Joint Latent Topic Models for Text and Citations

Ramesh Nallapati*
Computer Science Department
Stanford University
353 Serra Mall
Stanford, CA 94305
nmramesh@cs.stanford.edu

Amr Ahmed, Eric P. Xing and
William W. Cohen
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{amahmed,epxing,wcohen}@cs.cmu.edu

ABSTRACT

In this work, we address the problem of joint modeling of text and citations in the topic modeling framework. We present two different models called the Pairwise-Link-LDA and the Link-PLSA-LDA models.

The Pairwise-Link-LDA model combines the ideas of LDA [4] and Mixed Membership Block Stochastic Models [1] and allows modeling arbitrary link structure. However, the model is computationally expensive, since it involves modeling the presence or absence of a citation (link) between every pair of documents. The second model solves this problem by assuming that the link structure is a bipartite graph. As the name indicates, Link-PLSA-LDA model combines the LDA and PLSA models into a single graphical model.

Our experiments on a subset of Citeseer data show that both these models are able to predict unseen data better than the baseline model of Erosheva and Lafferty [8], by capturing the notion of topical similarity between the contents of the cited and citing documents. Our experiments on two different data sets on the link prediction task show that the Link-PLSA-LDA model performs the best on the citation prediction task, while also remaining highly scalable. In addition, we also present some interesting visualizations generated by each of the models.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database Applications—*data mining*

General Terms

Algorithms, Experimentation

Keywords

Topic models, LDA, PLSA, variational inference, hyperlinks, influence, citations

*This work was done while the author was at CMU.

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

1. INTRODUCTION

Proliferation of large electronic document collections such as the web, news articles, blogs and scientific literature in the recent past has posed several new, interesting challenges to researchers in the data mining community. In particular, there is an increasing need for automatic techniques to visualize, analyze and mine these document collections. In the recent past, latent topic modeling has become very popular as a completely unsupervised technique for topic discovery in large document collections. These models, such as PLSA [9] and LDA [4], exploit co-occurrence patterns of words in documents to unearth semantically meaningful probabilistic clusters of words called *topics*. These models also assign a probabilistic membership to documents in the latent topic-space, allowing us to view and process the documents in this lower-dimensional space.

Most of the models in this framework such as Dynamic topic models [5, 15], Pachinko Allocation [11], Correlated Topic Model [3], etc., model various aspects of document collections such as time, hierarchy of topics, correlations between topics respectively. However, all the above mentioned models ignore a rich feature that contains valuable information, namely, the citation or hyperlink structure. It is a known fact in information retrieval that a citation between two documents not only indicates topical similarity of the two documents but also authoritativeness of the cited document [10]. This idea has been exploited by algorithms such as PageRank [16] which are now *de facto* techniques in search engine technology.

In our work, we aim at addressing the problem of jointly modeling text and citations in the topic modeling framework. Our hope is that explicit modeling of citations captures the topicality of documents in the collection better, and thereby improves the predictive power of these models.

The rest of the paper is organized as follows. In section 2, we discuss some of the past work done on joint models of topics and citations in the framework of latent topic models. We introduce two new models in section 3 and their corresponding learning and inference techniques using variational approximations. In section 4, we describe the data sets, the tasks and evaluation we used for our experiments. Section 5 reports and analyzes the results of our experiments. We conclude the discussion in section 6 with a few remarks on directions for future work.

Note that in the rest of the paper, we use the terms 'citation', 'hyperlink' and 'link' interchangeably. Likewise, note that the term 'citing' is synonymous to 'linking' and so is

M	Total number of documents
$M_{\leftarrow}$	Number of cited documents
$M_{\rightarrow}$	Number of citing documents
V	Vocabulary size
K	Number of topics
$N_{\leftarrow}$	Total number of words in the cited set
d	A citing document
d'	A cited document
$\Delta(p)$	A simplex of dimension $(p-1)$
$c(d,d')$	citation from d to d'
L_d	Number of hyperlinks in document d
N_d	Number of words in document d

Table 1: Notation of some frequently occurring variables

For each document $d = 1, \cdots, M$
  Generate $\boldsymbol{\theta}_d \in \Delta(K) \sim \text{Dir}(\cdot|\alpha_\theta)$
  For each position $n = 1, \cdots, N_d$
    Generate $z_n \in \{1, \cdots K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_d)$
    Generate word $w_n \in \{1, \cdots, V\} \sim \text{Mult}(\cdot|\boldsymbol{\beta}_{z_n})$
  For each hyperlink $l = 1, \cdots, L_d$
    Generate $z_l \in \{1, \cdots K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_d)$
    Generate target doc. $d'_l \in \{1, \cdots M\} \sim \text{Mult}(\cdot|\boldsymbol{\Omega}_{z_l})$

Table 2: Generative process for the Link-LDA model

'cited' to 'linked'. The reader is also recommended to refer to table 1 for some frequent notation used in this paper.

2. PAST WORK

In one of the first efforts in applying topic models to modeling citation data, Cohn and Hoffman [6] built an extension to the PLSA [9] model, called PHITS. This model defines a generative process not only for text but also for citations (hyperlinks). The generation of each hyperlink in a document d is modeled as a multinomial sampling of the target document d' from the topic-specific distribution $\boldsymbol{\Omega}$ over documents. The model assigns high probability $\Omega_{kd'}$ to a document d' with respect to topic k, if the document is hyper-linked from several documents that discuss that topic. The authors showed that the document's representation in topic-space obtained from this model improves the performance of a document-classifier, compared to the representation obtained from text alone. Henceforth, we will refer to this model as Link-PLSA, for consistency of notation in this paper.

A similar model called mixed membership model was developed by Erosheva *et al* [8], in which PLSA was replaced by LDA as the fundamental generative building block. We will refer to this model as Link-LDA for notational consistency. The generative process for this model is shown in table 2 and the corresponding graphical representation is displayed in figure 1. As shown in the figure, the generative processes for words and hyperlinks are very similar and they share the same document-specific topic distribution $\boldsymbol{\theta}$ to generate their respective latent topics. Thus, this model (as well as Link-PLSA) captures the notion that documents that share the same hyperlinks and same words, tend to be on the same topic.

Both Link-PLSA and Link-LDA define hyperlinks as just values taken by a random variable (similar to words in the vocabulary). In other words, these models obtain probabilistic topical clusters of hyperlinks exactly the same way as the basic LDA and PLSA models discover topical clusters of words. Such methods fail to explicitly model the topical relationship between the text of the citing (linking) document and the text of cited (linked) document. One can hope to obtain better quality of topics by exploiting this additional information.

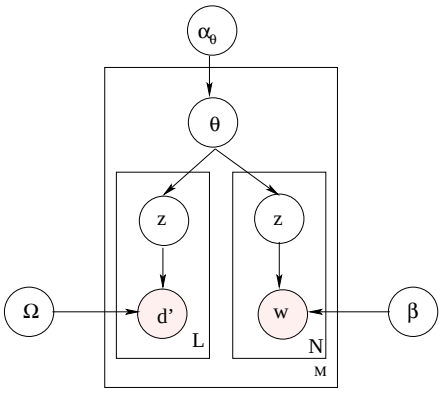

Figure 1: Graphical representation of the Link-LDA model: some of the subscripts are omitted for simplicity. Cf. table 2 for detailed description.

Recently, Dietz *et al* [7] proposed a new LDA based approach that allows flow of influence from the text of the cited documents to the text of the citing documents. In their approach, each citing document borrows topics from one of its citations in generating its own text. In choosing a citation to borrow topics from, the document uses its own distribution over its citations. This distribution is interpreted as the influence of each citation on the citing document. This model however does not *e*xplicitly model topicality of citations. In addition, this model assumes citations as input data, whereas in our work, we will propose models that can generate as well as predict citations for unseen documents.

3. TWO NEW MODELS

3.1 Pairwise Link-LDA

In this model, we combine the LDA model with the Mixed Membership Stochastic Block (MMSB) model [1], previously used in modeling protein-protein interactions. The MMSB model assigns probabilistic membership for proteins into topics based on their interactions as follows: for each pair of proteins (d, d'), we first draw a topic $z_{dd'}$ for protein d from its own distribution $\boldsymbol{\theta}_d$ over topics. Likewise, we also draw $z_{d'd}$ from $\theta_{d'}$. Then the presence or absence of an interaction between d to d' is generated as a binary random variable from a Bernoulli distribution whose parameter $\eta_{z_{dd'}, z_{d'd}}$ is specified by the topics sampled from the corresponding proteins for this particular interaction.

3.1.1 Generative Process

In this work, we extend this model to text by considering documents as analogous to proteins. Thus for each pair of documents, we generate the presence or absence of a citation represented by a Bernoulli random variable. The parameter of this distribution depends on the latent topics sampled from each of these documents. Note that protein-protein interaction was modeled in [1] as a symmetric interaction. However, a citation is directional and hence inherently asymmetric. To account for this, for each pair of documents (d, d'), we assign the directionality of the citation based on the time-stamps of the documents. For example, if d' is older than d, then we assume the citation is from d

For each document $d = 1, \cdots, M$
  Generate $\boldsymbol{\theta}_d \in \Delta(K) \sim \text{Dir}(\cdot|\alpha_\theta)$
  For each position $n = 1, \cdots, N_d$
    Generate $z_n \in \{1, \cdots, K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_d)$
    Generate $w_n \in \{1, \cdots, V\} \sim \text{Mult}(\cdot|\boldsymbol{\beta}_{z_n})$
For each document pair (d, d')
  Generate $z_{dd'} \in \{1, \cdots, K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_d)$
  Generate $z_{d'd} \in \{1, \cdots, K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_{d'})$
    Generate $c_{d'd} \in \{0, 1\} \sim \text{Bernoulli}(\cdot|\eta_{z_{dd'} z_{d'd}})$

Table 3: Generative process for Pairwise Citation LDA

to d'. In addition, if $z_{dd'}$ is the latent topic sampled from d for this interaction and $z_{d'd}$ is the corresponding topic from d' for the same interaction, the corresponding Bernoulli parameter used to generate the citation will be $\eta_{z_{dd'}, z_{d'd}}$ and not $\eta_{z_{d'd}, z_{dd'}}$. In other words, we allow the $\boldsymbol{\eta}$ matrix to be asymmetric ($\eta_{kk'} \neq \eta_{kk'}$), in order to capture the directionality of the citation.

The generative process for words remains same as that of LDA. Note that the document specific topic proportions $\boldsymbol{\theta}_d$, used in generating the words, is same as the one used in generating links for that document. A more detailed description of the generative process of the model is described in table 3. It is not possible to represent citations between all pairs of documents in the plate notation, hence we present a simplified graphical representation in figure 2, in which we show citations from one set of documents (citing set) to the other (cited set). It is clear from figure 2 that topicality of the cited document and the citing document are explicitly made dependent through the V-structure at the variable c, which is observed.

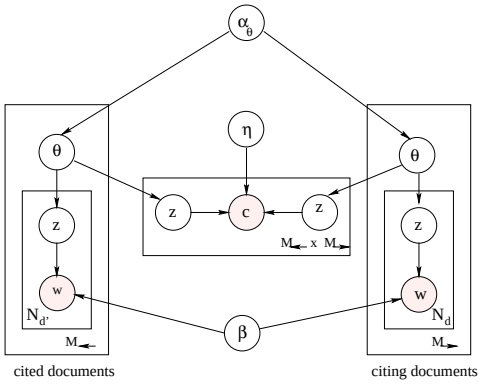

Figure 2: Graphical representation of the pairwise Citation LDA model: the representation simplifies the model into a set of cited documents and citing documents for ease of illustration. In reality, we have potential citations between all pairs of documents. Some of the subscripts are omitted for simplicity. Cf. table 3 for detailed description.

3.1.2 Variational inference

The log-likelihood of the observed data with respect to this model is as follows:

$$\begin{aligned}
&\log P(\mathbf{w}, \mathbf{c}|\alpha_\theta, \boldsymbol{\eta}, \boldsymbol{\beta}) \\
&= \log\Big(\int_{\boldsymbol{\theta}} \prod_{d=1}^{M} \Big\{ P(\boldsymbol{\theta}_d|\alpha_\theta) \prod_{n=1}^{N_d} \Big(\sum_k \theta_{dk} \beta_{kw_n}\Big) \Big\} \\
&\times \prod_{d,d'} \Big(\sum_{k,k'} \theta_{dk} \theta_{d'k'} (\eta_{k,k'})^{c_{dd'}} \\
&\quad (1 - \eta_{k,k'})^{1 - c_{dd'}}\Big) d\boldsymbol{\theta}\Big)
\end{aligned}$$

We use the following mean-field variational approximation for the posterior distribution:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \mathbf{z}) &= \prod_{d=1}^{M} \text{Dir}(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \Big(\prod_{n=1}^{N_d} \prod_{k=1}^{K} \text{Mult}(z^{dnk}|\boldsymbol{\phi}_{dn})\Big) \\
&\times \prod_{d,d'} \prod_{k=1}^{K} \prod_{k'=1}^{K} \text{Mult}(z_{dd'k}|\boldsymbol{\lambda}_{dd'}) \text{Mult}(z_{d'dk}|\boldsymbol{\lambda}_{d'd})
\end{aligned}$$

For reasons of space, we do not derive the steps involved in inference, but only present the final update equations below. The interested user may refer to [20, 4] for more details of the standard inference procedure.

$$\phi_{dnk} \propto \beta_{kw_n} \exp(\Psi(\gamma_{dk})) \tag{1}$$

$$\begin{aligned}
\lambda_{dd'k} &\propto \exp(\Psi(\gamma_{dk}) + \sum_{k'} \lambda_{d'dk'} (c_{dd'} \log(\eta_{kk'}) \\
&+ (1 - c_{dd'}) \log(1 - \eta_{kk'}))
\end{aligned} \tag{2}$$

$$\gamma_{dk} = \alpha_\theta + \sum_{n=1}^{N_d} \phi_{dnk} + \sum_{d'} \lambda_{dd'k} \tag{3}$$

$$\beta_{kv} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnk} \delta_v(w_n) \tag{4}$$

$$\eta_{kk'} = \frac{\sum_{d,d'} \lambda_{dd'k} \lambda_{d'dk'} c_{d,d'}}{\sum_{d,d'} \lambda_{dd'k} \lambda_{d'dk'}} \tag{5}$$

where $\Psi()$ is the digamma function, $\delta_v(w)$ is the delta function given by $\delta_v(w) = 1$ if $w = v$ and 0 otherwise.

In terms of implementation, first, we execute step (1) for each position in each document to compute the variational multinomial $\boldsymbol{\phi}$. Then, for each document pair (d, d'), we run step (2) for both documents alternatively until the parameters $\boldsymbol{\lambda}_{dd'}$ and $\boldsymbol{\lambda}_{d'd}$ converge. Next we update $\boldsymbol{\gamma}_d$, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, using steps (3) through (5) by using the sufficient statistics computed in steps (1) and (2). This process is repeated in an outer loop until the lower bound on the log-likelihood of the entire training set converges.

The implementation of inference for citing documents (which we will use in our experiments on log-likelihood) is slightly different. We follow steps (1), (2) and (3) as before, but we only update $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ for the citing documents, keeping the $\boldsymbol{\gamma}$'s for the cited documents fixed at the values learned during training. Also we skip steps (4) and (5) since, we are only performing inference.

3.1.3 Model limitations

The Pairwise Link-LDA model is a true generative model for text and citations, and is capable of modeling arbitrary link structure. However, since it requires explicit modeling of the presence or absence of links between each pair of documents, it becomes prohibitively expensive to model large

scale document collections. Hence, scalability is a big issue with this model.

3.2 Link-PLSA-LDA

As discussed above, Pairwise Link-LDA model is not very scalable. The Link-LDA model, on the other hand, models a citation as a multinomial sampling of the target document, and hence does not require comparing every pair of documents. As a result, it is more scalable. However, as we noted in section 2, this model fails to model the topical dependence between the cited and citing documents explicitly. As a compromise, we propose a new Link-PLSA-LDA model that combines the best properties of these two models. The new model follows the multinomial sampling process for generating citations, and thereby retains the scalability of the Link-LDA model. At the same time, it also explicitly models the topical dependence between the cited and the citing documents.

In order to achieve this objective, the new Link-PLSA-LDA model makes a simplifying assumption that the link structure in the corpus is a bipartite graph with all links emerging from the set of citing documents and pointing to the set of cited documents. In other words, we assume each document can either be cited or be a citing document, but not both.

3.2.1 Generative process

In this model, the generative process for the content and citations of the citing documents is the same as in Link-LDA. In addition, in order to explicitly model information flow from the citing document to the cited document, we defined an explicit generative process for the content of cited documents, that makes use of the same distribution $\mathbf{\Omega}$. In this new generative process, we view the set of cited documents as bins that are to be filled with words. We first associate a topic mixing proportions $\boldsymbol{\pi}$ for the entire set of cited documents. Then words are filled into the bins $N_{\leftarrow}$ times, where $N_{\leftarrow}$ is the sum total of the document lengths of the set of cited documents, as follows: each time, we first sample a topic k from the mixing proportions $\boldsymbol{\pi}$, then pick a bin d' from $\mathbf{\Omega}_k$ and fill a word occurrence from $\boldsymbol{\beta}_k$ into the bin. This process is exactly same as the symmetric parametrization[1] of PLSA as described in [9]. Since we used a combination of PLSA for cited documents and Link-LDA for citing documents to jointly model content and hyperlinks, we call this new model Link-PLSA-LDA.

The entire generative process is displayed step-by-step in table 4 and the corresponding graphical representation is shown in figure 3. One can see that dependencies propagate from the cited documents to the citing documents through the unobserved $\mathbf{\Omega}$, as per the d-separation principle in Bayesian networks [2].

3.2.2 Inference and Estimation

The likelihood of the observed data in this model is given

[1] $P(w,d) = \sum_z P(w|z)P(d|z)P(z)$ as opposed to the more common $P(w,d) = \sum_z P(w|z)P(z|d)$.

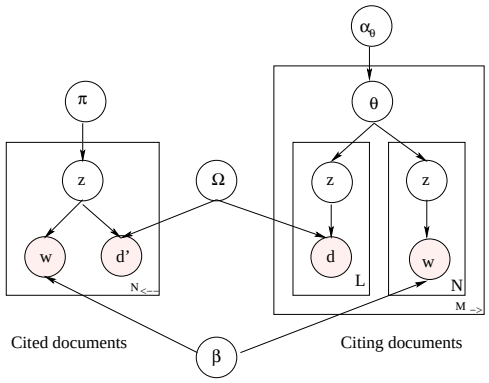

Figure 3: Graphical representation of the Link-PLSA-LDA model. Cf. table 4 for detailed description.

Cited documents:
For $i = 1, \cdots, N_{\leftarrow}$
  Generate $z_i \in \{1, \cdots K\} \sim \text{Mult}(\cdot|\boldsymbol{\pi})$
  Sample $d'_i \in \{1, \cdots, M_{\leftarrow}\} \sim \text{Mult}(\cdot|\mathbf{\Omega}_{z_i})$
  Generate $w_i \in \{1, \cdots V\} \sim \text{Mult}(\cdot|\beta_{z_i})$
Citing documents:
For each citing document $d = 1, \cdots, M_{\rightarrow}$
  Generate $\boldsymbol{\theta}_d \in \Delta(K) \sim \text{Dir}(\cdot|\alpha_\theta)$
  For each position $n = 1, \cdots, N_d$
    Generate $z_n \in \{1, \cdots K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_d)$
    Generate $w_n \in \{1, \cdots, V\} \sim \text{Mult}(\cdot|\boldsymbol{\beta}_{z_n})$
  For each citation position $l = 1, \cdots, L_d$
    Generate $z_l \in \{1, \cdots K\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}_d)$
    Generate $d'_l \in \{1, \cdots M_{\leftarrow}\} \sim \text{Mult}(\cdot|\mathbf{\Omega}_{z_l})$

Table 4: Generative process for the Link-PLSA-LDA model

as follows.

$$
\begin{aligned}
& P(\mathbf{w}, \mathbf{c}|\boldsymbol{\pi}, \alpha_\theta, \mathbf{\Omega}, \boldsymbol{\beta}) \\
&= \prod_{n=1}^{N_{\leftarrow}} (\sum_k \pi_k \Omega_{k d'_n} \beta_{k w_n}) \\
&\times \prod_{d=1}^{M_{\rightarrow}} \int_{\boldsymbol{\theta}_d} (P(\boldsymbol{\theta}_d|\alpha_\theta) \prod_{n=1}^{N_d} (\sum_k \theta_{dk} \beta_{k w_n}) \\
&\times \prod_{l=1}^{L_d} (\sum_k \theta_{dk} \Omega_{k c_l})) d\boldsymbol{\theta}_d
\end{aligned}
$$

where $\mathbf{w}$ is the entire text of cited and citing documents and $\mathbf{c}$ is the set of hyperlinks/citations.

As in the case of the Pairwise Link-LDA model, we will employ the mean-field variational approximation for the posterior distribution of the latent variables as shown below.

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \mathbf{z}, |\mathbf{w}, \mathbf{c}) = & \prod_{d=1}^{M_{\rightarrow}} (\text{Dir}(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \\
\times & \prod_{n=1}^{N_d} \prod_{k=1}^{K} \text{Mult}(z_{dnk}|\boldsymbol{\phi}_{dn}) \prod_{l=1}^{L_d} \prod_{k=1}^{K} \text{Mult}(z_{dlk}|\boldsymbol{\varphi}_{dl})) \\
\times & \prod_{n=1}^{N_{\leftarrow}} \prod_{k=1}^{K} \text{Mult}(z_{d'_n nk}|\boldsymbol{\xi}_{d'_n n})
\end{aligned}
$$

where d'_n is the document index of the n^{th} word in the cited set. Using standard variational inference procedure [20, 4],

we arrive at the following update equations:

$$\phi_{dnk} \propto \beta_{kw_n} \exp(\Psi(\gamma_{dk})) \quad (6)$$

$$\varphi_{dlk} \propto \Omega_{kd'_l} \exp(\Psi(\gamma_{dk})) \quad (7)$$

$$\gamma_{dk} = \alpha_\theta + \sum_{n=1}^{N_d} \phi_{dnk} + \sum_{l=1}^{L_d} \varphi_{dlk} \quad (8)$$

$$\xi_{d'_n nk} \propto \Omega_{kd'_n} \beta_{kw_n} \pi_k \quad \forall n = 1, \cdots, N_{\leftarrow} \quad (9)$$

$$\beta_{kv} \propto \sum_{n=1}^{N_{\leftarrow}} \xi_{d'_n nk} \delta_v(w_n) + \sum_{d=1}^{M_{\rightarrow}} \sum_{n=1}^{N_d} \phi_{dnk} \delta_v(w_n) \quad (10)$$

$$\pi_k \propto \sum_{n=1}^{N_{\leftarrow}} \xi_{d'_n nk} \quad (11)$$

$$\Omega_{kd'} \propto \sum_{n=1}^{N_{d'}} \xi_{d'nk} + \sum_{d=1}^{M_{\rightarrow}} \sum_{l=1}^{L_d} \varphi_{dlk} \delta_{d'}(d'_l) \quad (12)$$

These updates are performed iteratively in the same order as above, until convergence. Since the updates in steps (6) through (8) depend on each other, we also perform an inner iterative loop involving these equations, until they converge.

For performing inference only on the citing documents, we only iterate between steps (6) through (8) until convergence.

3.2.3 Model limitations

It is clear that one of the limitations of the model is the assumption of the bipartite link structure. This may appear as a very restrictive assumption, but it can be easily overcome in practice (see section 4.1 for more details).

Also, the Link-PLSA-LDA model defines the topical distribution for citations, $\mathbf{\Omega}$, over a fixed set of cited documents. This means that new documents can only cite documents within this fixed set. Hence this model is not fully generative, a weakness also shared by the PLSA and the Link-LDA models. We believe, in practice, it is not entirely unreasonable to assume that the set of cited documents is known at modeling time, and will not change. For example, the cited and citing documents could respectively correspond to previously published papers and currently submitted ones in the scientific domain; or last month's blog postings and current blog postings in a blog domain.

4. EXPERIMENTAL DESIGN

4.1 Data sets

For our experiments, we used two different types of linked data: scientific literature from *Citeseer* containing citations, and blog data containing hyperlinks.

Both data sets exhibit arbitrary link structure, but since the Link-PLSA-LDA model can only handle bipartite link structure, we transformed our link structure to a bipartite graph. This will allow us to compare all the models on equal footing. The way it is done is as follows: we assign each document to one of two disjoint sets called the cited set or the citing set based on whether they contain incoming links or outgoing links respectively. If a document contains both types of links, we create a duplicate of the document and assign one copy to each set such that only the incoming links are stored in the cited set and only the outgoing links are stored in the citing set. This reduces the original link graph to a bipartite one. In fact, this strategy has been successfully adopted by Dietz *et al*[7] in their work on modeling citation influences.

Since we made the graph bipartite, when we run the Pairwise Link-LDA model on this data set, we do not consider the existence of links between all pairs of documents, but only between all pairs (d', d) such that d' is in the cited set and d is in the citing set. This saves us some unnecessary computational effort.

4.1.1 Citeseer data

This data is a pre-processed subset of the larger, publicly available Citeseer collection[2] that was made publicly available by Lise Getoor's research group at University of Maryland[3]. There are 3312 documents in the corpus and the vocabulary size is 3703 unique words. We pruned this collection to include only those documents that cite or are cited by at least *two* other documents. We did this so that the computational costs of running the Pairwise Link-LDA model remain within reasonable limits. This reduced the corpus size to 1168 documents, of which only 186 documents (15.9%) have both incoming and outgoing links. We duplicated these documents to create a bipartite link structure between a cited set of 591 documents and a set of 763 documents[4]. Further, we split the set of 763 citing documents into 10 sets with 50-50 random splits for training and testing, to allow computation of variance.

4.1.2 Blog Data

The data set consists of 8,370,193 postings on the blogosphere collected by *Nielsen Buzzmetrics*[5] between 07/04/2005 and 07/24/2005. We processed this data set as follows. First, there are many postings that are mistakenly assigned their respective site-URLs as their permalinks. These non-unique identifiers make it difficult to disambiguate between their incoming hyperlinks. Hence, we filtered these postings out, which left us with 7,177,951 postings. Next, we pruned this graph until we are left with postings, each of which has at least 2 outgoing or 2 incoming hyperlinks. We are finally left with 2,248 postings with at least 2 outgoing links each and 1,777 documents with at least two incoming links each. Of these only 68 postings (3.8%) have both incoming links and outgoing links, which we duplicated as described above, to create a bipartite graph.

Next, we pre-processed and indexed these postings using *Lemur*[6] tool-kit employing the *Krovetz* stemmer and a standard stop-word list. We pruned the vocabulary of this data further by ignoring words that contain numerals, that are less than 3 characters long, or those that occurred in less than 5 documents. The vocabulary size of the resulting corpus is 13,506. We split the set of citing postings uniformly at random into two equal sets (which we call set I and set II) of 1,124 postings each for training and testing purposes.

4.2 Tasks and Evaluation

4.2.1 Log-likelihood of new data

In this task, we measure how well the models predict unseen data in terms of log-likelihood. The higher log-

[2] *http://citeseer.ist.psu.edu/oai.html*
[3] *http://www.cs.umd.edu/∼sen/lbc-proj/LBC.html*
[4] *http://www.cs.cmu.edu/∼nmramesh/citeseer.tar.gz*
[5] *http://www.nielsenbuzzmetrics.com.* Available with free license.
[6] *http://www.lemurproject.org*

likelihood the model assigns to unseen data, better is its predictive power and generalizability.

Note that the Link-PLSA-LDA model and the Link-LDA models can only generate new citing documents, while the Pairwise Link-LDA model has no such restriction. However, for comparison purposes, we restrict our experiments to measuring the likelihood of only unseen citing documents.

Our experimental set-up is as follows. We first train each model's parameters using the entire set of cited postings and one of the training splits of the citing set. Using these estimated model parameters, we perform inference on the corresponding test set of the citing documents. Using these inferred variational parameters, we compute the variational lower-bound on the cumulative log-likelihood of the citing test set (both text and citations included). We repeat this process for all the train-test splits of the data and report the average values.

4.2.2 Link Prediction

In this task, we use the learned model to predict hyperlinks for documents that are not seen in the training phase. The Pairwise Link-LDA model can predict both incoming links as well as outgoing links for any new document, but the Link-PLSA-LDA and Link-LDA models can only predict outgoing links (onto a fixed set of cited documents) for new documents. Hence, for a fair comparison, for all models, we only predict outgoing links for new documents onto the set of cited documents.

Our experimental design is very similar to that of subsection 4.2.1, but is described below for clarity. We first learn the parameters of each of the models using the entire set of cited documents and one of the training splits of the of citing documents. Then, providing only the text of the citing documents from the corresponding test set, we performed inference to obtain the posterior topic distribution for each citing document in this set using inference updates that use only text as evidence. For all the models, this would involve the following two equations, which are computed iteratively until convergence.

$$\begin{aligned} \phi_{dnk} &\propto \beta_{kw_n} \exp(\Psi(\gamma_{dk})) \\ \gamma_{dk} &= \alpha_\theta + \sum_{n=1}^{N_d} \phi_{dnk} \end{aligned}$$

Using these probabilities and the model parameters learned during the training phase, we can compute the conditional probability of citation to any document in the cited set $d' \in \{1, \cdots, M_{\leftarrow}\}$ given the content of the citing document $\mathbf{w}_d$. For the Link-LDA and Link-PLSA-LDA models, this probability would be as follows:

$$\begin{aligned} P(d'|\mathbf{w}_d) &= \sum_{k=1}^{K} P(d'|k) P(k|\mathbf{w}_d) \\ &\approx \sum_{k} \Omega_{kd'} E[\theta_{dk}|\mathbf{w}_d] \\ &= \sum_{k=1}^{K} \Omega_{kd'} \frac{\gamma_{dk}}{\sum_{k'} \gamma_{dk'}} \end{aligned}$$

For the Pairwise Link-LDA model, the probability of the existence of a citation is computed as follows:

$$\begin{aligned} P(c|\mathbf{w}_d, \mathbf{w}_{d'}) &= \sum_{k=1}^{K} \sum_{k'=1}^{K} \theta_{dk} \theta_{d'k'} \eta_{kk'} \\ &\approx \sum_{k=1}^{K} \sum_{k'=1}^{K} E[\theta_{dk}] E[\theta_{d'k'}] \eta_{kk'} \\ &= \sum_{k=1}^{K} \sum_{k'=1}^{K} \frac{\gamma_{dk}}{\sum_{i=1}^{K} \gamma_{di}} \frac{\gamma_{d'k'}}{\sum_{j=1}^{K} \gamma_{d'j}} \eta_{kk'} \end{aligned}$$

For each citing document, we use these conditional probabilities to rank the documents in the cited set. We measure the effectiveness of this ranking with respect to the ground truth, which is the set of actual citations of the citing document. Drawing analogy to an information retrieval scenario, we assume each citing document to be a query and the set of its true citations to be the set of relevant documents, and the set of all documents in the cited set to be the retrieved set. One of the standard metrics used in information retrieval to evaluate the quality of a ranked list against a true set of relevant documents is average precision. However, we believe this metric is not suited for the task of link prediction in blog domain for two reasons: (i) this metric assumes that the true set is exhaustive, i.e., we have the complete set of relevant documents and (ii) the metric assigns high importance to precision at the top of the ranked list. While this may be appropriate for a key-word based search engine, the scenario for citations is quite different: citations are not assigned to all topically relevant documents, but only to a few documents known to the author. Hence the set of true citations does not represent an exhaustive set of topically relevant documents and it does not make sense to assign high importance to the top of the ranked list. Instead, we should focus on how well the model ranks the true citations; in other words, the measure should be more recall oriented. Hence, we use the value of the rank at 100%-recall as our evaluation measure. In short, we call this *RKL*, an abbreviation for RanK of the Last relevant document. This measure looks at how high in the ranked list the model places all the true citations. Higher the rank (lower in terms of numerical value), the better is the performance of the model.

Note that there are other models for link prediction available in the literature [12, 19, 17, 18], but we limit our comparison to Link-LDA , Pairwise Link-LDA and Link-PLSA-LDA models, since they are our main interest in this paper.

4.3 Parameter settings

In the learning phase, for all the models, we terminate the outer iterations if the fractional decrease in the lower bound of the log-likelihood of the entire observed data, in two successive iterations, is less than 10^{-4}, or if the number of iterations exceeds than 100.

For the inner iterations involving variational parameters, the stopping condition is when the fractional decrease in the lower bound of the log-likelihood of the data involved in the loop (a document or a link etc.) is less than 10^{-6} in two successive iterations, or when the number of iterations exceeds 20. The same criteria are also applied at inference time, for all the models.

Lastly, for all the models, we could technically estimate the value of α_θ using an empirical Bayes technique, but we fixed it at 0.1 for simplicity.

5. RESULTS

5.1 Topic Visualizations

5.1.1 Pairwise Link-LDA model

For the visualization of Pairwise Link-LDA model, we ran a 10 topic model on the cited set and citing set I of the blog corpus. We selected 4 representative topics from the output and displayed in figure 4 the top words using the learned β_{kw} values and also the $\eta_{kk'}$ values in the form of edges, which indicate the likelihood of a citation from topic k to topic k'. As the figure indicates, the new model not only displays the contents of topics, but also the citation strength between them. For example, the figure indicates that within-topic citation probability is very high (indicated by the self-pointing arrows). In addition, we see some interesting patterns, such as strong citation probability between "Iraq War" and "London Bombings" (the discourse of the two topics is very similar), "Iraq War" and "Supreme Court Nominations"(both are of interest to the community of conservative bloggers). Also, there is almost zero citation probability between "London Bombings" and "CIA Leak" (indicated by the absence of an edge), which is not surprising, because they are completely unrelated.

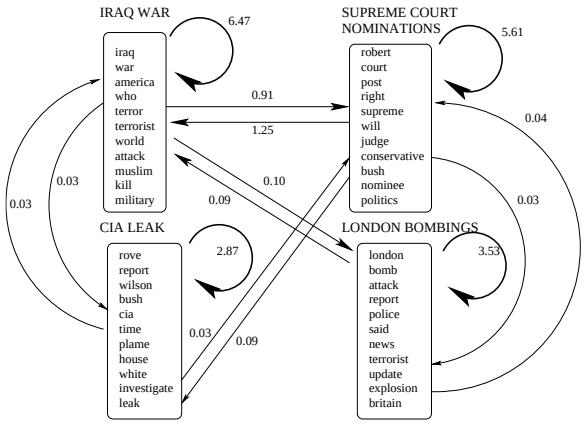

Figure 4: Visualization of the Pairwise Link-LDA model: the topics are hand-labeled. The weights of the edges correspond to the ratio of the probability of a citation between the corresponding topics to the prior probability of a random citation (0.0015). The thickness of the edges are roughly proportionate to these values. Although the graph is fully connected, we removed some edges whose probabilities are negligible.

5.1.2 Link-PLSA-LDA model

We ran the Link-PLSA-LDA model on set I of the citing postings and the cited postings with the number of topics K fixed at 25. We displayed 2 salient topics discovered by the Link-PLSA-LDA model in table 5. Like the previous model above, Link-PLSA-LDA tells us the most likely terms in each topic. For example, in the "CIA leak" topic, the model rightly identifies 'karl', 'rove', 'bush', 'plame' 'cia' and 'leak' as key entities in the topic. The name 'cooper' in the list refers to *Matt Cooper*, a reporter for the *Time* magazine, who testified in the CIA leak case. Similarly, the top terms in other topic are also equally illustrative of the topic content.

"CIA LEAK" **0.067**	"SEARCH ENGINE MARKET" **0.04**
TOP TOPICAL TERMS	
rove	will
his	search
who	new
time	market
cooper	post
karl	product
cia	brand
bush	permalink
know	time
report	yahoo
story	you
source	year
house	comment
leak	company
plame	business
TOP BLOG POSTS ON TOPIC	
billmon.org Whiskey Bar	edgeperspectives. typepad.com John Hagel
qando.net Free Markets & People	comparisonengines.com Comparison of Engines
captainsquartersblog .com, Captain's Quarters	blogs.forrester.com Charlene Li's Blog
coldfury.com The Light Of Reason	longtail.typepad.com The Long Tail
thismodernworld.com Tom Tomorrow	.searchenginejournal.com Search Engine Journal

Table 5: Visualization of the Link-PLSA-LDA model: topic titles are not part of the model. The numbers below the topic titles are the probability of each topic in the set of cited documents.

In addition, through the parameter $\Omega_{kd'}$, Link-PLSA-LDA also tells us the blog postings that are most influential in a topic k, as measured by both hyperlinks as well as by content. The most influential blogs for each topic are displayed at the bottom of table 5. As some of the titles of these blogs indicate, they seem topically very relevant. The blogs for the first topic are clearly political blogs. The second topic, "Search Engine Market", has all technology related blogs. The "CIA leak" topic has a mix of orientations (*billmon* and *tom tomorrow* are Democratic blogs, the others are Republican), hence the topic is most likely a mixture of argumentation back and forth between Democrats and Republicans.

The topic specific statistics described here are also learned by the Link-LDA model. There is, however, an additional statistic that the Link-PLSA-LDA model learns that is not directly learned by the Link-LDA model, namely the importance of topics (in terms of its occurrence in the set of cited postings), as measured by the parameter π. In table 5, we display the importance of each topic below its title. The numbers indicate that the "CIA-leak" topic is more popular than the "Search-Engine Market" topic as far as this corpus is concerned.

5.2 Log likelihood

Figure 5 compares the performance of the three models on log-likelihood evaluation (cf. section 4.2.1) on the Citeseer data, averaged over 10 runs. It is clear that both the new models are significantly better than the the Link-LDA baseline. Clearly, additional information of topical similarity between the text of the documents on either side of a link, leveraged by these models, is helping them predict new data better. Between the two models, it is apparent that Link-PLSA-LDA is clearly the better performer.

Surprisingly, the likelihood values for the Pairwise Link-LDA model decrease with increasing number of topics. We believe this is due to the fact that Pairwise Link-LDA con-

verges much slower than the other models due to its larger number of variational parameters. Hence the stopping criterion we used (cf. section 4.3) does not allow the model to reach its optimum state. Hence, it is unable to predict the test data as well as it should, at a higher number of topics. Despite this fact, it is appreciable that its performance is significantly better than that of the Link-LDA model.

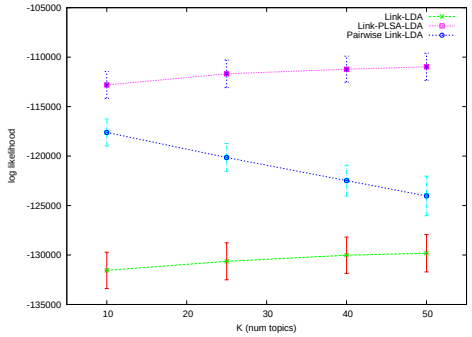

Figure 5: Likelihood performance of the three models on Citeseer data: higher is better. The error bar width is equal to two standard deviations as measured on 10 different runs.

We performed a similar evaluation on the blog data as well, but without the Pairwise Link-LDA model this time. The blog data is a larger corpus, and it takes exceedingly long time for this model to converge on this data, especially for higher number of topics. Hence we ignored it in these experiments.

In figure 6, we plotted the cumulative log-likelihood values for the Link-LDA and Link-PLSA-LDA models as a function of number of topics, using cross validation on the two sets of citing postings (cf. section 4.1.2). The plot again clearly shows that Link-PLSA-LDA predicts the data much better than the Link-LDA model.

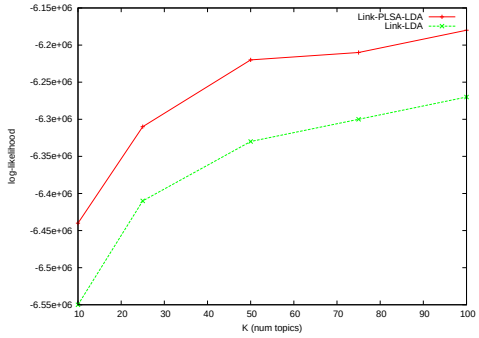

Figure 6: log-likelihood of blog data: higher is better

5.3 Link Prediction

Figure 7 compares the RKL performance of the three models as a function of number of topics K, on the Citeseer data, averaged over 10 runs. Again, Link-PLSA-LDA significantly outperforms the other models. However, Link-LDA significantly outperforms Pairwise Link-LDA this time. Since link prediction and log-likelihood have different objective functions, there is no reason to believe that models that do well on one task should perform well on the other too.

Figure 8 compares the performance of Link-PLSA-LDA with Link-LDA as a function of number of topics K, on the blog data, averaged over 2 runs. Again, we did not use the Pairwise Link-LDA model due to scalability issues. As was the case earlier, Link-PLSA-LDA again significantly outperforms Link-LDA at all values of K. Further, the performance only gets better as the number of topics is increased from 10 to 100.

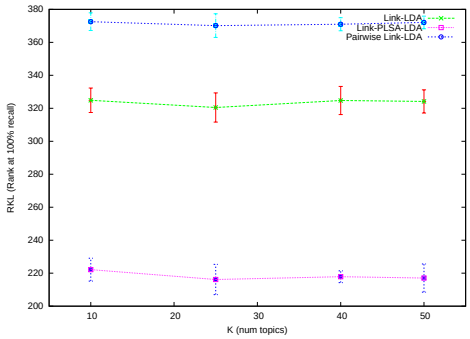

Figure 7: RKL (Rank at 100% recall) performance comparison of the three models on Citeseer data: lower is better. Error bars are 2 standard deviations wide.

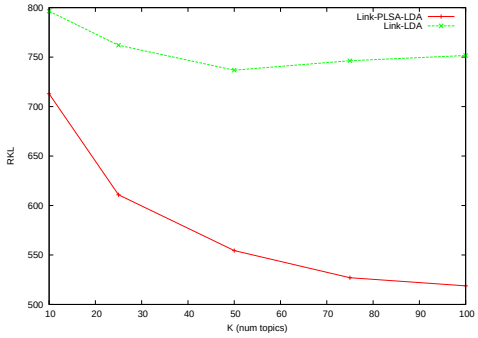

Figure 8: RKL (Rank at 100% recall) performance comparison of Link-PLSA-LDA with Link-LDA on blog data: lower is better

6. DISCUSSION

In this paper, we presented two novel topic models for joint modeling of links and text. We introduce the Pairwise Link-LDA for the first time, but limited preliminary results for Link-PLSA-LDA were presented recently in [14]. This work focused only on the single domain of blog data, and compared the Link-PLSA-LDA model to only the Link-LDA baseline. In this paper, we extend this work by performing more rigorous experiments on an additional data set and also by comparing the model to Pairwise Link-LDA , a new truly generative model for text citations.

Although the Pairwise Link-LDA is more expressive than Link-PLSA-LDA in terms of modeling arbitrary link structure, our experiments on Citeseer data and blog data show that the Link-PLSA-LDA model outperforms the former on both log-likelihood and link prediction. We believe the superior performance of the multinomial based Link-PLSA-LDA model compared to the multiple Bernoulli based Pairwise Link-LDA model has interesting parallels with the experience of Multinomial vs. Bernoulli distributions for text classification. It has been observed by McCallum *et al* [13] that modeling a document in terms of words that *occur* in the document using a multinomial distribution rather than

in terms of presence and absence of all the words in the vocabulary using a multiple Bernoullis yields superior performance. One of the reasons for this behavior could be because features that are present are much more important than features that are absent. The same reasoning could be extended to the behavior of the two models in the present context. Our hypothesis is that the Pairwise Link-LDA model suffers to due to significant expenditure of modeling effort on citations that are absent.

Apart from its superior performance, another attractive feature of the Link-PLSA-LDA model is its relative scalability to large document collections in comparison to the Pairwise Link-LDA model. To demonstrate its scalability, in figure 9, we plotted the average variational EM iteration (outer loop) run time for each model, as a function of number of topics. The measurements were made on a standard Linux machine with *Intel Pentium III* 0.7 GHz processor and 4GB memory. Clearly, there is a huge gap in computational time between Pairwise Link-LDA and the other two models. The run times of Link-PLSA-LDA and Link-LDA are comparable, but the former is the faster. The main reason for the prohibitive costs of the Pairwise Link-LDA model is that its computational complexity is quadratic in both the number of documents as well as number of topics. On the other hand, Link-PLSA-LDA and Link-LDA models are both linear in the number of documents and number of topics, making them more scalable.

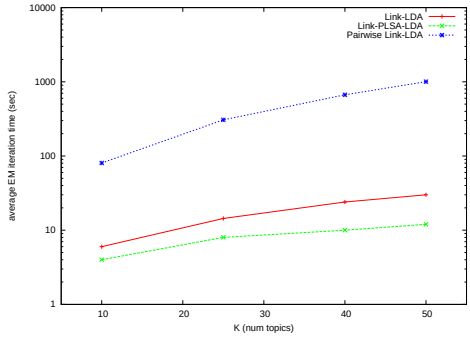

Figure 9: Comparison of average run times of the three models on Citeseer data: Y-axis is in log scale.

Despite the superior performance and scalability of the Link-PLSA-LDA model, we believe the Pairwise Link-LDA model is still very desirable owing to its superior semantics. Its generative process not only captures arbitrary link structure, but allows us to generate new cited as well as citing documents. Hence it may be more attractive in situations where capturing the true link structure is more important.

As part of our future work, we will try to build more tractable approximations to the Pairwise Link-LDA model.

non-scalability and restrictive assumptions respectively.

7. REFERENCES

[1] M. Airodi, D. Blei, E. Xing, and S. Fienberg. Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. In *International Biometric Society-ENAR Annual Meetings*, 2006.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, first edition, 2006.

[3] D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, 2006.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *International conference on Machine learning*, pages 113–120, 2006.

[6] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*, 2001.

[7] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *International Conference on Machine learning*, pages 233–240, 2007.

[8] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101:5220–5227, 2004.

[9] T. Hoffman. Probabilistic Latent Semantic Analysis. In *Uncertainty in Artificial Intelligence*, 1999.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[11] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *International conference on Machine learning*, pages 577–584, 2006.

[12] D. Liben-Nowell and J. Kleinberg. The link prediction problem in social networks. In *Conference on Information and Knowledge Management*, 2003.

[13] A. McCallum and K. Nigam. A comparison of event models for Naïve Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[14] R. Nallapati and W. Cohen. Link-LDA-PLSA: a new unsupervised technique for topics and influence in blogs. In *International Conference for Weblogs and Social Media*, 2008.

[15] R. Nallapati, J. Lafferty, W. Cohen, K. Ung, and S. Ditmore. Multiscale topic tomography. In *Conference on Knowledge Discovery and Data mining*, 2007.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *Technical report, Department of Computer Science, Stanford University*, 1998.

[17] B. Shaparenko and T. Joachims. Information genealogy: Uncovering the flow of ideas in non-hyperlinked document databases. In *Knowledge Discovery and Data Mining (KDD) Conference*, 2007.

[18] T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*, 2007.

[19] B. Taskar, Ming-FaiWong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Neural Information Processing Systems*, 2003.

[20] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. In *UC Berkeley, Dept. of Statistics, Technical Report*, 2003.