Increasing Source Documents of Rough Co-citation to Expand Co-citation Networks for Scientific Paper Searches

Masaki Eto
Gakushuin Women's College
Japan
masaki.eto@gakushuin.ac.jp

ABSTRACT

This study proposes to increase the source documents of "rough co-citation" when co-citation networks are expanded by rough co-citation to include new relevant documents. A rough co-citation relationship is a linkage between a pair of documents cited by two other documents in a similar citation context. The previous method (baseline) only employed documents that directly cited the seed as sources of rough co-citation to expand networks. The proposed methods used both sources adopted by the baseline method and the documents that cited documents having original or rough co-citation relationships with the seed. An information retrieval experiment was conducted to evaluate the search performances of networks expanded by the proposed methods. The experimental results indicated that the performances of the networks expanded by the proposed methods were better than those of the networks expanded by the baseline method.

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 information retrieval methods using citation linkage is to enlarge the citation networks for including new relevant documents.

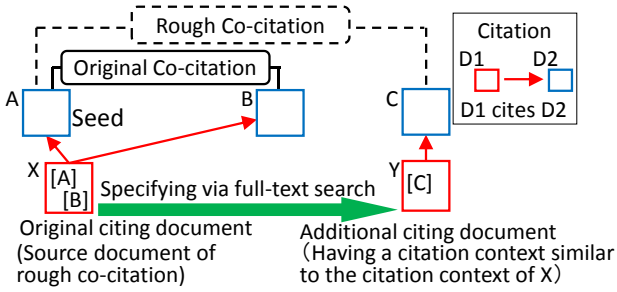

Figure 1. Schematic showing a rough co-citation.

For instance, Eto (2016) proposed the concept of "rough co-citation," which is a measure of relationship to expand co-citation networks. A rough co-citation relationship is a linkage between a pair of documents cited by two other documents in a similar citation context. An example of a rough co-citation is observed between documents A and C in Figure 1 wherein a similar citation context in both documents X and Y; document Y is identified via a full-text search using the title words of document X. The linkage strength of a rough co-citation relationship may be weaker than the original co-citation relationship because a rough co-citation relationship is determined by citations in two separate documents. However, rough co-citation linkages may yield new relevant documents that are not identified by the original co-citation linkages.

Eto (2016) only adopted documents that directly cited the seed as sources of rough co-citation to expand the network; only document X was adopted as source of rough co-citation in Figure 1. Therefore, his method may not maximize the potential of rough co-citation for the aim of searching documents similar to the seed by using the network. For example, incorporating the documents having rough co-citation relationships with document B into the network may enhance the search performances. This study proposes to increase the source documents of rough co-citation such that the expanded networks include additional new relevant documents.

EXPANDING THE NETWORK BY ROUGH CO-CITATION

Figure 2 shows an expanded co-citation network. Note that the square nodes indicating the citing documents are not included in the co-citation network. Documents i, j, and k represent the examples of sources of rough co-citation relationships.

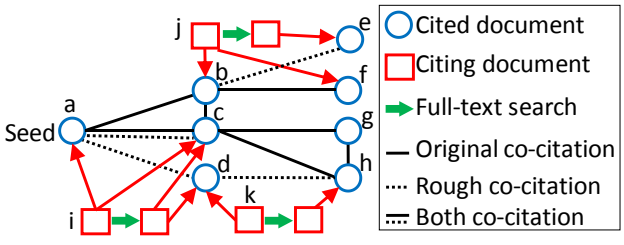

Figure 2. An expanded co-citation network.

This study compares three methods of expanding networks by rough co-citation: two proposed methods and a baseline method. The baseline method (Eto, 2016) uses documents (e.g., document i) that cite the seed as sources of rough co-citation. Method 1 uses both the sources adopted by the

baseline method and the documents (e.g., document j) that cite documents having original co-citation relationships with the seed. Method 2 uses both the sources adopted by Method 1 and the documents (e.g., document k) that cite documents having rough co-citation relationships with the seed.

In this network, the weight of an edge is computed as

$$w(v_1,v_2) = cociting(v_1,v_2) + rough\_cociting(v_1,v_2)\alpha \,, \quad (1)$$

where *cociting*(v_1,v_2) is the frequency of the original co-citation relationship between v_1 and v_2, *rough_cociting*(v_1,v_2) is the frequency of rough co-citation between v_1 and v_2, and α is a decay parameter for balancing the degrees of difference between the two co-citations.

RANKING THE DOCUMENTS IN THE NETWORK

To calculate document scores, the random walk with restart algorithm (Haveliwala, 2002) is applied to the expanded network. This algorithm iteratively investigates the entire network, and the similarity between a seed node and each node in the network is calculated. The long-term visit rate of each node is used as the document score; this study adopts these rates given by the steady state of

$$\vec{p} = 0.01\tilde{w}\vec{p} + 0.99\vec{s}. \quad (2)$$

Here, $\vec{p}$ is an n-dimensional vector (n is the number of nodes in the network), $\tilde{w}$ is an $n \times n$ transition probability matrix calculated using the edge weights, and $\vec{s}$ is an n-dimensional vector with 1 for the seed and 0 for the others.

EXPERIMENTAL SETUP

Additional citing documents were specified via a TF-IDF-based full-text search (Indri search engine) using the title words of the source document. The top-ranked N documents were adopted as additional citing documents (with N = 1, 5, and 10) per source citing document. The parameter α for Eq. (1) was set to the following six values: 0.01, 0.2, 0.4, 0.6, 0.8, and 0.99. Networks based on the three methods were created by using up to two hops from the seed; three or more hops were out of scope of this study.

To construct 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 under the condition that the document had at least one citation linkage with a document in the subset. The test collection comprised 100 seed documents that were randomly selected from all documents under the condition that each seed document had co-citation linkages with one 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; when nDCG was calculated, the experiment defined a relevance score of 3 for the documents whose JC was 0.3 or more, 2 for the documents whose JC was 0.2–0.3, and 1 for documents whose JC was 0.1–0.2.

The search runs for 100 query documents were executed by each method, after which the scores of nDCG@K per query document were computed. In the ranking process, when two or more documents had the same score, their ranks were randomly assigned for tie-breaking.

RESULTS

Table 1 lists the average scores of nDCG@K and a comparison between the baseline method and the two proposed methods with regard to the paired t-test scores. Note that this table only lists the scores of the best results using the aforementioned different α-values and N-values. The maximum scores of the three methods at each K are shown in bold.

All scores of Methods 1 and 2 were higher than those of the baseline method. In addition, the paired t-test shows statistically significant differences in the low-rank cases. The table summarizes Method 2 as an optimal method and suggests that the documents having original or rough co-citation relationships with the seed should also be adopted as sources of rough co-citation.

K	Baseline (α, N)	Method 1 (α, N)	Method 2 (α, N)
5	0.187 (0.8, 5)	0.200 (0.99, 1)	**0.203** (0.99, 1)
10	0.187 (0.6, 5)	**0.191** (0.6, 5)	0.189 (0.6, 1)
50	0.160 (0.8, 5)	0.166* (0.6, 5)	**0.167*** (0.8, 5)
100	0.144 (0.8, 5)	0.152** (0.4, 10)	**0.154**** (0.4, 10)

* P < 0.05, ** P < 0.01

Table 1. Average scores of nDCG@K.

CONCLUSION

This study proposed to increase the source documents of rough co-citation to expand the co-citation networks for the inclusion of additional new relevant documents. The experimental results indicated that the performances of the networks expanded by the proposed methods were better than those of the networks expanded by the baseline method, which is based on non-increased source documents.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP26730163.

REFERENCES

Eto, M. (2016). Rough Co-citation as a Measure of Relationship to Expand Co-citation Networks for Scientific Paper Searches, In *Proceedings of the 79th ASIS&T Annual Meeting*.

Haveliwala, T. H. (2002) Topic-sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web (WWW '02)*. ACM, New York, NY, USA, 517-526.