Weighting Citation Network Links Using 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

Citation networks are very important for understanding the interactions among publications and their relative importance to the field. In generating these networks, most of the past works have treated the citations of a paper as equally important. However, it is often the case that not all cited papers are equally important to the citing paper. In this project, we propose to learn the importance of cited papers using the cited document's textual semantic information and cross-document topic models in order to generate a weighted citation network. We evaluate our weighted network by predicting missing citations for a paper, given the paper's current citations.

1 Introduction

Existing citation networks only provide the links between citing and cited papers. Such networks can tell us that a given paper cites 10 others, but they provide no information about the relative importance of the cited papers to the citing paper. As a result, it is difficult to draw profound conclusions from these networks about the relationships between different papers in the network. If the links in the network are weighted according to their importance to the cited paper, then the network can easily be used to perform several tasks like automatic paper recommendation and missing citation prediction. 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 paper they are reviewing.

In practice, various techniques like random walks have been employed to infer information from a citation network, such as the popularity of an author or paper. These techniques work with a uniformly weighted network and utilize graph information like common neighbors and path length. 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 to obtain weights on the network links. 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.

Table 1: Results

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

2 Related Work

Nallapati et al. [**?**] use topic models created with Latent Dirichlet Allocation 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 [**?**] perform supervised random walks on the Facebook network to learn the weights on the links between facebook members. They use this weighted network to predict future links between the members of the social network. They learn the weights by minimizing the their error function which depends 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

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:**??**

5 Conclusion