A Study of Citation Recommendation by Fusing Heterogeneous Network Representation Learning and Attention Mechanisms

Jinzhu Zhang (zhangjinzhu@njust.edu.cn), Nanjing University of Science and Technology
Mengmeng Qiu (651484901@qq.com), Nanjing University of Science and Technology

1. INTRODUCTION

With the rapid increase of academic papers, researchers spend a lot of time on searching relevant papers for citation during paper writing (Kobayashi*, et al.*, 2018). To solve this problem, many approaches have been proposed for citation recommendation and are often classified into contentbased (Van Meteren and Van Someren, 2000), collaborative filtering-based (Yang*, et al.*, 2016) and network-based methods, such as p-CNN(Chen*, et al.*, 2017), WHIN-CSL (Chen*, et al.*, 2019), and PR-HNE (Ali*, et al.*, 2020).

Currently, citation recommendation is usually performed in homogeneous networks, which makes it difficult to take full advantage of the rich features and information of papers. Therefore, a heterogeneous network containing much more features needs to be further studied. In addition, most citation recommendation methods are based on shallow neural networks, leading to difficulties in revealing deep semantic and structural information. Finally, the differences in the effects of different relationships on citation recommendation are not reflected.

Therefore, we propose a model to learn higher quality feature representations in a citation network containing more types of nodes and edges, by fusing heterogeneous network representation learning and attention mechanisms. Firstly, we extract more types of nodes and introduce semantic links to build a more comprehensive and systematic heterogeneous network. Then we learn the weights of different relationships for citation recommendation based on the attention mechanism to generate a

deep semantic vector representation of citations and calculate the similarity for citation recommendation. Finally, we design ablation experiments to further explore the important influencing factors of citation recommendation.

2. DATA/METHODOLOGY

2.1 Data description

We retrieve the data through the Web of Science with the search formula as follows: (TS=(internet of things) AND PY=(2011-2018)) AND SILOID==("WOS") AND LA==("ENGLISH")) . We collect 5,194 papers after removing data with null values and withdrawn publication data. Then we extracted references from papers and generated a total of 4,044 citation relations, of which ninety percent are randomly used as the positive training set and the rest as the negative testing set.

2.2 Heterogeneous citation network construction containing semantic links

We construct a heterogeneous citation network containing multiple types of nodes and edges, as shown in Figure 1. In order to describe more complex semantic information in the network, we construct a collection of meta-paths, such as PP, PAP, PVP, PSP, PSP and PKP. Taking PAP as an example, it represents that two papers are written by the same author. In addition, we introduce semantic links to describe the semantic relationships between papers, represented as a special metapath PSLP. The semantic link exists between two papers if the similarity among them is greater than a threshold.

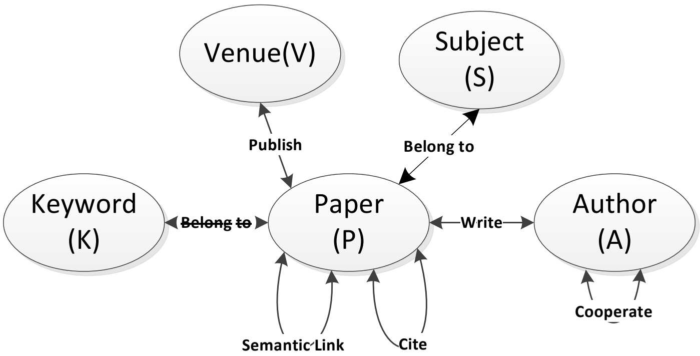

Figure 1 Features and relations of heterogeneous network for citation recommendation

2.3 Paper representation by fusing heterogeneous network embedding and attention mechanism

In this paper, we learn paper representation through a heterogeneous network embedding method considering the attention mechanism for citation recommendation (HAN-CR). It consists of nodelevel attention and semantic-level attention(Wang*, et al.*, 2019). Firstly, we compute vector representations based on specific meta-path and then learn the weights of the meta-paths by semanticlevel attention to obtain the final representations.

2.4 Citation recommendation based on the similarity between paper representations

In this paper, citation recommendation is considered as a binary classification problem. Firstly, we calculate the similarity between papers based on cosine similarity, which is treated as the possibility of recommendation. Then we determine two papers have a citation relationship when the similarity is greater than a threshold.

2.5 Ablation Experiments

In order to further explore the degree of influence of each factor in the citation recommendation network, we conduct an ablation experiment. A meta-path can be viewed as a feature, and an ablation experiment is to observe the performance after removing certain features of the model. If the indicator decreases it means that the corresponding node type has a positive impact on the recommendation effect, and vice versa it means that the corresponding node has a negative impact on the citation recommendation. If the indicator remains unchanged, it means that the node type has no effect on the results.

3 RESULTS AND DISCUSSION

The result shows that the proposed model HAN-CR performs better than node2vec and metapath2vec, as shown in Table 1. It shows that the proposed method HAN-CR performs best but with a relatively low AUC. This is because we only consider the citation relationships among 5,914 papers downloaded in the preliminary study. The ablation experiment shows that the author, subject, and semantic links have a positive effect on citation recommendation, while venue and keywords have the opposite effect, as shown in Table 2.

Table 1 Effectiveness of different methods

	HAN-CR	metapath2vec	node2vec
AUC	**0.52**	0.51	0.49
AP	**0.49**	0.48	0.49

Table 2 Results of ablation experiment

	all	-author	-venue	-keywords	-subject	-Semantic Links
AUC	0.52	0.50|**-0.2**	0.54|+0.02	0.54|+0.02	0.49|**-0.03**	0.51|**-0.01**
AP	0.49	0.48|**-0.1**	0.51|+0.02	0.51|+0.02	0.48|**-0.01**	0.49|0

4 CONCLUSION

In this paper, we extend the heterogeneous network by introducing semantic links and meta-path set, and apply the attention mechanism for better network embedding. The performance of this method performs best compared with other methods. The influence of different factors in the citation recommendation network is also further explored through ablation experiments. In the future, we hope to apply other deep learning methods for heterogeneous network embedding. In addition, we will supplement the reference metadata to generate more dense citation relationships for better performance.

5 REFERENCES

Ali, Z., Qi, G., Muhammad, K., Ali, B. and Abro, W. A. (2020), "Paper recommendation based on heterogeneous network embedding", *Knowledge-Based Systems*, Vol. 210**,** p.106438, doi: https://doi.org/10.1016/j.knosys.2020.106438.

Chen, J., Liu, Y., Zhao, S. and Zhang, Y. "Citation recommendation based on weighted heterogeneous information network containing semantic linking", *2019 IEEE international conference on multimedia and expo (ICME)*, IEEE, pp.31-36.

Chen, L., Jensen, C. S., Shahabi, C., Yang, X. and Lian, X. (2017), "Personalized Citation Recommendation via Convolutional Neural Networks", *Web and Big Data*, Vol. 10.1007 No. Chapter 23**,** pp.285-293.

Kobayashi, Y., Shimbo, M. and Matsumoto, Y. "Citation recommendation using distributed representation of discourse facets in scientific articles", *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pp.243-251.

Van Meteren, R. and Van Someren, M. "Using content-based filtering for recommendation", *Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop*, Vol. 30, pp.47-56.

Yang, Z., Wu, B., Zheng, K., Wang, X. and Lei, L. (2016), "A survey of collaborative filtering-based recommender systems for mobile internet applications", *IEEE Access*, Vol. 4**,** pp.3273-3287.