Can the strength of co-citation linkages be evaluated using context-aware citation network embeddings*

Masaki Eto

Gakushuin Women's College, Tokyo, Japan

Abstract

Co-citation linkages are commonly employed to measure implicit relationships between documents. However, traditional co-citation techniques equivalently treat the strengths of co-citation linkages within a single citing document, limiting technique effectiveness. This study explores whether context-aware citation network embeddings can be used to more accurately measure the strength of co-citation linkages within a single citing document. In particular, this study proposes a novel method that incorporates context-aware citation network embeddings into the co-citation approach and empirically evaluates its performance. Experimental results show that the proposed method outperforms a baseline based on traditional co-citation techniques. Thus, the proposed method can discriminate between weak and strong co-citation linkages.

Keywords

co-citation, network embedding, citation context

1. Introduction

In scientometrics or information retrieval, co-citation linkages are commonly used to measure implicit relationships between documents. However, traditional co-citation techniques equivalently treat the strengths of these linkages within a single citing document, presenting a notable limitation. Although some techniques attempt to distinguish between weak and strong co-citation linkages using surface structures like paragraphs [1][2], some challenges still exist. For example, if two similar documents are cited in analogous citation contexts but appear in separate paragraphs, the strength of their co-citation linkage may be inaccurately assessed as weak.

In recent years, techniques that utilize embeddings, such as bidirectional encoder representations from transformers (BERTs) [3], have enabled sophisticated analyses of citation contexts, potentially offering a solution to this problem. This study aims to investigate whether context-aware citation network embeddings can effectively measure the strength of co-citation linkages within a single citing document. To this end, this study proposes a novel method that applies context-aware citation network embeddings to co-citation analysis and empirically evaluates the performance of the proposed method.

2. Proposed method

The proposed method generates citation network embeddings that are aware of the citation context [4]. This technique fine-tunes the SciBERT model [5] in the masked paper prediction task. In this task, a cited document is masked (Figure 1), and the SciBERT model is trained to predict the cited document from the citing document and the citation context. A previous study [4] used this

* Joint Workshop of the 2th Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data (IMSC2024), Dec 20th, 2024, Hong Kong, China and Online

masaki.eto@gakushuin.ac.jp
0000-0003-2029-6745

technique for citation recommendation, which predicts additional documents that may be suitable as replacements for the masked document.

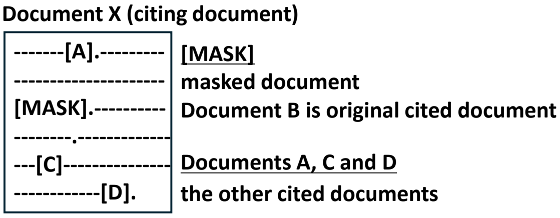

Figure 1: Example of a masked document and the other cited documents in a citing document.

This study employs this technique to measure the strength of co-citation linkages within a single citing document. The proposed method also predicts suitable documents, where the predicted documents are limited to other documents cited in the same citing document. As shown in Figure 2, the proposed method employs the context of the original citing document B and the cited documents (A, C, and D) to output a list of ranked documents that may be suitable as replacements for masked document B. This ranking also indicates the strength of the co-citation linkage with document B; the strongest linkage is between documents B and D.

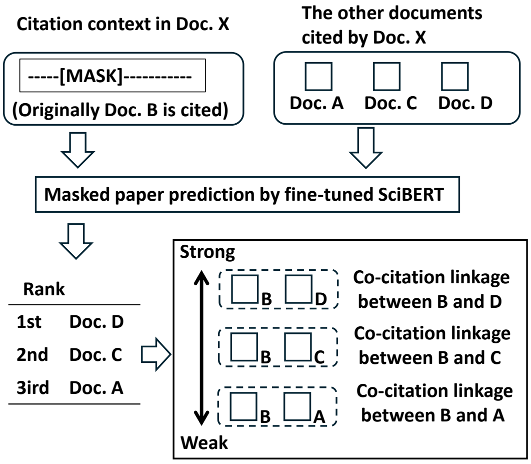
Citation context in Doc. X
-----[MASK]-----------
(Originally Doc. B is cited)
The other documents cited by Doc. X
Doc. A
Doc. C
Doc. D
Masked paper prediction by fine-tuned SciBERT
Rank
1st Doc. D
2nd Doc. C
3ird Doc. A
Strong
Weak
B
D
Co-citation linkage between B and D
B
C
Co-citation linkage between B and C
B
A
Co-citation linkage between B and A

Figure 2: Outline of the proposed method.

3. Experimental setup

This study conducted an experiment using the FullTextPeerRead dataset [6]. This dataset comprises 16,669 citation contexts, each comprising the texts preceding and following the citation position, the ID of the cited document, the title and abstract of the cited document, and the ID of the citing document. For each context, the texts were split into sentences using spaCy, which was loaded along with a biomedical text processing model (the en_core_sci_scibert model) [7]. This experiment used the two sentences preceding and following the citation position.

The aim of this experiment was to evaluate the effectiveness of the proposed method in terms of document ranking. In this evaluation, a document was considered relevant if it shared one or more indexing terms with the original cited document. As in a previous study [4], the names of the machine learning tasks defined in paperswithcode (https://paperswithcode.com/sota) were used as the indexing terms. These terms were searched for in the title and abstract of each cited document via string matching.

For the ranking tasks, 1,362 citation contexts were extracted from the dataset under the condition that the documents to be ranked contained at least one relevant document and one irrelevant document. The extraction process excluded tasks that could not discriminate between ranking performances.

The parameters used to fine-tune the SciBERT model were identical to those employed in a previous study [4]. In particular, the experiment used a batch size of 16, a learning rate of 5e−5, and

5 training epochs. The maximum length of the sub tokens preceding and following each citation position was set to 125. The sum of the token word/document, position, and token-type embeddings was used as the input representation. To fine-tune the model, the experiment used 15,307 citation contexts obtained by excluding the 1,362 contexts used for the ranking tasks from the 16,669 citation contexts.

4. Evaluation results

The experiment evaluated the ranking performance of the proposed method using two metrics: the precision of the top-ranked document (P@1) and the average precision (AP). Moreover, a baseline method based on traditional cocitation techniques was used for comparison. In particular, this baseline method ranks documents randomly because traditional co-citation techniques treat the strength of co-citation linkages within a single citing document as equivalent.

Table 1
Mean scores of P@1 and AP

	Mean P@1	Mean AP
Baseline	0.466	0.629
Proposed method	0.570	0.704

Table 1 compares the results of the proposed and baseline methods. As presented in this table, the proposed method outperformed the baseline method in both metrics. The paired t-test shows that there was a significant difference ($p \fallingdotseq 0.000$) between the two methods for each metric.

5. Conclusion

This study investigated whether context-aware citation network embeddings can effectively measure the strength of co-citation linkages within a single citing document. The experimental results demonstrated that the proposed method successfully differentiates between weak and strong co-citation linkages.

A potential direction for future research is to apply the proposed method to scientometrics. By accurately measuring co-citation strength using even a single citing document, the proposed method can enable the creation of co-citation maps with fewer citations. This applicability could be particularly valuable for mapping new knowledge domains where the number of citations is insufficient.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 23K11776.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] Bela Gipp and Joeran Beel. 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).

[2] Masaki Eto. 2013. Evaluations of context-based co-citation searching. Scientometrics, 94(2), 651-673.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

[4] Masaya Ohagi and Akiko Aizawa. 2022. Pre-trained Transformer-Based Citation Context-Aware Citation Network Embeddings. In The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22), June 20–24, 2022, Cologne, Germany.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

[6] Jeong, C., Jang, S., Park, E., & Choi, S. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. Scientometrics, 124, 1907-1922.

[7] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language

[8] Processing. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 319–327.