Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)

Utilizing Context in Generative Bayesian Models for Linked Corpus

Saurabh Kataria and **Prasenjit Mitra** and **Sumit Bhatia**
Pennsylvania State University
University Park, PA-16801

Abstract

In an interlinked corpus of documents, the context in which a citation appears provides extra information about the cited document. However, associating terms in the context to the cited document remains an open problem. We propose a novel document generation approach that statistically incorporates the context in which a document links to another document. We quantitatively show that the proposed generation scheme explains the linking phenomenon better than previous approaches. The context information along with the actual content of the document provides signicant improvements over the previous approaches for various real world evaluation tasks such as link prediction and log-likelihood estimation on unseen content. The proposed method is more scalable to large collection of documents compared to the previous approaches.

Introduction

Large collections of interlinked documents such as the World Wide Web, digital libraries of scientific literature, weblogs have given rise to several challenging problems, e.g., detecting latent structures like *topics*, present in a given corpus. These latent structures, inherently, tend to seek a clustering of *semantically* similar entities present in the collection. Probabilistic approaches such as LDA (Blei, Ng, and Jordan 2003) and PLSA (Hofmann 1999) model the co-occurrence patterns present in text and identify a probabilistic membership of the words and the documents in a lower dimensional space.

In a linked corpus, the link structure contains meaningful information about entities, e.g., documents, authors etc.; this information has been successfully utilized in web search (Brin and Page 1998). However, the content based *topic* models (Blei, Ng, and Jordan 2003; Hofmann 1999; Blei and Lafferty 2006) completely ignore this information. Recently, Dietz, et al. (Dietz, Bickel, and Scheffer 2007), Nallapati, et al. (Nallapati et al. 2008) and Cheng, et al. (Chang and Blei 2009) have shown that modeling the citation and the content together not only helps to better understand the latent structure present in the data, but also helps to understand certain aspects of a linked corpus such as novelty detection, influence prorogation, citation prediction etc.

Although current approaches look at what other documents influenced the content of a document, they overlook how those documents influenced the content of this document. In other words, the process of incorporation of the citation information ignores the context in which that citation appeared in the document.

In this work, we present a generative model of the content and citations in a document that belongs to a linked corpus. Our models incorporate context information while modeling content and citations jointly. We hypothesize that context information can help in improving the topic identification for words and, in turn, documents. We assume that the author of the citing document chooses a topic first, and then while writing the text of the document chooses the citation context to describe a citation. The citation context does not necessarily portray the entire content of the cited document, but, provides a description from the author's perspective in relation to the citing document's topic. The citation context contains words related to the chosen topic and these words can help identify the major topics in the cited document. On the other hand, the topic of the context words can be identified using the major topics of the cited document as well. On the world-wide-web, anchor text and words surrounding the anchor text represent the context of the hyper-linked document.

Our approach for modeling the textual context of the citation is general enough to be applied to any Bayesian latent variable model for linked corpuses. In this paper, we show how we adapt previous approaches: linked-LDA (Stephen, Fienberg, and Lafferty 2004) and link-PLSA-LDA (Nallapati et al. 2008) to propose cite-LDA, cite-PLSA-LDA. As a result, we show that the context information helps to improve upon various objective functions such as log-likelihood of the generation of the content and link prediction experimentally.

Related work

One of the earliest attempt at modeling text and citation together in a linked corpus was posed as an extension of probabilistic latent semantic analysis (Hofmann 1999) and was called PHITS (Cohn and Hofmann 2001). PHITS proposed a topical clustering of citations in a manner similar to the topical clustering of words proposed in PLSA (Hofmann 1999). The Bayesian version of PHITS was proposed as *mixed membership model* (Stephen, Fienberg, and Lafferty

2004) and *linked-LDA* (Nallapati et al. 2008) with dirichlet acting as conjugate distribution to the multinomial distribution for citation and word generation process in PHITS. Although PHITS and its Bayesian extenssions are quanitiatively successful in clustering the citations and words, the underlying generative process is too simplistic to explain various phenomenon related to linked structure of the corpus, e.g. influence propogation, associating words and links, etc.

Recently, Nallapati, et al., (Nallapati et al. 2008) proposed a more rigorous modeling of the content and links together, named link-PLSA-LDA, where the data is partitioned into two subsets of cited and citing documents[1] and both the subsets are modeled differently with the same global parameters. The cited set of documents is modeled using PLSA and the citing set of documents is modeled using the *linked-LDA* model. The underlying assumption behind the link-PLSA-LDA model is that there exist both a global topic-citations distribution according to which the citing document chooses its citations and a global word-topic distribution from which the words are generated. This bipartite representation approach was first proposed by Dietz, et al., (Dietz, Bickel, and Scheffer 2007) to impose an explicit relation between the cited and the citing text so that the two together can augment the information provided by the citation links, while modeling a linked corpus. The plate model representations of the linked-LDA and the link-PLSA-LDA models are depicted in Figure 1(a) and 1(b) respectively. Consequently, the only difference between link-PLSA-LDA and linked-LDA (or PHITS) is that the linked-LDA assumes the same generative process for the cited set of documents as that in the citing set of documents as evident in the Figure 1(a) and (b).

Modeling Citation Context

Notations: Let V, D and N_d denote the size of the word vocabulary, the number of documents and number of words in document d respectively. Let $D_{\leftarrow}$ be the number of documents that are cited by any other document in the corpus. Let T denote the number of topics and suppose there exist a $T \times V$ topic-word distribution matrix β that indexes a probabilistic distribution over words given the topic and a $T \times D$ topic-citation distribution matrix γ that indexes the probability of a document being cited given a topic. At the document level, we assume that the author chooses to mix the topics with θ_d as the mixing proportion for document d. We treat the context information explicitly as follows. First, we define a citation context for a cited document as a bag of words that contains a certain number of words appearing before and after the citation's mention in the citing document. In case a cited document is mentioned multiple times, we assimilate all the corresponding context words. The basic underlying assumption while incorporating this context is that given a topic with a sufficiently narrow sense, the choice of words and the cited documents are independent. Suppose the author has a topic in mind (i.e., a distribution over words), and she comes across multiple documents that she can cite related to this topic. Now, if she has sufficiently narrowed down the topic, then the choice of words from that topic do not depend upon the choice of the document that she would cite. Based upon this assumption, next we describe two models for a linked corpus.

The cite-LDA Model: cite-LDA is a generative model with the generation process described in Algorithm 1 and the corresponding graphical depiction is given in Figure 1(c).

Algorithm 1 The cite-LDA generation process

for each document $d \in (1, 2, .., D)$: **do**
  $\theta_d \sim Dir(.|\alpha_\theta)$.
  for each word in $w_n \in d$ that appears outside any citation context: **do**
    Choose a topic $z_n \sim Mult(.|\theta_d)$.
    Choose w_n from word-topic distribution, i.e. $w_n \sim Mult(.|z_n, \beta_{z_n})$.
  end for
  for each word in $w_n \in d$ that appears inside of any citation context: **do**
    Choose a topic $z_n \sim Mult(.|\theta_d)$.
    Choose w_n from topic-word distribution, i.e. $w_n \sim Mult(.|z_n, \beta_{z_n})$.
    Choose a document c_n *to link* from topic-citation distribution i.e. $c_n \sim Mult(.|z_n, \gamma_{z_n})$.
  end for
end for

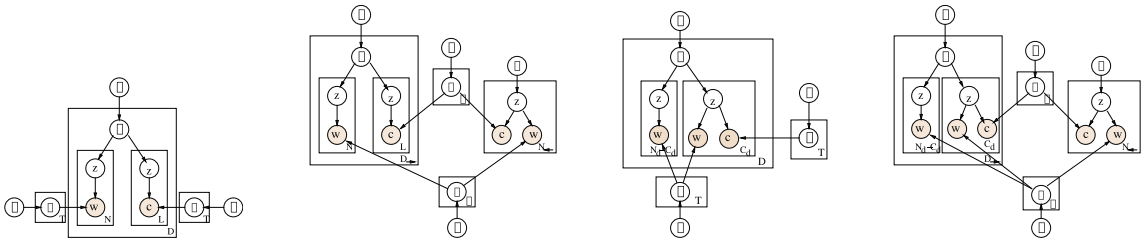
Formally, given the model parameters α, β and γ, the joint distribution of a topic mixture θ, the topic variables **z**, the document **w** and the citation context **c** can be written as:

$$
\begin{aligned}
&p(\theta, \mathbf{z,w,c}|\alpha, \beta, \gamma) \\
&= p(\theta|\alpha) \prod_{n=1}^{N_d - C_d} p(z_n|\theta)p(w_n|z_n, \beta) \prod_{n=1}^{C_d} p(z_n|\theta)p(w_n, c_n|z_n, \beta, \gamma) \\
&= p(\theta|\alpha) \prod_{n=1}^{N_d - C_d} p(z_n|\theta)p(w_n|z_n, \beta) \prod_{n=1}^{C_d} p(z_n|\theta)p(w_n|z_n, \beta)p(c_n|z_n, \gamma) \qquad (1) \\
&= p(\theta|\alpha) \prod_{n=1}^{N_d} p(z_n|\theta)p(w_n|z_n, \beta) \prod_{n=1}^{C_d} p(c_n|z_n, \gamma) \qquad (2)
\end{aligned}
$$

C_d is the total length of all citations contexts in the document d. The independence assumption allows us to factorize the joint distribution separately for the words and the citations. Intuitively, Eq. 1 implies that the author first picks the words from the topic and then citations from the topic or vice versa. The product $p(z_n|\theta).p(w_n|z_n)$ acts as the mixing proportions for the citation generation probability over the entire citation context of the corresponding citation. Therefore, one can expect that this explicit relation between citation generation probability and the word generation probability will lead to a better association of words and citations with documents than without utilizing the citation context explicitely.

The cite-PLSA-LDA Model: Similar to the link-PLSA-LDA (Nallapati et al. 2008) model, this model views the data as two separate sets of citing and cited documents as explained in previous section. cite-PLSA-LDA model assumes that the words and citations occurring in the citing documents generate from a smoothed (with a Dirichlet prior) topic-word and topic-citation multinomial distributions respectively. We model the generation of citation context by assuming the conditional independence of a word and a citation given the word. However, for cited documents, it is assumed that an empirical distribution of the topics is to be

[1] duplication is done for those documents that are both citing and cited in the corpus



(a) linked-LDA (b) link-PLSA-LDA (c) cite-LDA (d) cite-PLSA-LDA

Figure 1: Bayesian Network for (a) linked-LDA, (b) link-PLSA-LDA, (c) cite-LDA and (d) cite-PLSA-LDA

fitted that explains the generation of documents and words in cited set. Therefore, LDA (Blei, Ng, and Jordan 2003) and PLSA (Hofmann 1999) become the natural choice of frameworks for modeling the citing and the cited set respectively. The generation process assumed by the cite-PLSA-LDA model is described in Algorithm 2 and the corresponding graphical depiction is given in Figure 1(d).

Algorithm 2 The Cite-PLSA-LDA generation process

for each word w_n in cited set of documents: **do**
  Choose $z_i \sim Mult(.|\pi)$
  Choose $w_n \sim Mult(.|z_i, \beta_{z_i})$
  Sample $d_i \in 1, ...D_{\leftarrow} \sim Mult(.|z_i, \gamma_{z_i})$
end for
for each citing document $d \in (1, 2, .., D_{\rightarrow})$: **do**
  $\theta_d \sim Dir(.|\alpha_\theta)$
  for each word in $w_n \in d$ that appears outside any citation context: **do**
    Choose $z_n \sim Mult(.|\theta_d)$
    Choose w_n from word-topic distribution, i.e. $w_n \sim Mult(.|z_n, \beta_{z_n})$
  end for
  for each word in $w_n \in d$ that appears inside of any citation context: **do**
    Choose a topic $z_n \sim Mult(.|\theta_d)$
    Choose w_n from topic-word distribution, i.e. $w_n \sim Mult(.|z_n, \beta_{z_n})$
    Choose a document c_n *to link* from topic-citation distribution i.e. $c_n \sim Mult(.|z_n, \gamma_{z_n})$
  end for
end for

Formally, given the model parameters α, β, γ and π (the topic mixture for cited documents), the complete data likelihood can be obtained by marginalizing the joint distribution of a topic mixture θ for citing documents, the topic variable **z**, the document **w** and the citation context **c** and can be written as:

$$p(\mathbf{w},\mathbf{c}|\alpha,\beta,\gamma,\pi) = \prod_{n=1}^{N_{\leftarrow}}((\sum_k p(z|\pi)p(d_n|z)p(w_n|z)) \times \prod_d^{D_{\rightarrow}} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d}\sum_{z=1}^{K}(p(z_n|\theta_d)p(w_n|z_n,\beta)) \times \prod_{n=1}^{C}\sum_{z=1}^{K}(p(c_n|z_n,\gamma)))d\theta_d \quad (3)$$

Here, d_n indicates the document that word w_n belongs to.

Inference using Gibbs Sampling

The computation of the posterior distribution of the hidden variables θ and **z** is intractable for both cite-LDA and cite-PLSA-LDA model because of the pairwise coupling between θ, β and θ, γ. Therefore, we need to utilize approximate methods e.g. variational methods (Nallapati et al. 2008) or sampling techniques (Griffiths and Steyvers 2004) for inference. Considering that the markov chain monte carlo sampling methods such as Gibbs sampling come with a theoretical guarantee of converging to the actual posterior distribution and the recent advances in its fast computation capabilities over a large corpus (Porteous et al. 2008), we utilize Gibbs sampling as a tool to approximate the posterior distribution for both the models.

Inference Estimation for cite-LDA: According to Eq. 2, the joint probability distribution of the latent and the observed variables can be factorized as follows:

$$p(\mathbf{w},\mathbf{c},\mathbf{z}|\alpha,\beta,\gamma) = p(\mathbf{w}|\mathbf{z},\beta)p(\mathbf{c}|\mathbf{z},\gamma)p(\mathbf{z}|\alpha) \quad (4)$$

Let $n_k^{(c)}$ denote the number of times document c is observed with topic k. According to the multinomial assumption on occurences citations, we obtain:

$$p(\mathbf{c}|\mathbf{z},\varphi) = \prod_{i=1}^{C} p(c_i|z_i) = \prod_{k=1}^{K}\prod_{c=1}^{D} \varphi_{k,c}^{n_k^{(c)}}$$

$\varphi_{k,c}$ is proportional to the probability that document c to be cited with the topic k. The target posterior distribution for citation generation, i.e. $p(\mathbf{c}|\mathbf{z},\gamma)$, can be obtained by integrating over all possible values of φ:

$$p(\mathbf{c}|\mathbf{z},\gamma) = \int \prod_{z=1}^{K}\frac{1}{\Delta(\gamma)}\prod_{c=1}^{D}\varphi_{z,c}^{n_z^{(c)}+\gamma_c-1}d\varphi_z; \text{ where } \Delta(\gamma) = \frac{\prod_{i=1}^{dim(\gamma)}\Gamma(\gamma_i)}{\Gamma(\sum_{i=1}^{dim(\gamma)}\gamma_i)}$$
$$= \prod_{z=1}^{K}\frac{\Delta(\mathbf{n}_{\mathbf{z}\varphi}+\gamma)}{\Delta(\gamma)}; \text{ where } \mathbf{n}_{\mathbf{z}\varphi} = \{n_z^{(c)}\}_{c=1}^{D}$$

A similar derivation holds for $p(\mathbf{w}|\mathbf{z},\beta)$ and $p(\mathbf{z}|\alpha)$ leading to the expression (reader is referred to (Griffiths and Steyvers 2004) for further details) for joint distribution:

$$p(\mathbf{w},\mathbf{c},\mathbf{z}|\alpha,\beta,\gamma) = \prod_{z=1}^{K}\frac{\Delta(\mathbf{n}_{\mathbf{z}\phi}+\beta)}{\Delta(\beta)}\prod_{z=1}^{K}\frac{\Delta(\mathbf{n}_{\mathbf{z}\varphi}+\gamma)}{\Delta(\gamma)}\prod_{d=1}^{D}\frac{\Delta(\mathbf{n}_{\mathbf{m}}+\alpha)}{\Delta(\alpha)}$$

.

For Gibbs sampler, we need to derive $p(z_i = k|\mathbf{z}_{-i},\mathbf{w},\mathbf{c})$ where $\mathbf{z}_{-i}$ denote the entire state space of z except the i^{th} token and i iterates over each word in the corpus. With some algebraic manipulation, the updates for cite-LDA can be shown equivalent to Eq. (i) & (ii) in Table 1. Here, (**z,w**) implies that z is sample from outside the citation context whereas (**z,w,c**) inside the citation context.

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(t)}_{k,-i}+\beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i}+V.\beta} . \frac{n^{(k)}_{m,-i}+\alpha}{\sum_{k=1}^{K} n^{(k)}_{m,-i}+K.\alpha-1}; \text{ if } z_i \in (\mathbf{z}, \mathbf{w}). \quad \text{(i)}$$

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{c}) \propto \frac{n^{(t)}_{k,-i}+\beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i}+V.\beta} . \frac{n^{(c)}_{k,-i}+\gamma}{\sum_{c=1}^{D} n^{(c)}_{k,-i}+D.\gamma} . \frac{n^{(k)}_{m,-i}+\alpha}{\sum_{k=1}^{K} n^{(k)}_{m,-i}+K.\alpha-1}; \text{ if } z_i \in (\mathbf{z}, \mathbf{w}, \mathbf{c}) \quad \text{(ii)}$$

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{c}) \propto \frac{n^{(t)}_{k,-i}+\beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i}+V.\beta} . \frac{n^{(c)}_{k,-i}+\gamma}{\sum_{c=1}^{D} n^{(c)}_{k,-i}+D.\gamma} . \frac{n^{(.)}_{k}}{N_{\leftarrow}}; \text{ if } z_i \in (\mathbf{z}, \mathbf{w}^{\leftarrow}, \mathbf{c}) \quad \text{(iii)}$$

Table 1: Gibbs updates for cite-LDA(i,ii) and cite-PLSA-LDA(i,ii,iii)

Algorithm 3 Gibbs sampling for cite-PLSA-LDA model

```
while Not Converged do
    while Not Converged do
        for each word token w_n in citing documents: do
            if w_n appears in citation context of cited document c_n then
                sample z_i from p(z_i = k|z_{-i}, w, c) according to Eq.(ii), Table 3.
            else
                sample z_i from p(z_i = k|z_{-i}, w) according to Eq.(i), Table 3.
            end if
        end for
    end while
    for each word token w_n in cited documents: do
        sample z_i from p(z_i = k|z_{-i}, w, c) according to Eq.(iii) in Table 3.
    end for
end while
```

Inference Estimation for cite-PLSA-LDA: The joint distribution of the hidden topic variables $\mathbf{z}$, words $\mathbf{w}$ and the citations $\mathbf{c}$ can be written as:

$$p(\mathbf{w}, \mathbf{c}, \mathbf{z}|\alpha, \beta, \gamma, \pi) = p(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{c}|\mathbf{z}, \gamma)p(\mathbf{z}|\alpha)p(\mathbf{z}|\pi) \quad (5)$$

The derivation in previous section applies here which leads to following algebraic expression:

$$p(\mathbf{w}, \mathbf{c}, \mathbf{z}|\alpha, \beta, \gamma, \pi) = \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}_\phi^{\rightarrow}} + \beta)}{\Delta(\beta)} \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}_\varphi^{\rightarrow}} + \gamma)}{\Delta(\gamma)} \prod_{d=1}^{D_{\rightarrow}} \frac{\Delta(\mathbf{n}_{\mathbf{m}} + \alpha)}{\Delta(\alpha)} \times \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}_\phi^{\leftarrow}} + \beta)}{\Delta(\beta)} \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}_\varphi^{\leftarrow}} + \gamma)}{\Delta(\gamma)} \prod_{z=1}^{K} \pi_z^{n_z^{(.)}} \quad (6)$$

where $(\rightarrow)/(\leftarrow)$ indicates that the corresponding token was seen in citing/cited set and $n_z^{(.)}$ indicates the number of times topic z was observed in the cited set. The corresponding updates are obtained as given in Eq. (i), (ii) & (iii) in Table 3.

However, as we noted earlier, we intend to fit the topic distribution of words and citations learned from the citing set onto the cited set of documents. Therefore, a sequential scan over all the three partitions of the state space would be inappropriate. In other words, if we want to capture the conditional dependence based on topics between the citing set of documents and the cited set of documents, an iterative scheme of inference over citing documents and cited documents needs to be constructed. The corresponding Gibbs sampling update algorithm is depicted in Algorithm 3.

Experiments

We undertake two main tasks to evaluate cite-LDA and cite-PLSA-LDA model:(1) comparison of log-likelihood of words in the test set, (2) capability of predicting outgoing citations from the citing documents in the test set to the cited documents in the whole corpus.

Data Sets and Experimental Settings

We use two datasets: (1) scientific documents from the *CiteSeer* digital library, (2) web-pages from the *webkb* data set. These datasets have also been utilized by Nalapatti, et al. (Nallapati et al. 2008) for the two tasks.

Citeseer dataset: This dataset [2] was made publicly available by Lise Gatoor's research group at the University of Maryland and is a labeled subset of the CiteSeer [3] digital library. The data set contains 3312 documents belonging to 6 different research fields and the vocabulary size is 3703 unique words. There is a total of 4132 links present in the data set. We supplement the data set with the context information for each link. For each link, we add 60 words in the radius of 30 originating at the citation mention in the document. We vary the radius that proves to be crucial for the performance of the models (described later). For the pre-processing, we remove 78 common stop words and stem the words with porter stemmer which gives us 1987 unique words in the corpus. Further, we split the 1485 citing documents into 10 sets of 70-30 training and test split respectively. Since the link-PLSA-LDA and cite-PLSA-LDA model require bipartite structure for the corpus, we split the documents into two sets with duplication as suggested by Nalapatti, et al. (Nallapati et al. 2008).

Webkb dataset: This dataset [4] consists of web pages from the computer science department of various US universities. It includes faculty, staff, project and course web pages. The dataset consists of 2,877 different web pages with a vocabulary size of 102,927 words. After removing the stop words and stemming, the vocabulary size is 24,447 words. Note that the large vocabulary size as compared to CiteSeer dataset is due to the fact that a majority of pages contain unique nouns like faculty and staff names, project names etc. We found 1764 citations (hyperlinks) in the dataset with an average anchor text length of 3.02 words.

Loglikelihood Estimation on unseen text

This task quantitatively estimates the generalization capabilities of a given model over unseen data. In order to find the log-likelihood of words in the test set, we followed a similar approach taken by (Rosen-Zvi et al. 2004) where the inference algorithm is run exclusively on the new set of documents. We achieve this by extending the state of Gibbs sampler with the observation of the new documents. Before *sweeping* the test set, we first initialize the algorithm by randomly assigning topics to the words and the citation in the

[2] http://www.cs.umd.edu/ sen/lbc-proj/LBC.html

[3] http://CiteSeer.ist.psu.edu/

[4] http://www.cs.cmu.edu/ WebKB/

test set and then loop through the test set, until convergence, using following Gibbs sampling updates:

$$p(z_i^u|w_i^u = t, \mathbf{z^u_{-i}}, \mathbf{w^u_{-i}}) = \frac{n^{(t)}_{k,-i} + \beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i} + V.\beta} . \frac{n^{(k)}_{m^u,-i} + \alpha}{\sum_{k=1}^{K} n^{(k)}_{m^u,-i} + K.\alpha - 1} \quad (7)$$

Superscript $(.^u)$ stands for any unseen element. The sampling updates in Eq. 7 can be used to update the model parameters, $\Pi = (\theta, \phi, \varphi)$ for new documents as:

$$\theta_{m^u,k} = \frac{n^{(k)}_{m^u} + \alpha_k}{\sum_{k=1}^{K} n^{(k)}_{m^u} + \alpha_k}; \phi = \frac{{n^{(t)}_k}^u + n^{(t)}_k + \beta_t}{\sum_{t=1}^{V} {n^{(t)}_k}^u + n^{(t)}_k + \beta_t} \quad (8)$$

The predictive log-likelihood of a text document in the test set, i.e. $log(p(\mathbf{w^u}))$, given the model $\Pi = (\theta, \phi, \varphi)$ can be directly expressed as a function of the multinomial parameters of any given model:

$$p(\mathbf{w^u}|\mathbf{\Pi}) = \prod_{n=1}^{N_{m^u}} \sum_{k=1}^{K} p(w_n|z_n = k).p(z_n = k|d = m^u) \quad (9)$$

$$= \prod_{t=1}^{V} (\sum_{k=1}^{K} \phi_{k,t}.\theta_{m^u,k})^{n^{(t)}_{m^u}} \quad (10)$$

Fig. 2(a) & (b) show the comparison results on the two data sets. For both the data sets, we perform a 10-fold cross validation and report the average of the log-likelihood. Clearly, the cite-PLSA-LDA model outperforms all the other models on both of the data sets. The improvement in the performance is due to the fact that the association between citation and the words appearing in the context helps to identify the topic of the word. Also, the performance of cite-LDA and link-PLSA-LDA is comparable. We believe that this is because, for obtaining the topical association of words in citing document, the information provided by the link structure of the corpus and the context of links is as good as the content of cited document.

The improvements obtained for the CiteSeer data set is relatively larger than the web-kb data set. This is mainly because of the context of links in the web-kb data set is very noisy. There are very few instances where author of the web-page discuss a scientific project or his work and mention some links that are relevant to that discussion. In most cases, the links corresponds to class projects and department home-pages that do not have any context information close to the position of the link. On the other hand, the citations in CiteSeer data are always with a context and that context contains discussion relevant to the topic of the cited document.

Next, we discuss the effect of varying the context radius on perforance of the *cite* models. We measure the radius from the citation mention and vary it from 3 to 15 words. We observe a rapid increase in the log-likelihood function with the radius increasing from 3 to 10 words. After 10 words, the log-likelihood starts to stabilize and does not vary much after 14 words. This is mainly because after 10 words radius, the topic of discussion, generally, does not correlate much with the topics in the cited document. Also, for web-kb data, we observed this trend to appear only after 6 words of radius. Fig. 2(e) shows, for cite-PLSA-LDA, the change in log-likelihood with the change in the number of topics and the context radius. The same trend was observed for *link* models as well. The automatic selection of the appropriate radius for a given corpus will be of interest in future work.

Link Prediction

The experimental design for this task is very similar to the one in previous subsection. We first run the inference algorithm, descried in previous section, on training set for each model. Then we extend the Gibbs sampler state with the samples from the test set with following updates:

$$p(z_i^u|w_i^u = t, \mathbf{z^u_{-i}}, \mathbf{w^u_{-i}}) = \frac{n^{(t)}_{k,-i} + \beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i} + V.\beta} . \frac{n^{(k)}_{m^u,-i} + \alpha}{\sum_{k=1}^{K} n^{(k)}_{m^u,-i} + K.\alpha - 1}$$

$$p(z_i = k|\mathbf{z_{-i}}, \mathbf{w}, \mathbf{c}) \propto \frac{n^{(t)}_{k,-i} + \beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i} + V.\beta} . \frac{n^{(c)}_{k,-i} + \gamma}{\sum_{c=1}^{D} n^{(c)}_{k,-i} + D.\gamma}$$

$$. \frac{n^{(k)}_{m^u,-i} + \alpha}{\sum_{k=1}^{K} n^{(k)}_{m^u,-i} + K.\alpha - 1}; \text{ if } z_i \in (\mathbf{z}, \mathbf{w}, \mathbf{c}) \quad (11)$$

The parameters ϕ and θ can be obtained same as in Eq. 8, and the parameter φ can be obtained as:

$$\varphi = \frac{{n^{(c)}_k}^u + n^{(c)}_k + \gamma_t}{\sum_{c=1}^{D} {n^{(c)}_k}^u + n^{(c)}_k + \gamma_t}$$

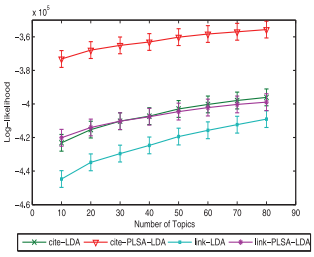
The probability $p(c|w_d)$, where c is the document to be cited and w_d is the citing document, can be expressed as:

$$p(c|w_d) = \sum_z p(c|z) \int p(z|\theta_d) d\theta_d \propto \sum_k \varphi_{c,k}.\theta_{k,d}$$

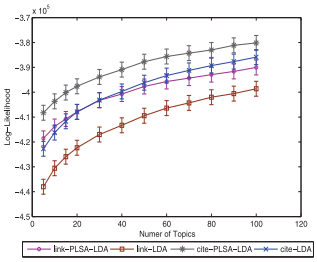
To evaluate the different models, we take a similar approach as taken by Nallapatti et al. (Nallapati et al. 2008). We label the actual citations of the document as relevant set for that citing document and evaluate our models based upon what rankings are given to these actual citations. In Fig.2 (c) & (d), we plot the average of the maximum rankings given to these relevant links. The plot contains average over all the test set. Clearly, the lower the rank assigned by the model, the better it performs. Cite-PLSA-LDA outperforms all the other model and cite-LDA and link-PLSA-LDA have comparable performance. The link-LDA model is again outperformed by the other models because of its over-simplicity.

Complexity Analysis

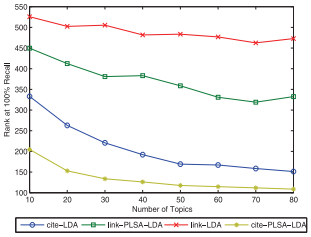
For link-LDA and link-PLSA-LDA, the time complexity of a single iteration of the Gibbs sampler grows linearly with the number of links present in the corpus. This can be prohibitive in the case of large corpuses such as the WWW where the links are in the order of 10^6. For cite-LDA and cite-PLSA-LDA, the modeling of citation variable is explicitly associated with the word variable, therefore, sampling from the posterior distribution of the topic variable does not depend upon the number of links and only grows linearly with the number of words in the corpus. The time complexity of one sampling iteration from the citing set for cite-PLSA-LDA and cite-LDA is $O(\sum_d \sum_n d_n * K)$ where K is the number of topics, d is the iterator over the documents



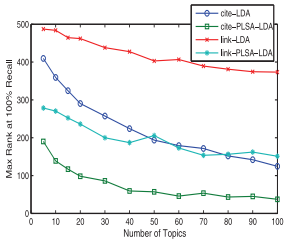
(a) Citeseer Dataset



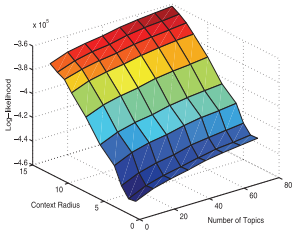
(b) Web-kb Dataset



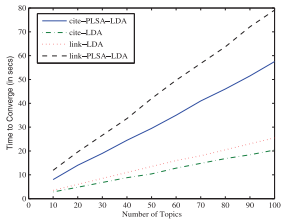
(c) Maximum rank at 100% recall for Citeseer Dataset



(d) Maximum rank at 100% recall for Web-kb Dataset



(e) context radius vs. number of topics vs. loglikelihood



(f) Convergence time on CiteSeer data

Figure 2: Comparision of Loglikelihood (a & b) and link prediction task (c & d) for the two proposed models with link-PLSA-LDA (Nallapati et al. 2008) and link-LDA (Stephen, Fienberg, and Lafferty 2004) on CiteSeer and web-kb datasets. (e) Effect of varying context length in CiteSeer data. (f) Comaprision of convergence time for the 4 models on CiteSeer data

and n is the iterator over the words in document d, whereas it is $O(\sum_d(\sum_n d_n * K + \sum_l d_l * K)$ for link-LDA and link-PLSA-LDA, where l is iterator over citations in document d. Fig.2(f) shows the convergence time for the 4 models on the CiteSeer data with varying number of topics. For cite-PLSA-LDA and link-PLSA-LDA model, we compare the performance of outer loop of Gibbs sampling until the model parameters reach convergence. The performance of cite-LDA and link-LDA is comparable whereas the performance of cite-PLSA-LDA and link-PLSA-LDA model is comparable, however, in both cases, the former performs better than the later.

Conclusion

We presented a framework that utilizes context information of citations in documents to model the generation process of documents and citations. Identifying the text from the context that describes the cited document is a challenging task. We show how to statistically model the citation context explicitly. Our model explains the generation process of the links and content both qualitatively and quantitatively. We utilize Gibbs sampling to perform inference on emission probabilities corresponding to citations and words given a topic and show significant improvement on various objective functions. We also utilize the models to find associations between citation context and the cited document.

Acknowledgement

This work was partially supported by grant HDTRA1-09-1-0054 from DTRA. We are thankful to Dr. Lise Gatoor and Dr. C. Lee Giles for making the Citeseer dataset publically available. We are also thankful to Dr. Frank Ritter for editing the final draft.

References

Blei, D. M., and Lafferty, J. D. 2006. Correlated topic models. In *NIPS 18*.

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

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, 107–117.

Chang, J., and Blei, D. 2009. Relational topic models for document networks. In *Proc. of Conf. on AI and Statistics (AISTATS'09)*.

Cohn, D., and Hofmann, T. 2001. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS 13*.

Dietz, L.; Bickel, S.; and Scheffer, T. 2007. Unsupervised prediction of citation influences. In *ICML 2007*, 233–240.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *In Proc of National Academy of Science U.S.A.* 5228–5235.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *UAI 1999*, 289–296.

Nallapati, R. M.; Ahmed, A.; Xing, E. P.; and Cohen, W. W. 2008. Joint latent topic models for text and citations. In *KDD 2008*, 542–550.

Porteous, I.; Newman, D.; Ihler, A.; Asuncion, A.; Smyth, P.; and Welling, M. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *KDD 2008*, 569–577.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *UAI 2004*, 487–494.

Stephen, E. E.; Fienberg, S.; and Lafferty, J. 2004. Mixed membership models of scientific publications. In *Proceedings of the National Academy of Sciences*, 2004.