Citation Recommendation based on Citation links Explanation

Yanping Zhang\textsuperscript{1*}, Dongyang Lin\textsuperscript{1}, Xi Chen\textsuperscript{1} and Fulan Qian\textsuperscript{1}

\textsuperscript{1} School of Computer Science and Technology, Anhui University, Hefei, Anhui, AH551, China

*Corresponding author’s e-mail: 87025@ahu.edu.cn

Abstract. With the rapid increase of scientific papers, it is difficult for researchers to obtain appropriate references. Due to the black-box property of neural network, the existing content-based citation recommendation methods are mainly embedding text into a low-dimensional space, and citation links are not interpretable (e.g., paper A cites paper B because of topic X). To solve this problem, we propose a Weighted Citation Network with Citation Links Explanation (WCN-CLE) algorithm. Because citation links mainly depends on the content, we need to explore textual reasons for citation links. Firstly, we extract word fragments through citation network information for each citation link as its explanation. Secondly, we construct a weighted HIN network with two kinds of vertex (paper and author). The weight of each edge is calculated by its word fragments. Then the query document is linked to the HIN network through author information and word fragments similarity with candidate papers. Finally, each vertex’s feature representation is obtained by network representation learning method, and recommend papers through vector similarity. Experimental results on two real-world datasets show that WCN-CLE has better performance due to its integration of structural information and semantic information.

1. Introduction

With the rapid development of information technology and the exponential growth of scientific publications, researchers cannot fully consult and digest the existing literature. Therefore, the efficient processing of scientific papers needs new technology. Citation recommendation is the key technology to solve this problem. It recommends a list of references related to researchers' information needs.

Most citation recommendation models are divided into two categories: local recommendation and global recommendation. Local recommendations are context-specific recommendations, and the global recommendations are recommendations for a given manuscript. We will focus on the latter: the global citation recommendation.

There are two kinds of information available in citation recommendation work: citation network structure and text content. In recent years, global recommendation methods mainly include content-based method \[1, 2\] and graph-based \[3, 4\] method. However, neural network has the black-box property \[5\]. The existing methods are mainly based on the similarity calculation after neural network embedding, which cannot explain the reason of similarity. So, users are confused about the returned recommendation list. They don't know why to recommend these papers to them.

To avoid black box property of citation recommendation work, we need to explore the reasons for citation in text level. For example, if the topic of a paper is about "machine learning", it is very likely that there will be the phase "machine learning" or related words/phrase in the content. The references to paper containing this word will also have a probability of being relevant topics. However, it is high
occasionality to find similarities only from the citing paper and cited paper. What if other papers have the same citation? This provides a way for us to explain the citation. As shown in Figure 1, papers [1, 2, 3, 5, 6, 7] refer to paper 4. For citation link paper 1 cites paper 4, we can find that 1/2 of the papers cite paper 4 contain "machine learning". Therefore, we can interpret this citation link as paper 1 cites paper 4 because "machine learning". We don't even need to think about the topic of the paper 4.

Figure 1. Topics included in the paper and their citation relationships.

WCN-CLE links the query documents into the network through word fragments. Then, the network representation learning method is used to map the vertices to the low dimensional embedding space, so as to calculate the similarity between the vertices in the network. Finally, recommend the most similar papers as recommendation list.

In this paper, we proposed a citation recommendation method (WCN-CLE) for scientific papers. The main contributions are summarized as follows:

1. Construct a weighted HIN network to represent the relationship between papers.
2. A new word fragments extraction method is proposed to explain each citation link.
3. Experiments on real-world datasets prove the effectiveness.

2. The WCN-CLE Framework

We present the definitions of tasks in this paper. Formally, given a query document q, the algorithm searches a list of documents most similar to q from a large number of published documents D, and its task is to find the most suitable reference of q in D.

2.1. Word Fragments Extraction and HIN Construction

We need to find the appropriate text for each citation link to explain. For each paper px, using word fragments in the text of the abstract and title as textual set Kx. The explanatory text K_{px, py} of citation link (px, py) is as follows:

\[ K_{px, py} = \bigcup_{k_i \in K_x} \left\{ k_i \ \bigg| \left| \frac{D_{ki} \cap N_{pi}}{N_{pi}} \right| > \alpha \right\} \]  \hspace{1cm} (1)

Where D_{ki} is all papers that contain k_i and N_{pi} is all papers that cites paper py. \( \alpha \) is hyperparameter, which limits the minimum probability of word fragments co-occurrence.

Obviously, not all word segments are useful, such as 'and', 'of'. These short word fragments have a high co-occurrence score, but it is useless. We calculate a score u_{ki} for each word segment to determine whether it is useful or not.

\[ u_{ki} = \frac{1}{|D_{ki}|^2} \sum_{px \in D_{ki}} \sum_{py \in D_{ki}} \left( \frac{|C_{px} \cap C_{py}|}{|C_{px}|} + \frac{|C_{px} \cap C_{py}|}{|C_{py}|} \right) \]  \hspace{1cm} (2)

Where C_{px} is references list to paper px. The higher the u_{ki} score, the more concentrated the distribution of papers where k_i words are located in the citation network. On the contrary, if the u_{ki} is lower, the more randomly the papers are distributed.
The word in Kx will be filtered by ukj. The sum of the u scores of the remaining words in Kpx, py becomes the weight of citation link (px, py).

\[ W_{px, py} = \sum_{k \in K_{px, py}} u_{kj} \]  

(3)

Where K’ is the textual set of word fragments after filtering.

All word fragments of the query document are used as a textual set. After filtering, the query document is connected with px, py through the similarity between the query’s textual set with citation link (px, py) explanatory text.

2.2. Network Representation Learning and recommendation

The goal of network representation learning is to map all the vertices to a low dimension R, and the space dimension is far less than the number of nodes. f is the mapping function of multitype vertex to feature representation. We use Node2vec [5] to learn f. Node2vec is a network representation learning algorithm, which maximizes the logarithm probability of Ns (v) in the network neighborhood according to the characteristic representation of node v \( v \in V \). It seeks to optimize the objective function, which is defined as follows:

\[ \max_f \sum_{v \in V} \log \Pr(N_s(v) | f(v)) \]  

(4)

When learning the low-dimensional representation of each node, the cosine similarity S is calculated. The top-k papers with the highest score are recommended. In addition, the reason why the query document is connected to the candidate paper network will also be returned.

\[ S_{q,p} = \frac{\langle f_q, f_p \rangle}{\| f_q \| \| f_p \|} \]  

(5)

3. Datasets and Experiments

3.1 Datasets

We experiment on the AAN (ACL Anthology Network) and DBLP dataset. The AAN dataset contains over 13K scientific articles, with an average of 5.5 citations per article. For DBLP dataset, we choose a subset instead of using full dataset. It contains over 18K scientific articles in the computer science domain, with an average of 7.1 citations per article. In both datasets, a document is accompanied by its title, abstract, venue (i.e. journal or conference where the document was published), authors, citations (i.e. other documents in the corpus that are referenced in the given document). We divide the data by year, where papers published in 2014 as target paper set for AAN and published in 2017 as target paper set for DBLP. For text preprocessing on abstract, we perform lemmatization and remove the stop words.

3.2 Baselines and Evaluation metrics

We compare our method with some baseline methods for citation recommendation: BM25, Doc2vec, Node2vec, WHIN-CSL.

BM25: BM25 is a text-based method, which computes similarity scores using only text information. We use BM25 as an IR-based baseline for the task of citation recommendation.

Doc2Vec [6]: Doc2vec is an unsupervised learning algorithm, which can learn the vectorization representation of different length documents. It is used to predict a vector to represent different documents.
Node2Vec [7]: Node2Vec is to maximize the probability of the occurrence of the nearest vertex given each vertex. Each node in the network is embedded into a low-dimensional space. We link the query document through the author and venue. Then train the network.

WHIN-CSL [8]: WHIN-CSL adds semantic linking to the citation network to capture semantic relations between objects. Then train the network and recommendation.

In all experiments, we use the abstract and title extracted from the dataset as context. We reported the results of standard measures for information retrieval and recommendation: Precise, Recall, F1 Score on the test set. Recall is computed as follow:

\[
\text{Recall}@k = \frac{1}{Q} \sum_{j=1}^{Q} \frac{R_p \cap T_p}{T_p}
\]  

Where Q is the number of target papers, and k is the recommendation numbers. For each target paper, Rp is the top-k paper list recommended based on a target paper p. Tp is the set of papers citing p.

3.3 Performance Comparison

We compare the proposed recommendation model WCN-CLE with other baselines in terms of citation recommendation performance. We compare the proposed methods with four different baselines using P@25, 50, Recall@25, 50 and F1@25, 50. Table 1 and Table 2 summarize the comparison results of DBLP and AAN datasets. In general, the proposed WCN-CLE method is superior to other methods in all metrics. Specifically, WCN-CLE obtains a 29.5% improvement in Recall@25 compared to the best baseline on the DBLP dataset. On the DBLP dataset, Recall@25 compared with WHIN-CSL, it increased by 11.4 percentage points.

Table 1. Performance comparison of different methods on AAN dataset.

| Metrics | P@25 | P@50 | Recall@25 | Recall@50 | F1@25 | F1@50 |
|---------|------|------|-----------|-----------|-------|-------|
| BM25    | 0.0515 | 0.0343 | 0.1912    | 0.2535    | 0.0811 | 0.0604 |
| Doc2Vec [6] | 0.0171 | 0.0125 | 0.0574    | 0.0863    | 0.0263 | 0.0218 |
| Node2Vec [7] | 0.0339 | 0.0127 | 0.1231    | 0.1694    | 0.0531 | 0.0236 |
| WHIN-CSL [8] | 0.0377 | 0.0293 | 0.1344    | 0.2020    | 0.0588 | 0.0511 |
| WCN-CLE | **0.0677** | **0.0441** | **0.2504** | **0.3136** | **0.1065** | **0.0773** |

Table 2. Performance comparison of different methods on DBLP dataset.

| Metrics | P@25 | P@50 | Recall@25 | Recall@50 | F1@25 | F1@50 |
|---------|------|------|-----------|-----------|-------|-------|
| BM25    | 0.2478 | 0.1570 | 0.4745    | 0.5917    | 0.3255 | 0.2481 |
| Doc2Vec [6] | 0.0866 | 0.0595 | 0.1700    | 0.2315    | 0.1147 | 0.0946 |
| Node2Vec [7] | 0.2316 | 0.1502 | 0.4473    | 0.5708    | 0.3051 | 0.2378 |
| WHIN-CSL [8] | 0.1960 | 0.1357 | 0.3802    | 0.5181    | 0.2586 | 0.2150 |
| WCN-CLE | **0.2826** | **0.1800** | **0.5461** | **0.6818** | **0.3724** | **0.2848** |

Citation recommendation work has achieved better recommendation results through citation links explanation. WCN-CLE can better ensure the stability of the original citation network than adding semantic links into the citation network. Similar to BM25, we did not embed words or sentences into a low-dimensional space. Keeping the original text, which is also the basis for the explaining citation links. However, compared with Doc2Vec and WHIN-CSL, which embed text in a low-dimensional space, the two methods of non-embedding text achieve better results.

4. Conclusions

In this paper, a new citation recommendation algorithm WCN-CLE is proposed. The algorithm captures the text features of citation relationship, and then makes citation recommendation. Experimental results show that compared with baseline algorithm, the proposed model can greatly improve the performance. We believe that it can open a door for text-based citation recommendation. We also want to use the full text of the paper to build text features, rather than just using abstracts and titles. There are also several
directions worthy of further exploration, such as the filtering method of word fragments and the extraction method of words.

Acknowledgments
This work is supported by:

- National Key Research and Development Program of China (2017YFB1401903).
- National Natural Science Foundation of China (Grants #61876001, #61602003 and #6167320).
- Major Program of the National Social Science Foundation of China (Grant No.18ZDA032).
- Provincial Natural Science Foundation of Anhui Province (#1708085QF156).

References

[1] Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015) A neural probabilistic model for context based citation recommendation. C. In Twenty-ninth AAAI conference on artificial intelligence.

[2] Dai, T., Zhu, L., Wang, Y., & Carley, K. M. (2019) Attentive Stacked Denoising Autoencoder With Bi-LSTM for Personalized Context-Aware Citation Recommendation. J. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 553-568.

[3] Ren, X., Liu, J., Yu, X., Khandelwal, U., Gu, Q., Wang, L., & Han, J. (2014). Cluscite: Effective citation recommendation by information network-based clustering. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 821-830.

[4] Guo, L., Cai, X., Qin, H., Guo, Y., Li, F., & Tian, G. (2019) Citation Recommendation with a Content-Sensitive DeepWalk Based Approach. C. In 2019 International Conference on Data Mining Workshops (ICDMW), 538-543.

[5] Ouarti, N., & Carmona, D. (2018) Out of the Black Box: Properties of deep neural networks and their applications. J. arXiv preprint arXiv:1808.04433.

[6] Le, Q., & Mikolov, T. (2014) Distributed representations of sentences and documents. C. In International conference on machine learning (pp. 1188-1196).

[7] Grover, A., & Leskovec, J. (2016) node2vec: Scalable feature learning for networks. C. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 855-864.

[8] Chen, J., Liu, Y., Zhao, S., & Zhang, Y. (2019) Citation recommendation based on weighted heterogeneous information network containing semantic linking. C. In 2019 IEEE International Conference on Multimedia and Expo (ICME), 31-36.