Augmenting context-aware citation recommendations with citation and co-authorship history

Anubrata Bhowmick[1], Ashish Singhal[2] and Shenghui Wang[3]

[1] a.bhowmick@student.utwente.nl [2] a.a.singhal@student.utwente.nl [3] shenghui.wang@utwente.nl
University of Twente, Drienerlolaan 5, 7522 NB Enschede, (The Netherlands)

Abstract

With the increasing number of research papers being published, it has become a challenge to search for the most suitable articles for accurate referencing. Many local citation recommendation systems have begun to locate the suitable candidates by using the texts accompanying the citation suffix, along with the metadata of the target documents. Previous research has shown the positive effects from the citation relationships on such recommendations, however, the influence from the co-authorship history has not been fully investigated. In this paper, we extend the model proposed by Jeong, Jang & Park (2020) by combining the context, citation history with co-authorship information into the recommendation system. We also propose to use more domain-specific embeddings to better capture the semantics in the context. Our experiments show the positive effect of co-authorship information on citation recommendations, and that our model based on the combination of domain-specifically embedded context, the citation and the co-authorship history significantly outperforms the basic context-based recommendation model.

Introduction

Citation Recommendation is defined as the task of recommending citations from a textual content. Due to the increasing number of scientific works in recent years, and the need of citing appropriate publications when writing scientific papers, citation recommendation has become one of the most important research topics. Depending on the length of the citation context that are used for recommendation, there are two categories of methods: 1) global citation recommendation where the entire text has been used to recommend citations or abstract can be used to understand the context, and 2) local or context-aware citation recommendation where a short-length window of words surrounding the citation, as shown in Figure 1, are used as context for recommending citations.

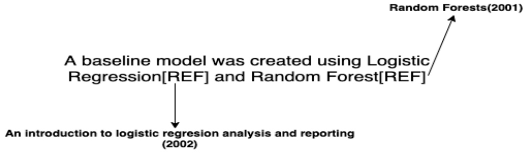
Random Forests(2001)
A baseline model was created using Logistic Regression[REF] and Random Forest[REF]
An introduction to logistic regresion analysis and reporting (2002)

Figure 1. Example of paper with citation mentions with [REF] placeholders.

There arises a vocabulary gap between the context and the corresponding cited publication which results in recommending not so high-quality publications. A publication's quality can be estimated by its previous citation relations and its authors' previous collaborations. A field breakthrough paper citing another publication speaks volumes about the cited paper quality. Similarly, a paper quality can also be determined individually by its authors, and with whom they have shared and gained knowledge by collaborating in the past. Currently existing state-of-the-art models have explored citation relations and metrics to recommend appropriate citations such as the BERT-GCN model proposed by Jeong, Jang & Park (2020). Very few have explored co-authorship relations due to the mixed views about the influence of co-authorship in this task. Ebesu & Fang (2017) employed co-author networks and found promising results and new direction for context-aware citation recommendation. In our paper, we extend the existing model of Jeong, Jang & Park (2020) with co-authorship network and domain-specific embeddings to understand if these may help in improving the performance of the existing model. By devising a way to use citation relations and co-authorship relations along with the domain-specific context embeddings, we observe a significant improvement over the existing model. Evidently, we managed to improve the Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) by 5% and Recall@k by almost 10%.

Related Work

Citation recommendation is about recommending citations given a piece of text. These systems are divided into global and local citation recommendation where global recommendation recommends the citations considering the whole publication text while local recommendation recommends citations based on local context of input text. There is a recommendation system which employs both local and global context together as described in (Muhsina & Naseer, 2019). Yang (2018) used forward and backward positions of the quotation marks of the citation in an LSTM based model to customize the context-aware citation recommendation by splitting the left and right sentences when encoding the citation context. It suggested a procedure for learning and combining LSTM cells with MLP to learn. The investigators have used the papers' author(s) and location details to encode the research model. There are other works such as Ebesu & Fang (2017) which used citation context along with authors network convolution to develop the Neural Citation Network for context-aware citation recommendation systems. This paper combined the cited and citing author network features in the citation context encoders. It becomes important to focus on co-authorship in scientific field as this dictates the quality of the research output. As per Kumar (2015), co-authorship networks generated from co-authorship meta information from various publications gives rise to a social network among the researchers which describes who is more willing to collaborate with whom influencing the publication quality. He also found that there is high correlation between co-authorships and citation behavior. It is found that co-authored papers are cited more than papers with single authors and Ding (2011) found in their study that more cited authors usually do not co-author with each other but cite each other. Lis, Yang & Ding (2009) in their study found that centrality measures from co-authorship network are highly correlated with citation counts. Biscaro & Giupponi (2014) in their study concluded that co-authorship networks' structure matters in scientific collaboration. Publications' bibliography information not only identifies the publication uniquely but also gives information about the quality of the publication. The bibliographic information such as citation count of publication, venue, year, author determines the quality of the work. These information are employed by the graph techniques which not only considers the citation count in the form edges in the graphs but also considers other bibliometric information as the nodes' features. Kipf & Welling (2016) proposed semi supervised classification with graph convolution networks to classify graph nodes by using the graph edges and nodes' features. They also introduced variational graph autoencoders to learn the latent representation of the graph.

Data Overview

For our work, we use the FullTextPeerRead dataset[1] proposed in Jeong, Jang & Park (2020). This dataset consists of peer reviews of submitted papers in top-tier venues of the ML/AI field, along with their bibliometric information. This dataset contains left and right context sentences

[1] https://bert-gcn-for-paper-citation.s3.ap-northeast-2.amazonaws.com/PeerRead/full_context_PeerRead.csv.

surrounding a citation[2] as well as the cited and the citing papers' meta information such as title, author, abstract, etc.

Table 1: Dataset description

Dataset Name	*FullTextPeerRead*[1]
Total number of papers	4,898
Total number of base papers	3,761
Total number of cited papers	2,478
Total number of citation context	17,247
Total number of unique authors	9,529
Years of papers published	2007-2017

This dataset has been split into the training and the test datasets based on the cut-off year of 2017. The training dataset consists of 3,411 papers and the test data consists of 2,559 papers. The left and right context sentences length has been trimmed to 50 words, such that sufficient information is taken from both sides of the citation.

Methodology

In order to check the effects of better context-embedding and co-authorship information, we extend the BERT-GCN model proposed in Jeong, Jang & Perk (2020) by incorporating the co-authorship network, as illustrated in Figure 2. The model consists of three components that gives embeddings for context sentences, co-authorship network and citation network. Here, the BERT transformer is used to learn the feature representation of context sentences surrounding the citations and individual Graph Convolution Networks (GCNs) are used to learn the network representations of citation network and co-authorship network. The code can be found here.[3]

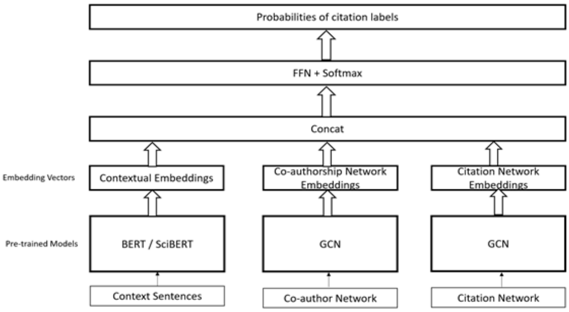

Figure 2: The BERT-GCN[2] model architecture

Context embedding

The context of a citation consists of the left and right text surrounding the cited references. We use the pre-trained BERT (Devlin, et. al., 2019) and SciBERT (Beltagy, et. al., 2019) to generate the embedding of such context. BERT in pretrained on Wikipedia articles and is capable to understand the meaning of its input due to its ability to read in both directions at once. SciBERT, an extension to original BERT is pretrained on multi-domain scientific publications and has the vocabulary which helps better understand scientific texts. As our main goal is to find suitable reference papers in computer science domain, SciBERT could be highly relevant in processing the input data from the dataset.

[2] This dataset does not consider multiple citation data where multiple papers are cited in a single sentence.

[3] https://github.com/anubratabhowmick/tf-BERT-GCN

Citation and co-authorship embedding

We generate the co-authorship network and citation network from the dataset. Two separate Graph Convolutional Network (Zhang, Hanghang & Jiejun, 2019) models are trained on these two different networks, generating two sets of embeddings for each author and each citation respectively. More specifically, for each network, we train a Variational Auto-encoder Graph Convolutional Network (VGAE) as proposed by Kipf & Welling (2017). Variational Graph Autoencoders works similarly as variational neural network auto encoders that generates stochastic based representation embeddings. We use VGAE GCN to generate the co-authorship and citation network embeddings.

Concatenation and training

The context around citations is the input to the BERT/SciBERT which on training learns these context feature representations and generates the contextual embedding. For each context sentence passed to BERT/SciBERT, the corresponding paper's citation embedding and the embedding of the last author, who is often the supervisor in the computer science domain, is concatenated with the BERT/SciBERT contextual embeddings output and passed to the feedforward network (FFN) layer. The output from this layer is then passed to the softmax activation function which then generates the probabilities of each publication being the citation for the context sentence passed as input. The BERT/SciBERT along with the feedforward layer is trained on the cross-entropy loss calculated by trying to predict the actual cited publication given the surrounding context, the corresponding citation and author embeddings as the input to the model.

Experiment Setting

The experiment has been done in eight variations, with the first four variants being trained with BERT and the last four with SciBERT, to understand how much better context embedding helps in the performance, and how much information does the respective citation and co-authorship network add to the models, in respect to the existing Jeong, Jang & Park, 2020 model. In the initial four variants, we have made a comparison of the BERT model exclusively and then, by adding the networks separately to BERT, and ending with the combined citation and co-authorship embedding added to BERT together. We repeated the same process in the next four variations by replacing the BERT with the SciBERT model.

BERT/SciBERT and GCN Setting

The context sentences are passed to BertTokenizer/SciBertTokenizer to tokenize the context sentences, respectively. The initial experiments, which only includes BERT and SciBERT separately have been trained by using these tokenized context sentences. BERT (and SciBERT) generates the contextual embeddings of the input at the [CLS] which is passed to the feedforward layer, followed by the SoftMax activation, giving the final output. The feedforward layer's input size is 768 which is equal to the output size of BERT's [CLS] token. The feedforward layer, followed by the SoftMax activation, generates the citation probabilities output of the size of 489 that is the total number of citation candidates, i.e. those that have been cited for at least 5 times in the training set. The BERT/SciBERT is being trained for 30 epochs with batch size set to 16. For training purposes, Adam optimizer (Kingma, & Ba, 2015) is used with the default parameters and the learning rate set to 2e-5.

For other settings where co-authorship networks' and citation networks' embeddings are being used, GCN has been trained for 200 epochs. The first hidden dimension in GCN for citation network training is 4837 which is total number of documents in the dataset and for co-authorship network it is 9529. The second hidden dimension in GCN is the same for both networks which is 768. The batch size for training for both networks is the total number of documents and the total number of authors in the dataset (full-batch gradient descent). For training, Adam optimizer is used with the learning rate of 0.01.

Results

We use similar metrics including Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Recall at k as mentioned in (Jeong, Jang & Park, 2020). MAP measures the average precision which reflects upon the rank position regarding the retrieval list. MRR reflects on the position of the actual result in the recommendation list. A higher MRR score indicates the higher ranks of the target citations in the recommendation list. Recall at k (R@k), where k is in [5,10,30,50,80], is used to measure the actual hit ratio in the top k recommendation results.

Table 2: MRR, MAP and R@k scores of experiments

Model Name	*MRR*	*MAP*	*R@5*	*R@10*	*R@30*	*R@50*	*R@80*
BERT Base*	0.415	0.415	0.480	0.520	0.593	0.637	0.689
BERT-GCN*(Citation)	0.418	0.418	0.486	0.529	0.604	0.649	0.699
BERT Base	0.432	0.432	0.504	0.548	0.621	0.668	0.717
BERT-GCN (Citation)	0.412	0.412	0.481	0.523	0.606	0.649	0.701
BERT-GCN (Co-Authorship)	0.439	0.439	0.505	0.556	0.632	0.677	0.728
BERT-GCN^2	0.443	0.443	0.516	0.561	0.643	0.689	0.735
SciBERT Base	0.467	0.467	0.542	0.593	0.675	0.722	0.766
SciBERT GCN (Citation)	0.467	0.467	0.547	0.592	0.677	0.722	0.772
SciBERT GCN (Co-authorship)	0.466	0.466	0.545	0.594	0.680	0.719	0.773
SciBERT-GCN^2	**0.468**	**0.468**	**0.548**	**0.598**	**0.685**	**0.733**	**0.781**

Table 2 shows our results and those (marked with *) reported by Jeong, Jang & Park (2020). Our self-implemented BERT base model outperforms the previously reported results, while the combined BERT and GCN with the citation network has slightly poor results than previously reported. This might be due to our parameter settings are not optimal. However, we can easily see the BERT-GCN with the co-authorship network performs better than both BERT-base and BERT-GCN with the citation network. A combination of both networks further improves the performance. Replacing BERT with SciBERT improves the performance significantly, even without the citation and the co-authorship network. The citation and co-authorship networks, when added to SciBERT show negligible increase in performance, but when both networks combined, the performance is again improved substantially. This suggests that co-authorship information has positive influence on context-aware citation recommendation. A combination of both the citation and co-authorship network to a base BERT/SciBERT model can certainly improve the recommendation performance.

Conclusion

Our proposed addition of the co-authorship network to the existing citation network, delivers a significant improvement in MAP, MRR and Recall@k over the existing model. This clearly shows that the co-authorship network manages to add meaningful information to the base model, and a combination with the citation network results in better citation recommendations. We can conclude that, in addition to the co-authorship network, better context embedding plays a significant role when it comes to model performance, and using proper context embeddings, along with the appropriate networks can go a long way in effective citation recommendation

There are also room for improvements in the model proposed in this paper. We have selected the last author's network embedding to be added to our model, but another viable way of using the co-author embeddings is separating the authors, increasing the dataset, and using all their embeddings separately. Although this would be computationally more expensive, there is a chance of improvement in performance over our proposed model. Another improvement could be using node features in the Graph Convolution Network, as encoding both the graph structure and the node features might improve the performance by a significant margin, while being computationally efficient, and could be worthwhile to investigate further.

References

Beltagy, I., Cohan, A. & Lo, K. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 3615-3620. http://dx.doi.org/10.18653/v1/D19-1371

Biscaro, C. & Giupponi, C. (2014). Co-authorship and bibliographic coupling network effects on citations. *PLoS ONE*, 9(6), e99502. https://doi.org/10.1371/journal.pone.0099502

Ding, Y. (2011). Scientific collaboration and endorsement: network analysis of Coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187-203. https://doi.org/10.1016/j.joi.2010.10.008.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of The North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* volume 1, 4171-4186.

Ebesu, T. & Fang, Y. (2017). Neural Citation Network for Context-Aware Citation Recommendation. *Proceedings of the 40th International ACM SIGIR Conference of Research and Development in Information Retrieval (SIGIR '17).* ACM, New York, NY, USA, 1093-1096.

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. https://doi.org/10.1007/s11192-020-03561-y

Kipf, T. N. & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *Neural Processing Letters*, 1 -12.

Kipf, T., & Welling, M. (2016). Variational Graph Auto-Encoders. NIPS Workshop on Bayesian Deep Learning. arXiv:1611.07308

Kingma, D.P. & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Kumar, S. (2015). Co-authorship networks: A review of the literature. *Aslib Journal of Information Management.* 67(1), 55-73. https://doi.org/10.1108/AJIM-09-2014-0116

Ma, Y., Hao, J., Yang, Y., Li, H., Jin, J. & Chen, G. (2019). Spectral-based Graph Convolutional Network for Directed Graphs. ArXiv, abs/1907.08990.

Muhsina, V. P. & Naseer, C. (2019). A Survey on Citation Recommendation System. *International Journal of Advanced Research in Computer and Communication Engineering, 8*(1), 85-91.

Yan, E.J. & Ding, Y. (2009). Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2017-2118. https://doi.org/10.1002/asi.21128

Yang, L., Zheng, Y., Cai, X., Dai, H., Mu, D., Guo, L. & Dai, T. (2018). A LSTM based model for personalized context-aware citation recommendation. *IEEE Access,* 6, 59618-59627. https://doi.org/10.1109/ACCESS.2018.2872730

Zhang, S., Tong, H., Xu, J. & Maciejewski, R. (2019). Graph Convolutional Networks: a Comprehensive Review. *Computational Social Network, 6, 11* https://doi.org/10.1186/s40649-019-0069-y