Unsupervised Prediction of Citation Influences

Laura Dietz DIETZ@MPI-INF.MPG.DE
Steffen Bickel BICKEL@MPI-INF.MPG.DE
Tobias Scheffer SCHEFFER@MPI-INF.MPG.DE
Max Planck Institute for Computer Science, Saarbrücken, Germany

Abstract

Publication repositories contain an abundance of information about the evolution of scientific research areas. We address the problem of creating a visualization of a research area that describes the flow of topics between papers, quantifies the impact that papers have on each other, and helps to identify key contributions. To this end, we devise a probabilistic topic model that explains the generation of documents; the model incorporates the aspects of topical innovation and topical inheritance via citations. We evaluate the model's ability to predict the strength of influence of citations against manually rated citations.

1. Introduction

When reading on a new topic, researchers need to get a quick overview about a research area. This can be provided by a meaningful visualization of the citation network that spins around some given publications. For example, consider researchers who read on a new topic with some references as starting points. The questions occur which other papers describe key contributions, how the various contributions relate to one another, and how the topic evolved over the past.

Clearly, Google Scholar, CiteSeer and other tools that allow to navigate in the publication graph provide a great support and have made this task much easier. Inspired by the work of Garfield (2004), we would like to create a bird's-eye visualization of a research area that complements in-depth navigation in the publication graph. But because the publication graph is linked densely, even a radius of two citations from a pivotal paper contains hundreds of publications. When examining the citations in detail, one finds that not all cited work has a significant impact on a citing publication. Papers can be cited as background reading, for politeness, fear of receiving an adverse review from an aggrieved reviewer, or as related work that was argued against. For a detailed list of motivations for citing publications see Trigg (1983), Sandor et al. (2006). By contrast, a bird's-eye visualization should show papers that significantly impact one another (as is depicted in Figure 3). This requires to measure the strength of a citation's influence on the citing work.

This paper starts by formalizing the problem setting, followed by a discussion of related work. Section 4 describes baseline algorithms based on Latent Dirichlet Allocation (Blei et al., 2003). In Sections 5 and 6 we introduce generative models that directly model the influence of citations in paper collections. We also derive a Gibbs sampler that learns the model from data. Section 7 empirically evaluates the different models. The paper concludes with a discussion and outlook.

2. Problem Statement

A universe of publications is given; publications consist of the full text or abstracts, as well as the citation graph structure. In the citation graph, vertices are publications, and a directed edge from node c to node d indicates that publication d cites publication c. In the following, $L(d)$ denotes the set of publications cited by d.

The goal is to find an edge-weight function γ measuring the strength of influence. Weights $\gamma_d(c)$ should correlate to the *ground truth* impact that c has had on d as strongly as possible. This ground truth is not observable. But for purposes of evaluation, we sample it by referring to the assessment of research scientists.

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

3. Related Work

Unsupervised learning from citation graphs has attracted much attention in the past. Bibliometric measures (Spiegel-Roesing, 1977) such as co-coupling are used in many digital library projects as a similarity measure for publications. In addition, graph based analyses have been introduced such as community detection (Flake et al., 2004), node ranking according to authorities and hubs (Kleinberg, 1999), and link prediction (Xu et al., 2005). Furthermore, progress has been made in studying how paper networks evolve over time (Newman, 2003).

Probabilistic latent variable models have a long tradition in this application area as well, such as the probabilistic formulation of HITS (Cohn & Chang, 2000), or stochastic blockmodels (Nowicki & Snijders, 2001) for the identification of latent communities. Those models became popular with pLSA and Latent Dirichlet Allocation (Hofmann, 2001; Blei et al., 2003) which learn hidden topics from text documents in an unsupervised manner. The topics are captured as latent variables that have a characteristic word distribution $\phi_t = p(w|t)$. The topics are inferred from the co-occurrences and automatically resolve synonyms and polysems.

The results of a Latent Dirichlet Allocation have been exploited in Mann et al. (2006) to extend bibliometric measures based on citation counts with topical information. On the downside, this approach is only feasible for publications which are cited reasonably often, which is not the case in specialized and evolving research fields. To address this, we suggest to do a probabilistic analysis on a fine-grained level (such as words).

Some efford has been made to include text and structure into the model on a fine-grained level. A combination of pHITS and pLSA for community analysis has been studied in Cohn and Hofmann (2000). Identifying topics of authors given publications and the author-of relationship has been evaluated in Rosen-Zvi et al. (2004), which also reveals the strength of influence for each author in conjointly written publications. To our knowledge, no one has included text and links into a probabilistic model to infer topical influences of citations.

4. Estimating the Influence of Citations with LDA

In this section, we derive a baseline model based on LDA.

Two independence assumptions lead to this model. The first is the Markovian assumption that publications with a strong impact are directly cited. In this case, ancestors to not provide additional information over the directly linked papers. The LDA model is additionally based on the assumption that the topic mix of each paper is chosen independently of the topic mix of the papers that it cites. This leads to the Latent Dirichlet Allocation model depicted in Figure 1a.

Latent Dirichlet Allocation associates each token in a document with a latent variable that chooses one of the underlying topics in the corpus. This is achieved by associating each document d with a multinomial distribution $\theta_d = p(t|d)$ over latent topic variables t (also referred to as the topic mixture). Likewise, each topic variable t is associated with a multinomial distribution $\phi_t = p(w|t)$ over words w.

In the LDA model, the strength of influence is not an integral part of the model (owing to the second independence assumption), but has to be determined in a later step using a heuristic measure. One heuristic (referred to as "LDA-JS") defines the strength of influence as the compatibility between the topic mixtures of citing and cited publication. The compatibility of two topic mixtures is measured by the Jensen-Shannon-Divergence. The definition of the weight function is given in Equation 1. $D_{KL}(.||.)$ is the Kullback-Leibler divergence.

$$\gamma_d(c) \quad = \quad \exp(-D_{JS}(\theta_d \| \theta_c)) \, , c \in L(d) \tag{1}$$

with $D_{JS}(\theta_d \| \theta_c) =$

$$\frac{1}{2} D_{\mathrm{KL}} \left(\theta_d \left\| \frac{\theta_d + \theta_c}{2} \right\| \right) + \frac{1}{2} D_{\mathrm{KL}} \left(\theta_c \left\| \frac{\theta_d + \theta_c}{2} \right\| \right)$$

A second heuristic (denoted by "LDA-post") uses the probability of a citation given the document $p(c|d)$ as strenght of influence (Equation 2). We refer to $p(d|t)$ as $p(c|t)$, if c is a cited document, $c \in L(d)$. With Bayes' rule and assuming a uniform prior $p(c)$ the posterior of a cited document given a topic can be written as $p(c|t) \propto p(t|c)$.

$$\gamma_d(c) = p(c|d) = \sum_t p(t, c|d) = \sum_t p(t|d) \cdot p(c|t) \tag{2}$$

5. Copycat Model

The Latent Dirichlet Allocation approach assumes that citations do not influence the underlying topics. In contrast to this, the copycat model approximates a citing document by a "weighted sum" of documents it cites. The weights of the terms capture the notion of the influence γ. In order to deal with synonyms and

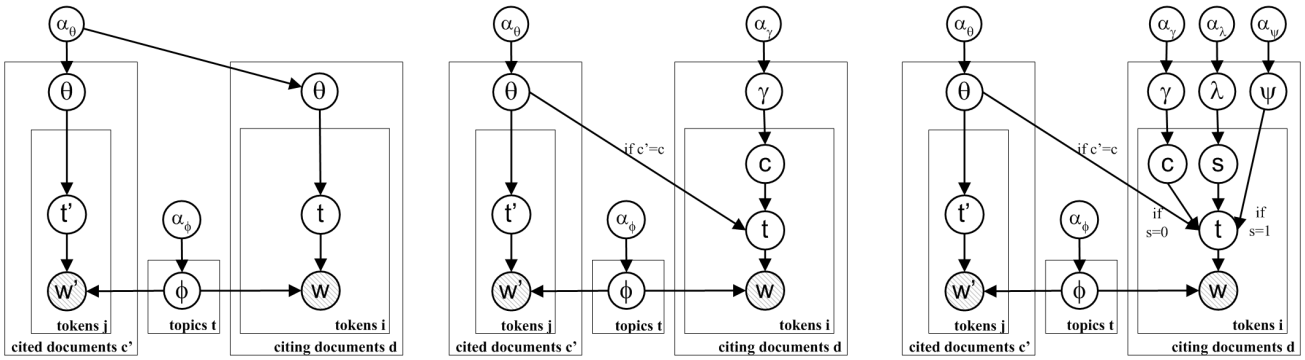

Figure 1. Approaches for estimating the influence of citations in plate notation. The citation influence model (c) combines Latent Dirichlet Allocation (a) and the copycat model (b) via the balancing parameter λ.

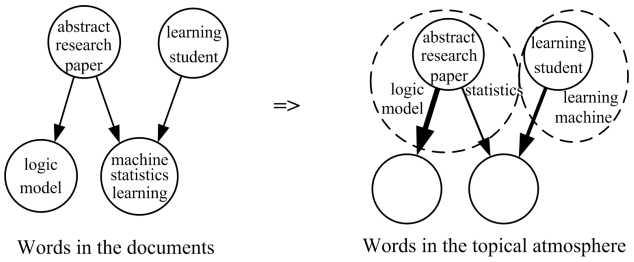

Figure 2. Association of words to the topical atmosphere (dashed circles) of cited publications.

polysems we use latent topic variables t on a word level, just as in Latent Dirichlet Allocation. Each topic in a citing document is drawn from one of the topic mixtures of cited publications. The distribution of draws from cited publications is modeled by a multinomial over citations with parameter γ. The model is depicted in plate notation in Figure 1b.

While learning the parameters of the copycat model, a sampling algorithm associates each word of each citing publication d with a publication it cites. For example let there be a cited publication c which is cited by two publications d_1 and d_2. The topic mixture θ_c is not only about all words in the cited publication c but also about some words in d_1 and d_2, which are associated with c (cf. Figure 2). This way, the topic mixture θ_c is influenced by the citing publications, which in turn influences the association of words in d_1 and d_2 to c. All tokens that are associated with a cited publication are called the topical atmosphere of a cited publication.

5.1. Bipartite Citation Graph

Following the Markov assumption that only the cited publications influence a citing publication, but not their ancestors, we transform the citation graph into a bipartite graph. The bipartite graph consists of two disjoint node sets D and C, where D contains only nodes with outgoing citation links (the citing publications) and C contains nodes with incoming links (the cited publications). Documents in the original citation graph with incoming and outgoing links are represented as two nodes $d \in D$ and $c \in C$ in the bipartite citation graph.

5.2. Mutual Influence of Citing Publications

The generative process determines citation influences independently for each citing document, but during inference of the model all citing and all cited papers are considered cojointly. The model accounts for dependencies between cited and citing papers, but also between co-cited papers (papers cited by the same document) and bibliographically coupled papers (papers citing the same document). Furthermore, the topics are learned along with the influences.

Because of bidirectional interdependence of links and topics caused by the topical atmosphere, publications originated in one research area (such as Gibbs sampling, which originated in physics) will also be associated with topics they are often cited by (such as machine learning).

On the downside, if words in a citing publication do not fit well to the vocabulary introduced by the cited literature, those words will be associated to one of the citations randomly. As an implication, this brings new words and associated topics to the topical atmosphere of the cited publication, that are not reflected by the cited text and should be considered as noise. This will lead to unwanted effects in the prediction of citation influences. In the following chapter we introduce the citation influence model, which addresses this issue.

Table 1. Variable description. Variables in the cited plate are denoted with prime.

Symbol	Description
c'	cited publication
d	citing publication
θ	topic mixture of the topical atmosphere of a cited publication
ψ	innovation topic mixture of a citing publication
ϕ	characteristic word distribution for each topic
γ	distribution of citation influences
λ	parameter of the coin flip, choosing to draw topics from θ or ψ
w', w	words in cited, citing publications respectively
t', t	topic assignments of tokens in cited, citing publications respectively
c	a cited publication, to which a token in a citing publication is associated
s	indicates whether the topic of a citing publication is drawn from inheritance or innovation
α	Dirichlet / beta parameters of the multinomial / Bernoulli distributions

6. Citation Influence Model

If the model enforces each word in a citing publication to be associated with a cited publication, noise effects are introduced to the model which may impede the prediction. In addition, it is not possible to model innovation (i.e., new or evolving topics) in the copycat model. The citation influence model depicted in Figure 1c overcomes these limitations. A citing publication may choose to draw a word's topic from a topic mixture of a citing publication θ_c (the topical atmosphere) or from it's own topic mixture ψ_d that models innovative aspects. The choice is modeled by a flip of an unfair coin s. The parameter λ of the coin is learned by the model, given an asymmetric beta prior $\vec{\alpha}_\lambda = (\alpha_{\lambda_\theta}, \alpha_{\lambda_\psi})$ which prefers the topic mixture θ of a cited publication.

The parameter λ yields an estimate for how well a publication fits to all its citations. In combination with γ, the relative influence of citations, $\lambda \cdot \gamma$ is a measure for the absolute strength of influence. The absolute measure allows to compare links from different citing publications. For the visualization, the citation graph can be thresholded according to the absolute measure.

6.1. Generative Process

Since the publication graph is given, the length of each document and bibliography is known. The citation influence model assumes the following generative process. In this process the influence of citations is directly modeled (captured by the model parameter γ).

- for all topics $t \in [1:T]$ do
  - draw the word distribution for each latent topic $\phi_t = p(w|t) \sim dirichlet(\vec{\alpha}_\phi)$
- for all cited documents $c' \in C$ do
  - draw a topic mixture $\theta_{c'} = p(t'|c') \sim dirichlet(\vec{\alpha}_\theta)$
  - for all tokens j do
    - draw a topic $t'_{c',j} \sim \theta_{c'}$ from the topic mixture
    - draw a word $w_{c',j} \sim \phi_{t'_{c',j}}$ from the topic specific word distribution
- for all citing documents $d \in D$ do
  - draw a citation mixture $\gamma_d = p(c|d)|_{L(d)} \sim dirichlet(\vec{\alpha}_\gamma)$[1] restricted to the publications c cited by this publication d
  - draw an innovation topic mixture $\psi_d = p(t|d) \sim dirichlet(\vec{\alpha}_\psi)$
  - draw the proportion between tokens associated with citations and those associated with the innovation topic mixture $\lambda_d = p(s = 0|d) \sim beta(\alpha_{\lambda_\theta}, \alpha_{\lambda_\psi})$
  - for all tokens i do
    - toss a coin $s_{d,i} \sim bernoulli(\lambda_d)$
    - if $s_{d,i} = 0$
      - draw a cited document $c_{d,i} \sim multi(\gamma_d)$
      - draw a topic $t_{d,i} \sim multi(\theta_{c_{d,i}})$ from the cited document's topic mixture
    - else ($s_{d,i} = 1$)
      - draw the topic $t_{d,i} \sim multi(\psi_d)$ from the innovation topic mixture
    - draw a word $w_{d,i} \sim multi(\phi_{t_{d,i}})$ from the topic specific word distribution

For a description of each variable see Table 1.

6.2. Learning the Model via Gibbs Sampling

Gibbs sampling (Gilks et al., 1996) allows to learn a model by iteratively updating each latent variable given fixed remaining variables.

[1] $\vec{\alpha}_\gamma$ is a symmetric prior of length $|L(d)|$

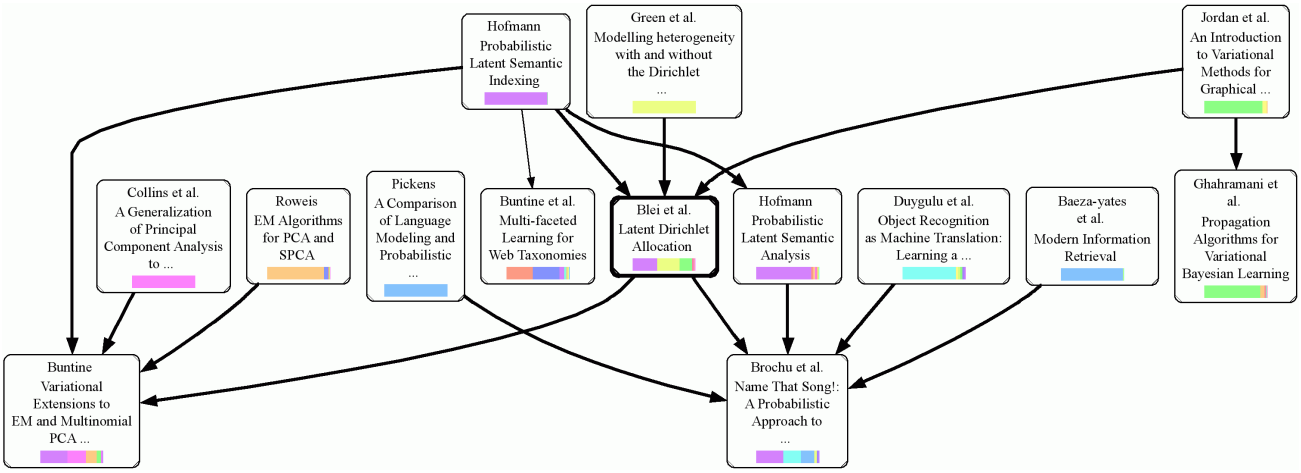

Figure 3. The filtered citation graph contains only edges which represent a significant influence.

Table 2. Update equations for the citation influence model.

$$p(c_i|\vec{c}_{\neg i}, d_i, s_i = 0, t_i \cdot) \tag{3}$$
$$= \frac{\mathcal{C}_{d,c,s}(d_i, c_i, 0) + \alpha_\gamma - 1}{\mathcal{C}_{d,s}(d_i, 0) + L(d_i)\alpha_\gamma - 1} \cdot \frac{\mathcal{C}_{c',t'}(c_i, t_i) + \mathcal{C}_{c,t,s}(c_i, t_i, 0) + \alpha_\theta - 1}{\mathcal{C}_{c'}(c_i) + \mathcal{C}_{c,s}(c_i, 0) + T\alpha_\theta - 1}$$

$$p(s_i = 0|\vec{s}_{\neg i}, d_i, c_i, t_i, \cdot) \tag{4}$$
$$= \frac{\mathcal{C}_{c',t'}(c_i, t_i) + \mathcal{C}_{c,t,s}(c_i, t_i, 0) + \alpha_\theta - 1}{\mathcal{C}_{c'}(c_i) + \mathcal{C}_{c,s}(c_i, 0) + T\alpha_\theta - 1} \cdot \frac{\mathcal{C}_{d,s}(d_i, 0) + \alpha_{\lambda_\theta} - 1}{\mathcal{C}_d(d_i) + \alpha_{\lambda_\theta} + \alpha_{\lambda_\psi} - 1}$$

$$p(s_i = 1|\vec{s}_{\neg i}, d_i, t_i \cdot) \tag{5}$$
$$= \frac{\mathcal{C}_{d,t,s}(d_i, t_i, 1) + \alpha_\psi - 1}{\mathcal{C}_{d,s}(d_i, 1) + T\alpha_\psi - 1} \cdot \frac{\mathcal{C}_{d,s}(d_i, 1) + \alpha_{\lambda_\psi} - 1}{\mathcal{C}_d(d_i) + \alpha_{\lambda_\theta} + \alpha_{\lambda_\psi} - 1}$$

$$p(t_i|\vec{t}_{\neg i}, w_i, s_i = 0, c_i \cdot) \tag{6}$$
$$= \frac{\mathcal{C}_{w,t}(w_i, t_i) + \mathcal{C}_{w',t'}(w_i, t_i) + \alpha_\phi - 1}{\mathcal{C}_t(t_i) + \mathcal{C}_{t'}(t_i) + V\alpha_\phi - 1} \cdot \frac{\mathcal{C}_{c',t'}(c_i, t_i) + \mathcal{C}_{c,t,s}(c_i, t_i, 0) + \alpha_\theta - 1}{\mathcal{C}_{c'}(c_i) + \mathcal{C}_{c,s}(c_i, 0) + T\alpha_\theta - 1}$$

$$p(t_i|\vec{t}_{\neg i}, w_i, d_i, s_i = 1, c_i \cdot) \tag{7}$$
$$= \frac{\mathcal{C}_{w,t}(w_i, t_i) + \mathcal{C}_{w',t'}(w_i, t_i) + \alpha_\phi - 1}{\mathcal{C}_t(t_i) + \mathcal{C}_{t'}(t_i) + V\alpha_\phi - 1} \cdot \frac{\mathcal{C}_{d,t,s}(d_i, t_i, 1) + \alpha_\psi - 1}{\mathcal{C}_{d,s}(d_i, 1) + T\alpha_\psi - 1}$$

The update equations of the citation influence model can be computed in constant time using count caches. The cache counts how often a combination of certain token assignments occurs. For random variables $\vec{var}_1$, $\vec{var}_2$, ..., $\vec{var}_n$, the notation $\mathcal{C}_{var_1,var_2,...,var_n}(val_1, val_2, ..., val_n) = |\{\forall i : var_{1,i} = val_1 \wedge var_{2,i} = val_2 \wedge ... \wedge var_{n,i} = val_n\}|$ counts occurrences of a configuration val_1, val_2, ..., val_n. For example, $\mathcal{C}_{d,c,s}(1, 2, 0)$ denotes the number of tokens in document 1 that are assigned to citation 2, where the coin result s is 0.

The update equations which are used to learn the citation influence model are given in Table 2, see Appendix A for details about the derivation.

After the sampling chain converges (i.e., after the burn-in phase), parameters that have been integrated out can be inferred from the count caches by averaging over the sampling chain after convergence. For example, γ is derived in Equation 8 with K denoting the length of the sampling chain (burn-in omitted).

$$\gamma_d(c) = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathcal{C}_{d,c,s}(d, c, 0)^{(k)} + \alpha_\gamma}{\mathcal{C}_{d,s}(d, 0)^{(k)} + |L(d)| \cdot \alpha_\gamma} \tag{8}$$

7. Experiments

The experiments are conducted with a subset of the CiteSeer data set[2], using abstract, title and citation information. In Section 7.1 we exemplarily analyze the citational vicinity of one research paper. In Section 7.2 the prediction performance is evaluated on influence labels provided by domain experts.

7.1. Narrative Evaluation

In order to explore the behavior of the citation influence model, we analyze the citational vicinity of the LDA paper (Blei et al., 2003) with the citation influence model. The input document collection consists of the paper along with two levels of citations in each direction. The model is trained with hyper parameters $\alpha_\phi = 0.01$, $\alpha_\theta = \alpha_\psi = 0.1$, $\alpha_{\lambda_\theta} = 3.0$, $\alpha_{\lambda_\psi} = 0.1$, $\alpha_\gamma = 1.0$ and 30 topics. The citation graph is filtered to only contain edges with an influence value $\gamma_d(c) > 0.05$. Figure 3 shows an extract of a visualization created by the graphviz tool *dot*[3].

In contrast to an unfiltered graph, the significantly in-

[2]Available at http://citeseer.ist.psu.edu/oai.html
[3]Available at http://www.graphviz.org

Table 3. Words in the abstract of the research paper "Latent Dirichlet Allocation" are assigned to citations. The probabilities in parentheses indicate $p(w, c|d, \cdot)$.

Cited Title	Associated Words	γ
Probabilistic Latent Semantic Indexing	text(0.04), latent(0.04), modeling(0.02), model(0.02), indexing(0.01), semantic(0.01), document(0.01), collections(0.01)	0.49
Modelling heterogeneity with and without the Dirichlet process	dirichlet(0.02), mixture(0.02), allocation(0.01), context(0.01), variable(0.0135), bayes(0.01), continuous(0.01), improves(0.01), model(0.01), proportions(0.01)	0.25
Introduction to Variational Methods for Graphical Methods	variational(0.01), inference(0.01), algorithms(0.01), including(0.01), each(0.01), we(0.01), via(0.01)	0.22

fluencing papers can be identified at the first glance.

The model correctly identifies the influencing work on Dirichlet processes and variational inference as well as the relatedness to pLSA, PCA, and variational methods. It also yields possible application areas such as querying music and taxonomy learning.

The topic proportions are included in the visualization via a topic spectrum bar. Each of the 30 topics is represented by a unique color.

Table 3 lists the three most influential cites and the words assigned to them[4].

7.2. Prediction Performance

For an evaluation against a ground truth we asked authors to manually label the strength of influence of papers they cited on a Likert scale[5].

In our experiments we compare four approaches: The

[4]The CiteSeer data set only contains four publications cited in "Latent Dirichlet Allocation". Missing cited publications are invisible to the model.

[5]Likert scale semantics used in the survey:
xx: "this citation influenced the publication in a strong sense, such as an approach that was build upon and refined"
x: "this citation influenced the publication, such as very strong related work"
o: "this citation did not have a strong impact on the publication, such as references basic research, other not especially related approaches, or other application domains"
oo: "this citation had no impact on the publication at all"
?: "I can not judge the influence"; or: "I have no opinion about this citation"

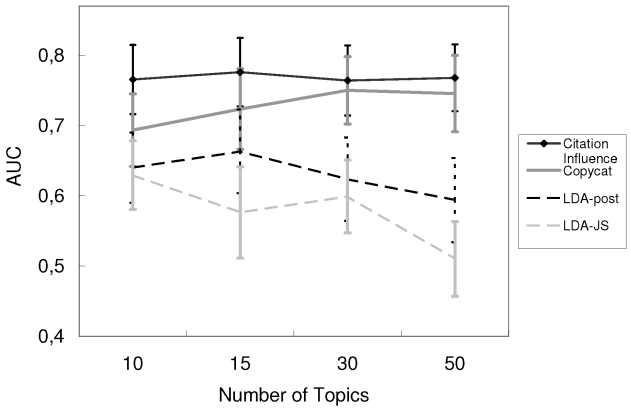

Figure 4. Predictive performance of the models. The error bars indicate the standard error of the AUC values averaged over the citing publications.

citation influence model, the copycat model, LDA-JS, and LDA-post. The models are trained on a corpus consisting of the 22 labeled seed publications along with citations (132 abstracts).

Experiments are conducted for 10, 15, 30, and 50 topics with the following hyper parameters tuned on a hold-out set.

- Citation influence model: $\alpha_\phi = 0.01$, $\alpha_\theta = \alpha_\psi = 0.1$, $\alpha_{\lambda_\theta} = 3.0$, $\alpha_{\lambda_\psi} = 0.1$, $\alpha_\gamma = 1.0$
- Copycat model: $\alpha_\phi = 0.01$, $\alpha_\theta = 0.1$, $\alpha_\gamma = 1.0$
- LDA-JS: $\alpha_\phi = 0.01$, $\alpha_\theta = 0.1$
- LDA-post: $\alpha_\phi = 0.01$, $\alpha_\theta = 0.1$

We also evaluated an approach based on the PageRank of cited nodes (with $\gamma_d(c) \propto PageRank(c)$) and an approach based on the cosine similarity of TF-IDF vectors of citing and cited publications ($\gamma_d(c) \propto \cos(\mathrm{TF\,IDF}(d), \mathrm{TF\,IDF}(c))$).

For each number of topics, we compare the citation influence and copycat model with the LDA-based approaches in terms of the ability to predict the influence of citations. The prediction performance of the γ parameter according to the manual labels of the authors on the Likert scale is evaluated in terms of the AUC (Area Under the ROC Curve) measure. The AUC values for the decision boundaries "xx vs. x, o, oo", "xx, x vs. o, oo", and "xx, x, o vs. oo" are averaged to yield a quality measure for each citing publication. Unlabeled citation links as well as those marked with "?" are excluded from the evaluation.

The predictive performance of the models is analyzed by averaging AUC values for the test set of 22 manually

labeled citing publications. The results are presented in Figure 4. Error bars indicate the standard error. We perform paired-t-tests for models with 15 topics with a significance level $\alpha = 5\%$. This reveals that the citation influence model is always significantly better than the LDA-post heuristic (even for 15 topics). In contrast to this, the citation influence model is not significantly better than the copycat model, and the copycat model is not significantly better than the LDA-post heuristic. For 30 and 50 topics, where LDA starts to degenerate, the copycat model shows a significant improvement compared to the LDA-post heuristic. Furthermore, the LDA-JS heuristic is always slightly below the LDA-post heuristic in predicting citation influences.

The approaches based on TF-IDF (AUC 0.45) and on PageRank (AUC 0.55) are not able to predict the strength of influence.

7.3. Duplication of Publications

The citation influence model treats the cited and the citing version of the same publication as completely independent from one another. The only coupling factor is the distribution over topic specific words ϕ.

We want to know, whether the model assigns similar topic mixes to the duplicated publications, despite this independent treatment. For all duplicated publications, we compare the topic mixtures of both versions via the Jensen Shannon Divergence. The average divergence between duplicated publications is 0.07, which is very low compared to the average divergence between other topic mixtures of 0.69.

8. Conclusion

We developed the copycat and the citation influence model; both model the influence of citations in a collection of publications.

The copycat model is derived from two instances of LDA by adding what we think is the minimum number of random variables necessary. The topic mix is shared between citing and cited papers, an influence parameter determines how the cited papers are blended into the citing document (the value of influence parameter γ is computed using Gibbs sampling). In the citation influence model, variables are added to accommodate two alternative ways in which each word's topic is generated. It is either inherited as in the copycat model, or is one of the paper's own innovative topics.

We compared the four models against manually annotated citations. Based on the AUC measure, we conclude that the citation influence model provides more accurate rankings than the baseline models.

In the future we want to extend the citation influence model by learning the number of topics and want to study the effect on other established topic model extensions such as time and author aspects. We see potential, that the strength of influence improves accuracy in methods originally working on unweighted graphs, such as identification of communities, authorities and hubs, or learning of distance metrics.

Acknowledgment

We gratefully acknowledge support from the German Science Foundation DFG. Thanks to Gerd Stumme, Peter Tandler, and Till Schümmer for labeling the strength of influence of their publications.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cohn, D., & Chang, H. (2000). Learning to probabilistically identify authoritative documents. *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 167–174).

Cohn, D., & Hofmann, T. (2000). The missing link - a probabilistic model of document content and hypertext connectivity. *NIPS '00: Advances in Neural Information Processing Systems*.

Doucet, A., de Freitas, N., Murphy, K. P., & Russell, S. J. (2000). Rao-blackwellised particle filtering for dynamic bayesian networks. *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* (pp. 176–183).

Flake, G. W., Tsioutsiouliklis, K., & Zhukov, L. (2004). Methods for mining web communities: Bibliometric, spectral, and flow. *Web Dynamics - Adapting to Change in Content, Size, Topology and Use*, 45–68.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*, 119–145.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain monte carlo in practice*. London, UK: Chapman & Hall.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177–196.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*, 604–632.

Mann, G., Mimno, D., & Mccallum, A. (2006). Bibliometric impact measures leveraging topic analysis. *JCDL '06: Proceedings of the Joint Conference on Digital Libraries*.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*, 1077–1087.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487–494).

Sandor, A., Kaplan, A., & Rondeau, G. (2006). Discourse and citation analysis with concept-matching. *International Symposium: Discourse and Document*.

Spiegel-Roesing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, *7*, 97–113.

Trigg, R. (1983). *A network-based approach to text handling for the online scientific community* (Technical Report).

Xu, Z., Tresp, V., Yu, K., Yu, S., & Kriegel, H.-P. (2005). Dirichlet enhanced relational learning. *ICML '05: Proceedings of the 22nd international conference on Machine learning* (pp. 1004–1011). New York, NY, USA: ACM Press.

A. Appendix Gibbs Sampling Derivation

We can derive a Rao-Blackwellized[6] version of the model by integrating out the multinomial distributions θ, ψ, ϕ, γ, and λ because the model uses only conjugate priors. For the derivation of Gibbs update equations we first derive the joint distribution of the remaining variables in Equation 9 from the generative process.

$$\begin{aligned} &p(\vec{w}, \vec{w}', \vec{t}, \vec{t}', \vec{c}, \vec{s}|\vec{\alpha}_\phi, \vec{\alpha}_\theta, \vec{\alpha}_\psi, \vec{\alpha}_\gamma, \vec{\alpha}_\lambda, \cdot) \qquad (9)\\ &= \int p(\vec{w}, \vec{w}'|\vec{t}, \vec{t}', \vec{\phi}) \cdot p(\vec{\phi}|\vec{\alpha}_\phi)\, d\vec{\phi} \cdot \int p(\vec{c}|\vec{\gamma}, L) \cdot p(\vec{\gamma}|\vec{\alpha}_\gamma, L)\, d\vec{\gamma}\\ &\cdot \int p(\vec{t}, \vec{t}'|\vec{s}, \vec{c}, \vec{\theta}, \vec{\psi}) \cdot p(\vec{\theta}|\vec{\alpha}_\theta) \cdot p(\vec{\psi}|\vec{\alpha}_\psi)\, d\vec{\theta}\vec{\psi} \cdot \int p(\vec{s}|\vec{\lambda}) \cdot p(\vec{\lambda}|\vec{\alpha}_\lambda)\, d\vec{\lambda} \end{aligned}$$

In the following, we exemplify the derivation of the update equation for c_i – equations for the other variables are derived analogously. The conditional of c_i is obtained by dividing the joint distribution of all variables by the joint with all variables but c_i (denoted by $\vec{c}_{\neg i}$) in Equation 10 and canceling factors that do not depend on $\vec{c}_{\neg i}$.

$$\begin{aligned} &p(c_i|\vec{c}_{\neg i}, w_i, s_i = 0, t_i \cdot)\\ &= \frac{p(\vec{w}, \vec{w}', \vec{t}, \vec{t}', \vec{c}, \vec{s}|\vec{\theta}, \vec{\psi}, \vec{\lambda}, \vec{\gamma}, \cdot)}{p(\vec{w}, \vec{w}', \vec{t}, \vec{t}', \vec{c}_{\neg i}, \vec{s}|\vec{\theta}, \vec{\psi}, \vec{\lambda}, \vec{\gamma}, \cdot)} \qquad (10)\\ &= \frac{\int p(\vec{c}|\vec{\gamma}, L) \cdot p(\vec{\gamma}|\vec{\alpha}_\gamma, L)\, d\vec{\gamma}}{\int p(\vec{c}_{\neg i}|\vec{\gamma}, L) \cdot p(\vec{\gamma}|\vec{\alpha}_\gamma, L)\, d\vec{\gamma}} \qquad (11)\\ &\cdot \frac{\int p(\vec{t}, \vec{t}'|\vec{s}, \vec{c}, \vec{\theta}, \vec{\psi}) \cdot p(\vec{\theta}|\vec{\alpha}_\theta) \cdot p(\vec{\psi}|\vec{\alpha}_\psi)\, d\vec{\theta}\vec{\psi}}{\int p(\vec{t}, \vec{t}'|\vec{s}, \vec{c}_{\neg i}, \vec{\theta}, \vec{\psi}) \cdot p(\vec{\theta}|\vec{\alpha}_\theta) \cdot p(\vec{\psi}|\vec{\alpha}_\psi)\, d\vec{\theta}\vec{\psi}} \end{aligned}$$

[6] Rao-Blackwellization (Doucet et al., 2000) is a procedure to reduce redundancy in a graphical model to improve the performance.

In the following, we derive the first fraction of Equation 11 (the second fraction is derived analogously). In Equation 12 we detail the derivation of the numerator; the derivation for the denominator is the same. If we assume that multinomial γ is given, then $p(\vec{c}|\vec{\gamma}, L) = \prod_{n=1}^{D} \prod_{l \in L(n)} \gamma_n(l)^{\mathcal{C}_{d,c,s}(n,l,0)}$. Note, that c_i is void if $s_i = 1$. Because we are only using conjugate priors, the multinomial-Dirichlet integral in Equation 12 has a closed form solution. It is resolved by the corresponding notation in Equation 13 using the count caches and pseudo counts α_γ. Here we assume that $\vec{\alpha}_\gamma$ is a symmetric Dirichlet parameter with scalar α_γ.

$$\begin{aligned} &\int p(\vec{c}|\vec{\gamma}, L) \cdot p(\vec{\gamma}|\vec{\alpha}_\gamma, L) d\vec{\gamma}\\ &= \prod_{n=1}^{D} \int \prod_{l \in L(n)} \gamma_n(l)^{\mathcal{C}_{d,c,s}(n,l,0)} \cdot p(\gamma_n|\vec{\alpha}_\gamma, L)\, d\gamma_n \qquad (12)\\ &= \prod_{n=1}^{D} \frac{1}{\frac{\prod_{l \in L(n)} \Gamma(\alpha_\gamma)}{\Gamma(\sum_{l \in L(n)} \alpha_\gamma)}} \cdot \frac{\prod_{l \in L(n)} \Gamma(\mathcal{C}_{d,c,s}(n,l,0) + \alpha_\gamma)}{\Gamma(\sum_{l \in L(n)} \mathcal{C}_{d,c,s}(n,l,0) + \alpha_\gamma)} \qquad (13) \end{aligned}$$

To yield the first fraction of Equation 11 we apply Equation 13 twice and reach Equation 14. The removal of the i'th token is expressed by the Kronecker delta $\delta(x, y)$ which is 1 iff $x = y$, 0 otherwise. In Equation 15 all factors for which δ is 0 are canceled. In Equation 16, we exploit that $\frac{\Gamma(n)}{\Gamma(n-1)} = n - 1$.

$$\begin{aligned} &\frac{p(\vec{c}|\vec{\gamma}, L) \cdot p(\vec{\gamma}|\vec{\alpha}_\gamma, L)}{p(\vec{c}_{\neg i}|\vec{\gamma}, L) \cdot p(\vec{\gamma}|\vec{\alpha}_\gamma, L)}\\ &= \frac{\prod_{n=1}^{D} \frac{\prod_{l \in L(n)} \Gamma(\mathcal{C}_{d,c,s}(n,l,0)+\alpha_\gamma)}{\Gamma(\sum_{l \in L(n)} \mathcal{C}_{d,c,s}(n,l,0)+\alpha_\gamma)}}{\prod_{n=1}^{D} \frac{\prod_{l \in L(n)} \Gamma(\mathcal{C}_{d,c,s}(n,l,0)-\delta(n,d_i)\delta(l,c_i)\delta(0,s_i)+\alpha_\gamma)}{\Gamma(\sum_{l \in L(n)} \mathcal{C}_{d,c,s}(n,l,0)-\delta(n,d_i)\delta(l,c_i)\delta(0,s_i)+\alpha_\gamma)}} \qquad (14)\\ &= \frac{\Gamma(\mathcal{C}_{d,c,s}(d_i, c_i, 0) + \alpha_\gamma)}{\Gamma(\mathcal{C}_{d,c,s}(d_i, c_i, 0) + \alpha_\gamma - \delta(0, s_i))} \cdot \frac{1}{\frac{\Gamma(\mathcal{C}_{d,s}(d_i,0)+L(d_i)\alpha_\gamma)}{\Gamma(\mathcal{C}_{d,s}(d_i,0)+L(d_i)\alpha_\gamma-\delta(0,s_i))}} \qquad (15)\\ &= \frac{\mathcal{C}_{d,c,s}(d_i, c_i, 0) + \alpha_\gamma - \delta(0, s_i)}{\mathcal{C}_{d,s}(d_i, 0) + L(d_i)\alpha_\gamma - \delta(0, s_i)} \qquad (16) \end{aligned}$$

Deriving $\frac{\int p(\vec{t},\vec{t}'|\vec{s},\vec{c},\vec{\theta},\vec{\psi}) \cdot p(\vec{\theta}|\vec{\alpha}_\theta) \cdot p(\vec{\psi}|\vec{\alpha}_\psi)\, d\vec{\theta}\vec{\psi}}{\int p(\vec{t},\vec{t}'|\vec{s},\vec{c}_{\neg i},\vec{\theta},\vec{\psi}) \cdot p(\vec{\theta}|\vec{\alpha}_\theta) \cdot p(\vec{\psi}|\vec{\alpha}_\psi)\, d\vec{\theta}\vec{\psi}}$ analogously and in combination with Equation 16, we are in the position to simplify the update equation for c_i from Equation 11 to yield the final update in Equation 17.

$$\begin{aligned} &p(c_i|\vec{c}_{\neg i}, w_i, s_i = 0, t_i \cdot)\\ &= \frac{\mathcal{C}_{d,c,s}(d_i, c_i, 0) + \alpha_\gamma - 1}{\mathcal{C}_{d,s}(d_i, 0) + L(d_i)\alpha_\gamma - 1}\\ &\cdot \frac{\mathcal{C}_{c',t'}(c_i, t_i) + \mathcal{C}_{c,t,s}(c_i, t_i, 0) + \alpha_\theta - 1}{\mathcal{C}_{c'}(c_i) + \mathcal{C}_{c,s}(c_i, 0) + T\alpha_\theta - 1} \qquad (17) \end{aligned}$$