Incorporating Satellite Documents into Co-citation Networks for Scientific Paper Searches

Masaki Eto

Gakushuin Women's College
Tokyo, Japan
masaki.eto@gakushuin.ac.jp

Abstract. To improve the search performance of retrieval methods using co-citation linkages, this study proposes a technique to enlarge a co-citation network by incorporating satellite documents. This technique specifies satellite documents via full-text searches for terms obtained from documents having co-citation linkages with a seed document; the appropriateness of each co-citation linkage is checked by using the strength of the co-citation context based on the results of parsing documents that cite the seed document. This study evaluates search performance using the proposed technique with IR experiments. Specifically, the random walk with restart algorithm, which can compute similarities between the seed document and each document in the network, is applied to the enlarged and initial networks. Scores of the normalized discounted cumulative gain (nDCG@K) were then compared. The results indicate that the search performance of the retrieval methods using the enlarged network outperforms those of a baseline method using the initial network.

Keywords: Co-citation, Context, TF-IDF, Random walk with restart

1 INTRODUCTION

In the field of scientific paper searches, citations are often used to measure implicit relationships between documents. One approach to improve the search performance of retrieval methods using citation linkages is to enlarge the citation networks by incorporating additional information. In the case of a network created using direct citation linkages, i.e., the linkages between the citing and cited documents, techniques to enlarge the network of citations on the basis of additional information, such as citing text [1] or user profiles [2], have been reported.

This study enlarges the networks connected by co-citations. A co-citation is defined as a linkage between a pair of documents concurrently cited by a third document. In the simplest retrieval method using co-citation, documents having a co-citation relationship with a given seed document that are known to be relevant are presented to the user under the assumption that documents co-cited with such a seed document tend to be topically similar to the seed document. Co-citation networks have been used in bibliometrics and can also be applied to scientific paper searches (e.g., [3]).

This study proposes a technique to enlarge the co-citation network by adding word-based linkages. When documents are detected by the co-citation linkage, it is possible to obtain more appropriate search terms from the document; such terms may not have been included in the original seed document. A set of new search terms may yield additional relevant documents that were not identified simply by the co-citation linkages or the user's original representation of his or her information needs. This study defines satellite documents as documents that are specified via full-text searches for new search terms. The purpose of the proposed technique is to incorporate these satellite documents into the initial network of documents, which is already connected by co-citation linkages.

In addition, the proposed technique attempts to reduce noise satellite documents incorporated into the initial co-citation network using the co-citation context. Some studies (e.g., [3] and [4]) have reported that using the contexts of co-citations has positive effects for reducing noise documents when co-citation networks are enlarged by additional co-citation linkages; therefore, it is feasible to use co-citation contexts when enlarging co-citation networks by adding word-based linkages.

This study empirically evaluates the search performance of retrieval methods using the proposed technique with IR experiments. Specifically, the random walk with restart (RWR) algorithm [5], which can compute similarities between the seed document and each document in the network, is applied to enlarged networks and initial networks, and the results are compared by computing scores of the cutoff version of the normalized discounted cumulative gain (nDCG@K).

2 PROPOSED TECHNIQUE

2.1 Specifying satellite documents

Figure 1 shows an initial network comprising document nodes connected by undirected co-citation linkages. In this network, a search query is a seed document that is known to be relevant to the information needs of a user. The weight of the edge, *w*, i.e., the strength of the co-citation linkage, is computed as

$$\mathrm{w}\ (d_1, d_2) = \mathrm{cociting}\ (d_1, d_2). \qquad (1)$$

Here, d_1 and d_2 are co-cited documents and cociting(d_1, d_2) denotes the total number of documents co-citing d_1 and d_2 in the target document set. Note that this study denotes a weighted edge between d_1 and d_2 as Edge (d_1, d_2, *w*).

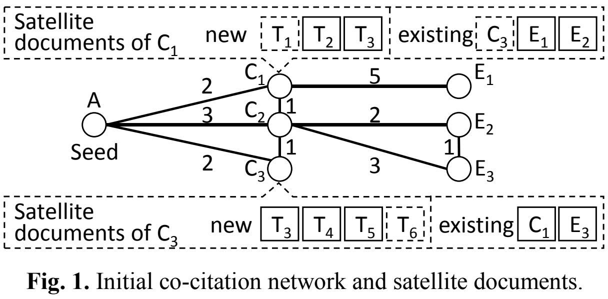

Fig. 1. Initial co-citation network and satellite documents.

The proposed technique specifies satellite documents by investigating documents one hop from the seed. This study defines host documents as source documents that are used to specify satellite documents. Using the title words of the host document as a query, the satellite documents are specified on the basis of a standard full-text search method; the seed document is excluded from the search target. For example, in Figure 1, Document C_1, a host document that is one hop from the seed, specifies six satellite documents. In the experiments in this study, the tf-idf retrieval function of the Indri search engine, which has been developed as part of the Lemur Project, was used. The top N documents ranked by this full-text search were adopted as satellite documents (e.g., $N = 10$).

In addition, as an optional process, the proposed technique attempts to check the appropriateness of each host document as a source because inappropriate host documents may yield noise. To check appropriateness, this technique uses the strength of co-citation context (see e.g., [6] and [7]) identified by parsing the full-text of documents that cite the seed and each host. More specifically, this technique examines reference positions within the text and if references to both the document and the seed appear within a paragraph in one or more citing documents, the document is selected as a host document because a seed and host co-cited in a strong context are expected to be closely related. For example, in Figure 1, if one or more documents cite Documents A and C_3 in the same paragraph, Document C_3 would be selected as a host document. Conversely, if no documents cite them in the same paragraph, Document C_3 would not be selected as a host document.

2.2 Incorporating satellite documents

If a satellite document is new, a new node is created with an undirected edge of weight 1 connecting the new node to its host. When two host documents share a new satellite document, one new node and two edges between the new node and each host node are created. In Figure 1, Document T_3 is specified by host Documents C_1 and C_3; therefore, a new node T_3, Edge (T_3, C_1, 1), and Edge (T_3, C_3, 1) are created. In addition, if a new document has co-citation linkages with documents already existing in the initial network or with other new documents, new edges are created and weights are assigned using Eq. (1).

When a satellite document already exists in a given network, the linkage between the satellite document and its host is used to create a new edge or recalculate the weight of a given edge. If the linkage is new for the network, an undirected edge of weight 1 is created between the satellite and its host. If the linkage already exists in the initial network, the weight of the existing edge is recalculated as

$$\mathrm{w}\ (d_1, d_2) = \mathrm{cociting}\ (d_1, d_2) + 1. \tag{2}$$

Some new linkages may be duplicated in the specified results. In such cases, the proposed technique treats them as one combined link and creates one new edge. For example, in Figure 1, Document C_1 has satellite document C_3 and vice versa; therefore, only Edge (C_1, C_3, 1) is incorporated into the network.

2.3 Ranking the documents in the network

To calculate document scores, the RWR algorithm is applied to the enlarged network. This algorithm iteratively investigates the entire network, and the similarity between a seed node and each node in the network is calculated (see, e.g., [3] and [8]). Specifically, the walker starts at a seed node and then either proceeds to the connected nodes on the basis of a probability calculated by weights or returns back to the seed node; these steps are repeated iteratively until convergence. The long-term visit rate of each node is used as a document score; these rates are given by the steady state of

$$\vec{p} = (1 - r)\tilde{w}\vec{p} + r\vec{s}. \tag{3}$$

Here, $\vec{p}$ is an n-dimensional vector (with n being the number of nodes in the network), $\vec{s}$ is an n-dimensional vector with 1 for the seed node and 0 for the others, and r is a return probability. This study uses the following 11 values of r in the experiments: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99. Also, $\tilde{w}$ is a transition probability matrix, and each transition probability between two nodes is the weight of an edge, which is normalized by the summation of the weights of the edges connected to the current node. In the case shown in Figure 1, the probability "A to C_1" is 0.286 given as 2/(2+3+2). Therefore, $\tilde{w}$ is an asymmetric matrix, i.e., one direction can be different from another direction, e.g., the probability "C_1 to A" is not equal to 0.286.

3 EXPERIMENTAL SETUP

As described in Section 2.1, the proposed technique has an optional process. Therefore, this study evaluates the search performance of two retrieval methods. First, *Proposed (all)* omits the optional process and simply identifies all documents one hop from the seed as host documents. Second, *Proposed (context)* selects host documents using the strength of the co-citation context. For both retrieval methods, the parameter N (i.e., the number of retrieved documents per host document) was set to 10 and 100. In addition, the study evaluates the search performance of a baseline method that applies the RWR algorithm only to the initial co-citation network. In this experiment, the three retrieval methods take up to two hops from the seed to create each initial co-citation network; three or more hops are out of scope.

To create a special test collection, the Open Access Subset of PubMed Central was used. The test collection was constructed by selecting approximately 152,000 documents from the subset with the condition that the document had at least one citation linkage with a document in the subset. The test collection contained 100 seed documents that were randomly selected from all the documents under the condition that each seed document had co-citation linkages with 10 or more documents.

In addition, this experiment adopted nDCG@K as a metric to evaluate the search performance (with $K = 5$, 10, 50, and 100). A document was considered relevant depending on the degree to which it shared MeSH Descriptors with the target seed document. More specifically, the Jaccard coefficient (JC) was used, i.e., when nDCG was calculated, the experiment used a relevance score of 3 for documents whose JC was 0.3 or more, 2 for documents whose JC was 0.2–0.3, and 1 for documents whose JC was 0.1–0.2.

4 RESULTS

Search runs for 100 seed documents were executed using each method.

4.1 Evaluation of incorporated documents

First, the experiment examined whether the newly incorporated documents were relevant (see Figure 1). Table 1 shows the average number of relevant incorporated documents; a document is relevant if the JC is 0.1 or more. Further, Table 1 lists the average ratio of the relevant documents, which is the total number of relevant documents over 100 search runs divided by the total number of new documents over the 100 search runs.

As shown in the table, the numbers of relevant documents were relatively large. For example, *Proposed (all)* incorporated more than 50 new relevant documents per seed. Therefore, the proposed technique has the potential to improve the search performance.

Further, the ratio of relevant documents for *Proposed (context)* was higher than that of *Proposed (all)*. This result indicates that the checking process using the co-citation context tends to exclude inappropriate host documents.

Table 1. Statistics of the incorporated documents.

	Propsed (all)		Proposed (context)	
N	10	100	10	100
Number of relevant documetns	50.23	265.36	7.38	44.50
Number of incorporated documents	298.53	2390.03	29.18	261.34
Ratio	0.168	0.111	0.253	0.170

4.2 Evaluation of the ranked retrieval results

Table 2 shows the average scores of nDCG@K and the results of the paired t-test between the baseline method and each retrieval method using the proposed technique. Note that this table shows only the scores of the best results ranked by Eq. (3) using the aforementioned 11 different r-values.

Table 2. Average scores of nDCG@K.

		Proposed N = 10		Propsed N = 100	
K	Baseline (*r*)	all (*r*)	context (*r*)	all (*r*)	context (*r*)
5	0.226 (0.9)	0.226 (0.99)	0.232* (0.99)	0.224 (0.9)	**0.234**** (0.9)
10	0.223 (0.99)	0.221 (0.99)	0.227** (0.99)	0.226 (0.99)	**0.230**** (0.99)
50	0.188 (0.99)	0.191* (0.99)	0.189** (0.99)	**0.197**** (0.99)	0.191 (0.99)
100	0.174 (0.99)	0.181** (0.99)	0.177* (0.99)	**0.188**** (0.99)	0.180** (0.99)

* P < 0.05, ** P < 0.01

In Table 2, the maximum scores of the five retrieval results at each *K* are shown in bold. These are the results of *Proposed (context)* and *Proposed (all)* with *N* = 100, with the paired t-tests showing statistically significant differences. Therefore, the retrieval methods using the proposed technique tended to outperform the baseline method.

Furthermore, the scores of *Proposed (context)* were higher than those of the baseline method in all cases, with the paired t-tests indicating a statistically significant difference in most cases. Conversely, some scores of *Proposed (all)*, i.e., with *N* = 10 at *K* = 10 and with *N* = 100 at *K* = 5, were lower than those of the baseline method. This suggests that the checking process had a stable and positive impact on improving the search performance.

5 CONCLUSION

This study proposed a technique to enlarge co-citation networks by incorporating satellite documents in scientific paper searches. Retrieval methods using the proposed technique tended to outperform the baseline method, which was based on the initial co-citation network.

6 ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP26730163.

7 REFERENCES

1. He, Q., Pei, J. Kifer, D., Mitra, P. and Giles. C. L. Context-aware citation recommendation. In Proceedings of the 19th International World Wide Web Conference (WWW2010), 421-430 (2010)
2. Sugiyama, K. and Kan, M. Exploiting Potential Citation Papers in Scholarly Paper Recommendation, In Proceedings of the 13th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2013), 153-162 (2013)
3. Eto, M. Document retrieval method using random walk with restart on weighted co-citation network, In Proceedings of the 77th ASIS&T Annual Meeting (2014)
4. Eto, M. Spread co-citation relationship as a measure for document retrieval. Proceedings of the fifth ACM workshop on Research advances in large digital book repositories and complementary media, 7-8 (2012)
5. Tong, H., Faloutsos, C. and Pan, J. 2008. Random walk with restart: fast solutions and applications. Knowledge and Information Systems, 14, 3, 327-346 (2008)
6. Gipp, B. and Beel, J. Citation proximity analysis (CPA) - A new approach for identifying related work based on co-citation analysis. In Proceedings of the 12th ISSI Conference. 2, 571-575 (2009)
7. Eto, M. Evaluations of context-based co-citation searching, Scientometrics 94, 2, 651-673 (2013)
8. Gori, M. and Pucci, A. Research paper recommender systems: A random-walk based approach. In Proceedings of IEEE/WIC/ACM Web Intelligence, 778-781 (2006)