Modeling Evolution of Ideas in the Web of Science

Laura Dietz and Steffen Bickel
Max Planck Institute for Computer Science
66123 Saarbrücken, Germany
{dietz,bickel}@mpi-inf.mpg.de

1 Problem Statement

When reading on a new topic researchers need to get a quick overview about a research area. Especially when reading a publication in an unfamiliar area, the reader is interested in how this publication is embedded in the web of science. The goal is to give the reader an overview of the evolution of ideas leading to the examined publication. We want to analyze this evolution by inferring and visualizing the strength of topical similarity between linked publications in the neighborhood of the initially examined work.

Digital libraries provide background knowledge about the universe of publications with titles, abstracts, and the underlying citation network. The publications examined by the user – referred to as seed publications – represent the input to the following problem. Given the seed publications $\mathcal{S}$, select the subgraph of the citation network that contains the seed publications along with approximately n additional publications $\mathcal{P}$, that have the strongest direct or indirect impact on the seed publications. For each selected publication the strength of influence on each seed publication is to be quantified. We call $\mathcal{S} \cup \mathcal{P}$ the relevant subgraph of the underlying citation network.

Note, that we do not make assumptions on the structure of the relevant subgraph. It may have a star-shaped structure if several topic areas influence the seeds; or it may consist of a few long chains if the seed contributes a technicality in the sequence of many modifications of original work. In contrast to our previous work [1] we predict evolution paths leading to the seed publications.

2 Approach: Seed Impact Model

In the domain of scientific publications, we assume that influencing work is in the citational vicinity. Work that is not directly cited is assumed to be linked via strong intermediate citation links. For instance, consider work that extends a special kind of Support Vector Machine. It may be that early prior work on SVMs is not cited, but has an implicit influence on the considered publication.

We follow the intuition of latent topics flowing over citation links from publication to publication. To address the problem statement, we examine topical flows draining into seed publications. The stronger the outgoing flow of a publication, the stronger is its absolute impact on the seeds. This impact measure acts as a ranking criterion for selecting the most relevant publications $\mathcal{P}$. Since we rely on short texts such as abstracts, it is necessary to take a step beyond syntactical similarity. Thus, we use an underlying probabilistic topic model based on Latent Dirichlet Allocation (LDA) [2].

The universe of publications is too large to be considered at once. We propose a sampling approach combining a generative model on fixed $\mathcal{P}$ with Metropolis-Hastings steps that alter the document set.

2.1 Generative Process

The generative model assumes words as atomic units which carry the topical flow through the citation graph.

Words in $\mathcal{P}$ that participate in a flow towards a seed publication, are modeled as nodes in a tree (cf. Figure 1). Tree edges may only connect words of publications that cite each other. Each word in a seed publication is the root of exactly one tree. Words in the seed publications are generated in an LDA-like manner, drawing a topic from the seed publication's topic mixture and generating the word from the topic's word distribution. The topic of a seed word is used as topic for the corresponding tree.

Words in a non-seed publication d are either assigned to a tree or modeled independently. First, a biased coin is flipped that decides whether the word is treated independently or associated with a tree. If not treated independently, the word is attached (by multinomial draw) to a tree that covers a publication citing d. The tree's topic is used to generate the word from the characteristic word distribution. The tree is extended with an edge towards the generated word. Words in non-seed publications that are not associated with a tree, are modeled via draws from d's own topic mixture as in LDA.

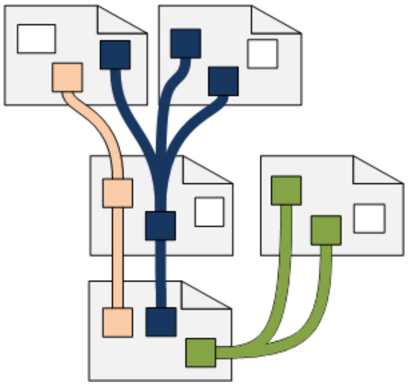

Figure 1: Three topical trees rooted in words of a seed document (bottom). Edges connect words (depicted by boxes) of the same topic in interlinked publications.

Each topical tree represents the evolution of an idea that leads to a seed publication.

2.2 Sampling

The model is sampled by interleaving Gibbs iterations and Metropolis-Hastings steps. The Gibbs iterations estimate parameters according to the generative model. In Metropolis-Hastings steps, we alter the document set $\mathcal{P}$ by either removing a publication from the set, or adding an adjacent publication. A document is proposed depending on the number trees it participates in.

$$\rho_{\text{add}} = \min\left[1; \frac{p(z^{new}|\mathcal{P}^{new}, \mathcal{S}) \cdot p(|\mathcal{P}^{new}|)}{p(z^{old}|\mathcal{P}^{old}, \mathcal{S}) \cdot p(|\mathcal{P}^{old}|)} \cdot \frac{q(\text{remove}||\mathcal{P}^{new}|) \cdot q_{\text{remove}}(d|\mathcal{P}^{new})}{q(\text{add}||\mathcal{P}^{old}|) \cdot q_{\text{add}}(d|\mathcal{P}^{old})}\right] \tag{1}$$

The acceptance probability for adding publication d to $\mathcal{P}^{old}$ is given in Equation 1. The first fraction represents the target densities which are divided into data likelihood given a fixed $\mathcal{P}$ and a Gaussian prior $p(|\mathcal{P}|)$ with mean n on the size. The remaining fraction involves the proposal density for an add operation and the probability for proposing document d.

The prior on the independence coin is tuned to give independent words a low probability. Thus, the data likelihood of a fixed document collection $\mathcal{P}$ increases the more words are associated to trees. In addition, the document proposal distributions prefer to select documents that are related to seeds.

3 Evaluation and Conclusion

The seed impact model is work in progress. We will conduct a survey asking authors to rate the influence of publications in the citational vicinity on their own work. We will use a Likert scale ($--$, $-$, $+$, $++$) for rating the strength of topical influence. For each decision boundary (e.g. $+$ vs. $++$) we calculate the area under the ROC curve (AUC). The AUC values are averaged to evaluate prediction performance.

We contribute a generative model that naturally models the evolution of ideas. Topics are inherited and shared along tree-structured subgraphs of the citation network. Although not detailed here, the approach is also capable of analyzing the impact of seeds on successive publications.

References

[1] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.