Diversifying Citation Contexts in Academic Literature for Knowledge Recommendation

Yunhan Yang
School of Information Management
Wuhan University
Wuhan, Hubei, China

Haihua Chen
Department of Information Science
University of North Texas
Denton, Texas, USA
haihua.chen@unt.edu

Wei Lu
School of Information Management
Wuhan University
Wuhan, Hubei, China

Brenda Reyes Ayala
Department of Information Science
University of North Texas
Denton, Texas, USA

ABSTRACT

Citation contexts of an article refer to sentences or paragraphs that cite that article. Citation contexts are especially useful for recommendation and summarization tasks. However, few studies have recognized the diversity of these citation contexts, thus leading to redundant recommendation lists and abstract [3]. To address this gap, we compared several strategies that can recommend a set of diverse citation contexts by re-ranking extracted citation contexts. Diversification was achieved by combining one of two semantic distance algorithms with one of two re-ranking algorithms. Experimenting with CiteSeerX dataset, our program produced a diverse list of 10 citation contexts that could be recommended to users. We evaluated the experiment results based on a user case study of 15 articles. The case study revealed that a diversity strategy that combined the "ESA" and "MMR" led to a better reading experience for participants compared to other diversity strategies. Our study provides insights to develop better automatic academic recommendation and summarization systems.

CCS CONCEPTS

• **Computing methodologies → Semantic networks**;

KEYWORDS

Citation context, Diversity, Knowledge recommendation

ACM Reference Format:
Yunhan Yang, Haihua Chen, Wei Lu, and Brenda Reyes Ayala. 2018. Diversifying Citation Contexts in Academic Literature for Knowledge Recommendation. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, Article 2, 2 pages. https://doi.org/10.1145/3197026.3203904

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203904

1 INTRODUCTION

Citation context is a well-studied topic in information science. To diversify citation contexts, or to provide a list of citation contexts with minimum overlaps helps to obtain a comprehensive view of how a work is cited, which will also reduce the difficulty of users' choice and save them a lot of time while citing. Existing methods on diversification can be divided into three categories: (1) contentbased, selecting items that are dissimilar to each other, (2) noveltybased, selecting items that contain new information when compared to what was previously presented to the user, and (3) semantic-based, selecting items that belong to different topics [4]. However, diversification performance is far from satisfactory. For example, CiteSeerX citation contexts still contain many duplicate items.

In this paper, we proposed and evaluated four different diversification strategies. We experimented with articles related to information retrieval with between 15 and 100 citations and put them together with their citation contexts from CiteSeerX. We then tested several diverse re-ranking strategies that combined both Explicit Semantic Analysis (ESA) and WordNet as semantic distance similarity algorithms with Maximal Marginal Relevance (MMR) and DivScore as diversification ranking algorithms to select the top 10 citation contexts as diversified results. Finally, we conducted a user case study to evaluate the results.

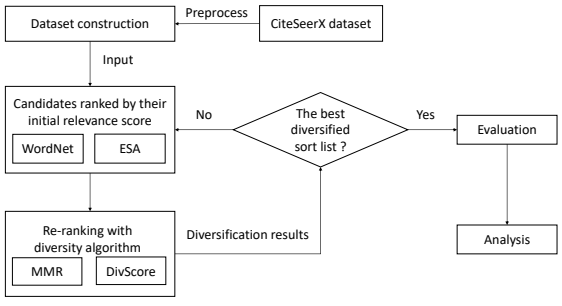

Figure 1: Figure 1: Research Design

2 METHODOLOGY

This study includes the following steps: dataset construction and pre-processing, semantic relevance calculation, ranking of candidates by initial relevance score, re-ranking, diversification evaluation, and results analysis. Figure 1 summarizes our research design.

2.1 Semantic distance calculation

In this study, we define a set of citation contexts of the cited article as $C = \{c_1, c_2, \cdots, c_n\}$. The semantic distance between them is calculated by their similarity $sim(c_i, c_j)$ based on the words using the semantic similarity algorithms ESA and WordNet. We used ESAlib tool to calculate ESA distance and NLTK to output the semantic scores of two sentences.

2.2 Re-ranking for diversification

We applied the MMR and Score Difference algorithms which belong to explicit and implicit diversification methods respectively. The former is used for iterative re-ranking, and the latter is used for single time re-ranking.

2.2.1 MMR re-ranking for diversification.

$$C_i = \begin{cases} \underset{c \in C}{argmax}[\lambda sim(c,q) - (1-\lambda)\max_{c_j \in S} sim(c,c_j)], i \neq j \\ \underset{c \in C}{argmax}[\mu * \lambda sim(c,q) - (1-\lambda)\max_{c_j \in S} sim(c,c_j)], i = j \end{cases}$$

In the above formula[1], C_i is the citation context with highest score in one round of iterative selection, S is the re-ranked list, λ is the coefficient: $\lambda \in [0,1]$, μ is the penalty coefficient. S is updated after every iteration, until completing the iteration. $sim(c,q)$ is the semantic distance between each citation context and the abstract of the cited article, and $sim(c,c_j)$ semantic distance between different citation contexts.

2.2.2 Score Difference based re-ranking for diversification. The Score Difference method has been widely used in document retrieval [2]. In this paper, we proposed a DivScore algorithm based on Score Difference as shown below:

ALGORITHM: DivScore Algorithm

for $1 \leq i \leq |R(q)|$ *do*

$$DivScore(C_i) = (1 - \frac{i-1}{N}) \times \mu * sim(C_i, q) + \frac{i-1}{N} \times DiffScore(C_i, C_{i-1})$$

end for

Sort C_i on $DivScore(C_i)$

where R(q) presents original citation context list of the cited paper and C_i is ranked by the correlation score. $sim(C_i, q)$ represents the correlation score between each citation context in the citing article and the abstract of the cited article. N denotes the number of citation contexts in the citation context list. $DiffScore(C_i, C_{i-1})$ represents the difference score between citation context C_i and C_{i-1} . It was calculated by $1 - sim(C_i, C_{i-1})$, penalty factor μ is set in front of $sim(C_i, q)$ when there is a similar citation context.

3 EXPERIMENTS

We randomly selected 36 cited papers with abstracts whose citation context counts were between 11 and 90 in information retrieval from CiteSeerX and pre-processed the data by removing all the punctuation and stop words. Then, the penalty coefficient of the "WordNet + MMR" strategy, the "ESA + DivScore" strategy, and the "ESA + MMR" strategy were set as 0.09, 0.45 and 0.20. We didn't consider the strategy "WordNet + DivScore" because of its poor performance. Together with the strategy using citation count only, there were four remaining diversification strategies to be evaluated.

In the evaluation phase, we presented the citation context lists generated by the four diversification strategies of 15 information retrieval articles to users. They are required to read each list and answer a few questions related to readability, diversity, and usefulness of the list to judge whether diversification is helpful in the academic writing process and which diversification strategy provided a better user experience.

4 RESULTS AND CONCLUSION

The Cohen's kappa score of the two annotators before conducting formal investigation was 0.676, was satisfactory according to Viera and Garrett's evaluation standard. We calculated the average of all respondents' score on the four strategies and three indicators (see table 1) and converted scores to 0-1 for the purpose of comparison.

Table 1: The evaluation results

Strat	Readability	Diversity	Usefulness	Overall
Strat1	0.2381	0.2857	0.0774	0.1897
Strat2	0.1905	0.2965	0.1607	0.1964
Strat3	0.3512	0.3840	0.2917	0.3371
Strat4	0.3690	0.4554	0.3452	0.3817

Notes: Strat1: Citation number (CiteSeerX); Strat2: WordNet + MMR; Strat 3: ESA + DivScore; Strat 4: ESA + MMR.

The experiment results showed that our proposed approach generated a more diverse citation context list than the original citation context list presented by CiteSeerX, which led to a better user reading experience. Moreover, among the four diversification strategies which combine "ESA", "WordNet" and "MMR", "DivScore", "ESA + MMR" performed the best.

5 ACKNOWLEDGMENTS

This work was partially supported by the NSFC 2017-2020 Projects 7167030644 and by the NSFC 2018-2020 Project 71704137.

REFERENCES

[1] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.

[2] Sadegh Kharazmi, Mark Sanderson, Falk Scholer, and David Vallet. 2014. Using score differences for search result diversification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 1143–1146.

[3] Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V Çatalyürek. 2015. Diversifying citation recommendations. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 55.

[4] Marialena Kyriakidi, Kostas Stefanidis, and Yannis Ioannidis. 2017. On Achieving Diversity in Recommender Systems. In *Proceedings of the ExploreDB'17*. ACM, 4.