Turkish Journal of Electrical Engineering and Computer Sciences

Volume 28 | Number 3 Article 22

1-1-2020

Combining metadata and co-citations for recommending related papers

SHAHBAZ AHMAD

MUHAMMAD TANVIR AFZAL

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 1519 – 1534

doi:10.3906/elk-1908-19

Combining metadata and co-citations for recommending related papers

Shahbaz AHMAD*, **Muhammad Tanvir AFZAL**
Department of Computer Science, Faculty of Computing,
Capital University of Science and Technology, Islamabad, Pakistan

Received: 05.08.2019 • **Accepted/Published Online:** 17.12.2019 • **Final Version:** 08.05.2020

Abstract: Identification of relevant documents is performed to keep track of the state-of-the-art methods and relies on research paper recommender systems. The proposed approaches for these systems can be classified into categories like content-based, collaborative filtering-based, and bibliographic information-based approaches. The content-based approaches exploit the full text of articles and provide more promising results than other approaches. However, most content is not freely available because of subscription requirements. Therefore, the scope of content-based approaches is limited. In such scenarios, the best possible alternative could be the exploitation of other openly available resources. Therefore, this research explores the possible use of metadata and bibliographic information to find related articles. The approach incorporates metadata with co-citations to find and rank related articles against a query paper. The similarity score of metadata fields is calculated and combined with co-citations. The proposed approach is evaluated on a newly constructed dataset of 5116 articles. The benchmark ranking against each co-cited document set is established by applying Jensen–Shannon divergence (JSD) and results are evaluated with the state-of-the-art content-based approach in terms of normalized discounted cumulative gain (NDCG). The state-of-the-art content-based approach achieved an NDCG score of 0.86 while the traditional co-citation-based approach scored 0.72. The presented method achieved NDCG scores of 0.73, 0.77, and 0.78 by incorporating the title, co-citation and title, and abstract, respectively, whereas the highest NDCG score of 0.77 was achieved by combining co-citations with metadata. However, better results are achieved by incorporating the title and abstract with NDCG score of 0.81. Therefore, it can be concluded that the proposed approach could be a better alternative in cases where content is unavailable.

Key words: Co-citation, cosine similarity, Jensen-Shannon Divergence, metadata relatedness, research paper recommender systems

1. Introduction

The immense growth of scientific literature on the web has hindered the process of finding relevant information. The scholarly community searches for relevant research documents for numerous purposes, such as to be aware of the trends in state-of-the-art work, to perform a comprehensive analysis of a particular topic, to write survey papers, and to start new research in a particular domain. However, the task has become tedious due to information overload and excess of scientific publications. Currently, the paradigm of relevant research papers' identification is dominant by research paper recommendation systems [1, 2] and incorporates artificial intelligence [3, 4]. However, when a user poses a query to well-known paper recommendation systems, millions of search results are provided to the user including relevant and irrelevant results. The user has to go through the endeavor

*Correspondence: shehbazsahi@live.com

of filtering out the most relevant results, which requires excessive human effort and there is a high probability that significantly related research articles are missed in this process. Therefore, the scientific community has been focusing on improving research paper recommendation systems in the past 17 years [5]. During this period, approximately 220 research articles and patents were published, suggesting over 80 approaches for building research paper recommender systems [5].

The existing approaches to finding relevant research papers can be classified into content-based, collaborative-filtering-based, stereotyping, bibliographic information-based, graph-based, global relevance, and hybrid approaches [5, 6]. Among these approaches, content-based approaches are widely used as 55% of them are based upon content. The motivation behind the content-based approaches is full-text as it holds the richness of features. Usually, the content contains details about motivation, methods, datasets, results, sections, citation context, multiple mentions of a citation, etc. Therefore, full-text options serve as a fundamental element of these approaches and are often provided in the form of key terms [7–10]. Key terms are words or phrases that are assumed to be more meaningful. These key terms serve as feature vectors and different mathematical and statistical measures are applied to find the (dis)similarity between feature vectors, which are further used to find similarity or dissimilarity between articles. Different models were proposed to extract key terms from papers and, among them, term frequency-inverse document frequency (TF-IDF) is most widely used [5]. Although the content-based approaches are widely used, the lack of full-text availability limits the applicability of these approaches. Renowned publishers like IEEE, ACM, Springer, and Elsevier require subscriptions to access the full-text of research articles, thus limiting open access to these research articles.

Content-independent approaches (such as collaborative filtering [11–13], graph-based [14], and hybrid approaches [7, 8, 15]) are also proposed in recommender systems. Collaborative filtering is based on the idea of "like-minded people like/dislike the same things". It creates a group of users with the same interests and recommendations are provided based on the preferences of peers. The core of collaborative filtering is the "user" as recommendations to a user are based on the preferences of other users of the same group. As a result of relying on users, this approach severely suffers from issues like cold start, sparsity, large computing time, scalability, and user intervention. The other approaches such as graph-based approaches build a connection between articles and articles with strong connections are assumed to be relevant and are recommended by the system. These connections are represented in the form of graphs and graph mining techniques are applied to excerpt the strong connection between nodes of the graph. The graph is built based on diverse relationships, such as citations, authors, co-authors, venues, and year of publication. This approach also significantly suffers from the coverage problem as only a fraction of the documents is cited. Therefore, it cannot be extensively used to find relevant papers. Some researchers have also focused on combining the best features of existing approaches; however, the problem is still there because the primary approach still dominates, and all the limitations of the primary approach are limitations of the new approach, as well.

Bibliographic information-based approaches are other oldest and widely used category of research paper recommender systems. These approaches build a similarity relationship between articles based on the citation [16, 17]. In general, the citation is assumed as the strongest relationship between citing and cited articles as this relationship is built by the author of the paper. These approaches assume that each "count" between citing and cited articles, whether using a direct citation, bibliographic coupling [16], or co-citation [17], has the same weight. Among bibliographic-based approaches, co-citation is the oldest and most widely used approach. State-of-the-art approaches incorporate content with co-citation and suggest that the actual similarity between co-cited documents is related to the proximity of citations in the text [7, 8, 18, 19]. However, the problem

with state-of-the-art approaches again arises when the content is not available as the main ingredient for these approaches is missing.

In scenarios wherein content is unavailable, the metadata of research articles can be incorporated with the co-citation. Metadata may include different fields like title, abstract authors, author-defined keywords, and publisher. Among these metadata fields, title and abstract are vital as these fields concisely summarize the research articles and contain a healthy proportion of key terms. Therefore, this research explores the possible use of metadata with co-citations for research paper recommender systems by presenting a novel approach that combines metadata similarity with co-citation scores to identify relevant documents.

This research incorporates metadata similarity with co-citation scores. This study utilizes the title and abstract from the metadata and the similarity scores of these metadata fields between each query and co-cited papers are calculated. The similarity score is calculated using a cosine similarity after performing standard preprocessing techniques. These scores are further combined with the co-citation score and we evaluate the effect of different combinations on the ranking of search results. For experimentation, 10 sets of co-cited articles are retrieved from CiteSeerX [20]. The co-citation strength for each pair of co-cited articles is calculated by retrieving the articles that have co-cited this pair of articles. This results in a dataset of 1616 distinct research articles, including 10 query papers, 90 co-cited articles, and 5116 citing articles. The metadata of all the queries and co-cited papers are extracted, and cosine similarity measures are applied to extract metadata to find the similarity between each query and co-cited article. Metadata similarity scores and co-citation strengths are further used to establish the ranking between each set of co-cited articles.

To evaluate the results, the gold-standard ranking for each set of co-cited articles is established by applying JSD on the content of the query and co-cited articles. The results are evaluated using the gold-standard dataset and employing the NDCG measure. The effectiveness of the proposed approach is compared with traditional co-citation and content-based approaches. The evaluation results show significant improvement in NDCG scores as compared to the scores attained by traditional co-citation.

2. Literature review

The most prominent approaches devised for research paper recommender systems can be broadly categorized into three broad classes: 1) content-based, 2) collaborative filtering, and 3) bibliographic information-based. These approaches are discussed in detail in the subsequent subsections.

2.1. Content-based approaches

Content-based approaches are the most widely used approaches when it comes to identifying relevant documents [5]. These approaches exploit the advantage of the full text of research articles as they may contain vital information regarding the research articles. They may contain information about datasets, results, adopted methodology, citation context, citation sentiments, sections, authors, author affiliations, venues, etc. Typically, words act as features and are often termed as key terms.

The extracted key terms are normally represented as feature vectors; however, some researchers also represented them as a graph for further processing. Typically, these feature vectors are provided for different statistical approaches such as Euclidean distance, Tanimoto coefficient, Jaccard index, Kullback-–Leibler divergence, and cosine to find the (dis)similarity between two research articles [21]. Among these approaches, cosine similarity is the most accepted and applied approach [5]. Some researchers extracted key terms from the title, while others extracted key terms from abstracts. The author-defined keywords in a research article are also

analyzed by researchers. Most of the researchers focused on extracting the key terms from whole body content [22], while some have analyzed the context of citations and sentiments.

Although content-based approaches are widely used and content carries essential information, the content is not always available. Various renowned publishers such as IEEE, ACM, Springer, and Elsevier require a subscription to access full-text documents. Therefore, these approaches cannot be applied if the content is unavailable. Moreover, these approaches suffer from sparse results. Huge vocabulary, synonyms, elusive nomenclature, and noisy PDF parsing are the main factors in the sparsity of the results. These approaches are also computationally expensive to implement as it requires many nontrivial subtasks such as PDF parsing, key term identification, and score calculations.

2.2. Collaborative filtering-based approaches

Collaborative filtering (CF)-based recommender systems predict user affinity by matching the user's interests to people of the same group. Recommendations are provided to a user by analyzing the ratings of the people of his/her group [23]. The core of CF-based systems is users. These systems do not rely on the content of research articles; instead, they rely on user interests and preferences. Thus, CF-based systems are content-independent systems.

Researchers followed different techniques based on CF to build research paper recommendation systems. The work in [23] allowed users to rate research articles, but users were too lazy to rate the articles. Afterward, the authors of [24] also tried to obtain ratings from users but faced the same problem, so they generated automated ratings for evaluation purposes. This highlights the inherent problem of CF systems, the user intervention. The users are the core of these systems and a healthy user contribution is required, but users are normally less active. This problem is known as the cold-start problem [25] and it can occur in the case of new users or new articles. If a new user rates no or few articles, systems are unable to find like-minded researchers. The same is the case with a new research article when a new article has no or few likes and it cannot be recommended. Some researchers such as the authors of [24] addressed this problem by gathering implicit ratings from user interactions. Researchers inferred ratings from the number of pages he/she read, the user's profile, bibliography, download patterns, and citation patterns.

There are a number of limitations and critiques of CF systems [26, 27]. The major critique these systems get is about autoinferring the user rating, which voids the core of CF systems. Another general problem faced by these systems is sparsity, as the number of research articles is very high and users are less in number. Thus, in this case, finding like-minded users is a difficult task. These systems also suffer from large computation times as they are generally less scalable and require more offline data processing.

2.3. Bibliographic-based approaches

Bibliographic information-based approaches use the citations provided in each document to analyze the relatedness or relationships between documents. Historically, researchers were solely relying on the little information provided in the citation. Some early used approaches were 'cited by' and 'reference list' [18]. 'Cited by' approaches consider two documents as relevant if they cite the same input document whereas reference list-based approaches consider two papers as relevant if they have been referred to by the input paper. In bibliographic-based approaches, best results are produced by bibliographic coupling and co-citation [19].

The prominent and oldest approach that used bibliographic information was proposed by Kessler, named bibliographic coupling [16]. Afterwards, Small and Kessler independently proposed a protuberant approach

called co-citation [17]. Co-citation analysis has most often been used to study the structure of science from the perspective of cited publications. Two papers are considered to have been co-cited if both are cited by one or more publications. The more the pair of papers has been co-cited, the more related they are expected to be. The concept of co-citation has been introduced in other knowledge domains such as co-cited authors and journals as well.

Citation analysis and content-based heuristics were also combined to recommend more related papers. In these models, researchers exploited the behavior of citation within the text [8, 15, 19]. Researchers analyzed the context of citations [18], number of repeated citations [28], and citations within the sections [29]. The majority of these approaches have investigated co-citation proximity [8, 15], while some of the approaches are based on bibliographic coupling.

However, the inherent problem of content-based approaches of full-text unavailability is also a problem of these approaches. These approaches cannot be applied in the case of content unavailability. In such scenarios, other content-like features must be explored that might be combined with co-citation to compete for the results of content-based approaches. One possible alternative of content-based features could be metadata.

3. Methodology

This section encompasses details about the proposed methodology. The main modules of the proposed methodology are dataset collection, building benchmark ranking, metadata extraction, metadata relevancy score calculation, a ranking of articles based on metadata relevancy scores, and evaluation of results. Figure 1 shows an overview of the adopted methodology. Subsequent subsections provide a detailed description of these modules.

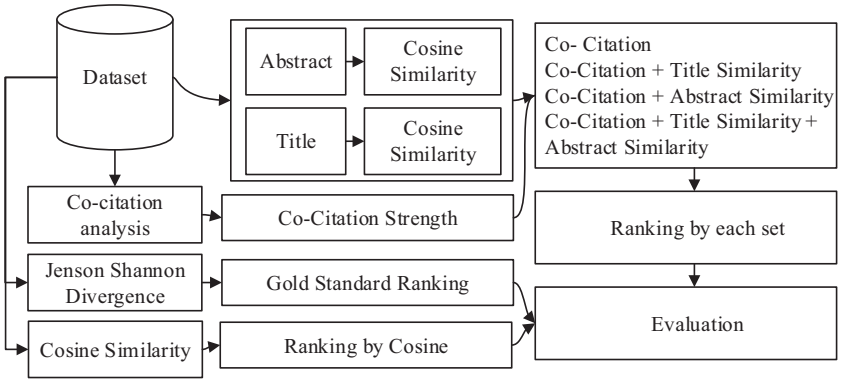

Figure 1. Overview of proposed methodology. The key components of the proposed methodology are dataset and gold-standard creation, metadata relevancy score calculations, and result evaluations.

3.1. Dataset

This research uses a dataset collected from CiteSeerX [20]. This research uses ten query papers to find co-cited documents. A total of 90 co-cited documents are found against ten query papers as CiteSeerX provides nine co-cited papers against a query paper. To attain the co-citation strength and verify and rank the co-cited papers against each query paper, the referenced list of each query paper is parsed. Citation identification and parsing is a challenging task due to different styles of references and citation anchors. However, the state-of-the-art approach proposed by Ahmad et al. [30] could effectively identify citation anchors and parse the references with a high F1 score. Therefore, we have employed this approach to identify the citations and parse the references.

Furthermore, we manually verified the random references and completed the missing information (if any) for the precise evaluation of our proposed approach. By applying this approach to the query papers, we were able to identify the 1516 distinct citing documents for each co-cited paper. Reference identification and extraction were further applied to these extracted co-cited documents to determine the strength of each co-cited article. These 1516 distinct citing documents cumulatively co-cited the papers 5116 times. The table 1 summarizes the statistics of the used dataset.

Table 1. Brief statistics of dataset.

Attribute	**Quantity**
Total distinct documents	1616
Query paper	10
Co-cited paper	90
Total number of citing documents	5116
Percentage of papers with recoverable abstracts	96
Percentage of papers with recoverable content	93

After retrieving the co-cited and citing documents, the full-text of query and co-cited documents are downloaded to establish a gold-standard ranking and apply content-based approaches. The content of 93 percent of the documents is successfully retrieved. The rest of the documents were either in the form of corrupted PDF or missing ones, and 95 percent of downloaded documents contain the abstract as the abstract of corrupted PDFs was manually typed into a database. After retrieving the PDFs, the full text, title, and abstract are extracted from the PDF by parsing, which is then stored in a database for further processing. To perform the task of PDF parsing and metadata extraction we used PDFx [31], which is a rule-based fully automated online-available tool that is capable of reconstructing the logical sections of a research article. PDFx inputs a PDF document and outputs the XML format with each logical section identified by XML mark-ups. We parsed this XML to extract the title, abstract, and full text of the research articles. This full text and the metadata are further used to establish a gold standard (explained in Section 3.7) and for evaluation of the proposed approaches. The overall process of dataset collection is shown in Figure 2.

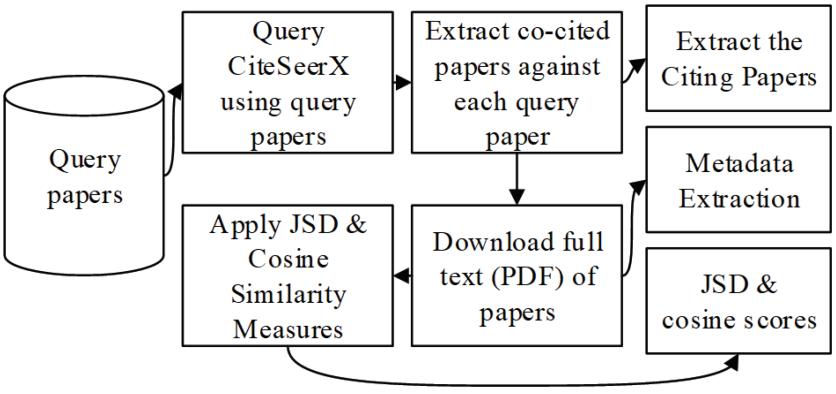

Figure 2. Dataset collection process. Key steps involved are querying the CiteSeerX using query papers, extracting co-cited papers along with their full-text, metadata extraction, and benchmark creation using JSD.

3.2. Metadata similarity score calculation

Metadata of research articles include title, abstract, author list, and keywords. The title and abstract of the articles present the most precise and descriptive summary of the entire research article. Therefore, we used the title and abstract to extract the key terms [32] that are further used to calculate cosine similarity [33] between documents. We ignored the author list because the co-cited documents are rarely authored by the same author(s) and do not significantly contribute towards finding the similarity.

3.2.1. Cosine similarity score calculation

Cosine similarity is the most widely used metric to find the similarity between documents [21, 34]. Mathematically, it measures the cosine of the angle between two nonzero vectors projected in a multidimensional space [21]. In the context of this research, the two vectors are arrays containing the key terms of the query and co-cited paper. Equation 1 depicts the mathematical form to calculate the cosine similarity between the query paper ($\overrightarrow{Q}$) and co-cited paper ($\overrightarrow{C}$). In this equation $\overrightarrow{S_Q}$ represents the array of key terms of $\overrightarrow{Q}$ and $\overrightarrow{S_C}$ depicts the array containing key terms of $\overrightarrow{C}$.

$$CS\left(\vec{Q},\vec{C}\right) = \frac{\vec{S_Q} \ . \ \vec{S_C}}{\| \vec{S_Q} \| \| \vec{S_C} \|} \tag{1}$$

Identification of key terms is the initial task to measure the cosine similarity. For this, we employ the frequently used approach TF-IDF for identification of key terms [5, 34]. Before applying TF-IDF, standard text preprocessing is performed. The preprocessing involved tokenization, stop word removal, and stemming. For stop word removal, a standard set of stop words provided by the NLTK toolkit is used. Afterwards, the Porter stemming algorithm is used to convert the terms into their roots using the NLTK toolkit [35]. After preprocessing, TF-IDF is used to extract the key terms. The top 30% of key terms of each document are used to construct the feature vectors to calculate cosine similarity.

3.3. Combination of metadata fields

The similarity scores for the title and abstract are calculated separately (as described in Section 3.2.1). After calculating separate scores for the title and abstract, the combined effect of these fields of data is also analyzed. Equation 2 depicts the mathematical way to combine the scores. The combined similarity score ($\text{Similarity}_{t+a\ ,\ (Q,C)}$) for title and abstract for query paper (Q) and co-cited paper (C) is calculated by averaging the cosine score of the title ($\text{CS}_{T,\ (Q,C)}$) and cosine score of the abstract ($\text{CS}_{A,\ (Q,C)}$) of Q and C. Using this similarity score ($\text{Similarity}_{t+a\ ,\ (Q,C)}$) the rankings are also established and analyzed.

$$Similarity_{t+a,(Q,C)} = \frac{CS_{T,(Q,C)} + CS_{A,(Q,C)}}{2} \tag{2}$$

3.4. Co-citation analysis

Against a query paper, 9 co-cited documents are retrieved. Furthermore, citing documents for each pair of co-cited documents are also retrieved from CiteSeerX. Citing documents are the research articles that co-cite the pair of co-cited documents [17]. Using the citing documents, the strength of co-cited pairs is calculated. In this dataset, the strength of co-citation ranges from 1 to 169. Similarity scores of the cosine similarity between title and abstract range from 0 to 1. Therefore, to combine the different parameters like co-citation

cosine similarities, the values must be on a common scale to eliminate the dominating factor. Therefore, the co-citation scores must be normalized from 0 to 1. There are different approaches available for this normalization purpose, like decimal scaling, min-max, and z-score. Z-score normalization cannot be effectively applied in this situation because the data are not normally distributed, whereas applying decimal scaling is also not suitable because the difference is minimum and the maximum value is not very high. Therefore, in this situation the better available and selected normalization technique is min-max normalization that could effectively scale the co-citation score between 0 to 1. The mathematical formula for min-max calculation is shown in Equation 3. In Equation 3, normalized co-citations strength ($\mathrm{ncs}_{s,\ d}$) of document D in set S is the fraction of the co-citation strength of document D in set S minus the minimum value of co-citation strength ($\min_s$) in set S over maximum co-citation strength ($\max_s$) minus minimum co-citation strength in set S.

$$\mathrm{NCS}_{\mathrm{S},\ \mathrm{D}} = \frac{\mathrm{CS}_{\mathrm{S},\ \mathrm{D}} - \mathrm{MIN}_{\mathrm{S}}}{MAX_S - MIN_S} \tag{3}$$

3.5. Combining co-citation and metadata

The individual scores for co-citation and metadata are further combined to test the effectiveness of metadata fields and co-citation. Three combinations, 1) co-citation and title, 2) co-citation and abstract, and 3) co-citation and title and abstract, are evaluated. Equations 4, 5, and 6 respectively denote the mathematical form to form these combinations. The similarity score for co-citation and title ($\mathrm{Similarity}_{CC+T\ ,\ (Q,\ C)}$) for query paper Q and co-cited paper C is calculated by averaging the co-citation score ($\mathrm{CC}_{(Q,C)}$) and cosine score ($\mathrm{CS}_{T,(Q,C)}$) for Q and C. Similarly, the average scores of CC ($\mathrm{CC}_{(Q,C)}$) and cosine similarity ($\mathrm{CS}_{A,(Q,C)}$) denote the similarity scores of co-citation and abstract $\mathrm{similarity}_{CC+A,(Q,C)}$ for papers Q and C. The average of CC $\mathrm{CC}_{(Q,C)}$, title$\mathrm{CS}_{T,(Q,C)}$, and abstract$\mathrm{CS}_{A,(Q,C)}$ scores provides the similarity score for CC, title, and abstract $\mathrm{similarity}_{CC+T+A\ ,\ (Q,\ C)}$ for co-cited paper C and query paper Q.

$$S_{CC+T\ ,\ (Q,\ C)} = \frac{CC_{\ (Q,C)} + CS_{T,(Q,C)}}{2} \tag{4}$$

$$S_{CC+A,(Q,C)} = \frac{CC_{(Q,C)} + CS_{A,(Q,C)}}{2} \tag{5}$$

$$S_{CC+T+A,(Q,C)} = \frac{CC_{T,(Q,C)} + CS_{T,(Q,C)} + CS_{A,(Q,C)}}{3} \tag{6}$$

3.6. Evaluation metric

The results are evaluated by employing normalized discounted cumulative gain (NDCG) [36]. The mathematical formula to calculate the NDCG is given in Equation 7. The NDCG for a set of co-cited documents S (NDCG_s) is evaluated as the fraction of discounted cumulative gain for set S (DCG_s) and ideal discounted cumulative gain (IDCG_s) for set S.

To calculate the DCG, documents in each set of co-cited documents are ranked according to their similarity scores. For ranking, we used "ordinal" ranking, which provides a distinct rank to each document in a set of co-cited documents. In each set of co-cited documents, these rankings are from 1 to 9 (as we have 9 documents in each set of co-cited documents), where 9 means highest rank and 1 means the lowest rank. Using these ranks, the DCG for each set of co-cited documents was calculated. The DCG (DCG_s) for a set of co-cited documents

S is calculated using Equation 8. In this equation, $\text{rank}_{S,\ i}$ shows the rank of the ith document in set S. IDCG is the DCG of ideal ranking. In our case, ideal ranking is the gold-standard ranking that is achieved by JSD (explained in Section 3.7). Using the gold-standard ranking scores for each set of co-cited documents, DCG is calculated, which is referred to as IDCG.

$$NDCG_s = \frac{DCG_s}{IDCG_s} \tag{7}$$

$$DCG_s = rank_{s,1} + \sum_{i=1}^{9} \frac{rank_{s,i}}{log_2(i+1)} \tag{8}$$

$$Total_{A,\ NDCG} = \frac{\sum_{i=1}^{N} NDCG_{i,\ A}}{N} \tag{9}$$

For all the proposed and contemporary approaches, the cumulative NDCG score is calculated. As given in Equation 9, cumulative NDCG ($\text{Total}_{A,\ NDCG}$) for an approach A is the average of its NDCG scores ($\text{NDCG}_{i,\ A}$) for all sets N of co-cited documents sets.

3.7. Evaluation and benchmark construction

To evaluate the effectiveness of the proposed approach, a gold-standard ranking for each set of co-cited documents is required but the authors of [5] concluded that no accepted benchmark is available. However, for this task, the literature suggests three approaches: user study and offline or online evaluation [5]. User study requires domain experts for possible evaluation and ranking and has been a useful way of evaluation. However, despite its usefulness, user study cannot be adopted in the case of large datasets as it becomes difficult for domain experts to evaluate a plethora of documents. Therefore, to avoid such issues, we preferred an automatic evaluation method, i.e. Jensen–Shannon divergence (JSD). JSD is also a frequently used metric to build a gold standard [1].

JSD is a measure of similarity between two probability distributions. JSD is based on Kullback–-Leibler divergence (KLD) with some key differences, such as it always has a finite value and is symmetric [21, 37]. As of this research, the key term distribution of the query paper forms one probability distribution and the key term distribution of the co-cited paper forms the second probability distribution. The mathematical form to calculate KLD and JSD is given in Equation 10 and 11, respectively, where Q is the key term distribution of the query paper and C is the key term distribution of the co-cited paper. The key-terms are identified using the TF-IDF approach.

$$JSD(Q||C) = \frac{1}{2}D(Q||M) + \frac{1}{2}D(C||M) - -WhereM = \frac{1}{2}(Q+M) \tag{10}$$

$$D_{KL}(Q||C) = \sum_{x \in X} Q(x)\ log\left(\frac{Q(x)}{C(x)}\right) \tag{11}$$

After measuring the JSD for each pair of query and co-cited papers, papers were ranked using "ordinal ranking" in each set of co-cited documents based on JSD scores. The pair of papers with the smallest JSD is top-ranked, as it has least divergence or maximum similarity. This ranking is assumed as the gold standard and is further used for evaluation purposes.

3.8. Solution architecture and software libraries

A brief overview of a solution architecture is given in Figure 3. The main components of the solution are web querying, data preprocessing, database handler, similarity finder, and result evaluations. The solution is implemented by using Python 3.7 and the MySQL database. For web querying and data fetching from CiteSeerX we used BeautifulSoup4 and the Requests library. The fetched data are preprocessed by the standard Natural Language Toolkit - NLTK 3.4.5. The preprocessed data are managed in the database by MySQL Connector/Python. TF-IDF is implemented by the standard library Sci-Kit TfidfVectorizer. The similarity measures, ranking, and evaluation were performed with the help of Sci-Kit, Pandas, and NumPy libraries.

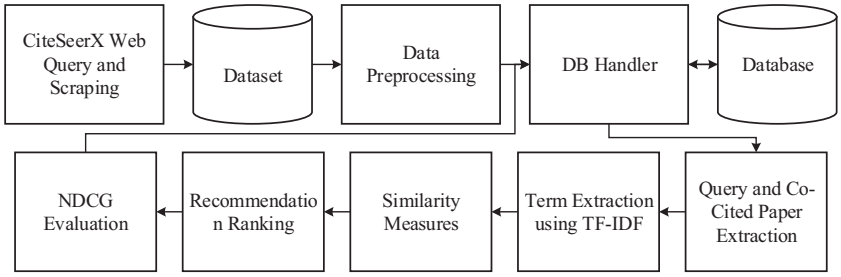

Figure 3. Overall solution architecture. The main components of software solutions such as web scrapping, database manipulation, TF-IDF, and similarity measures were implemented by standard Python libraries such as Sci-Kit, NumPy, and Pandas.

4. Results and discussion

The NDCG scores for each set of co-cited documents for traditional co-citation analysis and the content-based approach are shown in Figure 4. In this graph, the X-axis represents the set number of co-cited documents and the Y-axis shows the NDCG score. The content-based approach significantly outperformed in each set of co-cited documents when compared to traditional co-citation. The content-based approach achieved a mean NDCG score of 0.859 while traditional co-citation achieved a mean score of 0.716. It is observed that the content-based approach outperforms the traditional co-citation approach in all sets of co-cited documents. The dataset contains papers from different domains of computer science like digital libraries, collaborative filtering, and news recommendations. Therefore, initially, it can be concluded that content-based systems are independent of the research domains and generally outperform traditional co-citation approaches in each domain.

The traditional co-citation performed worse in "Set 2" of the co-cited documents. The titles of query and co-cited documents for this set are shown in Table 2. Critical analysis of the dataset depicts that this query paper discusses different generic concepts; therefore, it is co-cited in multiple papers and books which are related in general but not in specific. For instance, "Introduction to Algorithms" is a well-known book on algorithms that describes different algorithms; therefore, papers tend to co-cite this book and query paper. Also, it can be seen in the table that the paper co-cites different other generic papers like "Application of Dimensionality Reduction in Recommender System – A Case Study," which is not specifically related to the query paper. The traditional co-citation approach performed very well for "Set 7" with NDCG score of 0.81. The dataset snippet for this set is given in Table 3. It can be observed that all papers are closely related and discuss the collaborative filtering. Hence, by observing this pattern, it can also be concluded that traditional co-citation analysis performs much better for the query papers that discuss the precise concepts while it performs poorly for the query papers that discuss broad or generic concepts of any domain. The content-based approach showed similar behavior throughout the co-cited documents. This approach relies on the top 30% of key terms extracted using TF-IDF.

The query and the co-cited documents are generally from the same domain (as can be seen from Table 3) and therefore tend to have the same key terms. It can also be concluded that the content-based approach tends to perform with the same efficiency for a set of papers from the same domains regardless of coverage of concepts in the papers.

Table 2. Query and co-cited paper titles for Set 2.

Query paper – Set 2	**Explaining collaborative filtering recommendations**
Co-cited documents	Maximum likelihood from incomplete data via the EM algorithm
	Introduction to Algorithms
	An algorithmic framework for performing collaborative filtering
	Using collaborative filtering to weave an information tapestry
	Application of Dimensionality Reduction in Recommender System – A Case Study
	Item-based Collaborative Filtering Recommendation Algorithms
	Evaluating collaborative filtering recommender systems
	GroupLens: An Open Architecture for Collaborative Filtering of Netnews
	GroupLens: Applying collaborative filtering to Usenet news

Figure 5 represents the NDCG scores when metadata are incorporated to find the relatedness. In Figure 5, the set number of each co-cited pair and NDCG score is represented by the X-axis and Y-axis, respectively. In the majority of document sets, the abstract-based approach performed better than the title. When the title was incorporated, the achieved mean NDCG score was 0.727. By incorporating the abstract relevancy scores, significant improvement in terms of NDCG score was observed and the mean score was 0.783. The highest NDCG score is achieved by combining the scores of abstract and title, where the achieved NDCG score is 0.807.

Table 3. Query and co-cited paper titles for Set 7.

Query paper – Set 7	**Eigentaste: A constant time collaborative filtering algorithm**
Co-cited documents	Maximum likelihood from incomplete data via the EM algorithm
	An algorithmic framework for performing collaborative filtering
	GroupLens: An Open Architecture for Collaborative Filtering of Netnews
	Empirical Analysis of Predictive Algorithm for Collaborative Filtering
	Application of Dimensionality Reduction in Recommender System – A Case Study
	Item-based Collaborative Filtering Recommendation Algorithms
	GroupLens: Applying collaborative filtering to Usenet news
	Social Information Filtering: Algorithms for Automating Word of Mouth
	Learning Collaborative Information Filters

The abstract-based relatedness outperformed the title-based relatedness in each of the document sets except Set 7. The abstract of a research article is a brief summary of the entire content, while the title of a research article is a very precise description of the article and normally contains less than 15 words or key terms. Also, often authors tend to write very short and descriptive titles containing five to ten words. Therefore, abstracts contain much more information about articles in terms of key terms as compared to the title. Thus, abstract-based relatedness outperforms title-based relatedness. The title-based relatedness outperformed only in Set 7. The snippet of titles queried and co-cited papers is given in Table 3. Careful investigation of this set

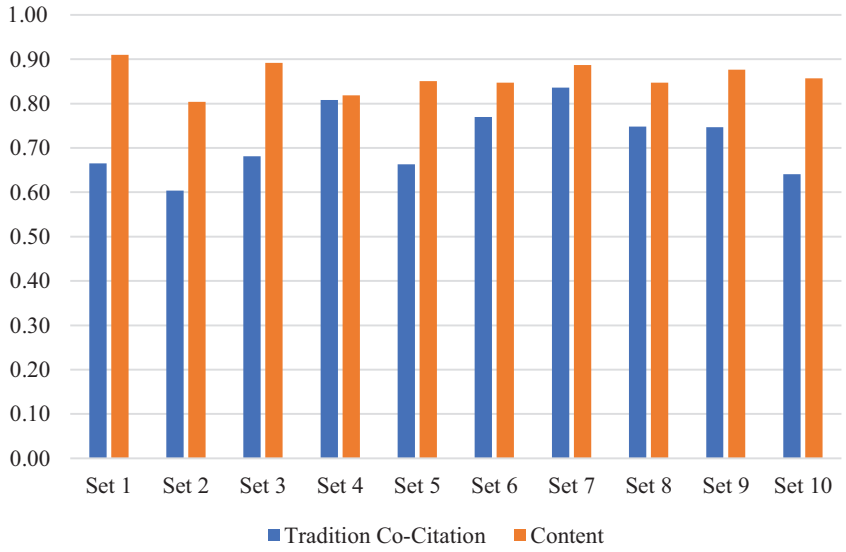

Figure 4. NDCG scores of traditional co-citation and content-based approaches. The content-based approach significantly outperformed in each set of co-cited documents when compared to traditional co-citation.

of documents revealed that the titles of all documents are very similar. For example, the term "collaborate" is present in 70% of documents and "algorithm" in 60% of documents. Hence, it can be concluded that in general abstract-based relatedness tends to perform better than title-based relatedness because abstracts tend to contain more information for machines than titles. However, if the titles match, then it depicts a strong relationship between the articles and therefore could perform better.

The experiments also showed that while incorporating the metadata, best results were achieved by combining the title and abstract score (as discussed in Section 3.3)). In this combination, strengths for both "title" and "abstract" were combined and therefore it produced the best results. As discussed earlier, the title and abstract contain the precise key terms that describe an article. As was observed, the abstract similarity tends to perform better and by combining the abstract with the title it was further strengthened. It also produced some consistent behavior throughout all co-cited sets.

Furthermore, the combined effect of CC and metadata is also evaluated (as discussed in Section 3.5). Figure 6 shows the NDCG score values against each combination for ten sets of co-cited documents. Co-citation, when combined with the title, achieved an average score of 0.766, while abstract and co-citation can reach a score of 0.761. When the co-citation is combined with title and abstract, it attained a 0.77 NDCG score. The outcomes show significant improvement compared to traditional co-citation. Mixed behavior of all the combinations is observed and experimental results show no significant difference between different combinations. However, it can be concluded that results of the traditional co-citation analysis can be significantly enhanced by incorporating metadata-based relevancy scores. This behavior is consistent throughout all sets of co-cited documents.

To evaluate the proposed approach for all sets of co-cited documents, the average score is calculated using Equation 12. The average score (Score_{ap}) for an approach (ap) is the average score for all the sets of co-cited documents (for sets 1 to 10). Figure 7 shows the average scores for each of the proposed approaches. The best-performing content-based approach has attained 0.86 NDCG scores. It is obvious from the literature also that content-based approaches perform best in this domain. The traditional co-citation approach achieved a score of 0.72. The results also correlate with the literature by favoring the argument that this approach needs

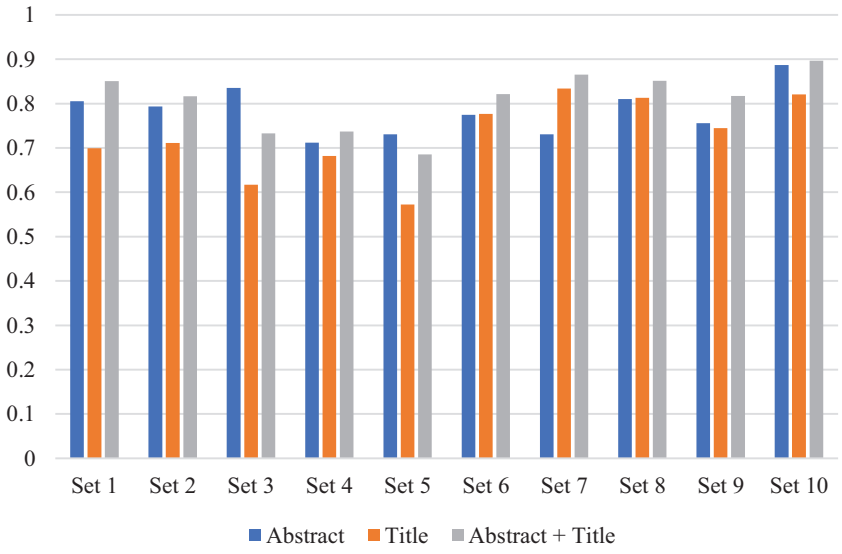

Figure 5. NDCG score for metadata-based relatedness for each set of co-cited documents. The highest NDCG score is achieved by combining the scores of abstract and title, where the achieved NDCG score is 0.807.

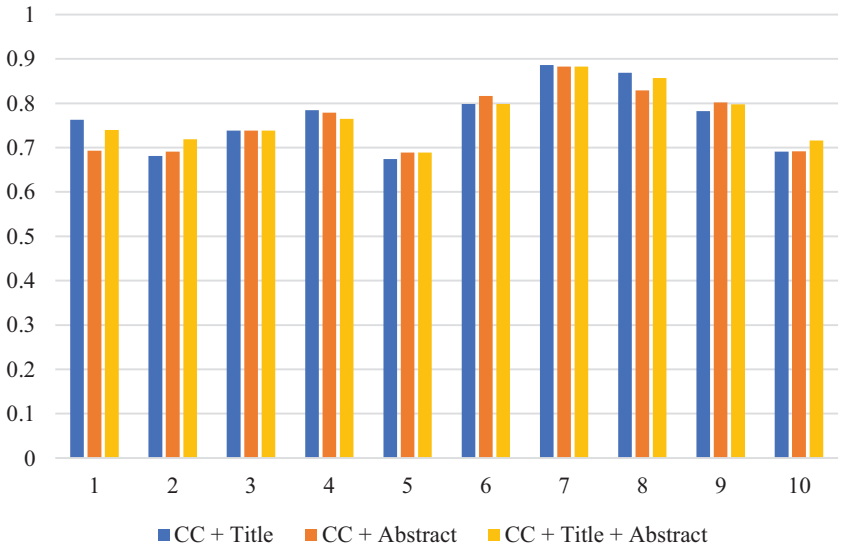

Figure 6. NDCG score of co-citation and metadata combinations. Superior results were achieved when co-citation was combined with title and abstract, with a 0.77 NDCG score.

content to improve accuracy. The results of the traditional co-citation approach are enhanced by incorporating the metadata relevancy scores. By incorporating title, abstract, and title + abstract, the achieved NDCG scores are 0.77, 0.76, and 0.77, respectively. The best results are achieved by combining title and abstract with an NDCG score of 0.81.

$$Score_{ap} = \frac{\sum_{i=1}^{10} Score_{ap,i}}{10} \tag{12}$$

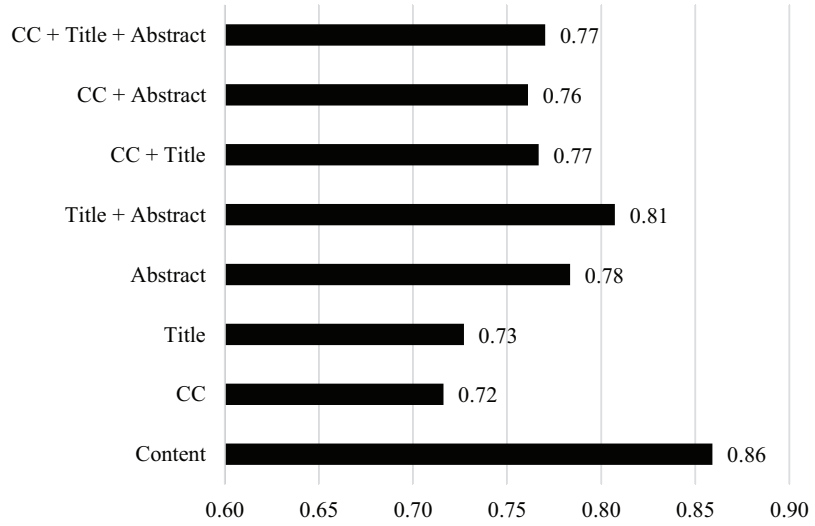

Figure 7. Mean NDCG score of traditional co-citation, content-based, and proposed approaches over all co-cited document sets. By incorporating title, abstract, and title + abstract, the achieved NDCG scores are 0.77, 0.76, and 0.77, respectively. The best results are achieved by combining title and abstract with an NDCG score of 0.81.

5. Limitations

The proposed approach is not immune to citation cartels and artificially generated citations, as highlighted in [38]. Moreover, the proposed approach is also dependent on the accuracy of citation identification and annotation algorithms.

6. Conclusion and future work

In the evaluation of the proposed approach, multiple findings were observed. 1) Abstract-based relatedness showed consistent behavior and outperformed title-based relatedness. The abstract remained the key component in the process as it contains much more information in terms of key terms as compared to the title. 2) Careful inspection of the dataset revealed that whenever titles of the documents matched in a set, it outperformed. The title normally contains a smaller number of key terms and therefore produces a smaller similarity score; however, whenever the titles of documents match, it is a high probability that the documents are related. It could also be concluded that key terms in titles carry more information than key terms in the abstract. 3) The results of traditional co-citation were significantly enhanced by merging the co-citation and metadata scores. A mixed behavior was observed by combining co-citation with title, abstract, and title plus abstract. In terms of co-citation and metadata combinations, it cannot be positively concluded which combination performed better in all cases but on average co-citation plus title and co-citation plus title plus abstract performed equally with NDCG scores of 0.77. 4). The combination of title and abstract outperformed by attaining the NDCG score of 0.81. The title contains fewer but more significant key terms. Hence, it can be concluded that the proposed approach with a combination of title and abstract can be applied to enhance the results of traditional co-citation measures in scenarios wherein content is unavailable.

In the future, we are interested to incorporate other metadata fields (such as author-defined keywords, author, and venue) to evaluate their effects on relatedness. Furthermore, we also intend to investigate the effect of other text similarity calculation techniques (such as Levenshtein distance, Hamming distance, and N-grams) in terms of finding relevant research articles.

References

[1] Habib R, Afzal MT. Sections-based bibliographic coupling for research paper recommendation. Scientometrics 2019; 119 (2): 643-656.

[2] Breitinger C, Gipp B, Langer S. Research-paper recommender systems: a literature survey. International Journal on Digital Libraries 2015; 17 (4): 305-338.

[3] Perc M, Ozer M, Hojnik J. Social and juristic challenges of artificial intelligence. Palgrave Communications 2019; 5 (1): 1-7.

[4] Helbing D, Brockmann D, Chadefaux T, Donnay K, Blanke U et al. Saving human lives: What complexity science and information systems can contribute. Journal of Statistical Physics 2015; 158 (3): 735-781.

[5] Beel J, Gipp B, Langer S, Breitinger C. Paper recommender systems: a literature survey. International Journal on Digital Libraries 2016; 17 (4): 305-338.

[6] Ricci F, Rokach L, Shapira B. Introduction to Recommender Systems Handbook. Recommender Systems Handbook; 2011. p. 1-35.

[7] Colavizza G, Boyack KW, van Eck NJ, Waltman L. The closer the better: similarity of publication pairs at different cocitation levels. Journal of the Association for Information Science and Technology 2018; 69 (4): 600-609.

[8] Gipp B, Beel J. Citation proximity analysis (CPA): a new approach for identifying related work based on co-citation analysis. In: ISSI'09: 12th International Conference on Scientometrics and Informetrics; Rio de Janeiro, Brazil; 2009. pp. 571-575.

[9] Martinčić-Ipšić S, Močibob E, Perc M. Link prediction on Twitter. PLoS One 2017; 12 (7): e0181079.

[10] Jalili M, Orouskhani Y, Asgari M, Alipourfard N, Perc M. Link prediction in multiplex online social networks. Royal Society Open Science 2017; 4 (2): 160863.

[11] Yin P, Zhang M, Li X. Recommending scientific literatures in a collaborative tagging environment. In: International Conference on Asian Digital Libraries; Berlin, Germany; 2007. pp. 478-481.

[12] Manouselis N, Verbert K. Layered evaluation of multi-criteria collaborative filtering for scientific paper recommendation. Procedia Computer Science 2013; 18: 1189-1197.

[13] Melville P, Mooney RJ, Nagarajan R. Content-boosted collaborative filtering for improved recommendations. AAAI/IAAI 2002; 23: 187-192.

[14] Huang Z, Chung W, Ong TH, Chen H. A graph-based recommender system for digital library. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries; Portland, OR, USA; 2002. pp. 65-73.

[15] Boyack KW, Small H, Klavans R. Improving the accuracy of co-citation clustering using full text. Journal of the Association for Information Science and Technology 2013; 64 (9): 1759-1767.

[16] Kessler MM. Bibliographic coupling between scientific papers. Journal of the Association for Information Science and Technology 1963; 14 (1): 10-25.

[17] Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the Association for Information Science and Technology 1973; 24 (4): 265-269.

[18] Liu C. The proximity of co-citation. Scientometrics 2012; 91 (2): 495-511.

[19] Liu S, Chen C. The effects of co-citation proximity on co-citation analysis. In: 13th International Conference of the International Society for Scientometrics and Informetrics; Durban, South Africa; 2011. pp. 474-484.

[20] Caragea C, Wu J, Ciobanu A, Williams K, Fernández-Ramírez J et al. Citeseer x: A scholarly big dataset. In: European Conference on Information Retrieval; Amsterdam, the Netherlands; 2014. pp. 311-322.

[21] Lin YS, Jiang JY, Lee SJ. A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering 2013; 26 (7): 1575-1590.

[22] Jomsri P, Sanguansintukul S, Choochaiwattana W. A framework for tag-based research paper recommender system: an IR approach. In: IEEE 24th International Conference on Advanced Information Networking and Applications Workshops; Perth, Australia; 2010. pp. 103-108.

[23] Naak A, Hage H, Aimeur E. A multi-criteria collaborative filtering approach for research paper recommendation in papyres. In: International Conference on E-Technologies; Les Diablerets, Switzerland; 2009. pp. 25-39.

[24] Yang C, Wei B, Wu J, Zhang Y, Zhang L. CARES: A ranking-oriented CADAL recommender system. In: 9th ACM/IEEE-CS Joint Conference on Digital Libraries; Austin, TX, USA; 2009. pp. 203-212.

[25] Schafer JB, Frankowski D, Herlocker J, Sen S. Collaborative filtering recommender systems. In: The Adaptive Web; Berlin, Germany; 2007. pp. 291-324.

[26] MacRoberts M, MacRoberts B. Problems of citation analysis. Scientometrics 1996; 36 (3): 435-444.

[27] Liu M. Progress in documentation the complexities of citation practice: a review of citation studies. Journal of Documentation 1993; 49 (4): 370-408.

[28] Shahid A. Recommending relevant papers using in-text citation frequencies and patterns. Capital University of Science and Technology, Islamabad, Pakistan, 2016.

[29] Khan AY, Khattak AS, Afzal MT. Extending co-citation using sections of research articles. Turkish Journal of Electrical Engineering & Computer Sciences 2018; 26 (6): 3345-3355.

[30] Ahmad R, Afzal MT. CAD: An algorithm for citation-anchors detection in research papers. Scientometrics 2018; 117 (3): 1405-1423.

[31] Constantin A, Pettifer S, Voronkov A. PDFX: Fully-automated PDF-to-XML conversion of scientific literature. In: ACM Symposium on Document Engineering; Florence, Italy; 2013. pp. 177-180.

[32] Zhang K, Xu H, Tang J, Li J. Keyword extraction using support vector machine. In: International Conference on Web-Age Information Management; Hong Kong; 2006. pp. 85-96.

[33] Bhowmik R. Keyword extraction from abstracts and titles. In: IEEE Southeast Conference; Huntsville, AL, USA; 2008; pp. 610-617.

[34] Chim H, Deng X. Efficient phrase-based document similarity for clustering. IEEE Transactions on Knowledge and Data Engineering 2008; 20 (9): 1217-1229.

[35] Bird S, Loper E. NLTK: The natural language toolkit. In: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions; Barcelona, Spain; 2004. pp. 31-37.

[36] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 2002; 20 (4): 422-446.

[37] Bigi B. Using Kullback-Leibler distance for text categorization. In: European Conference on Information Retrieval; Pisa, Italy; 2003. pp. 305-319.

[38] Fister I, Fister I, Perc M. Toward the discovery of citation cartels in citation networks. Frontiers in Physics 2016; 4: 49.