Strength of Co-citation Linkages Observed in Paragraphs with Many References

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

INTRODUCTION

A recent approach on co-citation linkage used in research is the measurement of co-citation strength through positions of references from the full text of citing documents. For example, if a paragraph contains references to two documents, co-citation is "strong" whereas if two references appear across two paragraphs, co-citation is "weak." Effects of this approach were reported in the research fields of bibliometrics (e.g., Boyack, Small, & Klavans 2013; Colavizza, Boyack, van Eck & Waltman 2018; Hsiao & Chen 2017) and information retrieval (e.g., Eto 2013, 2014; Gipp & Beel 2009).

However, such an approach has limitations because surface structures such as paragraphs and sentences are simply used to classify co-citation linkage strength. Eto (2013) has suggested that one of the problem patterns, i.e., the recent approach not working well, involves many documents being cited in one surface structure. He mentioned two possible reasons for this issue: (1) authors of the citing documents might aim to merely state that there are many related works and might not emphasize the relations between cited documents, and (2) there may be perfunctory citations that may not deeply refer to the content of cited documents (see also Krampen, Becker, Wahner, & Montada 2007).

The details of this problem remain unclear because Eto (2013) suggested this problem on the basis of a simple case study. Therefore, this study aims to clarify the likelihood of co-citation linkage identification in paragraphs with many references to improve this approach. Specifically, this study quantitatively examines the similarity scores of co-cited documents per number of references in a paragraph. It also compares the similarity scores of documents co-cited in paragraphs containing many references and those that do not contain as many.

METHOD

This study adopts the CITREC dataset (Gipp, Meuschke, & Lipinski 2015), which is based on the *PubMed Central Open Access Subset*, with articles related to biomedicine. The dataset contains citation information, paragraph IDs where references appear in articles, article IDs, MeSH descriptors assigned to each article in the dataset, etc.

The examination of this study specifies co-citation linkages observed in the same paragraphs by using citation information and paragraph IDs in the dataset; two references with same paragraph ID in the same document are referred to as a pair of documents with co-citation linkage in a paragraph. In cases where a pair of documents repeatedly appears within two or more different paragraphs in one citing document, this study treats them as having different co-citation linkages.

Co-citation linkage strength is evaluated through the similarity score of two co-cited documents, i.e., if the similarity score is high, co-citation linkage can be considered strong, whereas if the score is low, co-citation linkage can be considered weak. Similarity scores are calculated from MeSH descriptors assigned to each document and from Lin's computational procedure (Lin 1998), which are used in CITREC. Note that this study recalculates similarity scores according to the procedure because the scores are not provided for all document pairs in CITREC.

RESULTS

Figure 1 shows the mean and median similarity scores per number of references in a paragraph as seen in the first vertical axis. This figure also indicates the total number of co-citation linkages per number of references in a paragraph in the second vertical axis. As shown here, the similarity scores tended to be low when the number of references was high.

Furthermore, Figure 1 suggests around nine references as the boundary separating the nonstandard number of references in one paragraph, i.e., many references that might have problems, from the standard number of references in this dataset. The number of co-citation linkages observed in the paragraphs with nine or more references corresponded to about 10% of the total co-citation linkages. Since this is not a small percentage, solving the problem may improve the new approach using co-citation linkages.

Additionally, this study compared similarity scores of documents co-cited in paragraphs with standard number of references and those with the nonstandard ones. Table 1 shows results of the comparison, where the boundary is set at 8, 9, and 10 references. As shown in Table 1, the mean and median similarity scores of documents co-cited in paragraphs with the

nonstandard number of references are lower than those with the standard ones. This reveals that the degree of co-citation linkages in paragraphs with many references is weaker than that in paragraphs with relatively few references.

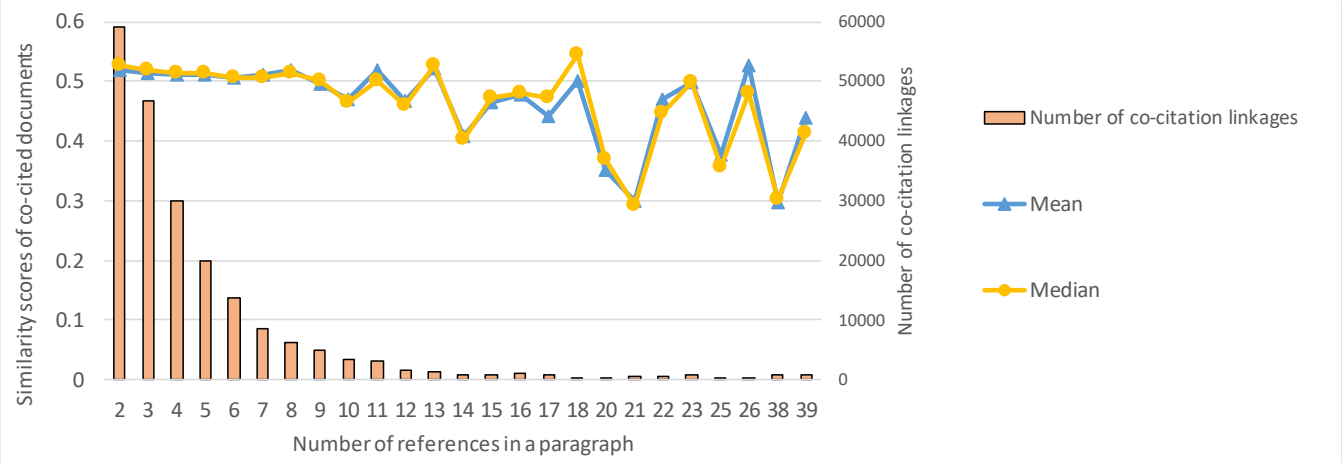

Figure 1. Similarity scores and total number of co-citation linkages per number of references in a paragraph.

Table 1. Similarity scores of documents co-cited in paragraphs with standard and nonstandard number of references.

		Boundary = 8		Boundary = 9		Boundary = 10	
	All	Standard	Nonstandard	Standard	Nonstandard	Standard	Nonstandard
Number	205,157	177,824	27,333	183,928	21,229	188,788	16,369
Mean	0.509	0.514	0.482	0.514	0.472	0.513	0.465
Median	0.513	0.519	0.480	0.519	0.473	0.519	0.462

CONCLUSION

This study quantitatively examined co-citation linkage strength in paragraphs with many references. It showed that (1) the number of such co-citation linkages is not small and that (2) the degree of such co-citation linkages is weaker than that observed in paragraphs with relatively few references. These results made it clear that improvement of the recent approach may be achieved through the identification of co-citation linkages in paragraphs with many references. It might be useful, for example, to count the number of references in one surface structure and consider the decay strength of such co-citation linkages as penalties when co-citation strength is measured through reference position.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP17K00455.

REFERENCES

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

Colavizza, G., Boyack, K. W., van Eck, N. J., & Waltman, L. (2018). The closer the better: Similarity of publication pairs at different cocitation levels. *Journal of the Association for Information Science and Technology*, 69(4), 600-609.

Eto, M. (2013). Evaluations of context-based co-citation searching. *Scientometrics*, 94(2), 651-673.

Eto, M. (2014). Document retrieval method using random walk with restart on weighted co-citation network, *Proceedings of the Association for Information Science and Technology*, 51(1).

Gipp, B., & Beel. J. (2009). Citation proximity analysis (CPA) - a new approach for identifying related work based on co-citation analysis. *Proceedings of the 12th International Conference on Scientometrics and Informetrics* (vol. 2, pp. 571–575).

Gipp, B., Meuschke, N., & Lipinski, M. (2015). CITREC: An evaluation framework for citation-based similarity measures based on TREC genomics and PubMed Central. *Proceedings of the iConference 2015*.

Hsiao, T. M., & Chen, K. H. (2017). Yet another method for author co-citation analysis: A new approach based on paragraph similarity. *Proceedings of the Association for Information Science and Technology*, 54(1) (pp. 170-178).

Krampen, G., Becker, R., Wahner, U., & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. *Scientometrics*, 71(2), 191-202.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning* (pp. 296-304).