Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 2708 – 2718

doi:10.3906/elk-1608-180

Paper recommendation using citation proximity in bibliographic coupling

Raja HABIB*, Muhammad Tanvir AFZAL
Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan

Received: 16.08.2016 • **Accepted/Published Online:** 13.10.2016 • **Final Version:** 30.07.2017

Abstract: Research paper recommendation has been a hot research area for the last few decades. Thus far, numerous different paper recommendation approaches have been proposed. Some of these include methods based on metadata, content similarity, collaborative filtering, and citation analysis, among others. Citation analysis methods include bibliographic coupling and co-citation analysis. Much research has been done in the area of co-citation analysis. Researchers have also performed experiments using the proximity of in-text citations in co-citation analysis and have found that it improves the accuracy of paper recommendation. In co-citation analysis, the similarity is discovered based on the frequency of co-cited papers in different research papers and those citing papers may belong to different areas. However, when proximity is used to calculate co-citation, the accuracy of recommendations improves significantly. Bibliographic coupling finds bibliographic coupling strength based on the common references between two papers. In bibliographic coupling, a large number of common references of two papers means that they belong to the same area, unlike co-citation analysis, in which there is a possibility that the citing papers may belong to different areas. Based on the observation that with the use of proximity analysis the accuracy in cases of co-citation analysis has improved, this paper investigates if the accuracy of paper recommendation can be further improved by using proximity analysis in bibliographic coupling. This paper proposes an approach that extends the traditional bibliographic coupling by exploiting the proximity of in-text citations of bibliographically coupled articles. The proposed approach takes into account the proximity of in-text citations by clustering the in-text citations using a density-based algorithm called DBSCAN. Experiments on a data set of research papers are presented to show that there is a substantial increase in accuracy of the recommendations produced by DBSCAN based on proximity analysis of in-text citations compared to traditional bibliographic coupling and content-based approaches.

Key words: Paper recommendation, bibliographic coupling, citation proximity analysis, DBSCAN

1. Introduction

Over the last few decades, research paper recommendation has emerged as a very hot research area with a wide range of applications such as citation recommendation and discovering relevant papers. A myriad of papers continue to be published on this topic [1]. Modern researchers find it tedious to get access to the required research papers in their particular fields, due to the information overload and overabundance of publications in conferences and journals. According to a study [2], there are almost 25 million freely available scholarly documents on the web. Research paper recommendation is one promising approach to tackle this information overload problem.

Different paper recommendation techniques have been proposed and implemented by the scientific com-

*Correspondence: r_habib_pk@yahoo.com

munity, considering the worth of research paper recommendation. These methods include metadata, content-based filtering, collaborative filtering, co-citations, and bibliographic coupling, among others.

The approaches centered on citation analysis have proven to be very significant. These include co-citation analysis, bibliographic coupling, and direct citations. Citations have long been referred to as a very productive and potentially fruitful source in many different areas of scientific research. The applications of citation analysis range from research evaluation to paper recommendation. Kuhn et al. [3] conducted comprehensive experiments and analyzed the inheritance patterns in citation networks to determine the memes in the scientific literature. Scientific memes are parts of the text that exist in a publication and get replicated in the citing papers. They discovered that there exists a relation between the occurrence of scientific memes and the degree to which they propagate along the citation graph. Similarly, Matja et al. [4] analyzed how the Matthew effect applies to the citation data. The Matthew effect refers to the phenomenon where "the rich get richer and the poor get poorer." They performed extensive experimentation to conclude that there is a preferential attachment in citation networks and that the publications that have a larger number of initial citations will receive many more in the future, as compared to the publications with a smaller number of the original citations. This preferential attachment can be used to recommend papers. The papers that acquire more citations tend to be favored more. In other research, Matja et al. [5] found that a power law can fit the citations distributions from the individual publications and that the citation data can also be used to evaluate the scientific research output.

Co-citation analysis [6] considers two papers similar if both have been cited by one or more common papers. Numerous techniques have been proposed that use the co-citation analysis. However, in co-citation analysis, papers are recommended based only on the fact that one or more common citing papers have cited the recommended articles. For example, if a paper 'A' cites two papers 'B' and 'C', then papers 'B' and 'C' are considered to be relevant or similar. However, the contents or any other features of the papers are not taken into account when determining their similarity. This feature can lead, thus, to inconsistencies or faulty recommendations. Figure 1 shows how co-citation analysis approach works.

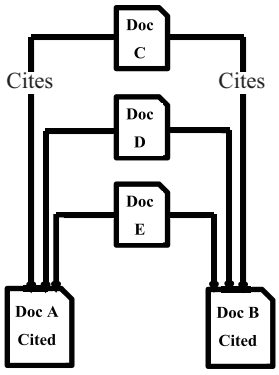

Figure 1. Co-citation analysis. Doc A and Doc B are co-cited and the co-citation frequency is 3, since both have been cited by 3 common documents Doc C, Doc D, and Doc E.

Another possible technique to discover related publications to a particular article is bibliographic coupling [7]. It uses citation analysis to determine the relationship between documents, for instance if two research papers both cite one or more common research papers. Bibliographic coupling uses coupling strength, which has, at the basis of its calculations, the number of common citations from both papers. For example, if papers 'A' and 'B' both cite 'C,' 'D', and 'E,' then papers 'A' and 'B' have a bibliographic coupling strength of three. The larger this number, the higher is the overlap between two papers' bibliography and the bibliographic coupling

strength between them. Unlike the co-citation approach, in bibliographic coupling, the references of the cited papers are taken into account while determining the similarity (Figure 2). In this technique, as well as the ones mentioned above, the logical structure of the paper and the occurrence of citations in the full text of the papers are ignored. Another problem with bibliographic coupling is that there are significant cases in which the references are included in the "References" section of the paper but are never used inside the full text of the paper. Shahid et al. [8] identified more than 10% such references in more than 16,000 references of the JUCS, which were included in the reference section but never in the main text of citing documents. Such citations are called false citations. Therefore using only bibliographic coupling strength may lead to incorrect results.

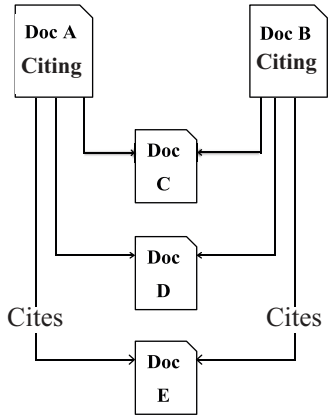

Figure 2. Bibliographic coupling. Doc A and Doc B are bibliographically coupled and the bibliographic coupling strength is 3 since they cite 3 common documents Doc C, Doc D, and Doc E.

This paper proposes a new paper recommendation approach that extends the traditional bibliographic coupling by exploiting the in-text citations and their proximities. In other words, this approach takes into account both the bibliographic coupling strength and the patterns of the in-text citations within the whole content of the paper. This approach focuses on exploring the citing patterns and the normalized proximity of in-text citation tags in the bibliographically coupled pairs. The proposed approach clusters the in-text citations using a density-based clustering approach called DBSCAN.

Experiments were conducted to show that when using DBSCAN for citation proximity analysis of bibliographically coupled papers, better recommendations are produced than with traditional bibliographic coupling and content-based approaches. A dataset including 320 bibliographically coupled pairs was used to evaluate the proposed approach. The results of the proposed method were compared to those of the bibliographic coupling approach and the content-based approach.

The results showed an increase in accuracy of paper recommendation from 20% (content similarity approach) and 45% (bibliographic coupling) to 55% (DBSCAN). Manual evaluation was also performed on a dataset of 50 bibliographically coupled research papers. The research papers were categorized as very strongly related, strongly related, or weakly related in the first step of the manual evaluation. Our proposed approach was then applied to these papers in the second steps. The results showed that there was an exact match in the categorization of research papers for 67% of the data set.

The remainder of this paper is structured as follows. We discuss the state-of-the-art literature review of the existing techniques in section 2. We discuss the DBSCAN algorithm in section 3. We discuss the details of our proposed approach in section 4. Section 5 discusses the results of our approach. At the end section 6 discusses the conclusion and future work.

2. Related work

According to a survey [1], 200 different approaches have been proposed for paper recommendation. Fifty percent of these approaches applied content-based filtering, 18% applied collaborative filtering, and 16% used graph-based techniques.

Research paper recommendation techniques can be placed into different categories based on the similarity measures. These include: (1) metadata-based approaches [9,10], (2) citation-based approaches [6,7,11], (3) content-based approaches [12,13], (4) collaborative filtering (CF)-based approaches [14], (5) user profile-based approaches [15,16], and (6) hybrid approaches. In the metadata-based approaches [9,10], the similarity between scientific articles is calculated by matching the metadata of papers, such as title, authors' names, journal, date of publication, and key words. The advantage of these approaches is that the metadata of the papers are available freely and openly. However, the metadata approaches may sometimes provide incorrect recommendations. For example, this may be the case when the same author may have published papers in different research areas.

Citation analysis is performed in different ways like bibliographic coupling [7] and co-citation analysis [6]. Research papers normally cite the papers that are related to them and so the relationships found using citations are usually meaningful. However, using citations alone, while ignoring the content of the research papers, may lead to incorrect results. For instance, some researchers cite a paper in the references section without actually using them in the main content of the paper. Such citations prove useless, thus. Similarly, the relevant papers that have not been cited may not be discovered from such approaches.

The content-based techniques use content similarity techniques to measure the relatedness of two papers [12,13]. Using content to determine the similarity between research papers provides much better results compared to using only the citations. The limitation of these approaches, however, is that the full content of papers is usually not available in certain digital libraries. Another limitation is that the processing of the whole content of the papers can prove to be very costly and time consuming.

Another approach to finding similarity between research papers is to use collaborative filtering [14]. This approach is used in many recommending systems [17,18]. In this technique, a user-item matrix is generated. In the case of scientific paper recommendation, the citation network is converted into a paper-citation matrix that is analogous to the user-item matrix. Collaborative filtering suffers from certain problems. One of these problems is the so-called "cold start problem". Moreover, papers are recommended to users based on their previous preferences. Thus, when a new user joins a recommending system, he/she needs to rate a certain number of papers before more can be recommended to other users based on his/her ratings. Similarly, when a new paper is added to the system, it needs to be rated by a certain number of users before it can be recommended to the other users. Another issue faced by the collaborative filtering based approaches is scalability. That is, this approach works particularly well when the number of users and papers is usually massive, but not otherwise. This leads to huge computation costs using collaborative filtering.

Another approach uses user profiles and access-log history to recommend scientific papers based on interests and usage behaviors [15,16]. These approaches are dependent upon the availability of usage profile information in digital libraries. Without the access to sufficient usage information, these approaches do not provide desirable results. For instance, when some digital libraries provide access only to a small or absolutely no part of the usage data.

The approaches based on citation analysis usually provide better results compared to the ones mentioned above. The two main approaches based on citation analysis are co-citation analysis and bibliographic coupling. These are discussed in the previous section.

Nassiri et al. proposed an approach based on citation analysis [19]. They proposed a technique called the normalized similarity index (NSI) to measure the similarity between two papers. In NSI, three types of citation relationships are considered: co-citations, bibliographic coupling, and longitudinal coupling. Longitudinal coupling refers to the indirect citations between two papers, i.e. the two papers are connected through some other interconnected papers. They calculated NSI for five different citation networks. They compared the results with the peer reviews. There was a high correlation between the two results. They also compared NSI with combined linkage (CL) and weighted direct citation (WDC) of those five networks. NSI provided much better results.

Gipp et al. proposed an approach called citation proximity analysis (CPA) to find the related papers [20]. Along with the citation analysis, the authors used the proximity between citations to discover the relatedness. They found out that the closer the citations are to each other, the more related the two papers are. For example, if two citations occur within the same sentence, they are considered to be more related to each other than if they occurred in two different paragraphs or two different sections. Using the position of citations, the type of paper can also be discovered, e.g., whether the paper is a state-of-the-art paper. This technique provides better results compared to other techniques that use traditional co-citation.

3. DBSCAN clustering algorithm

In our proposed algorithm we use a density-based clustering approach called DBSCAN to discover the clusters of in-text citations. DBSCAN generates clusters based on the density of the items in the item set and the algorithm uses the two parameters: Eps and minPts. Eps represents the radius and minPts represents the minimum points required within Eps. A point is called a core point if it has at least minPts number of points within its Eps. The points within the Eps of a core point are called directly density-reachable points. The point is called indirectly density-reachable if it does not lie within the Eps of a core point, but when there is a path of points from the core point of it. A point that is not reachable from the core point and is not a core point itself is called an outlier. The core points and the reachable points make the clusters. Figure 3 shows an example of the DBSCAN algorithm.

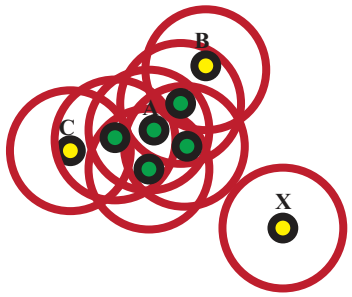

Figure 3. Clustering using DBSCAN. In this figure, minPts = 3. Green points are core points since they have at least 3 points within ε radius. All these points make a single cluster since they are all reachable from one another. Points B and C are not core points but are indirectly reachable from point A, and so they are also included in the same cluster. The point X, however, is an outlier since it is neither a core point nor is it density reachable.

4. Proposed approach

Figure 4 shows the block diagram of our proposed research paper recommender system. The main modules of our system are data acquisition, data normalization, DBSCAN clustering, and similarity score measuring.

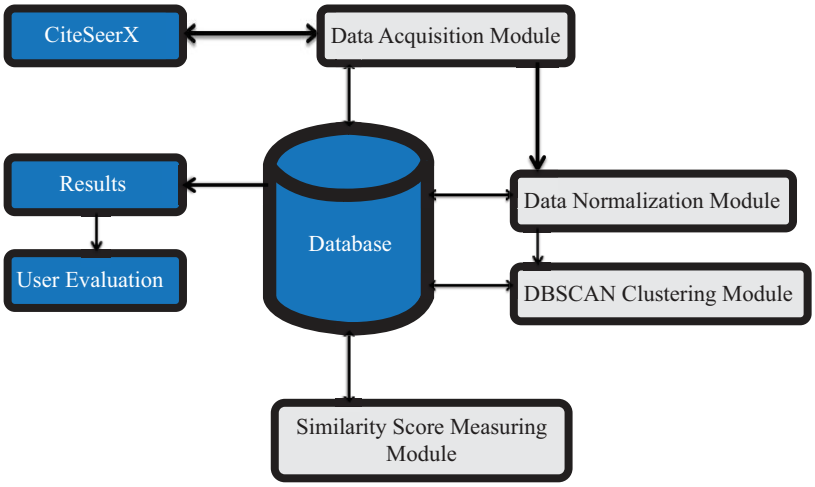

Figure 4. System overview. Proposed system consists of modules such as data acquisition, data normalization, DBSCAN clustering, and similarity score measuring.

In the data acquisition module, we gathered data from CiteSeerX, which is an online digital library. We used a focused web crawler to fetch the DOIs (digital object identifier) for all the research papers that showed in the list of results when we posed certain queries (key words such as data mining and computer architecture) on CiteSeerX. We determined the bibliographically coupled papers using the CIDs of papers. In the next step, we stored all the bibliographically coupled papers in an SQL database.

We then deployed the data acquisition module to collect 1150 documents for which we had previously collected the DOIs. These bibliographically coupled papers were then downloaded in PDF format with some user assistance. After downloading the papers, we extracted the proximities of all the in-text citations. Because the length of research papers varies, we first had to normalize the proximities of in-text citations. For this we used the min-max normalization, which is given by

$$New\_v = \left[\frac{v - MinX}{MaxX - MinX}\right](New\_MaxX - New\_MinX) + New\_MinX \tag{1}$$

In this equation, v represents the location of the citations. Table 1 shows the original values of in-text citations and the normalized valued of in-text citations.

Table 1. Data normalization.

Citation 1	Citation 2	Citation 3	Total words	Normalized proximity 1	Normalized proximity 2	Normalized proximity 3
16,439	8193	1427	22,238	739	369	65
12,283			13,987	878		
6881			8299	829		
7731			9598	806		
6662	228		14,866	449	16	
2238	1917	256	5792	387	332	45
1657			9523	175		
10,636	10,070	6693	11,728	907	859	571

In the next step, the normalized values of in-text citations were provided as input to the DBSCAN clustering module. We used WEKA to perform the clustering of in-text citations. We then imported the list of normalized in-text citations of the bibliographically coupled papers to WEKA. For this step, we considered minPts = 2 and $\varepsilon = 200$. As an output, we received the clusters.

Using these clusters, we subsequently calculated the similarity between research papers, based on the proximity of the in-text citations. Suppose the papers 'X1,' 'X2', and 'X3' cite papers 'A,' 'B,' and 'C' as shown in Table 2. A1 represents the first citation of paper 'A' in papers 'X1', 'X2', and 'X3'. A2 represents the second citation of paper 'A' from the papers 'X1', 'X2', and 'X3' and so on. In Table 2, the points (X1, A1), (X2, A1), (X3, A1), (X1, A2), (X3, A2), and (X2, A2) are core points with respect to paper 'A'. We obtained the following three clusters for paper 'A':

Table 2. DBSCAN clustering for citation proximity.

	A1	A2	A3	B1	B2	B3	C1	C2	C3
X1	100	600	2000	400			700		
X2	200	1800					400		
X3	500	800	1500	600	800	1500			

1. (X1, A1), (X2, A1)
2. (X3, A1), (X1, A2), (X3, A2) and
3. (X2, A2) and (X1, A3)

Similarly, for paper 'B', we obtained the following clusters:

1. (X1, B1), (X3, B1), (X3, B2)

The points (X3, A3), (X3, B3), (X1, C1), and (X2, C1) were considered outliers. Table 2 shows that in the clusters based on citations of paper 'A' there are more citations from paper 'X1' and 'X2' that are within the same proximity as compared to citations from paper 'X3'. Accordingly, 'X1' and 'X2' were considered more similar to each other with regards to paper 'A.' Similarly, based on citations of paper 'B,' we can see that papers 'X1' and 'X3' are more similar to each other as compared to paper 'X2'.

5. Results and analysis

There is no particular way of evaluating research paper recommenders. However, user studies have been conducted by some researchers to evaluate these [1,15,19]. We also used user studies and manual evaluation of our proposed approach.

To find and download these documents, we posed the queries to CiteSeerX such as "computer architecture," "distributed databases," "data mining," "information retrieval", and "intrusion detection." Using the crawler, we extracted the metadata of these research documents as discussed in the previous section. These metadata consist of information like the title of the paper, the list of authors' names, abstract, and references. In the next step, we fetched the in-text citations and their proximities. We then normalized the values of proximities of the in-text citations. Then we clustered the in-text citations using WEKA. Initially, we started with a dataset of 1150 research documents. However, some of these papers could not be adequately converted to

XML format. As a result, we ended up with a dataset containing 703 documents, which were bibliographically coupled.

We used a sample of 320 bibliographically coupled pairs. We divided these into sets of 10 papers. Each set consisted of 10 related papers.

We compared our approach with the traditional bibliographic coupling approach and with the content-based approach using a gold standard dataset. The dataset consisted of 320 bibliographically coupled pairs. Every paper was evaluated by two distinct users. For each paper, the interrater agreement was calculated between the users by using Spearman's correlation coefficient [21]. This agreed ranking by both users was considered to be the gold standard ranking. Figure 5 shows the interrater agreement between the two users for all 32 documents.

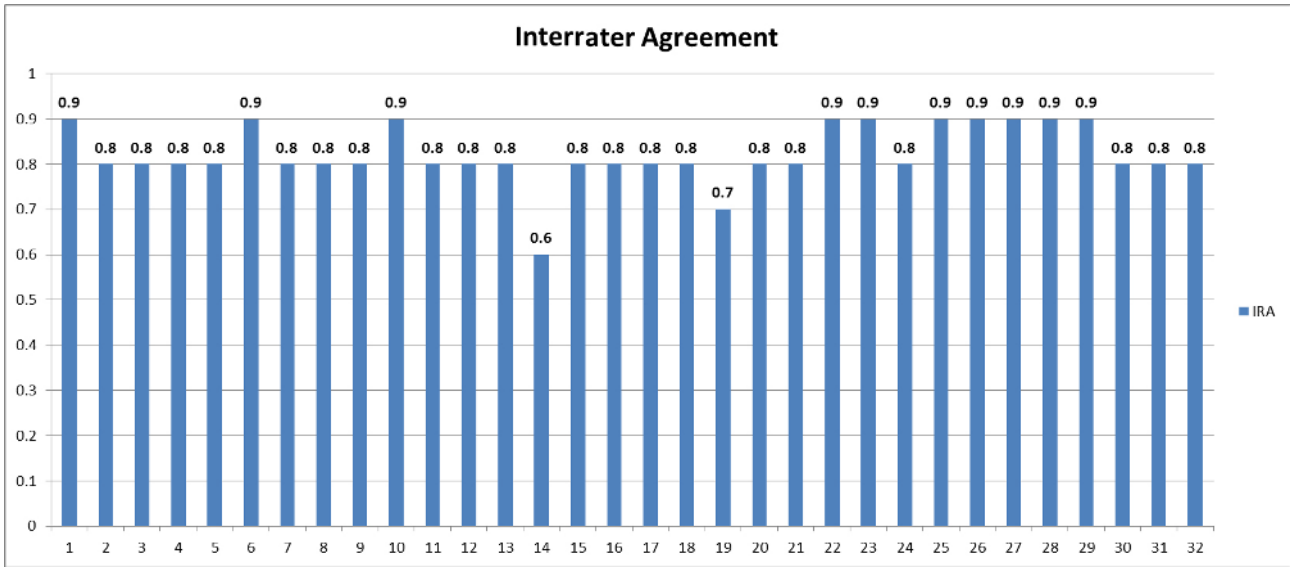

Figure 5. Interrater agreement. The analysis of the interrater agreement shows that the level of agreement between both users for 320 coupled pairs is on average at 81%.

Using this gold standard dataset, the proposed approach was compared with the bibliographic coupling approach and the content-based approach. This comparison was also done using Spearman's correlation coefficient. Figure 6 shows the comparison between the proposed approach and the bibliographic coupling approach. We found a higher correlation between the users' opinion and our proposed approach using DBSCAN as compared to the traditional bibliographic coupling approach for the majority of the documents used. Namely, average correlation improved by 22%.

We also compared the performance of our approach with the content-based approach. As shown in Figure 7, our proposed approach using DBSCAN performed better for the majority of the documents when compared to the content-based approach. The average correlation using DBSCAN with the gold standard ranking was 0.55, whereas the average correlation of the content similarity with the gold standard ranking was 0.20. Figure 8 shows the improvement in accuracy achieved by our proposed approach.

The proposed approach was evaluated manually too. A dataset composed of 50 bibliographically coupled research papers was used for manual evaluation. This dataset was divided into five subsets, each containing ten bibliographically coupled papers. These papers were manually categorized as very strongly related, strongly related, and weakly related, based on how much they were related to the query paper. Our proposed approach

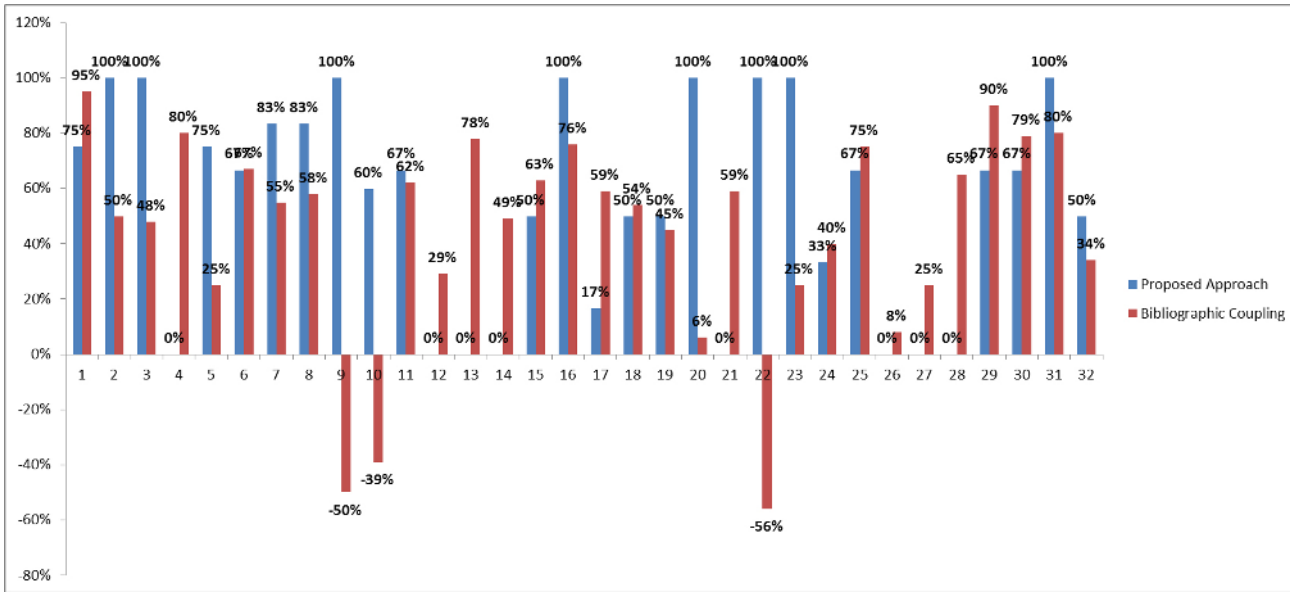

Figure 6. Proposed approach vs. bibliographic coupling. Agreement between users opinion and the proposed approach is better compared to bibliographic coupling for the majority of the documents out of 32 papers. ?

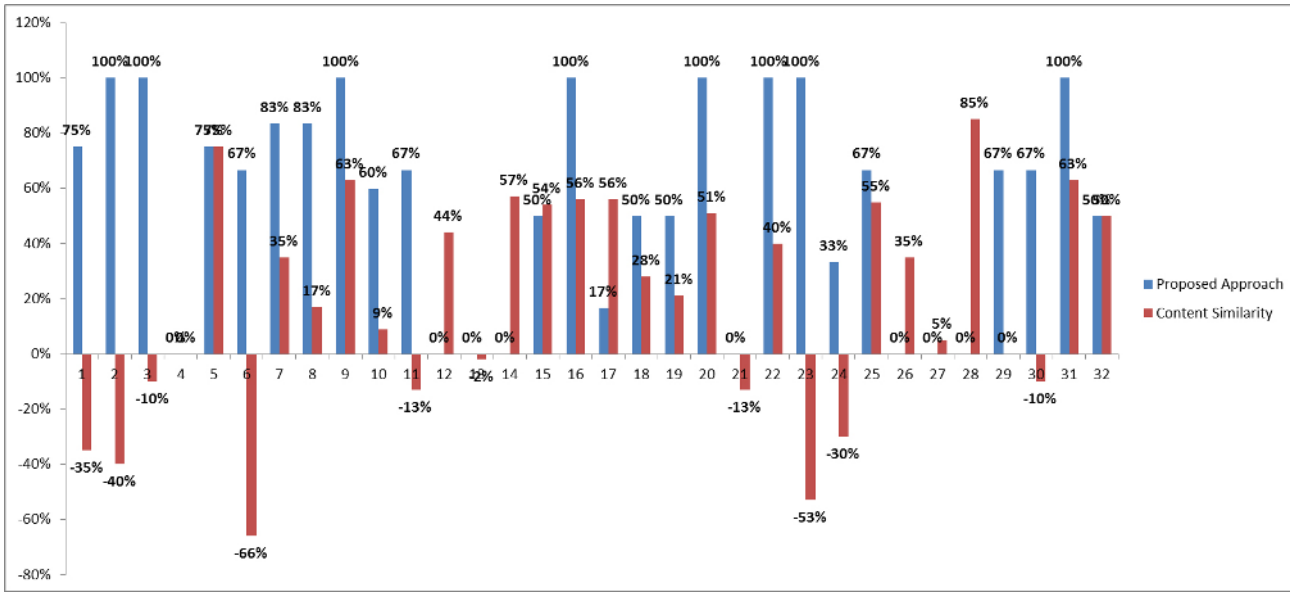

Figure 7. Proposed approach vs. content similarity. Agreement between users opinion and the proposed approach is better compared to content similarity for the majority of the documents out of 32 papers.

was then applied to these papers. Our approach ranked the papers in each dataset from 1 to 9. The papers with rankings 1–3 are categorized as very strongly related. The papers with rankings 4–6 are categorized as strongly related. The papers with rankings 6–9 are categorized as weakly related. The results of our proposed approach matched those of the manual evaluation for 30 out of the total 45 research papers. The accuracy of our proposed approach was 67% compared with manual evaluation.

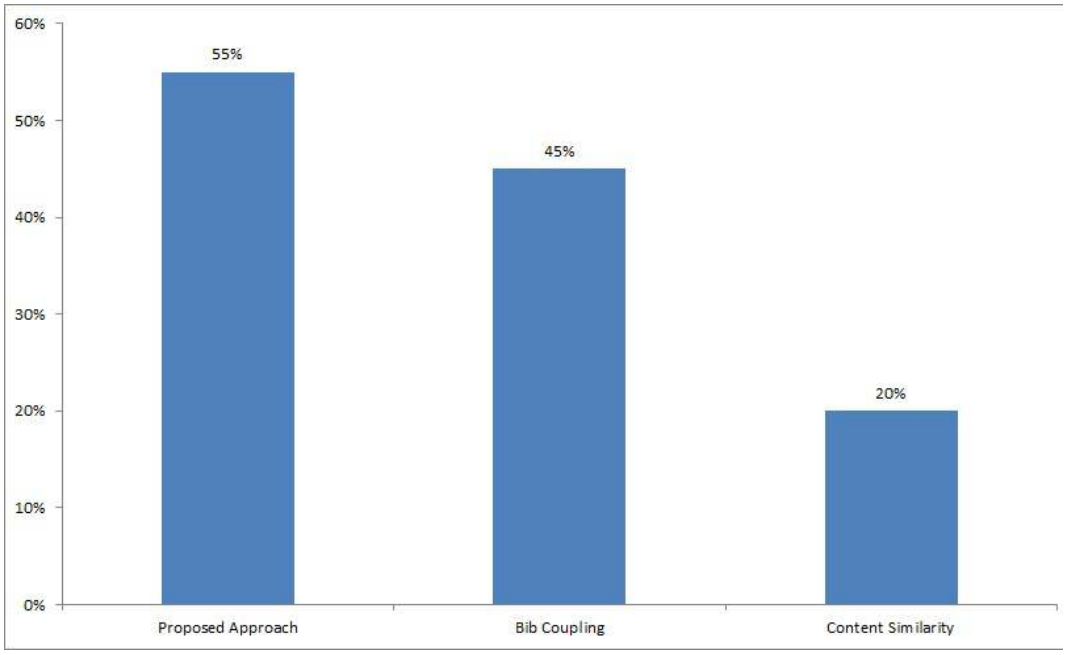

Figure 8. Improvement achieved with the proposed approach. The average correlation of the proposed approach with the gold standard ranking is 0.55, whereas the average correlation of the bibliographic coupling and content similarity with the gold standard ranking is 0.45 and 0.20, respectively.

6. Conclusion

Research paper recommender systems are becoming increasingly important for researchers due to the information overload. In this paper, we introduced a bibliographic coupling-based paper recommendation system that uses in-text citation proximity. We used the DBSCAN algorithm to cluster the in-text citations. We performed experiments on a dataset of bibliographically coupled research papers. Our experiments show that the accuracy of the recommendations, produced by DBSCAN-based proximity analysis of in-text citations, increased substantially, compared to traditional bibliographic coupling and content-based approaches. In the future, it would be interesting to see how other clustering algorithms would perform for these experiments, especially the cobweb clustering algorithm.

The proposed approach using the DBSCAN algorithm outperformed all main standard methods for calculating paper relatedness currently in use. At the same time, this algorithm circumvents some of the problems faced by other methods. The DBSCAN is thus a promising tool for a new generation of software and web platforms providing literature and reference recommendations to scientists and research across the world.

Acknowledgments

This research was supported by Higher Education Commission (HEC) of Pakistan. We thank the Capital University of Science and Technology Islamabad for assistance. We thank Ansar Mehmood for assistance with compilation of the gold standard dataset and for important recommendations and comments during this research and previous versions of the manuscript.

References

[1] Beel J, Langer S, Genzmehr M, Gipp B, Breitinger C, Nürnberger A. Research paper recommender system evaluation: a quantitative literature survey. In: Proc. International Workshop on Reproducibility and Replication in Recommender Systems Evaluation; 12 October 2013: ACM. pp. 15-22.

[2] Khabsa M, Giles CL. The number of scholarly documents on the public web. PloS one 2014; 9: e93949.

[3] Kuhn T, Perc M, Helbing D. Inheritance patterns in citation networks reveal scientific memes. Phys Rev X 2014; 4: 041036.

[4] Perc M. Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. J Informetr 2010; 4: 358-364.

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

[6] 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.

[7] Kessler MM. Bibliographic coupling between scientific papers. Am Doc 1963; 14: 10-25.

[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] Doerfel S, Jäschke R, Hotho A, Stumme G. Leveraging publication metadata and social data into folkrank for scientific publication recommendation. In: Proc. 4th ACM RecSys workshop on recommender systems and the social web; 9 September 2012; ACM. pp. 9-16.

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

[11] Garfield E. Citation analysis as a tool in journal evaluation. Science 1972; 178: 471-479.

[12] Ratprasartporn N, Ozsoyoglu G. Finding related papers in literature digital libraries. In: International Conference on Theory and Practice of Digital Libraries; 16 Sep 2007; Berlin, Germany: Springer. pp. 271-284.

[13] Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C. Content-based citation analysis: the next generation of citation analysis. J Assoc Inf Sci Technol 2014; 65: 1820-1833.

[14] McNee SM, Albert I, Cosley D, Gopalkrishnan P, Lam SK, Rashid AM, Konstan JA, Riedl J. On the recommending of citations for research papers. In: Proc. 2002 ACM conference on computer supported cooperative work; 16 November 2002: ACM. pp. 116-125.

[15] Lee J, Lee K, Kim JG. Personalized academic research paper recommendation system. In: arXivl; 19 April 2008; preprint arXiv:1304.5457.

[16] Sugiyama K, Kan MY. Scholarly paper recommendation via user's recent research interests. In: Proc. 10th annual joint conference on digital libraries; 21 June 2010: ACM. pp. 29-38.

[17] Goldberg K, Roeder T, Gupta D, Perkins C. Eigentaste: a constant time collaborative filtering algorithm. Inform Retrieval 2001; 4: 133-151.

[18] Hongyan P, Hongfei L, Jing Z. Collaborative filtering algorithm based on matrix partition and interest variance [J]. J China Soc Sci Tech Info 2006; 1: 008.

[19] Nassiri I, Masoudi-Nejad A, Jalili M. Moeini A. Normalized similarity index: an adjusted index to prioritize article citations. J Informetr 2013; 7: 91-98.

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

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