Scientometrics (2019) 119:643–656
https://doi.org/10.1007/s11192-019-03053-8

Sections-based bibliographic coupling for research paper recommendation

Raja Habib[1] · **Muhammad Tanvir Afzal**[1]

Received: 27 February 2018 / Published online: 2 March 2019

Abstract
Digital libraries suffer from the problem of information overload due to immense proliferation of research papers in journals and conference papers. This makes it challenging for researchers to access the relevant research papers. Fortunately, research paper recommendation systems offer a solution to this dilemma by filtering all the available information and delivering what is most relevant to the user. Researchers have proposed numerous approaches for research paper recommendation which are based on metadata, content, citation analysis, collaborative filtering, etc. Approaches based on citation analysis, including co-citation and bibliographic coupling, have proven to be significant. Researchers have extended the co-citation approach to include content analysis and citation proximity analysis and this has led to improvement in the accuracy of recommendations. However, in co-citation analysis, similarity between papers is discovered based on the frequency of co-cited papers in different research papers that can belong to different areas. Bibliographic coupling, on the other hand, determines the relevance between two papers based on their common references. Therefore, bibliographic coupling has inherited the benefits of recommending relevant papers; however, traditional bibliographic coupling does not consider the citing patterns of common references in different logical sections of the citing papers. Since the use of citation proximity analysis in co-citation has improved the accuracy of paper recommendation, this paper proposes a paper recommendation approach that extends the traditional bibliographic coupling by exploiting the distribution of citations in logical sections in bibliographically coupled papers. Comprehensive automated evaluation utilizing Jensen Shannon Divergence was conducted to evaluate the proposed approach. The results showed significant improvement over traditional bibliographic coupling and content-based research paper recommendation.

Keywords Paper recommendation · Bibliographic coupling · Citation proximity analysis · Logical sections

✉ Raja Habib
r_habib_pk@yahoo.com

[1] Capital University of Science and Technology, Islamabad, Pakistan

Introduction

In recent times, recommender systems for scientific papers have gained stature and importance, due to a colossal increase in the number of published research papers. As more and more papers are being published, the task of retrieving the relevant research papers is becoming more challenging. A study conducted by Khabsa et al. suggests that there are almost 25 million research papers that are freely available online (Khabsa and Giles 2014). While no doubt worthwhile, this leads to the problem of 'information overload': where a large number of results is returned to the researchers for their search queries, majority of which are usually irrelevant to them. Research paper recommendation has emerged as a promising solution to tackle this problem. Over past few decades, many researchers have proposed paper recommender systems (Beel et al. 2013a). These approaches make use of the metadata, content based filtering, collaborative filtering, co-citations and bibliographic coupling, among others.

Among these approaches, the ones based on citation analysis tend to be significant. These include bibliographic coupling, co-citation and direct citation. Co-citation considers two research papers relevant, if they have been cited by one or more common citing papers (Small 1973). Researchers have extended co-citation to include content analysis (Boyack et al. 2013; Gipp and Beel 2009). The incorporation of content analysis in co-citation resulted in an improvement in accuracy of research paper recommendation i.e. the co-citation approaches that used the content analysis recommended the papers that were more relevant to the query paper, as compared to those produced by tradition co-citation analysis. However, co-citation presents the relationship between two papers based on their co-occurrences in other papers, without considering the contents of the cited papers. Figure 1 shows how co-citation works.

Bibliographic coupling considers two research papers to be related if both of them cite one or more common research papers. The number of common citations between two research papers is called bibliographic coupling strength. The larger the value of the bibliographic coupling strength, the higher is the similarity between the papers. Figure 2 shows how bibliographic coupling works.

Unlike the co-citation approach, in bibliographic coupling, the references of the cited papers are taken into account while determining the similarity. Therefore, bibliographic coupling has inherited the benefits of recommending relevant papers; however, traditional

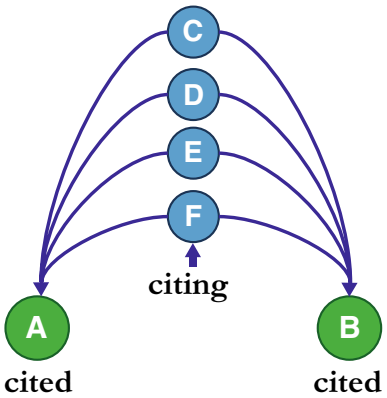

Fig. 1 Co-citation. Papers A and B are co-cited by papers C, D, E and F (Garfield 2001)

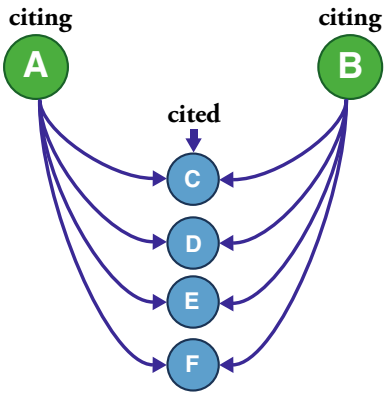

Fig. 2 Bibliographic coupling. Papers A and B are bibliographically coupled since both of them cite common papers C, D, E and F (Garfield 2001)

bibliographic coupling does not consider the citing patterns of common references in different logical parts of the citing papers. The significance of in-text citations in different logical sections of research papers has been proven by many researchers.

Studies have shown that almost all the authors follow a certain set of procedural standards when referencing other papers (Cronin 1984; Small 1976). For example, most relevant papers are cited in the methodology and result sections. Papers belonging to background knowledge are normally cited in the introduction or related work section. This makes the exploration of logical structure of research papers an area of interest for many researchers. In recent times, many researchers have shown interest in exploring the importance of position of in-text citations within the full content of research papers. The availability of full-text of research papers has made it possible for the researchers to develop innovative approaches for citation analysis, citation recommendation and paper recommendation (Bertin et al. 2013; Ding et al. 2013; Liu and Chen 2013). This full-text access to research papers has also provided possibilities for studying the distribution of in-text citations in the full content of research papers. Many researchers have shown interest in the localization of in-text citations in past as well. For example, Voos et al. conducted a manual study in order to find out if two citations can be given the same weight during citation analysis (Voos and Dagaev 1976). They used a very small dataset consisting of four research papers. McCain et al. proposed the idea for the first time that the section structure plays an important role in determining the function of in-text citations (McCain and Turner 1989). They studied and analyzed the in-text citations in different sections and proposed a scheme to assign different weights to the citations from different sections. Similarly Maričić et al. (1998) analyzed set of 357 research papers and concluded that location of in-text citations, along with their level and age, plays a vital role in citation analysis. They suggested that the in-text citations belonging to different sections have different values. Based on their analysis, they assigned different weights to different sections (Introduction: 10, Methods: 30, Results: 30, Discussion: 25).

Another study (Ding et al. 2013) highlights the fact that authors normally tend to prefer certain sections over the others while distributing the in-text citations. According to this study the Introduction section contains the largest number of in-text citations. The Literature Review section makes for the second most citing section followed by the Methodology section.

Studies show that the authors follow a set of norms and procedural standards when distributing the citations in the citing papers (Boyack et al. 2018; Cronin 1984; Small 1976).

According to a study, the highly cited papers get cited the most from the Introduction section (Ding et al. 2013). The Literature Review section makes for the second most citing section followed by the Methodology" section. Moreover, the citations from the Results and the Methodology sections are more important as compared to those from the Related Work section (Sugiyama and Kan 2013; Teufel 2009). The papers that are cited from the Results and the Methodology sections are usually more relevant to the citing papers. The authors usually cite the most relevant papers in these sections. However, the Related Work section and the Introduction may contain the citations to generic papers which may not be very relevant to the citing papers.

As shown above, a lot of work has been done in the past and in the recent time to show that authors follow a certain pattern when distributing the in-text citations. Different weighting schemes for sections have been proposed as well. However, not much research has been done to exploit the distribution of citations in sections in the context of citation analysis. In this paper, we propose a paper recommendation system approach that exploits the sections in bibliographically coupled papers to recommend relevant papers.

The rest of this paper is organized as follows. 'Related work' section discusses related work in this area. 'Methodology' section describes the architecture and different modules of proposed approach. 'Evaluation' section presents the evaluation of the results. Finally, 'Conclusion and future work' section provides the conclusion of this research with plans for future work.

Related work

Several research paper recommendation approaches have been proposed by researchers over past few decades. According to a study (Beel et al. 2013a), 200 different paper recommendation approaches have been proposed. These approaches can be classified as: (1) metadata-based approaches (Afzal et al. 2007; Doerfel et al. 2012), (2) citation-based approaches (Habib and Afzal 2017; Garfield et al. 1972; Kessler 1963; Liu and Chien 2017; Small 1973), (3) content-based approaches (Ratprasartporn and Ozsoyoglu 2007; Ding et al. 2014), (4) collaborative filtering (CF) based approaches (Amami et al. 2016; Hristakeva et al. 2017; McNee et al. 2002), (5) user profile-based approaches (Lee et al. 2013; Sahijwani and Dasgupta 2017; Sugiyama and Kan 2013), (6) data mining based approaches and (7) hybrid approaches.

There are numerous approaches to paper recommendation using citation analysis. Bibliographic coupling (Kessler 1963) and co-citation (Small 1973) are two citation analysis techniques that can help identify the closely related research papers (Smith 1981). Since the citations are mostly freely available on different digital libraries and are handpicked by the authors of the papers, these approaches make a good candidate for producing relevant papers. However, these approaches may not work in cases where authors cited a paper in the reference section but do not cite it in the full text of the papers (Shahid et al. 2011).

Recent research suggests that the accuracy of recommendations produced by the co-citation can be improved by using the proximity of in-text citations within the full text (Callahan et al. 2010; Elkiss et al. 2008; Gipp and Beel 2009; Liu and Chen 2012). Gipp et al. proposed an approach called citation proximity analysis (CPA) (Gipp and Beel 2009). They used the proximity of citations in full text to determine the strength of contextual co-citation among pairs of citations. CPA considers two citations more relevant to each other if they occur within the same sentence than if they occur within the same section.

Boyack et al. (2013) also presented an approach that uses the proximity of in-text citations for finding the related papers. This technique also uses the distance between the citations. But instead of using the sentence structure, the character or byte offset and centiles positions were used. 4 schemes (B, O, P1 and P2) were proposed for this purpose. Using the 1st scheme 'B', each co-citation pair is assigned a weight of 1. This scheme doesn't take the distance between the in-text citations into consideration. In the 2nd scheme represented by 'O', if the two in-text citations are within the same byte position, they are assigned a weight of 4. If references are within 375, 1500 and 6000 bytes, they are given weights of 3, 2 and 1 respectively. If the distance is more than 6000 bytes, a weight of 0 is assigned. In the 3rd scheme P1, the paper's text is divided into 20 equal parts which are considered as 5 centiles. The weights are assigned based on these centiles. In the 4th scheme P2, the byte range of centiles is changed. The similarity between the two papers is then discovered based on these weights.

Methodology

Figure 3 shows the system architecture for this approach. The important modules for this system are Data Acquisition, XML Conversion, Sections Extraction and Similarity Score Measuring. In the next sub-sections, we will discuss each of these modules in details.

Data acquisition

This module is used to collect two datasets for our experiment. There are many different digital libraries and online resources that offer the datasets. For example, PubMed provides access to almost 27 million citations for biomedical literature. Scopus is another huge repository of research papers. However, few of these repositories provide access to the datasets for free. Users have to pay for it. Another issue with some of these repositories is that it is a challenging task to extract the references from the papers. The process of downloading bibliographically coupled papers is complicated.

We used a digital library called CiteSeer to gather our dataset. CiteSeer is a huge repository that has around 2 million publications indexed. It provides access to the

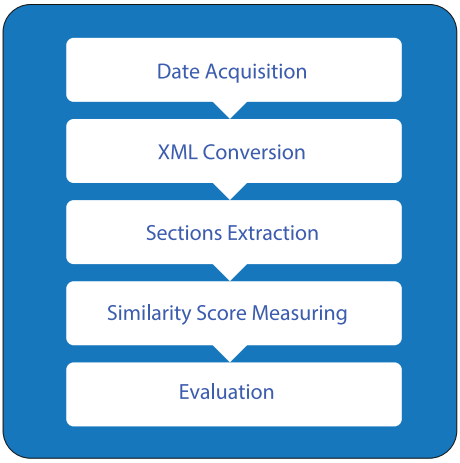

Fig. 3 System Architecture for section based bibliographic coupling

metadata (author's name, venue and year of publication, etc.) and the full texts of research papers. Researchers have used CiteSeer data in the past for various tasks, including text classification, collective classification and citation recommendation etc. Wang et al. (2016). There are two main reasons for using this digital library. The first is that it provides free access to the datasets, which can also be accessed in many different ways. The second is that it retains all the cited papers in a special table, and citing articles can be linked to them using a key attribute CID. In other words, CiteSeer simplifies the process of downloading datasets of bibliographically coupled papers.

We developed a focused crawler to download two different datasets. We used the first dataset for initial experiments, and the second for more extensive and comprehensive experiments. We called them dataset-1 and dataset-2. Initially, we collected dataset-1, containing 320 bibliographically coupled papers. Later, we collected the larger dataset-2, containing 5,000 bibliographically coupled papers from different domains.

In order to collect the dataset-1, We used the 7 queries mentioned in the Table 1. We chose these particular queries so that we could perform the initial experiments in diversified fields.

We used the 17 queries mentioned in the Table 2 to collect the dataset-2. These queries were chosen in order to provide a comprehensive and diversified dataset.

The dataset-1 consisted of 320 bibliographically coupled papers which were divided into 32 subsets. Each subset consisted of 10 papers that were bibliographically coupled based on a certain query paper. dataset-2 was divided into 226 subsets. These subsets were generated based on the combination of the search query used and the cited-paper-id. These subsets were later combined into 17 groups each representing a query.

XML conversion

Since the web crawler downloaded all the papers in PDF format, they needed to be converted into XML format in order to fetch the information related to sections and in-text citations. A freely available online tool called PDFx was used to convert our dataset of 5000 research papers in PDF format to XML format. PDFx is a tool designed specifically for conversion of scientific articles (Constantin et al. 2013). The converted XML files contain some very important elements such as section, ref and xref etc. The element xref with the attribute ref-type = 'bibr' represents the in-text citations and can be linked to the 'ref' tags through the attribute rid.

Table 1 Queries used for dataset-1

QID	Query
1	Network topology
2	Community detection
3	Face tracking
4	Microblogging usage
5	Defect detection
6	Co-citation analysis
7	Object-oriented databases

Table 2 Queries used for dataset-2

QID	Query
1	Social network
2	Information retrieval
3	Bayesian networks
4	Feature selection
5	Collaborative recommendation
6	Recommendation system
7	Content based filtering
8	Black box testing
9	Automatic generation
10	Regression testing
11	Query processing
12	Sensor networks
13	Wireless communications
14	Opinion mining
15	Subjectivity analysis
16	Online marketing
17	Graph theory

Section extraction

The XML documents from the previous module are passed on to the Section Extraction module. This module extracts the sections from the research papers using the special elements inside the XML documents denoted with the tag 'section'. This section element refers to all the sections inside the research paper. This element consists of a nested heading element denoted by 'h1'. This heading tag refers to the heading of each section. PDFx provides two more levels of heading element i.e. 'h2' and 'h3'. This module uses the Document Object Model (DOM) to traverse the XML files and to fetch the section headings.

Studies show that normally the research papers are organized in a standard way and contain specific sections. Studies show that most of the research papers contain certain sections (Golshan et al. 2012; Hengl and Gould 2002). These sections are given as follows:

1. Introduction
2. Related Work
3. Architecture/methodology
4. Results/comparisons
5. Conclusion/future work

These studies helped us to determine the main sections for our research too and we decided to use the same main sections as mentioned above. Using the section element, we fetched the sections from the research papers. In order to map these fetched sections to the sections mentioned above, we used the suggestions of a study conducted by Ding et al. (2013). Using this study, we can infer that the Introduction section contains the largest number of in-text citations followed by the Literature review section that contains the second largest number of in-text citations followed by the 'Methodology' section. The sections with the fourth and fifth largest number of citations are Results and Conclusion respectively. After

extracting the sections and the in-text citations from each section, we mapped the sections to the generic sections mentioned above, using the frequencies of the in-text citations.

In order to verify the section mapping of our system, we conducted a user study. We used the dataset-1 for this purpose. As we explained earlier, the dataset-1 consists of 32 different subsets with 10 bibliographically coupled papers in each subset. The dataset was assigned to two experts who have advanced experience and knowledge in the field of Computer Science. The two experts, we assigned the dataset-1 to, were pursuing their PhDs in the area of paper recommendation using citation analysis as well. This made them the perfect candidates for this user study, since they had the knowledge and hands on experience of the citations, research paper sections, and paper similarity.

The experts were assigned the task to manually map the sections of the papers in the dataset-1 to the generic sections that we mentioned above. This mapping produced by the experts was then compared with the mapping produced by our system. For this purpose we used the Spearman rank correlation coefficient. Other correlation coefficients like Pearsons coefficient and Kendalls Tau coefficient could have been used too. But we preferred to use the Spearmans coefficient because, unlike the other two above mentioned correlation coefficients, it doesnt need to make the assumption that the two variables are linearly related to each other. Moreover, it doesnt need the variables to be measure on interval scales (Hauke and Kossowski 2011).

The value of Spearmans correlation ranges between 0 and 1. Its value was 0.85 for the correlation between mappings produced by our system and those produced by the experts. According to Mukaka (2012), there exists a strong correlation if the value of Spearmans correlation coefficient is between 0.7 and 1.

Since there was high correlation between the mappings produced by our system and those by the experts in case of the dataset-1, we decided to use the same mapping criteria for the larger dataset i.e. dataset-2. Since the dataset-2 contains almost 5000 papers, it was not feasible to conduct user study for the dataset-2. However, we manually cross-checked randomly selected 100 papers from this dataset too, and found that the sections have been mapped with 90% accuracy. The sections were correctly extracted from these papers in 90% of the cases.

Similarity score measuoring

Many researchers have analyzed the distribution of in-text citations in research papers and their research suggests that the citations from different sections should be given different weights during citation analysis (Teufel 2009; Sugiyama and Kan 2013). These studies show that the citations from each section carry a different weight and have a different meaning. For example, the citations from Related Work and Introduction usually mean that the cited document might be a supporting document. The documents cited from the 'Methodology' and results sections, however, tend to be the most closely related ones. Similarly, the documents cited from the Related Work are considered to be the least important ones, since the Related Work may contain less related and more generic kind of citations too.

Considering the results of previous studies (Teufel 2009; Sugiyama and Kan 2013), the relation among the weights of different sections can be given by the following equation:

$$(weight(m) = weight(rs)) > weight(i) > weight(rw) \tag{1}$$

In Eq. (1), weight(m) denotes the weight of methodology section, weight(rs) denoted the weight of results section, weight(i) denotes the weight of introduction section and

Table 3 Weights of different sections

Section	Weight
Methodology	3
Results	3
Introduction	2
Related Work	1

Table 4 Weights for in-text citation pairs from cross sections

Paper B	Paper A			
	Introduction	Related Work	Methodology	Results
Introduction	2	0.5	0.24	0.24
Related Work	0.5	1	0.24	0.24
Methodology	0.24	0.24	3	3
Results	0.24	0.24	3	3

weight(rw) denotes the weight of related work section. As is obvious from the above equation, the in-text citations from the methodology and results section are given more weight than those from the introduction section. And the in-text citations from the related work carry the least weight. We determined the weights for different sections in two steps.

In the first step, we used the JensenShannon divergence (JSD) to rank the papers in dataset-1. JSD is based on the Kullback Leibler divergence. JSD finds the distance between two probability distributions. In the case of research papers, the word distribution of individual research papers forms one probability distribution and the word distribution of the entire cluster forms the second probability distribution. In this case, the clusters refer to the subsets of the dataset-1 based on the queries as mentioned in "Data acquisition" section.

In the second step, we generated the rankings for the dataset-1 using our system. For this purpose, we initialized the weight of Related work section with a value of 1 and changed the weights of other sections by increasing the value by 0.5 for same sections and 0.2 for cross sections. We used the Spearmans coefficient to determine the correlation between our rankings and the rankings produced by the JSD for all the different weights of the sections. We found out that the weights mentioned in the Table 3 produced the best results. The value of correlation for these values of weights was 0.8.

Table 3 represents the weights for the citations from the same sections. For example if paper 'A' and paper 'B' cite a common paper from the Methodology chapter, the weight will be 3. The weights for citations from the cross sections were calculated in the same way as mentioned above. The weights for same section and cross sections citations are shown in Table 4.

In Table 4, the sections mentioned along the Y-axis represent the sections of paper 'A' and the sections mentioned along the X-axis represent the sections of paper 'B'. If paper 'A' and paper 'B' cite a common paper from sections Introduction and Results respectively, the value of weight will be 0.24.

The weights mentioned in Tables 3 and 4 represent the weights for in-text citation pairs. Therefore the weights for multiple in-text citations can also be determined from these, by summing up the weights of each pair.

Evaluation

In order to evaluate the accuracy of our proposed approach we compared its performance with the content based paper recommendation and the traditional bibliographic coupling approach. There hasn't been much research in the area of research paper recommendation using the bibliographic coupling. Therefore, we decided to compare our approach with the traditional bibliographic coupling. However, we decided not to compare our approach with the CPA approach because the datasets that we used were bibliographically coupled and using co-citation analysis on these datasets would not have produced correct results.

Research paper recommendation approaches can be evaluated using user study, online evaluation or offline evaluation (Beel et al. 2013b). User studies have been useful way of evaluating paper recommendation systems (Sugiyama and Kan 2013; Beel et al. 2013b). Despite being useful, user studies have certain limitations as well. Conducting a user study for a large dataset is not feasible since it requires many experts who are willing to evaluate such a large dataset. Since the dataset-2 had almost 5000 research papers, conducting a user study was not the preferred method of evaluation for this dataset. Therefore, we decided to use the automatic method of evaluation i.e. Jensen Shannon Divergence (JSD). JSD finds the distance between two probability distributions. In the case of research papers, the word distribution of individual research papers forms one probability distribution and the word distribution of the entire cluster forms the second probability distribution.

JSD produced the rankings for the bibliographically coupled papers automatically. Then we used the Spearman's correlation coefficient to determine the correlation between the results of our approach and those produced by the JSD.

We also compared the results of our approach with those of the traditional bibliographic coupling and the content similarity. This comparison was done using Spearman's correlation coefficient. Figure 4 shows the comparison between the proposed approach and the bibliographic coupling approach. We found a higher correlation between the JSD results and our proposed approach as compared to the traditional bibliographic coupling approach for the majority of the documents used. We also compared the performance of our approach with the content-based approach. As shown in Fig. 5, our proposed approach performed better for the majority of the documents when compared to the content-based approach.

Figure 6 represents the average of the correlations for all the queries. The X-axis represents the three approaches and the Y-axis represents the average of correlations for all the queries. As we can see from this figure, our proposed approach has an average correlation

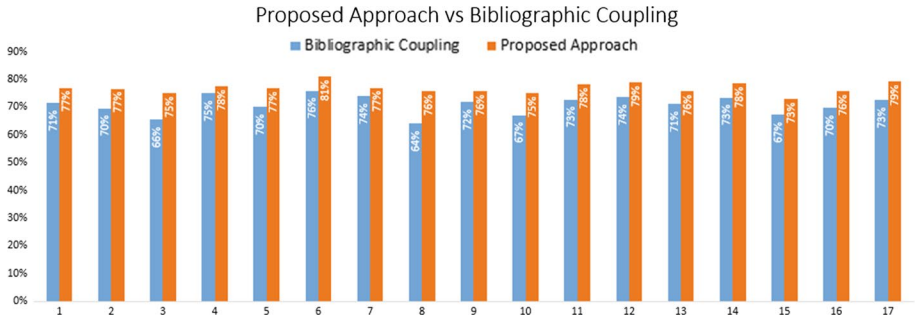

Fig. 4 Proposed approach vs. bibliographic coupling. Agreement between JSD results and the proposed approach is better compared to bibliographic coupling for the majority of the queries out of 17 queries

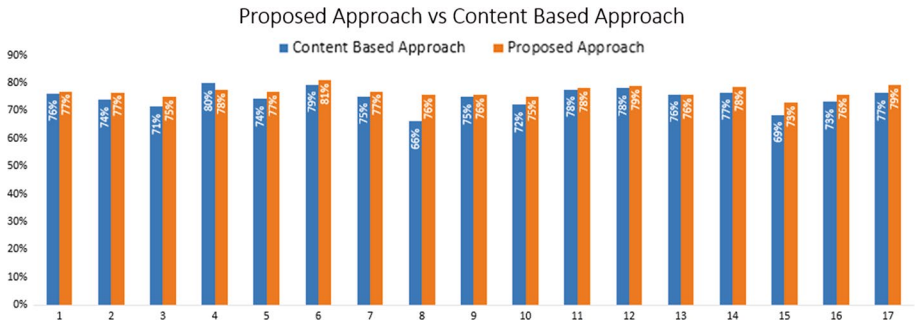

Fig. 5 Proposed approach vs. content based approach. Agreement between JSD results and the proposed approach is better compared to content based approach for the majority of the queries out of 17 queries

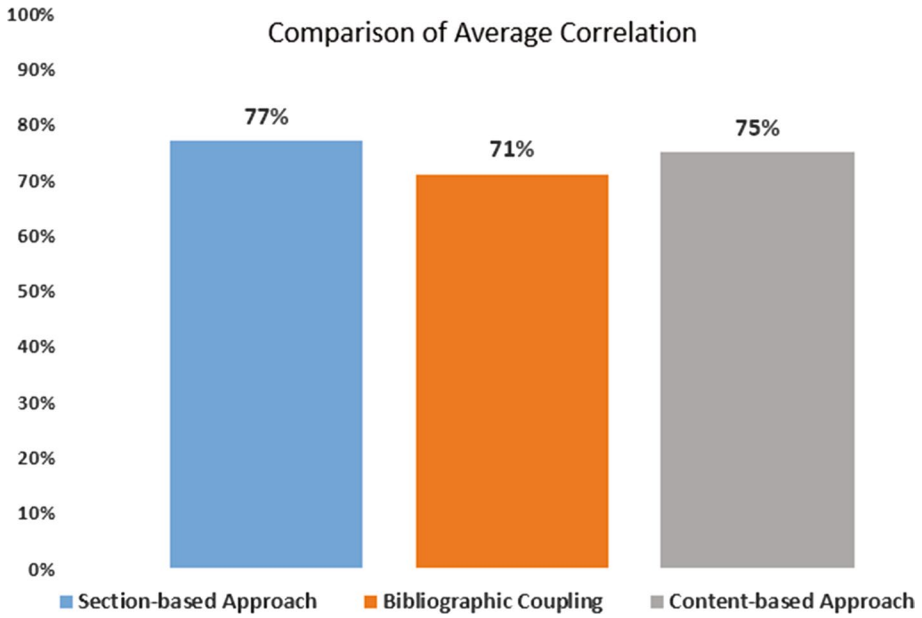

Fig. 6 Average correlations of all queries

of 0.77 with the results of JSD. The average correlation of our proposed approach is higher than the content based approach and the bibliographic coupling approach. The average increase in correlation for our approach is 8.5% and 2.7% as compared to bibliographic coupling and content based approaches respectively.

Conclusion and future work

Research paper recommendation systems have emerged as a revolutionary concept to help researchers who face the strenuous job of gaining access to the relevant research papers, due to information overload and the over-abundance of publications in conferences and

journals. Over the last few decades, many researchers have shown interest in proposing and developing innovative paper recommendation systems. In this paper, we proposed a new approach for paper recommendation that extended the traditional bibliographic coupling by incorporating the analysis of in-text citations and their existence in the logical sections of the research papers. This approach arose from an intuitive sense that authors follow certain standards when they distribute the in-text citations in their papers and that in-text citation from certain sections carries more weight than the others. Comprehensive experiments on a dataset of bibliographically coupled research papers show that the proposed approach using the logical sections outperformed the content based and bibliographic coupling based paper recommendation approaches. This research might also prove to be useful for researchers working in other areas such as identification of important citations, evaluation of h-index etc.

However, our research has following limitations: First, we only used CiteSeer to acquire the two datasets. Although using CiteSeer has certain benefits, but if the research had been extended to cover other digital libraries, the results might have been different. Second, our research was conducted based on certain queries that we mentioned in the Data Acquisition module. If other queries had been included, the results would have been more diverse.

In the future, we intend to work on an automatic way of assigning weights to different sections in order to improve the results. Neural networks can be used to assign weights automatically to different sections. We also intend to include other digital libraries and a diverse set of queries for data acquisition.

Acknowledgements This research was supported by Higher Education Commission (HEC) of Pakistan. Website: http://www.hec.gov.pk/.

References

Afzal, M. T., Kulathuramaiyer, N., & Maurer, H. A. (2007). Creating links into the future. *Journal of Universal Computer Science*, *13*(9), 1234–1245.

Amami, M., Pasi, G., Stella, F., & Faiz, R. (2016). An lda-based approach to scientific paper recommendation. In *International conference on applications of natural language to information systems* (pp. 200–210). Springer.

Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., & Nürnberger, A. (2013). Research paper recommender system evaluation: A quantitative literature survey. In *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation, RepSys '13* (pp. 15–22). New York: ACM. https://doi.org/10.1145/2532508.2532512.

Beel, J., Langer, S., Genzmehr, M., & Nürnberger, A. (2013). Introducing docear's research paper recommender system. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 459–460). ACM.

Bertin, M., Atanassova, I., Lariviere, V., & Gingras, Y. (2013). The distribution of references in scientific papers: An analysis of the IMRAD structure. In *Proceedings of the 14th ISSI conference* (Vol. 591, p. 603).

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the Association for Information Science and Technology*, *64*(9), 1759–1767.

Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual cocitation: Augmenting cocitation analysis and its applications. *Journal of the Association for Information Science and Technology*, *61*(6), 1130–1143.

Constantin, A., Pettifer, S., & Voronkov, A. (2013). PDFX: fully-automated PDF-to-xml conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering* (pp. 177–180). ACM.

Cronin, B. (1984). The citation process: The role and significance of citations in scientific communication. T. Graham London.
Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, *7*(3), 583–592.
Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833.
Doerfel, S., Jäschke, R., Hotho, A., & Stumme, G. (2012). Leveraging publication metadata and social data into folkrank for scientific publication recommendation. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web* (pp. 9–16). ACM.
Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the Association for Information Science and Technology*, *59*(1), 51–62.
Garfield, E. (2001). *From bibliographic coupling to co-citation analysis via algorithmic*. Griffith: A citationist's tribute to Belver C.
Garfield, E., et al. (1972). Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science.
Gipp, B., & Beel, J. (2009). Citation proximity analysis (cpa): a new approach for identifying related work based on co-citation analysis. In *ISSI09: 12th international conference on scientometrics and informetrics* (pp. 571–575).
Golshan, B., Lappas, T., & Terzi, E. (2012). Sofia search: a tool for automating related-work search. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 621–624). ACM.
Habib, R., & Afzal, M. T. (2017). Paper recommendation using citation proximity in bibliographic coupling. *Turkish Journal of Electrical Engineering & Computer Sciences*, *25*(4), 2708–2718.
Hauke, J., & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87.
Hengl, T., & Gould, M. (2002). Rules of thumb for writing research articles. Enschede, September.
Hristakeva, M., Kershaw, D., Rossetti, M., Knoth, P., Pettit, B., Vargas, S., & Jack, K. (2017). Building recommender systems for scholarly information. In *Proceedings of the 1st workshop on scholarly web mining* (pp. 25–32). ACM.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology*, *14*(1), 10–25.
Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS ONE*, *9*(5), e93949.
Lee, J., Lee, K., & Kim, J. G. (2013). Personalized academic research paper recommendation system. arXiv preprint arXiv:1304.5457.
Liu, S., & Chen, C. (2012). The proximity of co-citation. *Scientometrics*, *91*(2), 495–511.
Liu, S., & Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology*, *64*(3), 627–639.
Liu, X.Y., & Chien, B.C. (2017). Applying citation network analysis on recommendation of research paper collection. In *Proceedings of the 4th multidisciplinary international social networks conference on ZZZ* (p. 30). ACM.
Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the Association for Information Science and Technology*, *49*(6), 530–540.
McCain, K., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, *17*(1–2), 127–163.
McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., & Riedl, J. (2002). On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 116–125). ACM.
Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71.
Ratprasartporn, N., & Ozsoyoglu, G. (2007). Finding related papers in literature digital libraries. Research and Advanced Technology for Digital Libraries pp. 271–284.
Sahijwani, H., & Dasgupta, S. (2017). User profile based research paper recommendation. arXiv preprint arXiv:1704.07757.
Shahid, A., Afzal, M., & Qadir, M. (2011). Discovering semantic relatedness between scientific articles through citation frequency. *Australian Journal of Basic Applied Sciences*, *5*, 1599–1604.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology*, *24*(4), 265–269.
Small, H. G. (1976). Structural dynamics of scientific literature. *International Classification*, *3*(2), 67–74.
Smith, L. C. (1981). Citation analysis. *Library Trends*, *30*(1), 83–106.
Sugiyama, K., & Kan, M. Y. (2013). Exploiting potential citation papers in scholarly paper recommendation. In *Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries* (pp. 153–162). ACM.
Teufel, S. (2009). Citations and sentiment. In *Workshop on text mining for scholarly communications and repositories*, University of Manchester.
Voos, H., & Dagaev, K. S. (1976). Are all citations equal? or, did we op. cit. your idem? *Journal of Academic Librarianship1*(6), 19–21.
Wang, Y., Zhang, H., Li, Y., Wang, D., Ma, Y., Zhou, T., & Lu, J. (2016). A data cleaning method for citeseer dataset. In *International conference on web information systems engineering* (pp. 35–49). Springer.