Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 3345 – 3355
© TÜBİTAK
doi:10.3906/elk-1703-260

Extending co-citation using sections of research articles

Arjumand Yar KHAN[1,*], **Abdul SHAHID**[2], **Muhammad Tanvir AFZAL**[1]
[1]Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan
[2]Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan

Received: 21.03.2017 • **Accepted/Published Online:** 15.10.2018 • **Final Version:** 29.11.2018

Abstract: The excessive amount of digital information has made it crucial to extract the relevant information. This hinders researchers in finding documents pertaining to their research. There exist various state-of-the-art techniques, such as co-citation, bibliographic coupling, and their recent extensions like citation proximity analysis and citation order analysis, that recommend the relevant documents against the posed query. Most of these approaches are statistical in nature and thus can further be extended by incorporating some semantics to enhance the results. In this paper, we present an extension of a co-citation-based technique to identify the most relevant documents to co-cited document(s). The proposed system explores in-text citation frequencies and in-text citation patterns of co-cited documents within the different logical sections of cited-by papers. Furthermore, we have evaluated the proposed approach with the co-citation approach and citation proximity analysis (CPA) approach on a dataset acquired from CiteSeer. The outcomes revealed that most of the time the proposed approach outperformed other state-of-the-art techniques. The average correlation of the proposed approach is increased by 68% as compared to the co-citation-based approach. In comparison with CPA approach, the average correlation of the proposed approach is increased by 39% with respect to gold-standard rankings.

Key words: Recommending relevant papers, co-citations, citation proximity analysis, in-text citation frequencies and patterns, citation analysis

1. Introduction

The digital information over the World Wide Web is rapidly being increased [1]. This bulk of information hinders users in retrieving relevant information [2–4]. In this regard, various efforts have been made in the form of citation indexes systems, such as CiteSeer, Google Scholars, etc. However, these systems return a plethora of documents, requiring users to make cognitive efforts to ascertain the relevant documents.

Such scenarios cause exasperation and most often users end up missing the most relevant documents. For instance, let us consider a scenario wherein a user intends to find relevant state-of-the-art papers using Google Scholar. The paper is:

Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inform Sci 1973; 24: 265-269.

To search for relevant papers for this focused paper, Google Scholar provides the user with two search options: 1) the user can either search for relevant documents using the "cited-by" feature or 2) by using the "related document" feature. Clicking on the "cited-by documents" feature, Google Scholar returns about 1800

[*] Correspondence: arjumandykhan87@gmail.com

research documents, whereas the "related documents" feature provides 101 research documents. Two documents may be relevant in different senses. For example, documents that are topically related may be less relevant as compared to the documents that have focused on the same problem. Similarly, if one document discusses an extended version of a technique proposed in an earlier study, then these two documents may be considered highly relevant as compared to documents that are only topically related.

Let us contemplate the returned lists (i.e. cited-by documents and related documents) provided by Google Scholar for one of the important techniques, i.e. citation proximity analysis. It is an extended work of the query paper found on page number 35 or at the 352nd position in the citations of the cited-by documents list:

Gipp B, Beel J. Citation proximity analysis (CPA)-A new approach for identifying related work based on co-citation analysis. In: Proceedings of the 12th International Conference on Scientometrics and Informetrics; 2009. pp. 571-575.

In the "related document" lists, however, it was completely ignored.

All of these deficiencies in the existing works have motivated us to overcome them so that most widely known systems can be improved. In the literature, various approaches such as citation-based approaches [2,5–8], content/text-based approaches [9,10], collaborative filtering-based approaches [11,12], and metadata-based approaches [13,14] have been proposed.

The previous studies did not perform analysis of citations at a deep level (i.e. content level), and hence the results are not refined. Therefore, considering in-text citations, details can play a paramount role in finding more relevant documents as compared to the surface level (i.e. just bibliography details). In this study, we propose an extension of the co-citation-based approach that exploits in-text citation frequencies and the role of citation patterns (in-text citations across the logical sections) in co-cited documents. The study has a close relevance to co-citation and CPA-based techniques. Therefore, we have compared our results with these techniques to evaluate the proposed study.

The rest of the paper is organized as follows. Section 2 discusses various state-of-the-art approaches (related work). Section 3 presents a detailed discussion about the proposed approach. Section 4 shows the analysis and experimental results, followed by the conclusion and future work in the next section.

2. Related work

To identify relevant papers, the scientific community has proposed various approaches using different data sources, e.g., content (text), metadata, user collaboration data, or citations. Text-based approaches compare the textual content of a research paper for finding relatedness between two or more documents. In recent year, a content-based approach has been proposed that creates the summaries of the research documents by extracting their important keywords. Afterwards, the similarity between generated summaries is calculated to determine relevance among them [10]. One study assigned publications to prespecified ontology-based contexts and computed context similarities between them [9]. Another performed term similarity between documents by extracting terms from title, abstract, and body components of a research paper [15].

Similarly, some contemporary schemes have utilized the metadata of research papers for finding relevant documents. One study exploited abstract, title, keywords, and reference list parameters, which are freely available and could be harnessed to recommend relevant documents [13]. Another extracted author(s), title(s), venue information, keywords, ACM top (if any), and other citation components from citation entries using the template-based information extraction using a rule-based learning approach (TIERL) for relevant document discovery [14].

Furthermore, similarities among users have also been considered in the literature for discovering relevant documents. Such schemes are called collaborative filtering approaches. Among them, a renowned technique was proposed by Zhang and Li [12]. Based on the documents viewed by a user, the proposed approach first creates a user profile, which is then used to create a concept (terms) tree for each user. Afterwards, the tree-edit distance is used to compute the correlation strength between users. Similarly, another study computed the user's interest vectors by focusing on the tags assigned by the user to a literature, the keywords of the tagged literatures, and their citations for recommending relevant documents [15].

Another branch of contemporary studies has focused on the citation network to recommend the relevant documents. For example, one study [16] used citation network information to analyze the life of scientific memes. The citation-based approaches have an edge because authors already provide a kind of relationship when they cite previously published related work. The importance of these citations in exploring the relationships between research documents has been realized in the literature by different researchers. Among them, a popular citation-based approach is known as bibliographic coupling, proposed by Kessler [5]. In bibliographic coupling, two papers P_1 and P_2 are considered similar if they share some common references in their bibliographic sections. These common references define the bibliographic strength between two or more research documents. However, bibliographic coupling depends on the references contained in the coupled documents; therefore, it is static in nature [5].

Later on, a dynamic citation-based technique was proposed by Small [6]. This approach is known as co-citation approach. It is primarily based on finding pairs of most cited papers. For example, two papers A and B do not directly cite each other, but both have been cited by paper C, which indicates that papers A and B are related. However, this technique ignores the distribution of citations within the document text [2].

The propagation of citations holds the Matthew effect, i.e. "For to all those who have, more will be given" [17]. This means that citations will be increased over a period of time, and this will create hindrance in retrieving the most relevant documents. Therefore, it is important to propose a technique that can discover the most important citations from the bulk of citations. For that matter, a citation-based approach known as citation proximity analysis (CPA) was proposed [2]. In recent years, a citation-based approach that uses a density-based clustering approach called DBSCAN has also been proposed. The proposed technique finds relevant documents by exploring in-text citation proximity of bibliographic coupled documents [8].

The proposed technique here is an extended version of the co-citation approach and has focused on some of the limitations of the co-citation approach. It considers the relationship between two or more documents if they are co-cited closely to each other in a cited-by document. The closer the citation tags of two or more documents are, the higher the possibility that they are more related.

3. Proposed technique

In the previous sections, it was highlighted that a significant amount of research work has already been done to identify document relationships using different data sources. Among these data sources, citation remains an important approach (area) for recommending relevant documents. A recent citation-based approach known as CPA is an extended version of the co-citation analysis approach.

Although CPA performs better in identifying related articles as compared to the co-citation approach, it cannot completely replace the co-citation approach [2]. Furthermore, much effort is required for calculating CPA results. Similarly, documents that are co-cited in research articles other than in the same sentence or paragraph or documents acquire zero score, and so a few citations may not be considered relevant in CPA-based approaches.

Considering the limitations of the CPA approach, we believe that considering in-text citations along the logical sections (e.g., Introduction, Related work, Methodology, Results, etc.) of a research document can be helpful in recommending relevant documents. In the literature, the importance of these sections for discovering relevant documents has already been explained [18,19]. The critical analysis of the state of the art has led us to formulate the following hypothesis: Co-citation-based relevant document ranking mechanisms can be improved by exploiting in-text citation frequencies and in-text citation patterns of the co-cited documents in the different logical sections of cited-by documents.

To address this hypothesis, we present a comprehensive methodology. The architecture diagram for the proposed system is shown in Figure 1.

3.1. Document fetcher

A document fetcher module was used to download cited and cited-by documents from CiteSeer.

3.2. Xml conversion

All cited-by documents were converted to XML format using the PDFx tool [20]. The PDFx tool is specifically designed for converting research articles into XML format. This tool makes use of well-defined sets of ontologies such as DOCO (document component) ontology and DEo (discourse element ontology).

3.3. Document preprocessing

In this step, some manual processing was performed to solve issues caused by accented characters (metadata), etc.

3.4. Bibliographic units extractor

In this step, bibliographic units of the documents converted into XML format are identified. For this, the citation tags were extracted by parsing the reference lists of each research document.

3.5. Citation frequency calculator

Frequencies of all in-text citations were found using xPath and xQuery solutions.

3.6. Section identifier

Different logical sections of cited-by documents such as the introduction, related work, methodology, and results were mapped manually.

3.7. Assigning section weights

Considering the importance of assigning weight to each logical section, as explained in [18], weights are assigned to these logical sections as shown in Figure 2. For example, usually a paper cited in the related work/introduction section presents the related state-of-the-art approach, whereas a document cited in methodology/results sections has a close relation with the citing document [18,19]. Using Eq. (1), weights were assigned to different logical sections.

$$WMeth/WRes > WIntro > WRe \tag{1}$$

Here, $WMeth$ and $WRes$ represent the weights of methodology and result sections, whereas $WIntro$ and WRe represent the weights of introduction and related work sections, respectively.

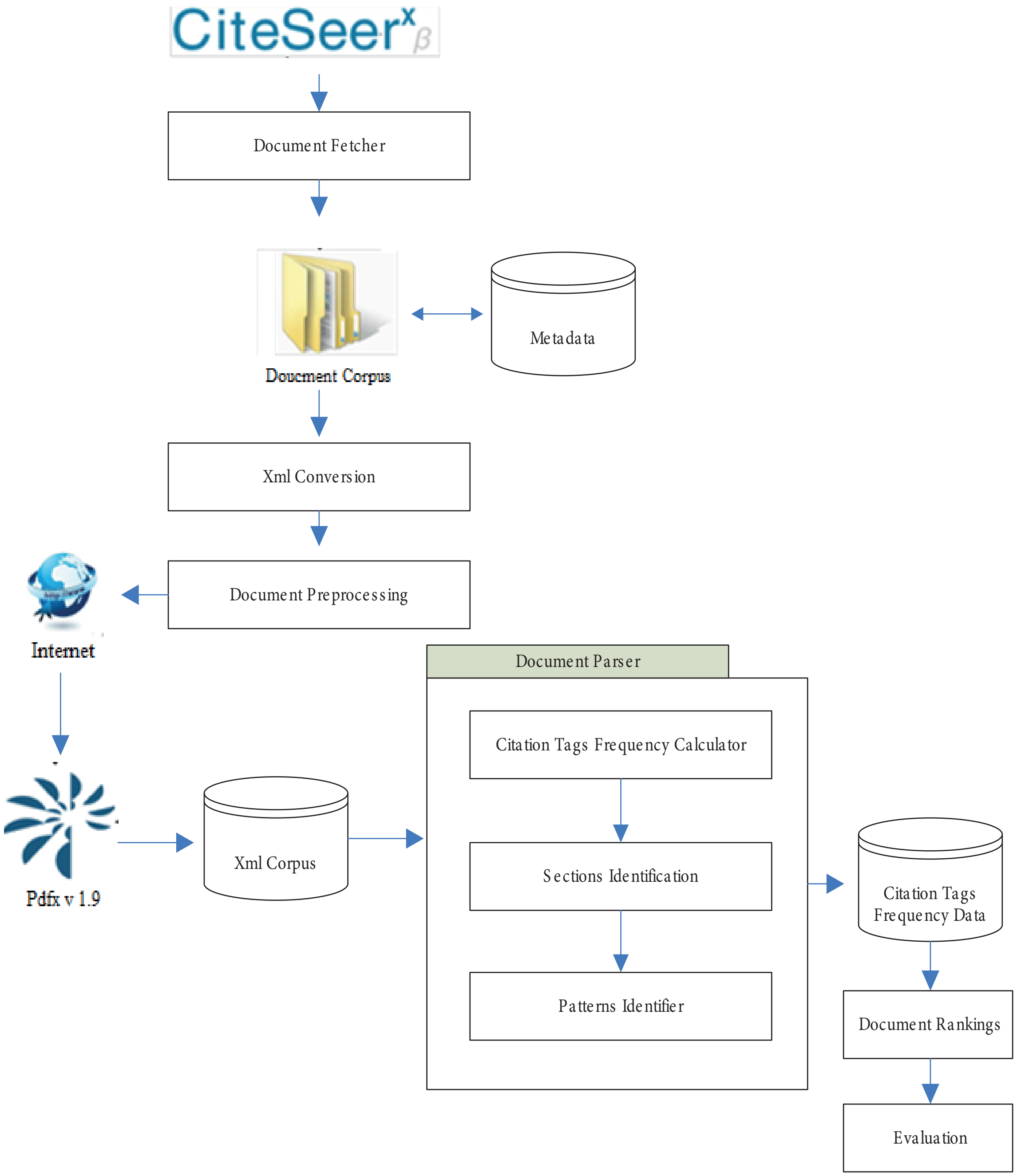

Figure 1. The architecture diagram of the proposed system is shown. First of all, research articles and their metadata are collected from CiteSeer. The XML conversion module with the help of online application PDFx creates XML versions of the acquired PDF files. The prepared XML versions of the articles are then parsed for extraction of citation tags, in-text citation frequencies, and sections. Finally, the data are kept in a database for further detailed analysis.

3.8. Pattern identifier

Citation frequency information of co-cited documents was mapped onto the logical sections of cited-by documents.

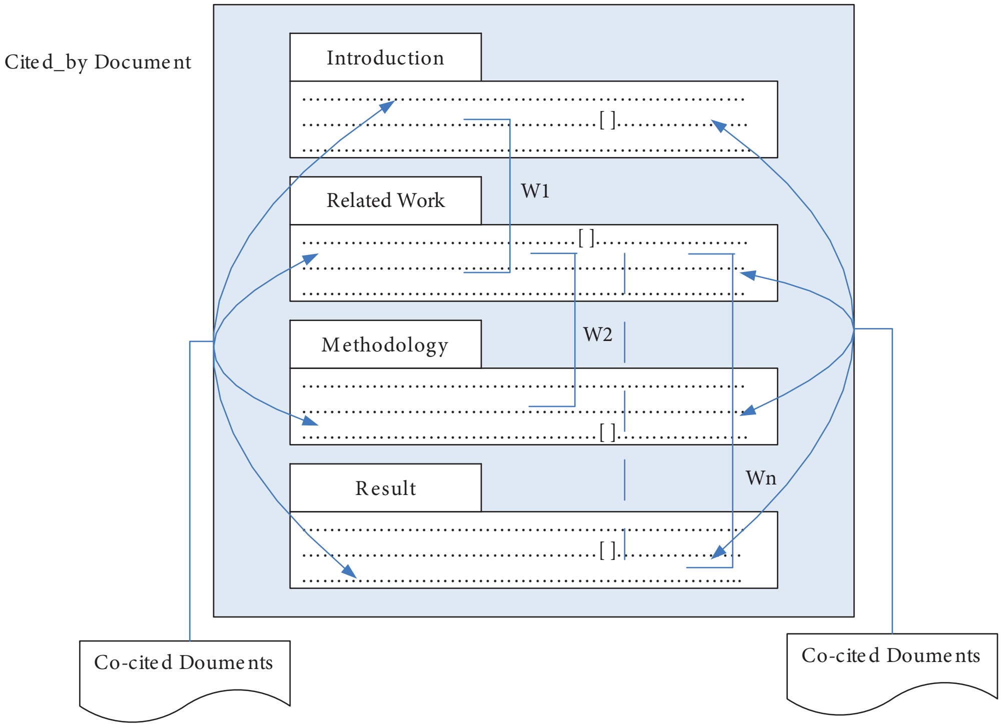

Figure 2. Logical sections with their corresponding weight mechanisms are explained. Different weights, e.g., W1, W2, and Wn, are assigned to the in-text citation evidence of co-cited documents in the citing document.

3.9. Document rankings

In the final step, the documents were ranked by combining section weights with number of occurrences (frequencies) of co-cited (two) documents. In this way, the total score of two or more co-cited documents in cited-by document(s) was computed using Eqs. (2) and (3). A comprehensive user study was then conducted to evaluate the proposed system/approach rankings with the document rankings obtained using the co-citation approach and CPA approach.

$$Score\left(Pa,Pb,Pk\right)=\sum_{Ki=1}^{MaxSection}\ \sum_{Kj=1}^{MaxSection}\ \left(Min(Fp_{ki}pa,Fp_{kj}pb)*SecWeights\left(P_{ki},P_{kj}\right)\ \right) \qquad (2)$$

Here, Pa and Pb represent co-cited documents and Pk represents a cited-by document. Similarly, $Fp_{ki}\,pa$ represents the frequency of document Pa in the ith section of Pk paper, whereas $Fp_{kj}\,pb$ represents the frequency of Pb document in the jth section of Pk document.

$$RScore\left(a,b\right)=\sum_{k=1}^{MaxCitedbyPapers} Score\left(Pa,Pb,Pk\right)\quad whereas\quad Pk\neq Pa,Pb \qquad (3)$$

4. Results

To evaluate the proposed approach, a total of 672 documents were extracted from CiteSeer. Out of these 672 documents, 304 documents were co-cited in 368 cited-by research documents. These co-cited documents were ranked based on their total scores in cited-by document(s) computed using Eqs. (2) and (3). A total of 17 different rankings were generated from 368 co-cited documents. Further, the co-citation [6] and CPA [2] were applied on this dataset; thus, two different sets of 17 ranked document(s) lists (one list by applying co-citation and one list by applying CPA) were obtained. These rankings were then compared with the document(s) rankings generated by the proposed approach.

4.1. Building a gold standard

To compare the proposed approach with the co-citation approach and CPA, a gold-standard dataset was required as a benchmark for comparison. Therefore, a comprehensive user study was conducted wherein 6 university professors, 11 PhD scholars, and 34 postgraduate students participated to develop a benchmark dataset. This sample of users was then divided into 17 different groups called "sets" based on their expertise about the domain in such a way that university professors and PhD scholars were placed in different sets (groups). Some of the PhD scholars were grouped with postgraduate students. Thus, 17 different sets were formed, each containing three members. The details of the set creation are illustrated in the Table.

Table. Users distribution among different sets (groups).

Users	Set's detail
Professors (6)	Set: 7, 17
PhD students (9)	Set: 11, 13, 16
PhD students (2) + MS student (1)	Set: 12
MS students (33)	Set: 1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 14, 15

Each of the three members in specific groups was given a list (table) mentioning a source document (target document) and some co-cited documents (candidate papers). This list consisted of abstracts, author information, and downloadable links of all candidate documents as well as the target document.

Users were asked to assign rank numbers to these candidate papers based on their relatedness with respect to a specific target paper assigned to them. The participants were briefed about relatedness. For example, extension of a target paper is more relevant than a paper that has cited the target paper just for background study. In the first experiment, interrater agreement between the three users in a specific set was found using the Spearman correlation coefficient using Eq. (4). The bar graph in Figure 3 represents interrater agreement values between each set.

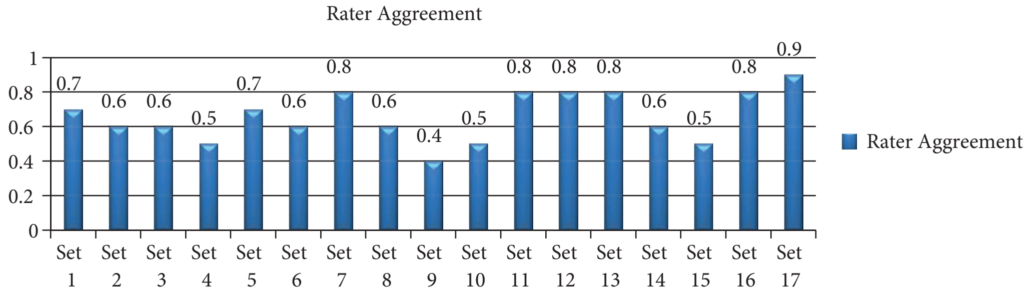

Figure 3. Interrater agreement is shown. As there were 17 rankings, 17 ranked lists were given to the annotators to compute ideal rankings. Afterward, interrater agreement was computed between their lists.

$$\rho = 1 - 6\frac{\sum di^2}{n\left(n^2 - 1\right)} \tag{4}$$

It was found that the average agreement value of all 51 users on the total dataset was approximately 64%, as shown in Figure 3. To construct a gold-standard dataset, only those document rankings were considered wherein all of the three raters (users) in a specific set gave similar rankings, whereas the rest of the document rankings were removed from the dataset. After removing such papers, the dataset contained 499 research documents, out of which 220 research documents were co-cited in 279 cited-by documents.

4.2. Comparing results of proposed approach, co-citation, and CPA approach

In the second experiment, the rankings obtained from the proposed approach, co-citation approach, and CPA approach were compared with the gold-standard document rankings. Since the data obtained from all of the above-mentioned approaches for comparison were discrete, another important correlation coefficient known as Spearman correlation coefficient [21] was used to compute interrater agreement as shown in Figure 4.

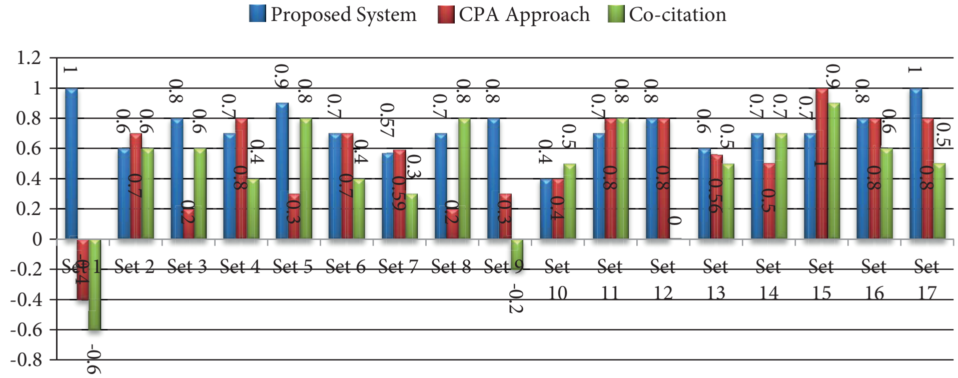

Figure 4. In this figure, the agreement of user sets (benchmark rankings) with the proposed approach, co-citation, and CPA approach is shown. Spearman correlation was computed for the mentioned three schemes with benchmark ranked lists. The proposed scheme outperformed the previous state-of-the-art approaches.

From Figure 4, it is obvious that rater agreement with the proposed approach was better as compared to the co-citation approach and CPA approach in most cases. It was found that the average agreement of the proposed approach with gold-standard ranking was 0.74, whereas the average agreement of the co-citation approach with the gold-standard ranking was 0.44. This indicates that the average correlation has been increased by 68% in the proposed approach as compared to the co-citation approach. Similarly, the average agreement of the CPA approach with the gold-standard ranking was observed as 0.53, which indicates that correlation of the proposed approach increased by 39% as compared to the CPA approach.

5. Discussion

The obtained results illustrate that the proposed system performed well in terms of qualitative results (i.e. retrieval of most relevant documents) as compared to its counterpart. For example, as stated earlier, the proposed system has produced 68% and 39% better results as compared to the co-citation and proximity-based approach, respectively. We think this improvement could be due to incorporation of semantic features. In our case, it is structural information of in-text citation frequencies in the body text of the citing paper. These encouraging results point toward other semantic features that could also be beneficial to improve the results further.

The results of the proposed system are encouraging, but still there are some apprehensions that should be taken into consideration, such as identification and calculation of in-text citations, identification of logical sections, etc.

The first and foremost important one is the conversion of PDF to XML/text with sufficient details. For example, how many times and where has the cited paper been referred to in the body text of the citing paper? It is not a trivial task to accomplish this; rather, it is a separate research problem [22]. We have utilized PDFx, which is specifically designed and developed for converting research articles into XML format [20].

The second issue is the precise mapping of article sections on logical sections. The logical sections are not apparent in any article. We have developed a dictionary and template-based technique that can map research article section instances over logical sections [23]. However, the important point is that the efforts to extract logical sections and mapping citation frequencies onto these sections are considerable and should be taken into account. Apart from the mapping of sections, section weights were considered based on the importance of sections highlighted in previous research. We believe an artificial neural network can be designed to fine-tune section weights in order to produce more quality results.

6. Conclusion and future work

Due to the increased volume of documents on the web, recommending relevant research documents has become the primary concern of the scientific community. In this regard, researchers around the globe have proposed different approaches. However, the existing studies recommend a bulk of documents containing relevant and irrelevant documents.

In this research, we have proposed a citation-based approach, which helps the scientific community in retrieving the most relevant research documents. The proposed approach is an extension of the co-citation technique that considers in-text citations across the logical sections of cited-by articles to compute the most relevant documents. The proposed approach was tested on the CiteSeer dataset and the results were compared with two state-of-the-art approaches known as co-citation and CPA approaches. The overall results indicated high percentage gains of 39% and 68% over CPA and co-citation, respectively. However, this approach has only been validated on a dataset of computer science. There is a need to evaluate it on different diversified datasets to comprehensively determine the effectiveness of the technique.

This research can be extended in various directions. For example, it can be used for exploring the evolution of specific topics in scientific collaboration network environments such as studies conducted by various researchers [24,25].

References

[1] Kaplan D, Lida R, Tokunaga T. Automatic extraction of citation contexts for research paper summarization: a co reference-chain based approach. In: Proceedings of the https://aclanthology.coli.uni-saarland.de/volumes/proceedings-of-the-acl-ijcnlp-2009-student-research-workshop 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries; 7 August 2009; Singapore. pp. 88-95.

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

[3] El-Arini K, Guestrin C. Beyond keyword search: discovering relevant scientific literature. In: Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining; 21–24 August 2011; San Diego, CA, USA. New York, NY, USA: ACM. pp. 439-447.

[4] Huang Z, Chung W, Ong TH, Chen H. A graph-based recommender system for digital library. In: Proceedings of the 2002 Joint Conference on Digital libraries; 14–18 July 2002; Portland, OR, USA. New York, NY, USA: ACM/IEEE-CS. pp. 65-73.

[5] Kessler MM. Bibliographic coupling between scientific

[6] papers. Am Doc 1963; 14: 10-25.

[7] Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inform Sci 1973; 24: 265-269.

[8] Shahid A, Afzal M, Qadir M. Discovering semantic relatedness between scientific articles through citation frequency. Aust J Basic Appl Sci 2011; 5: 1599-1604.

[9] Habib R, Afzal MT. Paper recommendation using citation proximity in bibliographic coupling. Turk J Elec Eng & Comp Sci 2017; 25: 2708-2718.

[10] Ratprasartporn R, Po J, Bani-Ahmad S, Ozsoyoglu G. Context based literature digital collection search. Int J VLDB 2009; 18: 277-301.

[11] Singla S, Duhan N, Kalkal U. A novel approach for document ranking in digital libraries using extractive summarization. Int J Comput Appl 2013; 74: 25-31.

[12] Yin P, Zhang M, Li X. Recommending scientific literatures in a collaborative tagging environment. In: Proceedings of the 2007 International Conference on Asian Digital Libraries: Looking Back 10 years and Forging New Frontiers; 10–13 December 2007; Hanoi, Vietnam. New York, NY, USA: ACM. pp. 478-481.

[13] Zhang Z, Li L. A Research paper recommender system based on spreading activation model. In: Proceedings of the 2nd International Conference on Information Science and Engineering; 4–6 December 2010; Hangzhou, China. New York, NY, USA: IEEE. pp. 928-931.

[14] Chen YL, Wei JJ, Wu SY, Hu Y. A similarity-based method for retrieving documents from the SCI/SSCI database. J Inf Sci 2006; 32: 449-464.

[15] Afzal MT, Kulathuramaiyer N, Maurer H. Creating links into the future. J UCS 2007; 13: 1234-1245.

[16] Yoon SH, Kim SW, Kim JS, Hwang WS. On computing text-based similarity in scientific literature. In: Proceedings of the 20th International Conference Companion on World Wide Web; 28 March–1 April 2011; Hyderabad, India: ACM. pp.169-170.

[17] Kuhn T, Perc M, Helbing D. Inheritance patterns in citation networks reveals scienti

[18] fic memes. Phys Rev X 2014; 4: 041036.

[19] Perc M. The Matthew effect in empirical data. J R Soc Interface 2014; 11: 20140378.

[20] Sugiyama K, Kan MY. Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries; 22–26 July 2013; Indianapolis, IN, USA. New York, NY, USA: ACM/IEEE-CS. pp. 153-162.

[21] Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing; 22–23 July 2006; Sydney, Australia. New York, NY, USA: ACM. pp. 103-110.

[22] Constantin A, Pettifer S, Voronkov A. PDFX: Fully-automated PDF to-XML conversion of scientific literature. In: 2013 Symposium on Document Engineering; 10–13 September 2013; Florence, Italy. New York, NY, USA: ACM. pp. 177-180.

[23] Spearman C. The proof and measurement of association between two things. Am J Psychol 1904; 15: 72-101.

[24] Shahid A, Afzal MT, Qadir MA. Lessons learned: the complexity of accurate identification of in-text citations. Int Arab J Inf Technol 2015; 12: 481-488.

[25] Shahid A, Afzal MT. Section-wise indexing and retrieval of research articles. Cluster Computing (in press).

[26] Lužar B, Levnajić Z, Povh J, Perc M. Community structure and the evolution of interdisciplinarity in Slovenia's scientific collaboration network. PLoS One 2014; 9: e94429.

[27] Hâncean MG, Perc M, Vlăsceanu L. Fragmented Romanian sociology: growth and structure of the collaboration network. PLoS One 2014; 9: e113271.