Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs

Ramesh Nallapati and **William Cohen**
{nmramesh@cs.cmu.edu} {wcohen@cs.cmu.edu}
Machine Learning Department,
Carnegie Mellon University,
5000 Forbes Ave., Pittsburgh, PA 15213, USA

Abstract

In this work, we address the twin problems of unsupervised topic discovery and estimation of topic specific influence of blogs. We propose a new model that can be used to provide a user with highly influential blog postings on the topic of the user's interest.

We adopt the framework of an unsupervised model called Latent Dirichlet Allocation(Blei, Ng, & Jordan 2003), known for its effectiveness in topic discovery. An extension of this model, which we call Link-LDA (Erosheva, Fienberg, & Lafferty 2004), defines a generative model for hyperlinks and thereby models topic specific influence of documents, the problem of our interest. However, this model does not exploit the topical relationship between the documents on either side of a hyperlink, i.e., the notion that documents tend to link to other documents on the same topic. We propose a new model, called Link-PLSA-LDA, that combines PLSA (Hoffman 1999) and LDA (Blei, Ng, & Jordan 2003) into a single framework, and explicitly models the topical relationship between the linking and the linked document.

The output of the new model on blog data reveals very interesting visualizations of topics and influential blogs on each topic. We also perform quantitative evaluation of the model using log-likelihood of unseen data and on the task of link prediction. Both experiments show that that the new model performs better, suggesting its superiority over Link-LDA in modeling topics and topic specific influence of blogs.

Introduction

Proliferation of blogs in the recent past has posed several new, interesting challenges to researchers in the information retrieval and data mining community. In particular, there is an increasing need for automatic techniques to help the users quickly access blogs that are not only informative and popular, but also relevant to the user's topics of interest.

Significant progress has been made in the recent past, towards this objective. For example Java *et al* (Java *et al.* 2006) studied the performance of various algorithms such as PageRank, HITS and in-degree, on modeling influence of blogs. Kale *et al* (Kale *et al.* 2006) exploited the polarity (agreement/disagreement) of the hyperlinks and applied a trust propagation algorithm to model the propagation of influence between blogs.

The above mentioned papers address modeling influence in general, but it is also important to model influence of blogs with respect to the topic of the user's interest. This problem has been addressed by the work of Haveliwala (Haveliwala 2002) in the context of key-word search. In this paper, PageRanks of documents are pre-computed for a certain number of topics. At query time, for each document matching the query, its PageRanks for various topics are combined based on the similarity of the query to each topic, to obtain a topic-sensitive PageRank. The author shows that the new PageRank results in superior performance than the traditional PageRank on key-word search. The topics used in the algorithm are, however, obtained from an external repository.

Ideally, it would be very useful to mine these topics automatically as well. The problem of automatic topic mining from blogs has been addressed by Glance *et al* (Natalie S. Glance & Tomokiyo 2006), where the authors used a combination of NLP techniques, clustering and heuristics to mine topics and trends from blogs. However, this work does not address modeling the influence of blog postings with respect to the topics discovered.

In our work, we aim at addressing both these problems simultaneously, *i.e.*, topic discovery as well as modeling topic specific influence of blogs, in a completely unsupervised fashion. Towards this objective, we employ the probabilistic framework of latent topic models such as the Latent Dirichlet Allocation (Blei, Ng, & Jordan 2003), and propose a new model in this framework.

The rest of the paper is organized as follows. In section , we discuss some of the past work done on joint models of topics and influence in the framework of latent topic models. We describe our new model in section . In section , we report the results of our experiments on blog data. We conclude the discussion in section with a few remarks on directions for future work.

Note that in the rest of the paper, we use the terms 'citation' and 'hyperlink' interchangeably. Likewise, note that the term 'citing' is synonymous to 'linking' and so is 'cited' to 'linked'. The reader is also recommended to refer to table 1 for some frequent notation used in this paper.

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'
$\text{Dir}(\cdot\|\alpha)$	Dirichlet distribution with parameter α
$\text{Mult}(\cdot\|\boldsymbol{\beta})$	Multinomial distribution with parameter $\boldsymbol{\beta}$
L_d	Number of hyperlinks in document d
N_d	Number of words in document d
β_{kw}	Probability of word w w.r.t. topic k
$\Omega_{kd'}$	Probability of hyperlink to document d' w.r.t. topic k
π_k	Probability of topic k in the cited document set.

Table 1: Notation

Past Work

Latent topic modeling has become very popular as a completely unsupervised technique for topic discovery in large document collections. These models, such as PLSA (Hoffman 1999) and LDA (Blei, Ng, & Jordan 2003), 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.

In (Cohn & Hofmann 2001), the authors built an extension to the PLSA (Hoffman 1999) model, called PHITS, that also simultaneously models the topic specific influence of documents. 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. Therefore, $\Omega_{kd'}$ can be interpreted as topic specific influence of the document d' with respect to topic k. 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* (Erosheva, Fienberg, & Lafferty 2004), 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. As in Link-PLSA, Link-LDA assigns high topic specific probability $\Omega_{kd'}$ to a document d' if it is frequently hyperlinked from documents that discuss the topic k. Therefore, we can interpret $\Omega_{kd'}$ as the influence of document d' with respect to topic k.

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: please refer to table 1 for explanation of the notation.

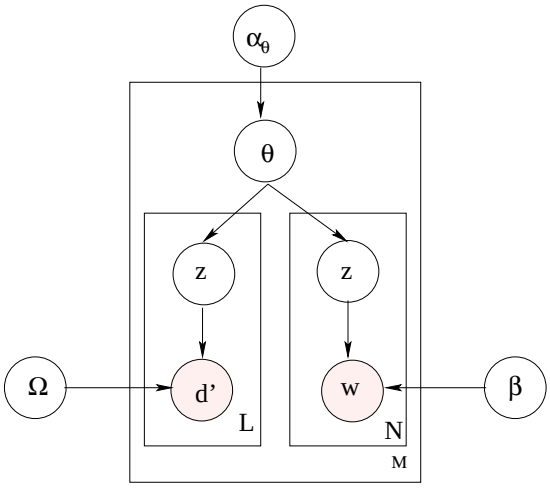

Figure 1: Graphical representation of the Link-LDA model

It is important to note that 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 hyperlinked documents exactly the same way as the basic LDA and PLSA models discover topical clusters of words. Thus, in effect they only exploit the co-occurrence of hyperlinks in documents, but they fail to explicitly model the topical relationship between the contents of the citing (linking) document and the cited (linked) document.

It is reasonable to expect that if a document d links to another document d', then both d and d' should be topically related. One can hope to obtain better quality of topics and influence by exploiting this additional information. More recently, Dietz *et al* (Dietz, Bickel, & Scheffer 2007) proposed

a new LDA based approach that allows flow of topic information from the cited documents to 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 captures only general influence of citations, but does not explicitly model topic specific influence of citations. In the next section, we will describe our new model that will address these issues.

New model: Link-PLSA-LDA

In this section, we describe the new model, which we call Link-PLSA-LDA , in detail. Subsection presents the generative process, while subsection describes how the model captures topic specific influence of blogs. In subsection , we discuss some limitations of the model while subsection presents the mathematical details of the inference and estimation of model parameters using variational approximations.

Generative process

In our work, we retained the approach of Link-LDA (Erosheva, Fienberg, & Lafferty 2004) and Link-PLSA (Cohn & Hofmann 2001), in which citations are modeled as samples from a topic-specific multinomial distribution $\mathbf{\Omega}$ over the cited documents. Thus, the generative process for the content and citations of the citing documents is 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 of PLSA as described in (Hoffman 1999). 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 3 and the corresponding graphical representation is shown in figure 2. One can see that information flows from the cited documents to the citing documents through the unobserved nodes $\boldsymbol{\beta}$ and $\mathbf{\Omega}$, as per the D-separation principle in Bayesian networks (Bishop 2006).

Modeling topic specific influence of blogs

As in Link-PLSA and Link-LDA, we can interpret $\Omega_{kd'}$ as the influence of document d' in topic k. Unlike in Link-PLSA and Link-LDA, where this influence arises solely by virtue of the document d' being cited by documents that discuss topic k, the new model also takes into account the content of d' in computing the topical influence of d'. This is a direct consequence of the fact that $\mathbf{\Omega}$ is employed in generating the text of the cited documents too. In addition, the parameter π_k in the new model can be interpreted as the importance or popularity of each topic in the data. Thus the new model offers us an additional statistic compared to the Link-LDA model.

The output of the model can be used to provide the user with highly influential blogs related to the topic of user's interest as follows. Let $Q = (q_1, \cdots, q_n)$ be the user's query that represents his/her topic of interest. One could return most influential blogs on this topic, ranked according to the following probability:

$$\begin{aligned} P(d'|Q) &= \sum_{z=1}^{K} P(d'|z)P(z|Q) \\ &\propto \sum_{z=1}^{K} P(d'|z)P(Q|z)P(z) \\ &= \sum_{z=1}^{K} \Omega_{zd'}(\prod_{i=1}^{N} \beta_{zq_i})\pi_z \end{aligned} \quad (1)$$

While $\Omega_{zd'}$ represents the topic specific influence of the document with respect to topic z, the term $\prod_{i=1}^{n} \beta_{zq_i}$ represents the similarity of the topic z to the user's topic of interest, while π_z can be interpreted as the prior importance of the topic z in the cited document set.

Model limitations

Since we generate cited documents and citing documents differently, a single document cannot both have citations and be cited. Thus, the model assumes a bipartite graph of citations from the citing set to the cited set. Although this is a serious modeling limitation, this can be easily overcome in practice: if a document has citations and is also cited, one can duplicate the document, retain only outgoing citations in one copy and incoming citations in the other and place them in their respective sets. In fact, this strategy has been successfully adopted by (Dietz, Bickel, & Scheffer 2007) in their work on modeling citation influences, which suffers from a similar limitation.

Also, note that 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 that is shared also by the PLSA model as well as the Link-LDA model. 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.

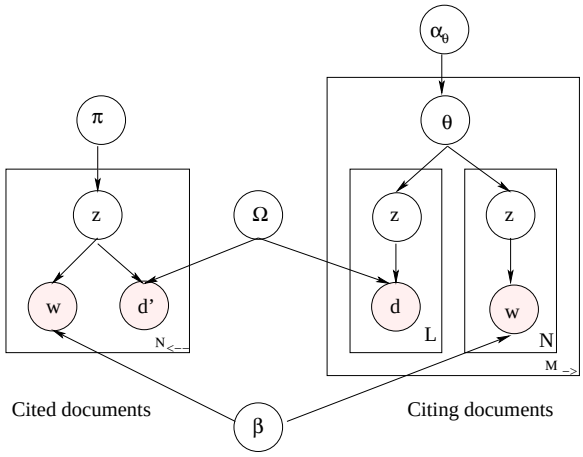

Figure 2: Graphical representation of the Link-PLSA-LDA model

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

Table 3: Generative process for the Link-PLSA-LDA model: please refer to table 1 for explanation of the notation.

Inference and Estimation

The likelihood of the observed data in this model is given as follows.

$$
\begin{aligned}
P(\mathbf{w}, \mathbf{c} \quad | \quad & \boldsymbol{\pi}, \alpha_\theta, \boldsymbol{\Omega}, \boldsymbol{\beta}) \\
= & \prod_{d'=1}^{M_{\leftarrow}} \prod_{n=1}^{N_{d'}} (\sum_k \pi_k \Omega_{kd'} \beta_{kw_n}) \\
\times & \prod_{d=1}^{M_{\rightarrow}} \int_{\boldsymbol{\theta}_d} (P(\boldsymbol{\theta}_d|\alpha_\theta) \prod_{n=1}^{N_d} (\sum_k \theta_{dk} \beta_{kw_n}) \\
\times & \prod_{l=1}^{L_d} (\sum_k \theta_{dk} \Omega_{kc_l})) d\boldsymbol{\theta}_d \qquad (2)
\end{aligned}
$$

where $\mathbf{w}$ is the entire text of cited and citing documents and $\mathbf{c}$ is the set of hyperlinks/citations. In general, this model is intractable for parameter estimation and inference, due to the pairwise coupling of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. Hence, researchers usually employ approximate techniques such as Gibbs sampling (Andrieu *et al.* 2003) or variational approximations (Wainwright & Jordan 2003). In this work, we will employ the mean-field variational approximation for the posterior distribution of the latent variables, as shown in figure 3. The graphical representation corresponds to the following parametric form:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \mathbf{z}, |\mathbf{w}, \mathbf{c}) = & \prod_{d=1}^{M_{\rightarrow}} (P(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \\
\times & \prod_{n=1}^{N_d} \prod_{k=1}^{K} ((\phi_{dnk})^{z_{dnk}}) \prod_{l=1}^{L_d} \prod_{k=1}^{K} ((\varphi_{dlk})^{z_{dlk}})) \\
\times & \prod_{d'=1}^{M_{\leftarrow}} \prod_{n=1}^{N_{d'}} \prod_{k=1}^{K} (\xi_{d'nk})^{z_{d'nk}}
\end{aligned}
$$

where the variational parameters have the following meaning:

- γ_{dk} is proportional to the posterior probability that the citing document d discusses topic k,
- ϕ_{dnk} is the posterior probability that the word at the n^{th} position in document d is generated from topic k,
- φ_{dlk} is the posterior probability that the l^{th} hyperlink in document d is generated from topic k and,
- $\xi_{d'nk}$ is the posterior probability that the n^{th} word in the cited document d' is generated from topic k.

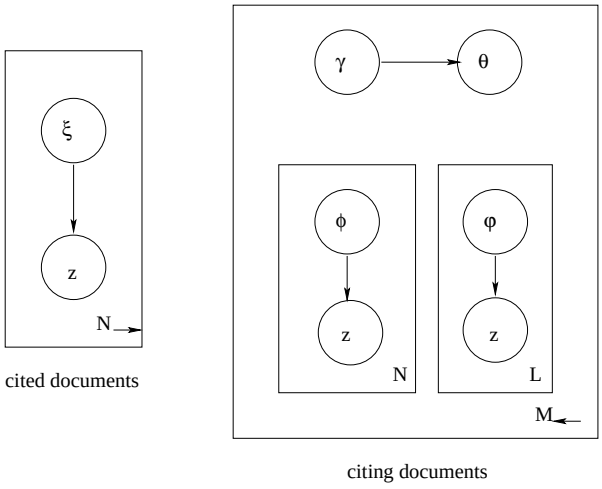

Figure 3: Graphical representation of the mean field approximate posterior distribution for the Link-PLSA-LDA model

Using Jensen's inequality, it can be shown that this approximation provides us with the following tractable lower-bound on the observed data log-likelihood:

$$
\begin{aligned}
& \log P(\mathbf{w}, \mathbf{c}|\boldsymbol{\pi}, \boldsymbol{\Omega}, \alpha_\theta) \\
\geq & \sum_{d'=1}^{M_{\leftarrow}} \sum_{n=1}^{N_{d'}} (\sum_k (\xi_{d'nk} (\log \pi_k + \log \beta_{kw_n} \\
+ & \log \Omega_{kd'})) + H(\boldsymbol{\xi}_{d'n}))
\end{aligned}
$$

$$+ \sum_{d=1}^{M_{\rightarrow}} (E_{\gamma}[\log P(\boldsymbol{\theta}_d|\alpha_{\theta})] + H(P(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d))$$

$$+ \sum_{n=1}^{N_d} (\sum_{k=1}^{K} \phi_{dnk}(E_{\gamma}[\log \theta_{dk}] + \log \beta_{kw_n}) + H(\boldsymbol{\phi}_{dn}))$$

$$+ \sum_{l=1}^{L_d} (\sum_{k=1}^{K} \varphi_{dlk}(E_{\gamma}[\log \theta_{dk}] + \log \Omega_{kd'_l}) + H(\boldsymbol{\varphi}_{dl})) \quad (3)$$

where $H()$ is a short notation for the entropy of the distribution in its argument, and $E_p[\,]$ represents the expectation of its argument with respect to the distribution parametrized by its subscript $_p$. It can be shown that the difference between the left hand side and the right hand side of eq. (3) above is equal to the KL-divergence between the variational posterior and the true posterior of the latent variables (Wainwright & Jordan 2003). Hence maximizing the lower bound is equal to finding a variational approximation that is closest to the true posterior in terms of the KL-divergence distance.

Differentiating the lower bound above with respect to each of the parameters and equating the resulting expression to zero, yields the following parameter updates:

$$\phi_{dnk} \propto \beta_{kw_n} \exp(\Psi(\gamma_{dk})) \quad (4)$$

$$\varphi_{dlk} \propto \Omega_{kd'_l} \exp(\Psi(\gamma_{dk})) \quad (5)$$

$$\gamma_{dk} = \alpha_{\theta} + \sum_{n=1}^{N_d} \phi_{dnk} + \sum_{l=1}^{L_d} \varphi_{dlk} \quad (6)$$

$$\xi_{d'nk} \propto \Omega_{kd'} \beta_{kw_n} \pi_k \quad (7)$$

$$\beta_{kv} \propto \sum_{d'=1}^{M_{\leftarrow}} \sum_{n=1}^{N_{d'}} \xi_{d'nk} \delta_v(w_n) + \sum_{d=1}^{M_{\rightarrow}} \sum_{n=1}^{N_d} \phi_{dnk} \delta_v(w_n) \quad (8)$$

$$\pi_k \propto \sum_{d'=1}^{M_{\leftarrow}} \sum_{n=1}^{N_{d'}} \xi_{d'nk} \quad (9)$$

$$\Omega_{kd'} \propto \sum_{n=1}^{N_{d'}} \xi_{d'nk} + \sum_{d=1}^{M_{\rightarrow}} \sum_{l=1}^{L_d} \varphi_{dlk} \delta_{d'}(d'_l) \quad (10)$$

These updates are performed iteratively until convergence. Since the updates in eq. (4) through (6) depend on each other, we also perform an inner iterative loop involving these equations, until they converge. Equations (4) through (7) are referred to as inference steps, since they involve the variational parameters. On the other hand, equations (8) through (10) are referred to as estimation steps, since they involve estimating the model parameters $\beta, \boldsymbol{\Omega}$ and $\boldsymbol{\pi}$.

Lastly, the value of α_{θ} can also be estimated by the same process described above, but for simplicity, we fixed it a value of 0.1, in both the models.

Experiments

In this section, we describe the experiments we performed on blog data. In subsection , we describe the details of the corpus we used. In subsection , we display the output of topics and the topical influence of blogs, obtained by the Link-PLSA-LDA model. In subsections and , we present quantitative evaluations of the Link-PLSA-LDA model and compare it with the performance of the Link-LDA model.

	Cited Postings	Citing Postings	
		set I	set II
# documents	1,777	1,124	1,124
Avg. Doc. Len.	160.274	217.46	221.73
Avg. # links/Doc	3.62	2.90	2.83
Max. # links/Doc	53	21	20
Vocabulary size	13506		

Table 4: Summary statistics of the corpus

Data

The data set consists of 8,370,193 postings on the blogosphere collected by *Nielsen Buzzmetrics*[1] 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 constructed a graph of hyperlinks that originate from and point to postings within this corpus. We then pruned this graph until we are left with postings, each of which has at least 2 outgoing or 2 incoming hyperlinks. The size of the pruned corpus is quite small compared to the original corpus, since most postings have hyperlinks that point to external sources such as news stories. 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 have both incoming links and outgoing links. We duplicated these 68 postings such that one set consists of only incoming links and the other set consists of only outgoing links. This allowed us to define a perfect bipartite graph of postings that contain 1,777 cited postings (with incoming links) and 2,248 citing postings (with outgoing links). Next, we pre-processed and indexed these postings using *Lemur*[2] 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. 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 purposes we will describe later. The statistics of the corpus are summarized in table 4.

Topical influence of blog postings

We ran the model on set I of the citing postings and the cited postings with the number of topics K fixed at 25 and α_{θ} fixed at 0.1. We displayed 4 salient topics discovered by the Link-PLSA-LDA model in table 5. Like the regular

[1] http://www.nielsenbuzzmetrics.com

[2] www.lemurproject.org

LDA model, 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*, who was a reporter for the *Time* magazine, that testified in the CIA leak case. Similarly, the top terms in other topics are also equally illustrative of the topic content. In addition, Link-PLSA-LDA also tells us the blog postings that are most influential in those topics, 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 three topics are clearly political blogs with the exceptions of *wizbangblog.com* and *the sharpener*, which are multi-editor community blogs. The last topic, "Search Engine Market", has all technology related blogs. The "Iraq war" and "Supreme court" topics have mostly Republican blogs associated with them. 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 so far 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 topics are also arranged in the descending order of their importance from left to right.

The topical influence analysis of the Link-PLSA-LDA model presented above seems to make intuitive sense, but it is not clear how good is the quality of output, compared to the Link-LDA model. One way to compare would be to hire human experts and make them evaluate the quality of the topic influence rankings of both the models. However, this is both cost and time intensive. Instead, we adopted two indirect, but cheaper and quicker ways to evaluate the performance of these two models. We present these two tasks in subsections and below.

Log-likelihood

In this subsection, we measure how well the models predict unseen data in terms of log-likelihood. The higher log-likelihood the model assigns to unseen data, better is its predictive power and generalizability. Our experimental set-up is as follows. We first train the model's parameters $\mathbf{\Omega}$ and π using the entire set of cited postings and set I of citing postings. Using these estimated model parameters, we perform inference on the set II of citing documents using eq. (4) through (6). Using these inferred parameters, we can compute a lower-bound on the cumulative log-likelihood of citing set II, as shown in eq. (3)[3]. Similarly, we repeat the same process by training on set II of citing postings and performing inference on set I. We report the total cumulative log-likelihood of the entire set of citing postings, by summing up the values obtained in each experiment. Note that we use the same experimental setup for both the models. In figure 4, we plotted the cumulative log-likelihood values for both the models as a function of number of topics. The plot clearly shows that Link-PLSA-LDA predicts the data much better than the Link-LDA model, indicating that the former is a better model for modeling topics and influence.

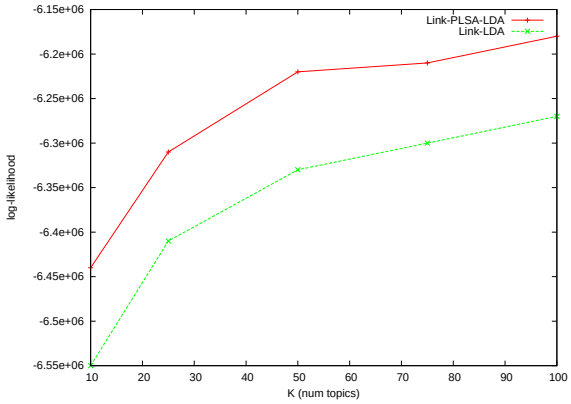

Figure 4: Cumulative log-likelihood of the citing postings with respect to Link-PLSA-LDA and Link-LDA: higher is better

Link Prediction

In this task, we use the learned model to predict hyperlinks for postings that are not seen in the training phase. The intuition is that if a model produces better quality topical influence analysis, then it should also be a better predictor of hyperlinks, since hyperlinks are known to be closely related to influence. There are other models for link prediction available in the literature (Liben-Nowell & Kleinberg 2003; Taskar *et al.* 2003; Shaparenko & Joachims 2007; Strohman, Croft, & Jensen 2007), but we confine ourselves to only the Link-PLSA-LDA and Link-LDA models, since they are our main interest in this paper. Note that we do not claim to, nor do we intend to outperform the best available model on this task.

Our experimental design is very similar to that of subsection , but is described below for clarity. We first learned the parameters of the Link-PLSA-LDA model using the entire set of cited postings and set I of citing postings. Then, providing only the text of the citing postings set II, we performed inference to obtain the posterior topic distribution γ_d for each citing document in this set using the following inference updates that use only text as evidence:

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

Using these probabilities and the model parameter $\mathbf{\Omega}$ learned during the training phase, we can compute the conditional

[3]Note that the first two lines in the RHS of eq. (3) correspond to the log-likelihood of the cited data, which we ignore in this computation.

Topic 21 "CIA LEAK" **0.067**	Topic 7 "IRAQ WAR" **0.062**	Topic 16 "SUPREME COURT NOMINATIONS" **0.06**	Topic 20 "SEARCH ENGINE MARKET" **0.04**
TOP TOPICAL TERMS			
rove	will	robert	will
his	war	court	search
who	attack	bush	new
time	iraq	his	market
cooper	terrorist	supreme	post
karl	who	john	product
cia	world	nominate	brand
bush	terror	judge	permalink
know	muslim	will	time
report	america	conservative	yahoo
story	one	right	you
source	people	president	year
house	think	justice	comment
leak	bomb	nominee	company
plame	against	senate	business
TOP BLOG POSTS ON TOPIC			
billmon.org Whiskey Bar	willisms.com Iraq what might	themoderatevoice.com The Moderate Voice	edgeperspectives. typepad.com John Hagel
qando.net Free Markets & People	instapunk.com InstaPun***K	blogsforbush.com Blogs for Bush	.comparisonengines.com Comparison of Engines
captainsquartersblog .com, Captain's Quarters	jihadwatch.org Jihad Watch	michellemalkin.com Michelle Malkin	blogs.forrester.com Charlene Li's Blog
coldfury.com The Light Of Reason	thesharpener.net The Sharpener	captainsquartersblog.com Captain's Quarters	longtail.typepad.com The Long Tail
thismodernworld.com Tom Tomorrow	thedonovan.com Jonah's Military	wizbangblog.com Wizbang	.searchenginejournal.com Search Engine Journal

Table 5: Topic display generated by 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.

probability of any cited posting $d' \in \{1, \cdots, M_{\leftarrow}\}$ given the content of the citing posting $\mathbf{w}_d$, as shown below:

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

This probability, $P(d'|\mathbf{w}_d)$, is the probability that the citing posting d is topically relevant to the cited posting d'. Hence one could assume that the higher this probability is, more is the chance that the author of the citing posting d would create a hyperlink to the cited posting d'. For each citing posting, we use these conditional probabilities to rank the entire set of cited postings. We measure the effectiveness of this ranking with respect to the ground truth, which is the set of actual citations of the citing posting.

We performed two-fold cross validation similar to the experiments in subsection , *i.e.*, we did another experiment in which we train the model on citing postings set II and cited postings, and evaluate on citing postings set I. We report the average performance of the models on sets I and II.

As a baseline, we used the Link-LDA model. The experimental design for this model is exactly same as that of the Link-PLSA-LDA model.

Evaluation Drawing analogy to an information retrieval scenario, we assume each citing posting to be a query and the set of its true citations to be the set of relevant documents, and the set of all cited postings 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 in blog data is quite different. In blogs, citations are not assigned to all topically relevant documents, but only to a few postings that the author has cared to read. Hence the set of true citations does not represent an exhaustive set of topically relevant documents. For the same reason, there is no guarantee that the most topically relevant postings are necessarily hyperlinked to. Hence 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 rates the postings that are actually hyperlinked. In particular, we look at the worst case scenario: how well the model ranks its most poorly ranked true citation. We call this the rank of the last relevant document or RKL in short. The lower the value of the rank is, the better is the performance.

Figure 5 compares the RKL performance of Link-PLSA-LDA with Link-LDA as a function of number of topics K. The RKL values are averaged over the entire set of citing postings, using a two-fold cross-validation technique described earlier. It is clear from the plot that Link-PLSA-LDA 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.

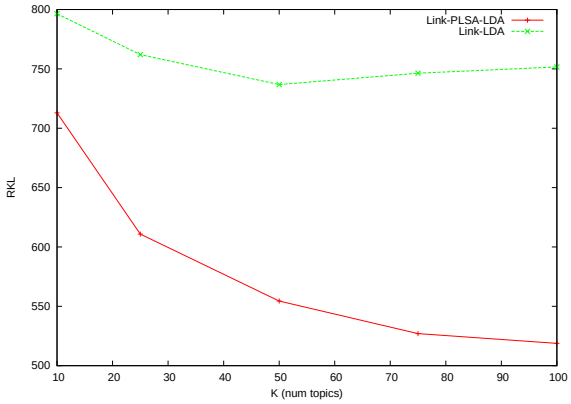

Figure 5: Performance comparison of Link-PLSA-LDA with Link-LDA: RKL is the rank of the last relevant document; lower is better

This performance again suggests that Link-PLSA-LDA is a better model than Link-LDA according to the proposed measure. We do however, caution the reader that additional experiments need to performed, especially using human judgments, to establish this conclusively.

Conclusions

In this work, we proposed a new model that discovers topics as well as models topic specific influence of blogs in a completely unsupervised fashion. Our experiments demonstrate that the new model is superior to the existing Link-LDA model on two different quantitative evaluations.

As part of our future work, we intend to perform experiments in key-word search (as described in section) to evaluate the new model's performance in providing the user with highly influential blog postings on the topic of his/her interest. Procuring labeled blog data for evaluation also forms part of future plans.

As discussed in section , one of the shortcomings of the Link-PLSA-LDA model is that it is not fully generative. In other words, the world of the hyperlinked documents is fixed and it is not possible to link to a new document under this model. In addition, the model restricts the hyperlink graph to be bipartite. At present, we are in the process of constructing a new model that is truly generative, one that allows arbitrary hyperlink structure.

References

Andrieu, C.; de Freitas, N.; Doucet, A.; and Jordan, M. 2003. An introduction to mcmc for machine learning. In *Machine Learning*.

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

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of machine Learning Research* 3:993–1022.

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

Dietz, L.; Bickel, S.; and Scheffer, T. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, 233–240.

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

Haveliwala, T. H. 2002. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference*.

Hoffman, T. 1999. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*.

Java, A.; Kolari, P.; Finnin, T.; and Oates, T. 2006. Modeling the spread of influence on the blogosphere. In *Workshop on weblogging ecosystem, World Wide Web conference*.

Kale, A.; Karandikar, A.; Kolari, P.; Java, A.; Finin, T.; and Joshi, A. 2006. Modeling trust and influence in the blogosphere using link polarity. In *International Conference on Weblogs and Social Media*.

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

Natalie S. Glance, M. H., and Tomokiyo, T. 2006. Modeling trust and influence in the blogosphere using link polarity. In *World Wide Web conference*.

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

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

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

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