UDC 004.9+519.816 https://doi.org/10.20535.2786-8729.5.2024/316521

THE ALGORITHM FOR SELECTING PUBLICATIONS ON A GIVEN TOPIC CONSIDERING KEYWORD PRIORITIES

Olha Suprun *
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0009-0006-9165-3446

Oksana Zhurakovska
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
https://orcid.org/0000-0002-2804-5556

*Corresponding author: olha.suprun.w@gmail.com

The article investigates the problems that exist in existing search engines for scientific publications. The search algorithms used in various search engines for scientific publications are described. The aim of the article is to develop a method for selecting publications on a given topic based on assessing the relevance of keyword sets. A review of the literature that was analyzed during the research is presented. Among the publications studied were materials related to the theory of set similarity, namely the use of the Jacquard coefficient and editing distance. A measure for determining the similarity of keyword sets is presented, which is based on the Jacquard coefficient taking into account the weighting coefficients of keywords. An algorithm is presented that can be used to determine the degree of similarity of publications to a user's search query based on keyword sets with weighting coefficients. The algorithm is based on the measure of similarity presented by us and the editing distance presented by us. The algorithm can be used to rank search results in search engines for scientific publications, as well as to compare the efficiency of different search engines, assess the quality of the results they return. The algorithm can also be used in book and film recommendation systems based on user preferences. The article provides the pseudocode of the algorithm. It is demonstrated on a limited data set how the measure calculated by the algorithm changes depending on the distribution of keyword weights in the user's query and the number of keywords.

Key words: search of scientific publications, similarity of sets, Jaccard criterion, edit distance.

1. Introduction

Searching for scientific publications takes a lot of time, forcing you to visit many different Internet resources in order to find the desired information. Often, when researching a topic, it is necessary to review several dozen. Therefore, the fact that the information system has the ability to form sets of publications according to specified query conditions is a function that will be in demand by users. Also, the researcher may need to take a break for a while and then return to the study, which is why he has to search for the same queries and review the same publications several times. The solution to this is the ability of the system to save the formed sets of publications for the user.

In addition, another problem for users in modern search engines is that the necessary publications are not always at the top of the search results. It happens due to the peculiarities of the work of search engines for scientific materials. Publications with more citations, higher authority of authors and published in more authoritative publications will be higher in the search results. In fact, the content may be less relevant to the user's query than the results placed below. All this forces the researcher to spend more time on research.

The article presents existing systems for searching scientific publications, highlights their features and proposes a method for improving search in such systems.

This article highlights the use of the Jacquard coefficient to assess the degree of correspondence of an article to a search query using a set of keywords, which can be used to rank search results and

increase search relevance. A modified coefficient based on the Jacquard coefficient is presented and takes into account the presence of keyword weights.

This coefficient is the basis of the algorithm used to assess the degree of correspondence of publications to a search query. The calculated scores are used to rank search results.

The article is devoted to the issue of searching and selecting publications according to the user's request. The issue of determining the degree of proximity of publications by topic, in particular by sets of keywords, taking into account their weight coefficients, is considered. This problem is especially relevant in the process of conducting scientific research at the stage of analyzing the state of the subject area. At this stage, it is very important to form a relevant sample of scientific research from the area under study to analyze existing solutions and identify unresolved issues. Analysis of literary sources showed that the proposed existing solutions for searching for relevant publications are still imperfect and the development of methods and algorithms that increase the efficiency of such a search remains relevant.

2. Literature review and problem statement

Currently, there are many search engines that allow you to find scientific literature: Google Scholar, CORE (an aggregator of open access documents), BASE, Arxiv, and others. Some of them have an open application programming interface (API) that allows you to search and obtain metadata about a publication. Others provide a list of links to publications, and to obtain data from links to a page, you need to extract metadata using "scraping". Unfortunately, this method of obtaining metadata about a page is not always possible or a rather complicated and time-consuming process, since some sites prohibit such access.

In fact, the above projects use different means of collecting publications and information about them. Web crawling or "crawling" is an approach that is usually used by search engines when bot programs index pages on the network and collect information about them [1]. Downloading and parsing html of web pages, sites and semantic analysis of available information. Web scraping is similar to web crawling, but imitates human actions on the site, due to which it collects more information. All three of these methods are quite difficult to implement. The reason for this is that all pages have a different structure, different selectors, classes and their names, require semantic analysis, the use of machine learning. Therefore, they were not used to develop the algorithm in this article.

Another approach is to use APIs that provide access to metadata and publications in already collected repositories or journals, such as DOAJ API, CORE API, Arxiv API – the listed applications provide access to open data. Arxiv API has the largest number of resources among the listed, so it was used to test the algorithm.

Specifically, in this case, it was also necessary to extract keywords for the publication from its description, since the Arxiv API does not provide keywords in the search results in metadata.

The creation of search engines for the selection of scientific publications and their features have been described in the studies of other scientists.

For example, there is a study [2], aimed at improving search algorithms for scientific publications by applying the concept of "Entity Set Search" and an unsupervised ranking algorithm. The search in the presented algorithm is carried out not simply by sets of keywords, but by sets of related terms representing the subject of scientific research, which in the article are called entities. The algorithm does not take into account the frequency of occurrence of certain keywords or simply coincidences of terms and takes into account only the presence of entities, thereby increasing the relevance of the search. The presented algorithm does not require pre-labeled data to work. This approach is useful in the case when there is not enough annotated data in the database of scientific publications.

The publication [3] describes methods for optimizing search engines. The publication describes the difference between conventional search engines and academic search engines (search engines for scientific publications), the difference in indexing of conventional web pages and scientific publications. The article draws attention to the problem that not all scientific publications can be indexed because some of them are located in closed databases, to which the robot-workers performing

the indexing of publications do not have access. Work with such databases is carried out only by agreement. The article also describes the features of ranking search results and what structure, format and metadata publications should have for correct indexing, which will increase the chances of publications appearing in search results.

In the article [4], a method of searching and ranking results based on the number of citations in scientometric databases and the presence of links to a group of key articles related to the search query is described. In another article by the same author, a method of searching for publications based on citation links is proposed [5].

A similar study [6] examines the impact of the context of coherent citations and normalization by citation frequency on the efficiency of search methods.

Another publication describes the interaction between information retrieval and bibliometrics to improve searching in scientific databases [7]. It describes the use of bibliometric methods to select the most relevant publications, which can help improve the accuracy and efficiency of searching in the scientific literature.

The publication explores the effectiveness of using citation networks to search for scientific evidence [8].

In fact, search engines are not ideal, and there is even a publication that is dedicated to finding errors and limitations in 42 search engines for academic materials [9]. There is also a publication that is dedicated to evaluating the work of digital libraries [10].

Based on the analysis of publications and existing systems, it can be concluded that the problem of selecting publications on a given topic, taking into account the priorities of keywords, remains unsolved. The article is devoted to solving this problem, namely, selecting publications that are relevant to the user's query based on a set of keywords and their weighting factors.

3. The aim and objectives of the study

The purpose of the study: to develop a method for selecting publications on a given topic based on assessing the relevance of sets of keywords.

To achieve the goal, it is necessary to perform the following tasks:

– to determine a measure that allows assessing the degree of proximity of two publications by sets of keywords and weight coefficients of keywords;

– to develop an algorithm for selecting publications on a given topic.

4. The study materials and methods

4.1. A measure for assessing the degree of proximity of sets

In order to determine how well a search result matches what is desired, we use a formula that combines the Jaccard coefficient and the keywordi importance coefficients provided by the user.

The Jaccard coefficient is the ratio of the intersection of two sets to their union [11]. The purpose of this coefficient is to calculate the similarity measure between sets – the larger it is, the more similar the sets are considered:

$$J = \frac{|A \cap B|}{|A \cup B|}, \quad (1)$$

where A is the first set of elements, B is the second set of elements, and the straight brackets indicate the cardinality of the set.

The use of the Jacquard coefficient in set similarity theory is described in great detail in the publication [12].

This method is used to determine the similarity of texts, search for plagiarism, search for web page mirrors, and can also be used in a rather non-trivial way in determining ratings and collaborative filtering, for predicting purchases, etc. [12].

The use of other variations of the Jacquard coefficient is also described in other publications [13, 14].

4.2. Algorithm for determining the degree of similarity of publications by sets of keywords

To determine whether elements of keyword sets are similar to each other, we use the edit distance. This is a measure that shows the similarity of strings based on the number of insertions, deletions, and permutations to make the elements the same [15]. We need this measure because the elements of the sets we compare are strings that can have different forms but have the same meaning. Words can be singular or plural, have different cases or tenses. In this case, they will still be considered similar elements of the sets.

Here is the pseudocode of the method we use to calculate the edit distance:

```
Function editDistance(string a, string b) returns number:

Create an empty matrix named matrix
   For each i from 0 to the length of string b:
      Set matrix[i][0]= i
   For each j from 0 to the length of string a:
      Set matrix[0][j]= j
   For each i from 1 to the length of string b:
      For each j from 1 to the length of string a:
         If characters b[i-1]and a[j-1]are the same:
            Set matrix[i][j]= matrix[i-1][j-1]
         Else:
            Set matrix[i][j]= minimum of:
               - matrix[i-1][j-1]+ 1 (substitution)
               - matrix[i][j-1]+ 1 (insertion)
               - matrix[i-1][j]+ 1 (deletion)
   Distance = matrix[length of b][length of a]
   MaxLength = maximum(length of a, length of b)
   Return Distance / MaxLength.
```

The edit distance is normalized and falls within the range from 0 to 1. Overall, the algorithm scheme for calculating the degree of similarity of keyword sets will consist of two methods: a method to calculate the degree of similarity for each set of keywords representing the publications being searched and a second method for ranking the obtained results based on the calculated degree of similarity for each publication.

Steps for the algorithm to calculate the degree of similarity for all publications:

1. Obtain the set of keywords with weight coefficients from the user's search query.
2. Obtain the set of publications represented by keyword sets.
3. Iterate through all publications (let each publication in this array be called item).
4. For each item, perform the following actions:
   4.1 Initialize a variable to count the number of keywords that match between the item and the user query keywords.
   4.2 Initialize a variable to sum the weight coefficients of the matching keywords.
   4.3 For each pair of keywords from item and the user query, calculate the edit distance. If the distance is less than the predefined threshold, consider the keywords as matching and increment the variables initialized in steps 4.1–4.2.
5. Once all pairs of values from the keyword sets of item and the user query have been compared, calculate the degree of relevance of the publication to the user's query.
6. The result will be an array containing metadata about the publications and similarity scores for each of them based on the user query.

5. Results of investigating

5.1 Measure for calculating the proximity of a user's query publication

Let us introduce a notation to describe the measure.

$K = \{K_1, K_2, K_3, \ldots, K_n\}$ – a set of keywords by which we search, keywords are provided by the user in the search query.

$W = \{w_1, w_2, w_3, \ldots, w_n\}$ – a set of weight coefficients (importance coefficients) of the keywords we search for. The user can leave them the same or determine which keywords he considers more important in the search and which value of the coefficient should be increased. Each weight coefficient lies in the range from 0 to 1 and they are normalized – their sum is equal to 1.

The set of publications to search among $A = \{A_i\}, i = \overline{1, N}$, where A_i is a set of keywords for the publication.

It is proposed to calculate the degree of proximity of two publications represented by sets of keywords based on the Jaccard coefficient (1) taking into account the weight coefficients of keywords, which is reflected in the modified coefficient:

$$J'_i = \frac{\left|A_i \cap K\right| \cdot \sum\limits_{j \mid K_j \in K \cap A_i} w_j}{\left|A_i \cup K\right|}, \quad (2)$$

Where A_i – is the set of keywords of the publication, i – is the serial number of the publication, K – is the set of keywords from the user's query, w_j – is the weight coefficient of the keyword K_j.

In the numerator of formula (2) is the product of the power of the set denoting the intersection of the set of publication keywords and the set of query keywords, of the sum of the weight coefficients of those keywords from the user query that are included in this intersection. In the denominator is a number denoting the power of the set of the union of the sets of publication keywords and from the user search query.

5.2 Algorithm for assessing the degree of similarity of publications and ranking search results

For each result, a score is calculated using this formula, and the results are ranked – results with the highest relevance scores appear at the top of the publication overview.

In order for our algorithm to determine whether keywords are similar, we chose an edit distance of 0.4; if the distance between the word combinations is less than or equal to this value, we consider that the word combinations coincide and the sets of keywords have a similar element.

Here is pseudocode that displays the scoring algorithm for each position in the publication search results:

```
Create an empty list named newData.
   For each item in data:
      Set numberSimilar = 0.
      Set weights = 0.
      If item.keywords is not empty:
         For each keywordQuery in keywordsQuery:
            For each keyword in item.keywords:
               Calculate distance = editDistance(keyword, keywordQuery.keyword).
               If distance ≤ DISTANCE_SAME_KEYWORDS:
                  Increment numberSimilar by 1.
                  Increment weights by keywordQuery.priority.
                  Continue to the next iteration.
         Calculate value = (numberSimilar * weights) / (length of keywordsQuery + length of
item.keywords - numberSimilar).
         Add item with the calculated value to newData.
Return newData
```

To assess the degree of similarity in the algorithm, formula (2) is used, using which we calculate value.

The input data is the *data* array, which contains metadata for all publications, including sets of keywords, which are either defined by the author of the publications or selected from the annotation or text using the algorithm.

We use the *DISTANCE_SAME_KEYWORDS* value to assess whether a pair of keywords is the same. If the author defined the keywords in the metadata, then this value may be less than 0.3. If the keywords were extracted from the text using the algorithm, then they are more likely to be found in different temporal forms, in different times, in different plural forms. On such datasets, in order to return a larger number of results, you can increase the value of the *DISTANCE_SAME_KEYWORDS* constant to 0.4.

Array *keywordsQuery* is an array that we receive from the user from a search query. It contains objects that store keywords and their priority.

At the output, we get an array *newData* in which metadata about publications and estimates of the degree of similarity of each of them to the user's search query are stored.

This algorithm can be used both for ranking search results and for evaluating the performance of various search engines for scientific publications.

For example, let's take the IEEE scientific publication search engine. For a query with the keywords "mock testing", "integration testing", we received 94 results. 5 results did not have keywords, so we can evaluate 89 results using the algorithm.

The distribution of the obtained estimates indeed showed that more relevant results are at the beginning of the search results. In Figure №1, the graph shows the calculated relevance estimates according to the presented algorithm in a quantitative distribution by intervals. In the algorithm, we used weight coefficients of 0.5 for both keywords.

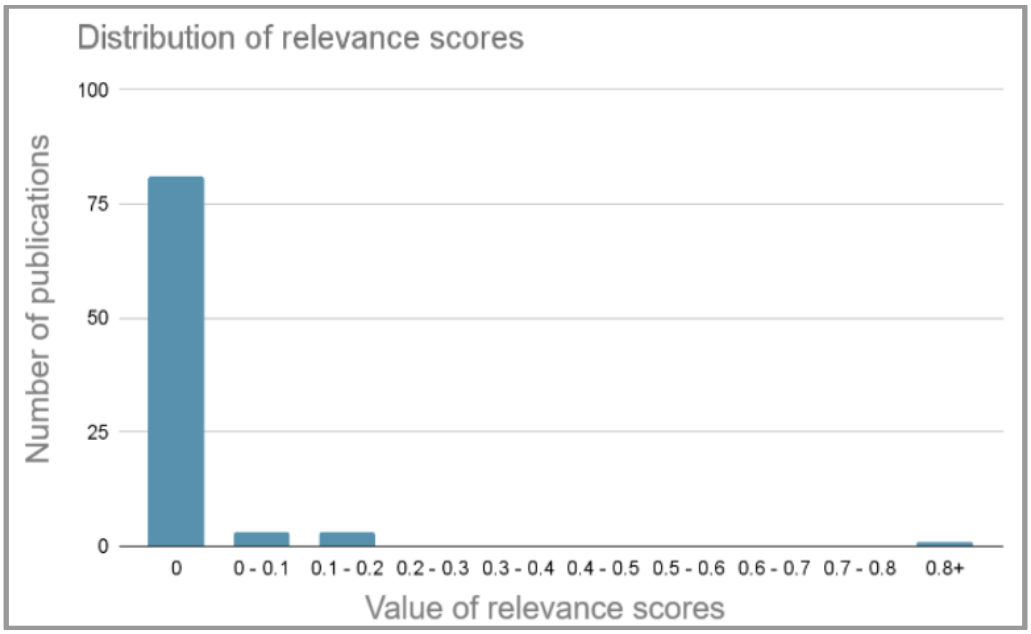

Fig.1. Distribution of relevance scores

Now, using the same search results, we will demonstrate how the distribution of relevance scores will look if we change the priorities of keywords in the search query.

We leave the set of publications the same, the set of keywords the same "mock testing", "integration testing". We set the importance coefficients to 0.6 and 0.4, respectively. As a result, we get the distribution as in Figure №2.

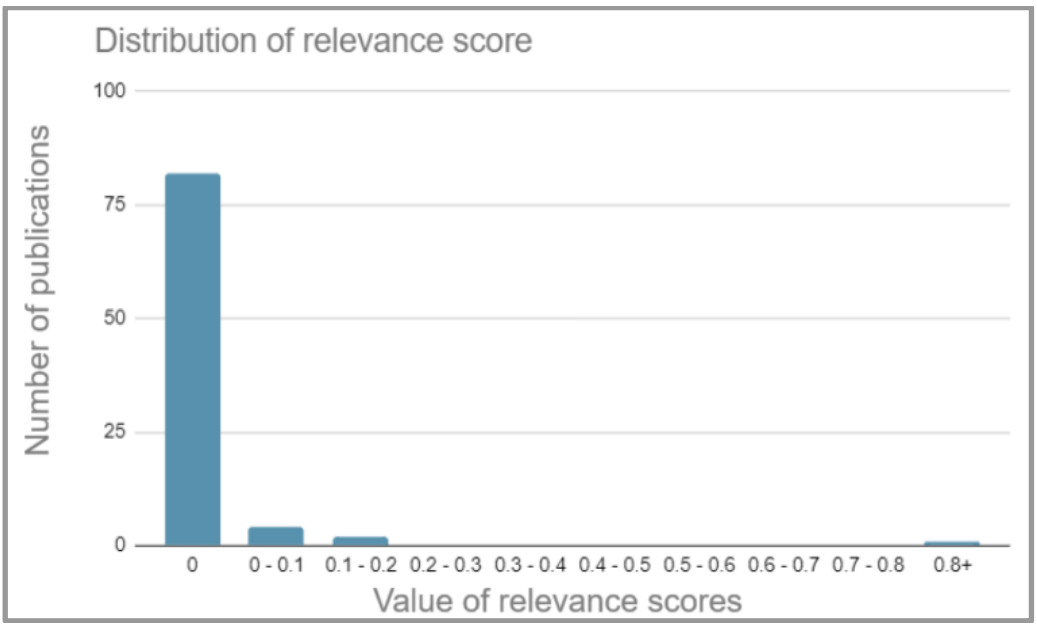

Fig. 2. Distribution of relevance scores

As we can see, the distribution of scores has changed from 0 to 0.2. Now let's try adding another keyword to the query. Now the set of keywords in the search query looks like this: "mock testing", "integration testing", "security". The importance coefficients are 0.333 for all keywords. The set of publications remains the same.

By adding a new keyword, the graph changed as follows, as shown in Figure №3. A value in the range 0.5–0.6 was added, a value in the range 0.8+ disappeared. The number of ratings in the range 0–0.1 increased.

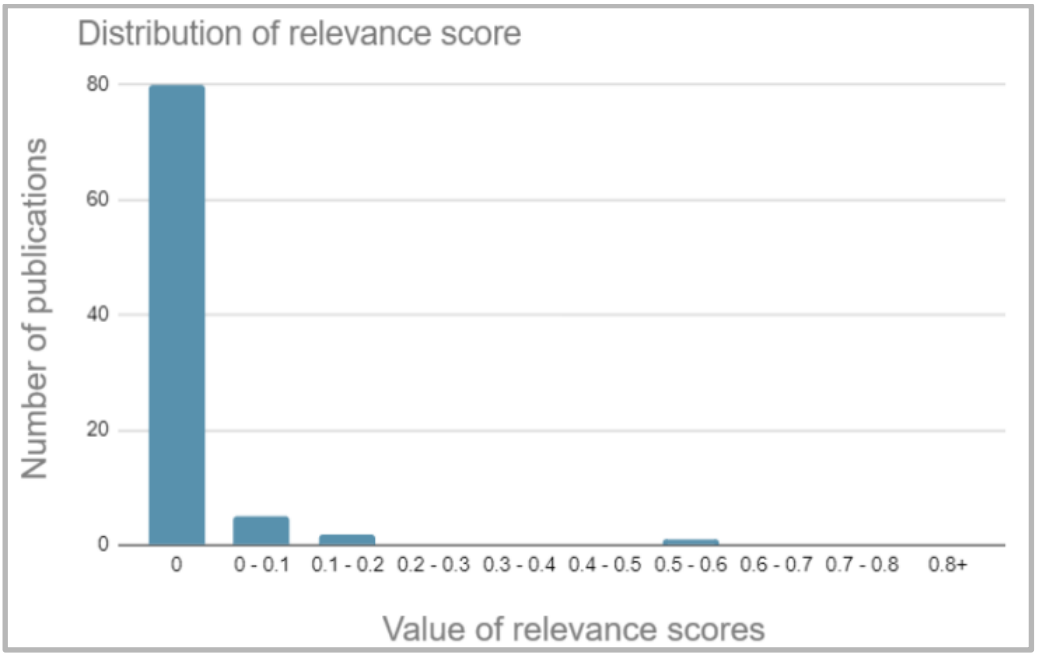

Fig. 3. Distribution of relevance scores

Now, with this same set of keywords in the search query and with this same set of publications, we will run the algorithm with keyword weights of 0.2, 0.2, 0.6, respectively.

Evaluating the results of the last query, shown in Figure №4.

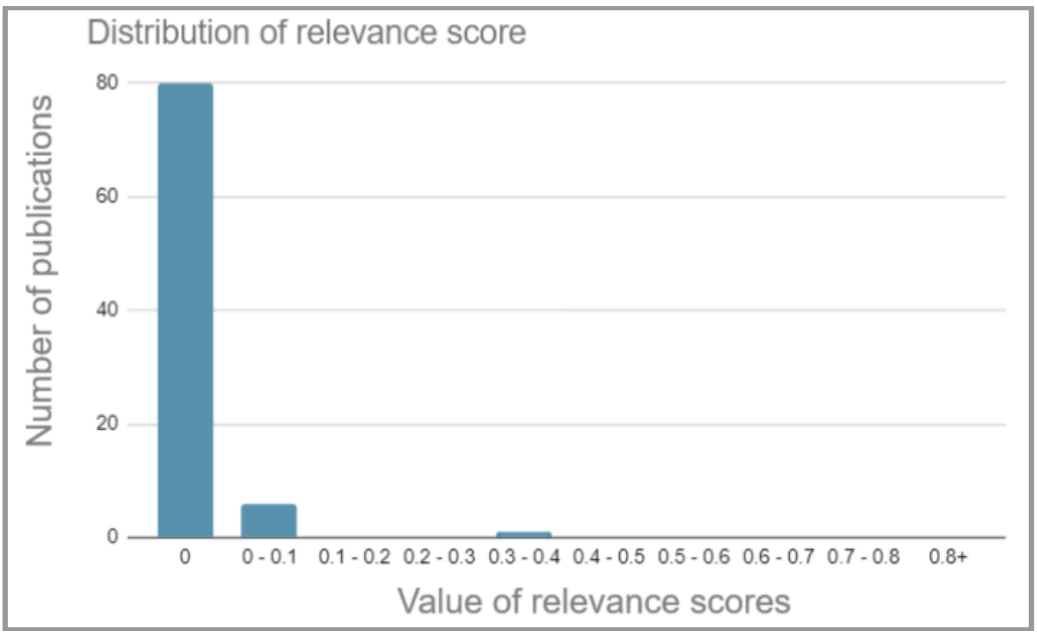

Fig. 4. Distribution of relevance scores

In Figure №4 shown that overall the values of the obtained estimates have decreased and their distribution has shifted to an interval with smaller values.

6. Discussion of results

6.1. Discussion of a measure for calculating the proximity of a user's query publication

As a result of the study, we were able to observe how adding weighting factors, in the modified Jaccard coefficients, affects the distribution of relevance scores of search results with sets of keywords.

Now we can conclude that this modification has an impact on the ranking of search results and will help to show at the very beginning those results in which the keywords were more important for the user.

This modification, taking into account the importance factors of keywords or elements of the set, can be used not only in creating search engines, but also for creating recommendation systems for films, books, news, etc.

Another way to use it can be to evaluate the quality control of filters in search engines.

6.2. Discussion of the algorithm for assessing the degree of similarity of publications and ranking search results

The problems encountered during development are as follows: not all APIs provide keywords for a publication in the metadata. This requires the creation of a mechanism that, if necessary, will extract keywords from the available information about the publication. The approach that was used for this in the development of the basic algorithm requires further improvement.

The developed algorithm shows the best results in cases where the sets of keywords are comparable. This case best demonstrates the goal of this algorithm: those publications that have keywords with higher importance coefficients rise to the highest positions in the overview – this means that the user is more likely to find a publication with the accents he needs.

If the user were to manually scroll through or even examine all these publications in detail, it would take him an extremely long time. This algorithm can be a simple solution for implementing a system that significantly saves time by returning a selection of the most relevant publications to the user for review.

The algorithm still needs improvement, but has already shown its effectiveness.

This algorithm allows us to solve the problem of returning irrelevant results. We can cut off some of the results for which the similarity measure is 0 or close to 0. In this way, we will generally improve the relevance of all search results, although this will reduce their number.

The algorithm that we presented in the article can be used as the basis of a search engine for scientific publications. We presented ideas for creating one of such systems in the abstracts of the report of the V International Scientific and Practical Conference of Young Scientists and Students "Software Engineering and Advanced Information Technologies SoftTech-2023 [16].

This algorithm can also be used to compare the efficiency of different search engines, assessing the relevance of the results to the search query based on a comparison of sets of keywords of the query results and the keywords of the query itself.

7. Conclusions

7.1. Defining a measure for assessing the degree of proximity of publications by sets of keywords and keyword weights

This publication describes a measure that can be used to assess the degree of similarity of sets of keywords. The method of calculating the measure is based on the Jaccard coefficient. The presented measure, described by formula (2), is its modification and takes into account the weight coefficients of keywords, which represent their priority for the user.

7.2. Development of an algorithm for selecting publications on a given topic

This article describes an algorithm for creating selections of scientific publications by given topics and keywords. In our subjective opinion, keywords are often underestimated as a search query in literature search, and this algorithm shows that this approach can be quite effective and convenient during the search. The algorithm uses a measure to assess the degree of proximity of publications by sets of keywords, which is described by formula (2).

The algorithm can be used to develop search engines for scientific publications and to compare the quality of search results of different search engines.

References

[1] M. Yadav and N. Goyal, "Comparison of Open Source Crawlers – A Review," Int. J. Sci. & Eng. Res., vol. 6, pp. 1544–1551, 2015, https://www.ijser.org/researchpaper/Comparison-of-Open-Source-Crawlers--A-Review.pd.

[2] J. Shen, J. Xiao, X. He, J. Shang, S. Sinha, and J. Han, "Entity Set Search of Scientific Literature: An Unsupervised Ranking Approach," Proc. 41st Int. ACM SIGIR Conf. Research & Development in Information Retrieval, Association for Computing Machinery, Ann Arbor, MI, USA, pp. 565–574, 2018, https://doi.org/10.1145/3209978.3210055.

[3] J. Beel, B. Gipp, and E. Eilde, "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.," J. Scholarly Publ., vol. 41, no. 2, pp. 176–190, 2010, https://doi.org/10.1353/scp.0.0082.

[4] C.W. Belter, "A relevance ranking method for citation-based search results," Scientometrics, vol. 112, pp. 731–746, 2017, https://doi.org/10.1007/s11192-017-2406-y.

[5] C.W. Belter, "Citation analysis as a literature search method for systematic reviews," J. Assn. Inf. Sci. Tech., vol. 67, pp. 2766–2777, 2016, https://doi.org/10.1002/asi.23605.

[6] O.V. Mazurets, O.V. Kozenko, M.O. Molchanova, and O.V. Sobko, Using cosine similarity metrics and Jaccard index for intellectual analysis of semantic similarity of text documents. Collection of scientific papers based on the materials of the XV All-Ukrainian scientific and practical conference "Actual problems of computer sciences APKN-2023". Khmelnytskyi, pp. 146–147, 2023.

[7] E. Ristad and P. Yianilos, "Learning String Edit Distance", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp. 522–532, 1998. https://doi.org/10.1109/34.682181.

[8] J. Leskovec, A. Rajaraman, and J.D. Ullman, Mining of Massive Datasets, 2nd ed., Cambridge: Cambridge University Press, pp. 92–98, 2014, https://doi.org/10.1017/CBO9781139924801.
[9] Z. Li and A. Rainer, "Academic Search Engines: Constraints, Bugs, and Recommendation," pp. 1–8, 2022, https://doi.org/10.48550/arXiv.2211.00361.
[10] E. Kelly, "Assessment of Digitized Library and Archives Materials: A Literature Review," pp. 1–34, 2016, https://doi.org/10.6084/M9.FIGSHARE.3206038.
[11] S. Varma, S. Shivam, A. Thumu, A. Bhushanam, and D. Sarkar, "Jaccard Based Similarity Index in Graphs: A Multi-Hop Approach," 2022 IEEE Delhi Section Conf. (DELCON), New Delhi, India, 2022, pp. 1–4, https://doi.org/10.1109/DELCON54057.2022.9753316.
[12] M. Eto, "Evaluations of context-based co-citation searching," Scientometrics, vol. 94, pp. 651–673, 2013, https://doi.org/10.1007/s11192-012-0756-z.
[13] P. Mayr and A. Scharnhorst, "Scientometrics and information retrieval: weak-links revitalized," Scientometrics, vol. 102, pp. 2193–2199, 2015, https://doi.org/10.1007/s11192-014-1484-3.
[14] K.A. Robinson, A.G. Dunn, G. Tsafnat, and P. Glasziou, "Citation networks of related trials are often disconnected: Implications for bidirectional citation searches," J. Clin. Epidemiol., vol. 67, no. 7, pp. 793–799, 2014, https://doi.org/10.1016/j.jclinepi.2013.11.015.
[15] J. Santisteban and J. Tejada-Cárcamo, "Unilateral Weighted Jaccard Coefficient for NLP," 2015 Fourteenth Mexican Int. Conf. on Artificial Intelligence (MICAI), Cuernavaca, Mexico, 2015, pp. 14–20, https://doi.org/10.1109/MICAI.2015.9.
[16] O.V. Suprun, O.S. Zhurakovska, "V International Scientific and Practical Conference of Young Scientists and Students 'Software Engineering and Advanced Information Technologies SoftTech-2023. INFORMATION SYSTEM FOR SEARCHING SCIENTIFIC PUBLICATIONS," pp. 310–313, Dec. 19–21, 2023.

УДК 004.9+519.816

АЛГОРИТМ ПІДБОРУ ПУБЛІКАЦІЙ ЗА ЗАДАНОЮ ТЕМАТИКОЮ ІЗ УРАХУВАННЯМ ПРІОРИТЕТІВ КЛЮЧОВИХ СЛІВ

Ольга Супрун
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0009-0006-9165-3446

Оксана Жураковська
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
https://orcid.org/0000-0002-2804-5556

В статті досліджено проблеми пошукових систем наукових публікацій. Описано алгоритми пошуку, які використовуються у пошукових системах наукових публікацій. Мета статті полягає в розробці методу підбору публікацій за заданою тематикою на основі оцінки подібності множин ключових слів. Викладено огляд літератури проаналізований під час виконання дослідження. Серед досліджених публікацій були матеріали, що стосувалися використання коефіцієнту Жаккарда та відстані редагування. Представлено міру для визначення подібності множин ключових слів, що базується на коефіцієнті Жаккарда з урахуванням вагових коефіцієнтів ключових слів. Представлено алгоритм, що може бути використаний для визначення ступеню подібності публікацій пошуковому запиту користувача на основі множин ключових слів з ваговими коефіцієнтами. В основі алгоритму лежать представлена нами міра та відстань редагування. Алгоритм може бути використаний для ранжування результатів пошуку у пошукових системах наукових публікацій, а також для порівняння ефективності роботи різних пошукових систем, оцінки якості результатів, що вони повертають. В статті наведено псевдокод алгоритму. Продемонстровано на обмеженому наборі даних як змінюється підрахована алгоритмом міра в залежності від розподілу вагових коефіцієнтів ключових слів та в залежності від кількості ключових слів.

Ключові слова: міра подібності множин, відстань редагування, теорія подібності множин, коефіцієнт Жаккара, система підбору наукових публікацій.