2019) 09-17.

[6] P. Smeros, C. Castillo, K. Aberer, Scilens: Evaluating the quality of scientific news articles using social media and scientific literature indicators, in: The World Wide Web Conference, 2019, May, pp. 1747-1758.

[7] N. Salimi, Quality assessment of scientific outputs using the BWM, Scientometrics, 112(1) (2017) 195-213.

[8] New Scopus Article Metrics: A better way to benchmark articles, 2015. URL: https://blog.scopus.com/posts/new-scopus-article-metrics-a-better-way-to-benchmark-articles.

[9] Retraction guidelines, 2019. URL: https://publicationethics.org/retraction-guidelines.

[10] M. Thelwall, Artificial intelligence, automation and peer review, JISC, Bristol, 2019.

[11] Y. Hlavcheva, O. Kanishcheva, The Reference Analysis as a Quality Characteristic of a Scientific Article, in: CEUR Workshop Proceedings, 2020 PhD Symposium at ICT in Education, Research, and Industrial Applications, Kharkiv, 2020, pp. 7-18.

[12] Step by Step Guide to Reviewing a Manuscript, Wiley. URL: https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html.

[13] PLOS ONE: accelerating the publication of peer-reviewed science. URL: https://journals.plos.org/plosone/s/reviewer-guidelines.

[14] Report for Publons State of Peer Review 2018 Survey, 2018. URL: https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf.

[15] 'Friendly' reviewers rate grant applications more highly, 2019. URL: https://www.nature.com/articles/d41586-019-01198-3.

[16] Publons. URL: https://publons.com/about/home/.

[17] WOS И PUBLONS: Yzmerenye Nauky, 2020. URL: https://openscience.in.ua/publons.html.

[18] JA. Teixeira da Silva, A. Al-Khatib, The ClarivateTM Analytics acquisition of Publons – an evolution or commodification of peer review?, Research Ethics 15(3-4) (2019) 1-11. doi:10.1177/1747016117739941.

[19] S. Price, P. A. Flach, Computational support for academic peer review: a perspective from artificial intelligence, Communications of the ACM, 60(3) (2017) 70-79.

[20] Reviewer Finder, SpringerNature. URL: https://www.springernature.com/gp/editors/resources-tools/reviewerfinder.

[21] Artificial Intelligence to help meet global demand for high-quality, objective peer-review in publishing, 2020. URL: https://blog.frontiersin.org/2020/07/01/artificial-intelligence-peer-review-assistant-aira/.

[22] David M. Blei, Y. Ng Andrew, Michael I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 Jan (2003) 993-1022.

[23] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimedia Tools and Applications (2018). doi:10.1007/s11042-018-6894-4/.