Citation Prediction with Weighted Citation Network Links and Contextual Semantic Information

Jonathan C. Barker **Kartik Goyal** **Zhengzhong Liu**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{jcbarker, kartikgo, liu}@cs.cmu.edu

Abstract

In this work, we address the problem of missing citation prediction. Current approaches to this problem includes making prediction through the citation network structure and identifying them through semantic similarity. In this work, we attempt to incorporate both these information to help make better prediction. We spot the phenomenon that not all cited papers are equally important to the citing paper. Thus we adopt a supervised random walk method to learn the appropriate weights on these links using textual semantic information and cross-document topic models. We evaluate our weighted network by predicting missing citations for a paper, given the paper's current citations.

1 Introduction

Identifying relevant literature from electronic collection is becoming difficult with the rapid growth of the collection. This makes it very challenging for researchers to conduct efficient and comprehensive literature review. Thus automation of such process become necessary. We consider the problem as a traditional network link prediction problem. Existing citation networks only provide the links between citing and cited papers. Such networks can provide us binary information about whether a link exists between two papers, but fail to encode information about the relative importance and relevance. As a result, it is difficult to draw profound conclusions from these networks about the relationships between different network nodes. We address this problem by assigning weights to network links according to their importance and relevance to the cited paper in order to mine deep information. Tasks like automatic paper recommendation and missing citation prediction can be made possible. Authors can use it to detect any papers they might not have cited in their initial draft, while journal reviewers can be aided in detecting any relevant works which are not cited by the submitted papers.

In practice, various techniques like random walks [8, 9] have been employed to find node relevance in networks. These techniques generally work with a uniformly weighted network and utilize information encoded in graph structure. Our belief is that these techniques can be used more effectively if the network graph were to contain weighted links. The weights on the links would represent the similarity between citing and cited papers according to their semantic correlation, citing patterns, shared authors and temporal proximity. Using these weights should provide more meaningful results for tasks like missing citation prediction when the existing network based techniques are applied.

The major challenge in this approach is figuring out how various sources of information can be combined with the network structure. In doing so we must determine the relative importance of these sources/features with respect to the relationship score of two papers. When have finally synthesized this information to create our weighted network we aim to show that we can use it to produce more meaningful results for tasks such as missing citation prediction.

2 Related Work

Nallapati et al. [6] use topic models created with Latent Dirichlet Allocation(LDA) [**?**] to predict the existence of a citation link between two papers. Apart from generating words, they also generate the reference links from the topic model, thus creating a topical dependency between cited and citing papers.
Backstorm and Leskovec [1] perform supervised random walks on networks to learn the weights on the links between community members. They use this weighted network to predict future links between the members of the social network. The weights are learnt by minimizing an error function on the probability of predicting incorrect and correct future links.
LDA can be used for determining semantic correlations between two documents. Celikyilmaz et.al describe some of the similarity measures based upon LDA which they used in the QA systems. These similarity measures used both, topic proportions and importance of similar words for estimating similarity.

3 Method

A major challenge of this work is how to learn the relevant weights of features of according, we aim to learn a function that can properly assign weights based on the dataset. Here we adopt a method similar to [1] which optimize a function that can assign higher random walk score to real missing nodes versus other nodes. The network we will attempt to create weights for is the ACL Anthology Network (AAN). Since this citation network does not have weights on links, intrinsic evaluation is not feasible. Instead we will perform an extrinsic evaluation of our weighted network by predicting missing citations.

The weights on the network are determined by various features which depend on citation patterns, semantic correlation and the network metadata. These features are defined for a citing and cited paper:

1. Number of citances for the cited paper. (”citance” refers to a citing sentence in a paper)
2. Co-citance score. This is defined as the number of times another paper is cited in the same citance as the target cited paper.
3. The number of common papers which both citing and cited papers cite.
4. Topical similarity between the citing and cited paper, calculated using topic proportions from Latent Dirichlet Allocation(LDA).
5. Number of common authors.
6. The difference between time of publication, measured in years.

To extract features 3, 5 and 6, we used the metadata from the AAN. To determine 1 and 2 we used regular expressions to extract citances and match them with the cited papers. Mallet toolkit was used for running LDA, which uses gibbs sampling for parameter inference, on the raw text of all the papers. The topic proportion vectors were created and were used to calculate cosine similarity between documents.

4 Experiment

Here are the results:1

5 Conclusion

References

[1] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM ’11, pages 635–644, New York, NY, USA, 2011. ACM.

Table 1: Results

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

[2] David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] Asli Celikyilmaz, Dilek Hakkani-Tur, and Gokhan Tur. Lda based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, SS '10, pages 1–9, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[4] Aaron Clauset, Cristopher Moore, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[5] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002.

[6] Ramesh Nallapati, Amr Ahmed, Eric P. Xing, and William Cohen. Joint latent topic models for text and citations. In *KDD*, pages 542–550, 2008.

[7] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, page 542, New York, New York, USA, 2008. ACM Press.

[8] Purnamrita Sarkar and Andrew W. Moore. Fast dynamic reranking in large graphs. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 31, New York, New York, USA, 2009. ACM Press.

[9] Hanghang Tong, Christos Faloutsos, and JY Pan. Fast random walk with restart and its applications. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 613–622. Ieee, December 2006.

[10] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *KDD*, 2009.