A Citation-Based Recommender System for Scholarly Paper Recommendation

Khalid Haruna[1,2](✉), Maizatul Akmar Ismail[1],
Abdullahi Baffa Bichi[2], Victor Chang[3], Sutrisna Wibawa[4],
and Tutut Herawan[4,5,6]

[1] Department of Information Systems,
Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, Malaysia
[2] Department of Computer Science,
Faculty of Computer Science and Information Technology,
Bayero University, Kano, Nigeria
kharuna.cs@buk.edu.ng
[3] IBSS Xi'an Jiaotong Liverpool University, Suzhou, China
[4] Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[5] Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
[6] Politeknik Negeri Malang, Malang, Indonesia

Abstract. Several approaches have been proposed to help researchers in acquiring relevant and useful scholarly papers from the enormous amount of information (information overload) that is available over the internet. The significant challenge for those approaches is their assumption of the availability of the whole contents of each of the candidate recommending papers to be freely accessible, which is not always the case considering the copyright restrictions. Also, they immensely depend on priori user profiles, which required a significant number of registered users for the systems to work effectively, and a stumbling block for the creation of a new recommendation system. This paper proposes a citation-based recommender system based on the latent relations connecting research papers for the scholarly paper recommendation. The novelty of the proposed approach is that unlike the existing works, the latent associations that exist between a scholarly paper and its various citations are utilised. The proposed approach aimed to personalise scholarly recommendations regardless of the user expertise and research fields based on paper-citation relations. Experimental results have shown significant improvement over other baseline methods.

Keywords: Contextual information · Paper-citation relations
Publicly available metadata · Recommender system

1 Introduction

Results of various academic findings are disseminated in the forms of journal articles, conference proceedings, seminars, symposia, theses and etcetera [1], to serve as guidelines for the use of future generations. However, the voluminous amount of this information makes information seeking process very much wearisome [2, 3].

O. Gervasi et al. (Eds.): ICCSA 2018, LNCS 10960, pp. 514–525, 2018.
https://doi.org/10.1007/978-3-319-95162-1_35

The use of the generic search engines when searching for related information over the internet has become the most common and convenient method among researchers [4]. A reasonable level of expertise needs to be achieved to locate relevant and promising information efficiently [5]. Additionally, researchers follow the list of references to the papers they have already possessed for more explorations [6]. However, the coverage of this approach is insufficient and cannot trace the papers that are published after the possessed documents [4].

On the other hand, digital libraries such as IEEE, ScienceDirect, and SpringerLink, can provide proactive systems capable of recommending scholarly papers that match researcher's interests in a timely fashion [7]. Fortunately, they require considerable attention from the users to explicitly state their interests, which is tedious and take up much of researcher's valuable time.

To solve the above problems, research paper recommender systems have been proposed [6–15], to recommend scholarly papers to individual researchers proactively. The challenge is to provide relevant papers to the right researchers in the right way [4]. However, the vital concern to these approaches is that they presumed the whole contents of each of the candidate recommending papers to be freely accessible, which is not always the case considering the copyright restrictions. Furthermore, the approaches largely depend on priori user profiles, which required some registered users for the systems to work effectively, and a stumbling block for the creation of new research paper recommender system.

While there are lots of approaches based on citation-relations for scholarly paper recommendations [16–20], they do not leverage the latent relations across research papers, instead employed direct relations such as the co-citation relations presented in [20]. Identifying and incorporating the latent relations across research papers could play a significant role and improve the recommendation performance.

An initial approach to solving the above problems has been proposed in [6]. The authors mined the hidden relation between a target paper and its references to present utile recommendations. Differently, the hidden association between a target paper and its citation relations is leveraged in this paper. The novelty of the proposed approach are twofold;

a. Firstly, an independent research paper framework that utilises public contextual metadata to personalise scholarly papers regardless of the user expertise and research field is proposed.
b. Secondly, the proposed approach does not require a priori user profile.

The remaining sections of this paper are as follows. Section 2 presents some related work on recommending research papers. Section 3 presents the proposed citation-based recommender system for the scholarly paper recommendation. Section 4 describes the experimental setup and discusses the experimental results. Section 5 concludes the paper.

2 Related Works

The pattern of information seeking behaviour among different researchers has been reviewed in [21, 22]. Their findings reveal that expert researchers are more proficient in using search engines as compared to novice researchers. While [23] discussed the processes of identifying researchers' information need and in [24], a positive step in associating researcher's information seeking behaviour with the design of an ideal system has been reported.

Research paper recommender systems have also been utilised [7–14, 16] to ease the tasks of information seeking process, by suggesting relevant scholarly papers based on some information that is more elaborate than a few keywords. Different researchers have proposed different use of user-provided information. To be specific, [16] explored the use of collaborative filtering approach to recommend scholarly papers to a researcher from the set of citations to one of his/her papers. The aim was to test the ability of the collaborative filtering approach in recommending some set of citations that would be much significant as additional references to a target paper. The experimental results across six different algorithms reveal that the choosing algorithm affects the recommendation results. Also, some algorithms provide either very novel recommendations or very much relevant recommendations, but no single algorithm achieved both.

A citation-network has been explored in [16], to enhance the recommendation performance. However, the approach generates sparsity problem in the paper-citation network and thereby making the recommendation process very much tricky. In alleviating the sparsity problem, [12] applied the concept of the collaborative filtering approach to identify potential citation papers from the list of papers authored by a researcher. The experimental results show that recommendations after discovering the potential citation papers are more effective than collaborative filtering with binary or similarity values. Still, the approach generates poor prediction results for multidisciplinary scholars that work on several research topics. The research was later extended in [25] to cater the problem of multidisciplinary problems by proposing an adaptive neighbour selection. Also, the authors investigate the different sections of scholarly papers to find a better and adequate representation. The best result was achieved by considering the full paper text and conclusion and thus, can serve as a better representation of scholarly papers.

In the above systems, those initial information provided by the users are used to represent their interests in user profiles, and the system searches for similar items to make recommendations. The main weakness of those methods is that they presumed the whole contents of each of the candidate recommending papers to be freely accessible, which is not always the case considering the copyright restrictions.

Different from the above researches, [11] proposed a framework that generates potential queries using terms from only publicly available metadata, title and abstract of a target paper. The approach then applies the content-based approach to rank recommending papers that are more related to the target paper. The authors in [6] have also utilised the only publicly available contextual metadata using the concept of a collaborative approach to mine the hidden relations between a target paper and its references to present essential recommendations. While [6] utilised paper-reference relations, in this paper, the latent associations that exist between paper-citations relations are leveraged to personalise recommendations regardless of the user expertise and research field.

Based on a depth study of existing related works above, the problem is defined as follows:

Given a target paper p_i as a query, extract all the set of citations Cf_j of the target paper p_i. For each of the citations Cf_j, retrieve all other papers p_{ri} that reference. Measure the extent of similarity $W^{p_i \to p_{ri}}$ between p_{ri} and p_i, recommend the top- N most similar papers.

```
Algorithm:  A Citation-Based Recommender System
Input: Target Paper
Output: Top-N Recommendation

Given a target paper $p_i$ as a query,
    (1) Extract the set of its citations $Cf_j$.
    (2) For each citations $Cf_j$, retrieve all other papers $p_{ri}$ that $Cf_j$ referenced.
    (3) Measure the extent of similarity $W^{p_i \to p_{ri}}$ between $p_{ri}$ and $p_i$.
    (4) Recommend the top-N papers.
```

Algorithm 1: A Citation-Based Recommender System

3 Proposed Citation-Based Recommender System for Scholarly Paper Recommendation

The proposed approach starts by transforming the corpus into a paper-citation matrix. Rows of the matrix represent the candidate papers, and columns denote the citations (see Table 1). A target paper (*Pi*) is defined as the paper to which a researcher has possessed and wants to receive other recommendations similar to it. Upon receiving the user's query, the proposed approach identifies the target-paper from the paper-citation matrix and Algorithm 1 is then applied. The algorithm extracts the target paper's citations, and for each citation, it retrieves from the web other papers that referenced any of the target papers citations. Equation 1 is then used to measure the extent of similarity between the target paper and each of the retrieved papers. Finally, it recommends the top-N papers to the researcher.

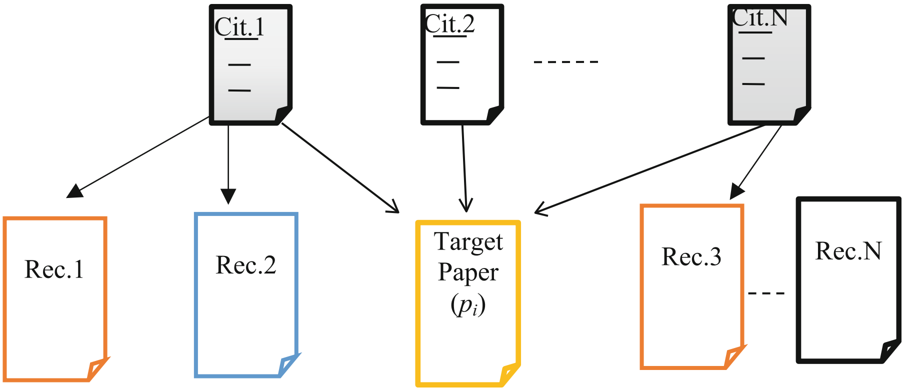

Fig. 1. Proposed citation-based recommendation scenario

To understand the proposed approach clearly, Fig. 1 portrays a target-paper (p_i) with citations (*Cit.1* to *Cit.N*), in which each of the citations has referenced some set of other papers (*Rec.1* to *Rec.N*). The goal is to measure the extent of similarity ($W^{p_i \rightarrow \mathrm{Re}\,c.j}$) between the target-paper (*Pi*) and each of the co-referenced papers ($R_{ec.j}$). To do this, the contextual relations between the target paper and its neighbouring papers are mined to transform the paper-citation matrix into a relational matrix to represent the target paper (*Pi*) concerning each of its neighbouring papers ($R_{ec.j}$). The top-*N* recommendation list from the results of these associations is then presented to the user.

Table 1. Paper-citation relation matrix

Paper/citation	c_1	c_2	c_3
p_i	1	1	1
p_1	1	1	–
p_2	–	–	1
p_3	–	1	1
p_4	–	–	1
p_5	1	–	–

For illustration, assume that a target paper (*Pi*) is identified from the user's query and arrived at the paper-citation relations matrix depicted by Table 1 after extracting all other references ($R_{ec.j}$) to the target paper's citations. To get the relationship between the target paper (*Pi*) and each of the neighbouring papers ($R_{ec.j}$), a single-role relational matrix is obtained from the double-role relational matrix as depicted in Table 2. For simplicity, two papers are considered significantly co-occurring if both have at least a common cited-paper. Additionally, a binary value of one (1) or zero (0) is used to state the co-occurrence or otherwise between two citing papers. Equation (1) is then applied to measure the extent of similarity ($W^{p_i \rightarrow \mathrm{Rec}.j}$) between (*Pi*) and ($R_{ec.j}$).

Table 2. Pair-wise paper similarity matrix

	p_i	p_1	p_2	p_3	p_4	p_5
p_i	1	1	1	1	1	1
p_1	1	1	0	1	0	1
p_2	1	0	1	1	1	0
p_3	1	1	1	1	1	0
p_4	1	0	1	1	1	0
p_5	1	1	0	0	0	1

3.1 Similarity Measure

In identifying similar research papers to the target paper (*Pi*), it becomes imperative to not only consider how similar the candidate recommending papers are to the target paper (*Pi*) but also how they deviate. It is therefore felt as in [26], that Jaccard similarity coefficient *J* given by Eq. (1) is more suitable for measuring the similarity and diversity between (*Pi*) and ($R_{ec.j}$).

$$J = W^{P_i \rightarrow \mathrm{Re}\,c.j} = \frac{Z_{11}}{Z_{01} + Z_{10} + Z_{11}} \tag{1}$$

Where Z_{11} is the total attributes where both X and Y are having a value of 1. Z_{01} is the total attributes where X is 0 and Y is 1 and Z_{10} is the total attributes where X is 1 and Y is 0.

4 Experimental Setup

4.1 Dataset

Similar to the works presented in [4, 25], the publicly available dataset presented in [12] has also been utilised in this paper. Some statistics of the utilised dataset is presented in Table 3.

Table 3. Statistics of the utilized dataset

Total number of researchers	50
Average number of researchers' publications	10
Average number of citations of each researchers' publications	14.8 (max. 169)
Average number of references to each researchers' publications	15.0 (max. 58)
Total number of recommending papers	100,351
Average number of citations of the recommending papers	17.9 (max. 175)
Average number of references to the recommending papers	15.5 (max. 53)

4.2 Experimental Evaluation

To measure the quality and effectiveness of the proposed approach, 5-fold cross-validation is performed to each of the target paper's citations by selecting 20% as a test set. Mean average precision (MAP) and mean reciprocal rank (MRR) given by Eqs. 2 and 3 respectively are used to measure the system's ability in recommending essential papers at the top the recommendation list. This is important because users usually browse only top-ranked recommendations [27]. Precision, recall, and F1 measures given by Eqs. 4, 5 and 6 respectively, are also used to assess the general performance of the proposed approach. These formulas are related and have been used in similar work [4, 6]. The recommendation results obtained are then compared with two (2) other baseline methods presented in [6] and [16].

$$MAP = \frac{1}{I}\sum_{i \in I}\frac{1}{ni}\sum_{k=1}^{N}P(R_{ik}) \tag{2}$$

$$MRR = \frac{1}{N_p}\sum_{i \in I}\frac{1}{rank(i)} \tag{3}$$

$$precision = \frac{\sum(relevant\_papers) \cap \sum(retrieved\_papers)}{\sum(retrieved\_papers)} \tag{4}$$

$$recall = \frac{\sum(relevant\_papers) \cap \sum(retrieved\_papers)}{\sum(relevant\_papers)} \tag{5}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

5 Results and Discussions

The aggregate results obtained by the proposed approach from the publication lists across the 50 researchers using the said dataset is presented in this section. Figures 2 and 3 demonstrate the results comparisons based on (MAP) and (MRR) respectively.

As can easily be seen from Fig. 2 that the proposed approach has tremendously and unanimously outperformed the baseline methods for all *N* recommendations values based on mean average precision (MAP). Co-Citation performs the worst of the three results, while as expected, the performance of the proposed approach decreases as the number of N increases. This is because as the number of N increases, the tendency of retrieving irrelevant results also increases and thereby affecting the cumulative MAP results. However, the highest results based on (MAP) is obtained when N = 10 (N@10).

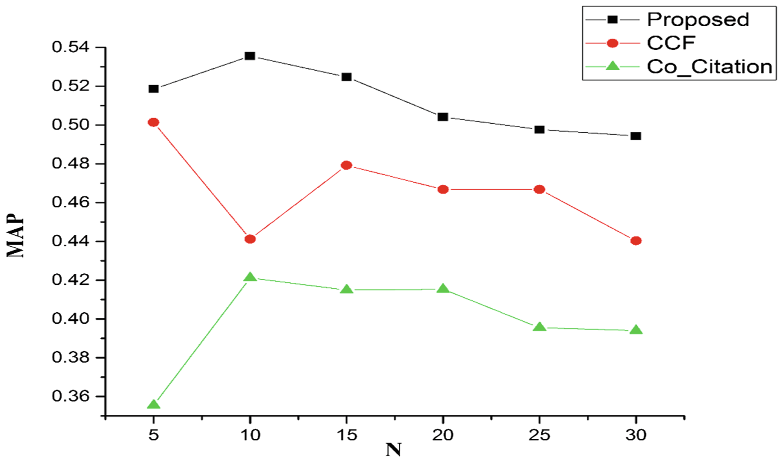

Fig. 2. Mean average precision (MAP)

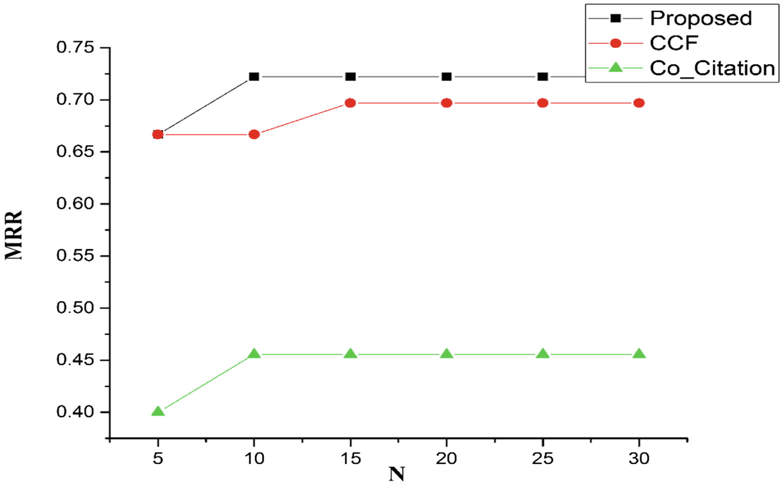

Fig. 3. Mean reciprocal rank (MRR)

On the other hand, the results comparison based on (MRR) is depicted in Fig. 3. The results difference between the proposed method and the CCF is not much significant. However, both the two approaches have significantly outperformed the Co-Citation method. This is because, the two approaches can leverage the latent associations that exist between a scholarly paper and its various citations, and different from the Co-Citation method that only uses the common citations relations.

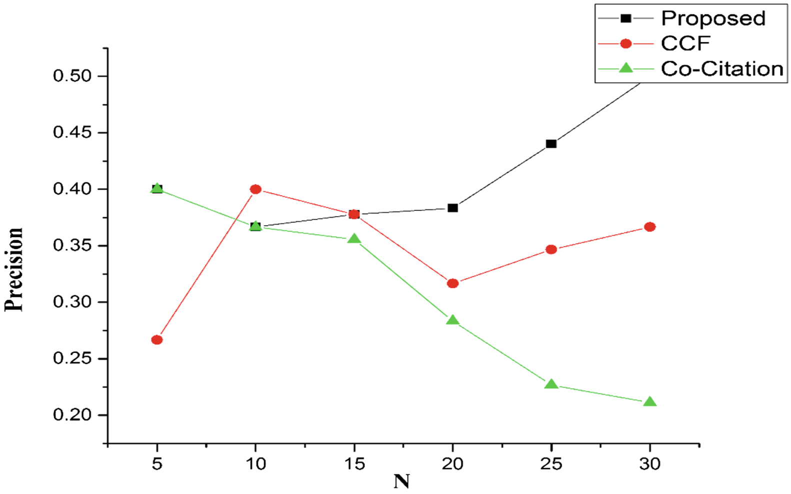

Fig. 4. Precision performance on the dataset

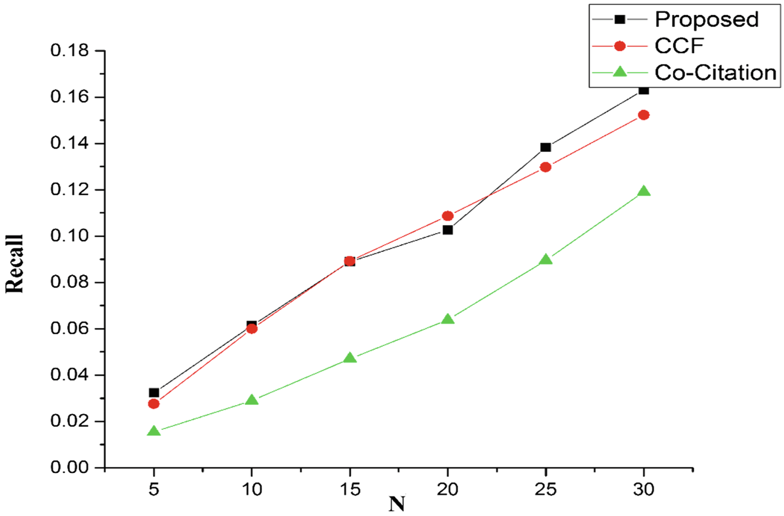

Fig. 5. Recall performance on the dataset

Figures 4 and 5 respectively represent the results comparisons based on Precision and Recall. From Fig. 4, the Precision result of the proposed approach is significant over the baseline methods for all *N* recommendations values. However, the CCF approach outperformed the proposed approach when N = 10 (N@10). The improvement of the proposed approach over the other baseline methods becomes outstandingly significant when the number of recommendations (*N)* is higher than 15. The Co-Citation results start with encouraging results specifically when N = 5 (N@5), but becomes less significant as the number of recommendations (*N)* increases.

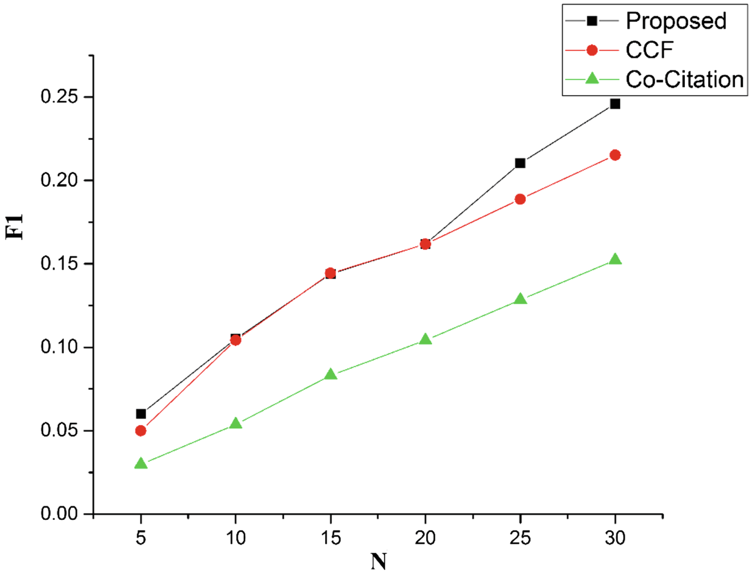

Fig. 6. F1 performance on the dataset

The results comparisons based on recall is depicted in Fig. 5. The CCF method performs better than the proposed approach when $N = 20$ (N@20). Averagely, the performance difference between the proposed and CCF approaches is not much significant. However, both approaches have statistically outperformed the Co-Citation method based on recall.

Figure 6 provides the results comparison based on the F1 measure. Similar to the results of recall depicted in Fig. 5, the performance difference between the CCF and the proposed approach based on the F1 measure is not much significant. However, results of the proposed approach start to be significant as the number of N increases, especially when N is above 20. Also, both the proposed and CCF approaches have shown significant improvement over the Co-Citation method based on the F1 measure.

In conclusion, as can easily be deduced from the presented results (Figs. 2, 3, 4, 5 and 6) that identifying and incorporating the latent relations across research papers plays a significant role in scholarly paper recommendations, and could result to improved recommendation performance. Furthermore, while the results difference between the proposed approach and CCF is not much significant, both the two results have unanimously outperformed the Co-Citation method for all N recommendations values. This is attributed to the direct relations employed by the Co-Citation method.

Additionally, while the proposed approach does not show much significant improvement over CCF method based on the simulated experiments using the static dataset, it is asserted that the proposed approach would sufficiently and statistically outperform the CCF method when a live user study with the real participant is conducted.

6 Conclusion and Future Work

Considering the challenge researchers faced, in acquiring relevant and useful scholarly papers from the enormous amount of information (information overload) that is available over the internet, this paper has successfully proposed a citation-based recommender system for the scholarly paper recommendation. The proposed approach has utilised the latent associations that exist between a scholarly paper and its various citations to personalise recommendations based on paper-citation relations.

Using a publicly available dataset, the proposed approach has improved the baseline methods in recommending useful and utile recommendation based on (MAP) and (MRR). The proposed approach has also shown significant improvement over the other baseline methods in assessing the general recommendation performance based on precision, recall and F1 measures.

One advantage of the proposed approach is its ability to leverage the latent associations that exist between a scholarly paper and its various citations. The next target is to add more strict rules in measuring the relativity between a target paper and the recommending papers to improve the recommendation utility.

Acknowledgement. This research is supported by collaborative research fund from Universitas Negeri Yogyakarta, Indonesia.

References

1. Robson, C., McCartan, K.: Real World Research. Wiley, Hoboken (2016)
2. Haruna, K., Ismail, M.A.: Scholarly paper recommendation using publicly available contextual metadata: conceptual paper. In: Seminar on Information Retrieval and Knowledge Management (SIRKM 2017), 19 July 2017. Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (2017)
3. Haruna, K., Ismail, M.A., Suhendroyono, S., Damiasih, D., Pierewan, A.C., Chiroma, H., et al.: Context-aware recommender system: a review of recent developmental process and future research direction. Appl. Sci. **7**, 1211 (2017)
4. Haruna, K., Ismail, M.A., Damiasih, D., Sutopo, J., Herawan, T.: A collaborative approach for research paper recommender system. PLoS One **12**, e0184516 (2017)
5. Kai-Wah Chu, S., Law, N.: The development of information search expertise of research students. J. Librariansh. Inf. Sci. **40**, 165–177 (2008)
6. Liu, H., Kong, X., Bai, X., Wang, W., Bekele, T.M., Xia, F.: Context-based collaborative filtering for citation recommendation. IEEE Access **3**, 1695–1703 (2015)
7. Sugiyama, K., Kan, M.-Y.: Scholarly paper recommendation via user's recent research interests. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, pp. 29–38 (2010)
8. Agarwal, N., Haque, E., Liu, H., Parsons, L.: Research paper recommender systems: a subspace clustering approach. In: International Conference on Web-Age Information Management, pp. 475–491 (2005)
9. Gori, M., Pucci, A.: Research paper recommender systems: a random-walk based approach. In: IEEE/WIC/ACM International Conference on Web Intelligence 2006, WI 2006, pp. 778–781 (2006)

10. Gipp, B., Beel, J., Hentschel, C.: Scienstein: a research paper recommender system. In: Proceedings of the International Conference on Emerging Trends in Computing (ICETC 2009), pp. 309–315 (2009)
11. Nascimento, C., Laender, A.H., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, pp. 297–306 (2011)
12. Sugiyama, K., Kan, M.-Y.: Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 153–162 (2013)
13. Beel, J., Langer, S., Genzmehr, M., Nürnberger, A.: Introducing Docear's research paper recommender system. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 459–460 (2013)
14. Haruna, K., Ismail, M.A.: An ontological framework for research paper recommendation. Int. J. Soft Comput. **11**, 96–99 (2016)
15. Haruna, K., Ismail, M.A.: Evaluation techniques for context-aware recommender systems: a systematic mapping. J. Inf. Retrieval Knowl. Manage. **3**, 23–35 (2017)
16. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., et al.: On the recommending of citations for research papers. In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 116–125 (2002)
17. An, Y., Janssen, J., Milios, E.E.: Characterizing and mining the citation graph of the computer science literature. Knowl. Inf. Syst. **6**, 664–678 (2004)
18. Price, D.J.D.S.: Networks of scientific papers. Science **149**, 510–515 (1965)
19. Newman, M.E.: The structure of scientific collaboration networks. Proc. Nat. Acad. Sci. **98**, 404–409 (2001)
20. Small, H.: Co-citation in the scientific literature: a new measure of the relationship between two documents. J. Assoc. Inf. Sci. Technol. **24**, 265–269 (1973)
21. Catalano, A.: Patterns of graduate students' information seeking behavior: a meta-synthesis of the literature. J. Doc. **69**, 243–274 (2013)
22. Lazonder, A.W.: Exploring novice users' training needs in searching information on the WWW. J. Comput. Assist. Learn. **16**, 326–335 (2000)
23. Ismail, M.A.: Identifying how novice researchers search, locate, choose and use web resources at the early stage of research. Malays. J. Libr. Inf. Sci. **3**, 67–85 (2011)
24. Ismail, M.A.: Support system for novice researchers (SSNR): usability evaluation of the first use. Int. Arab J. Inf. Technol. **9**, 361–367 (2012)
25. Sugiyama, K., Kan, M.-Y.: A comprehensive evaluation of scholarly paper recommendation using potential citation papers. Int. J. Digit. Libr. **16**, 91–109 (2015)
26. Leydesdorff, L.: On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. J. Assoc. Inf. Sci. Technol. **59**, 77–85 (2008)
27. Hildreth, C.R.: Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. Inf. Res. **6**(2) (2001)