Context Sensitive Topic Models for Author Influence in Document Networks*

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea and C. Lee Giles
The Pennsylvania State University
College Park, PA-16801
{skataria, pmitra, ccaragea, giles} @ist.psu.edu

Abstract

In a document network such as a citation network of scientific documents, web-logs, etc., the content produced by authors exhibits their *interest* in certain *topics*. In addition some authors *influence* other authors' interests. In this work, we propose to model the influence of cited authors along with the interests of citing authors. Moreover, we hypothesize that apart from the citations present in documents, the context surrounding the citation mention provides extra topical information about the cited authors. However, associating terms in the context to the cited authors remains an open problem. We propose novel document generation schemes that incorporate the context while simultaneously modeling the interests of citing authors and influence of the cited authors. Our experiments show significant improvements over baseline models for various evaluation criteria such as link prediction between document and cited author, and quantitatively explaining unseen text.

1 Introduction

The popularity of Web 2.0 applications has resulted in large amounts of online text data, e.g. weblogs, digital libraries of scientific literature, etc. These data require *effective* and *efficient* methods for their organization, indexing, and summarization, to facilitate delivery of content that is tailored to the interests of specific individuals or groups. Topic models such as Latent Dirichlet Allocation (LDA) [Blei *et al*., 2003] and Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 1999] are generative models of text documents, which successfully uncover hidden structures, i.e., *topics*, in the data. They model the co-occurrence patterns present in text and identify a probabilistic membership of words and documents into a much lower dimensional space compared to the original term space. Since their introduction, many extensions have been proposed.

One such line of research aimed at modeling the interests of authors to answer important queries about authors, e.g., who produced similar work [Rosen-Zvi *et al*., 2004], who belongs to the same research community [Liu *et al*., 2009; Wang *et al*., 2005] and who are the experts in a domain [Tu *et al*., 2010]. However, another fundamental question about the attribution of topics to authors still remains not answered: who influences the generation of new content in a particular topic of interest? In this work, we propose generative models that take the linkage between authors of citing and cited documents into consideration and explore various qualitative and quantitative aspects of this question.

Another line of research aimed at modeling topics for content and citations together to quantify the influence of citations over the newly generated content [Dietz *et al*., 2007; Nallapati *et al*., 2008; Chang and Blei, 2009; Kataria *et al*., 2010]. However, these statistical methods for parameterizing the influence of a document cannot easily quantify the influence of authors because one document often has multiple authors.

In this work, we exploit the complementary strengths of the above lines of research to answer queries related to authors' influence on topics. Specifically, we present two different generative models for inter-linked documents, namely the author link topic (ALT) and the author cite topic (ACT) models, which simultaneously model the content of documents, and the interests as well as the influence of authors in certain topics. As in the author topic model (ATM) [Rosen-Zvi *et al*., 2004], ALT models a document as a mixture of topics, with the weights of the mixture being determined by the authors of the document. In order to capture the influence of cited authors, ALT extends ATM to let the set of cited authors in a document be represented as a mixture of topics and again the weights of the topics are determined by the authors of the document.

Moreover, we hypothesize that the context in which a cited document appears in a citing document indicates how the authors of the cited document have influenced the contributions by the citing authors. ACT extends ALT to explicitly incorporate the citation context, which could provide additional information about the cited authors. Kataria et al. [2010] have previously used the citation context while jointly modeling documents and citations (without authors) and have shown that a fixed-length window around a citation mention can provide improvements over context-oblivious approaches. Unlike Kataria et al. [2010], we model the authors of the doc-

*To appear in Proceedings of 22nd International Joint Conference on Artificial Intelligence (IJCAI'11), Barcelona, Spain, July 16-22, 2011

ument along with the content and argue that a fixed-length window around a citation mention can provide either limited or erroneous information in cases where the context spans are larger or smaller, respectively, than the length of the window. Hence, we dynamically select an adaptive-length window around a citation that is statistically more likely to explain the cited document than a fixed-length window.

In summary, our contributions are as follows:

- We propose generative models for author-author linkage from linked documents conditioned on topics of interest to authors. Our models are able to distinguish between authors' interests and authors' influence on the topics.
- We utilize the context information present in the citing document explicitly while modeling the cited authors and obtain significant benefits on evaluation metrics on real world data sets. Moreover, we dynamically select the length of context surrounding the citation mention and circumvent the erroneous context inclusion by a fixed window approach.

2 Related work

One of the earliest attempts at modeling the interests of authors is the author topic model (ATM) [Rosen-Zvi *et al*., 2004], where the authors and the content are simultaneously modeled with coupled hyper-parameters for the interests of authors and the themes present in text (shown in Fig. 1(a)). The (latent) topics represent the shared dimensions among the interest of authors and the themes. Bhattacharya and Getoor [2006] extended ATM to disambiguate incomplete or unresolved references to authors. Another stream of author centric modeling deals with expert finding [Fang and Zhai, 2007; Balog *et al*., 2009; Tu *et al*., 2010] where an expert is defined as a person *knowledgeable* in the field. We define an *expert/interested* author as someone who has produced several contributions in a particular field whereas an influential author as someone who has certain key contributions in that field and gets cited more often. Therefore, given a field, an influential author is not necessarily an expert in that field, however, her key contributions have led several interested authors to contribute to that field. However our main goal is to model the influence of authors along with the interest of authors.

Linking to external content or entities is an important ingredient of social content such as citation graph of academic documents, asynchronous communications such as weblogs, e-mails, etc. The *mixed membership model* [Erosheva *et al*., 2004], also referred as *linked-LDA* [Nallapati *et al*., 2008], extended LDA to model links among documents with an additional parameter that governs link generation from citing documents to cited documents. Further extensions of *linked-LDA* analyzed the association between words and hyperlinks [Nallapati *et al*., 2008; Gruber *et al*., 2008; Chang and Blei, 2009], influence propagation [Dietz *et al*., 2007], community of links detection [Liu *et al*., 2009], context-sensitive citation and text modeling [Kataria *et al*., 2010]. To model the authors in an inter-linked corpus of documents, Tu et al. [2010] proposed an extension of the author topic model to inter-linked documents. In contrast to our approach, they consider the entire citing document as the context of the citation, which, as explained in § 4.2, can easily be considered as a special case of our approach. In addition, it performs inferior to dynamically selecting the context length.

Topic models have also been extended to social networks of entities where entity-entity relationships conditioned upon topics are explored. Mccallum, et al., [2007] extended the basic ATM to cluster the entity pairs based upon topic of conversation in e-mail corpus. Their approach assumes that the sender and the recipient both decide the entire topic of conversation. This assumption is not applicable in our setting because only the author of the citing document decides the topic of the document and every cited authors may not share the interest in all the topics discussed in citing document. Newman et al. [2006] and Shiozaki et al. [2008] proposed other entity-entity relationship models for named-entities in news articles where documents are modeled as mixture of topics over both entities and words.

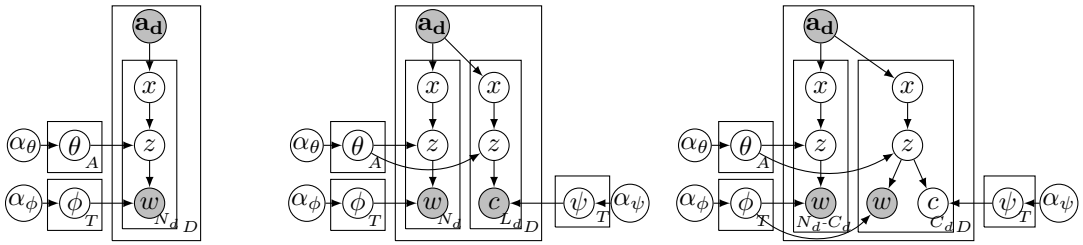
3 Models

Before presenting our models, we introduce some useful notations. Let V, D, A, $\mathbf{a_d}$ and N_d denote the size of the word vocabulary, the number of documents, the number of authors, a set of authors and the number of words in document d respectively. Let T denote the number of latent topics, i.e., the latent variable z (see Fig. 1) can take any value between 1 and T inclusively. Suppose there exists a $T \times V$ topic-word distribution matrix ϕ that indexes a probabilistic distribution over words given a topic and a $T \times A$ topic-author distribution matrix θ that indexes the probability with which an author shows interest in a topic. The corresponding hyper-parameters for distributions ϕ and θ are α_ϕ and α_θ respectively.

3.1 Author Link Topic Model

Citations among documents exhibit the biases of citing authors towards certain influential authors who have key contributions in the topic of discourse. We quantify the influence of an author given a topic by the probability, denoted by φ_{cz}, that the author c's work gets cited when there is a mention of the topic z in a citing document. Since the Author Topic Model (ATM) does not model the citations among the documents, it is not possible to estimate the influence of an author given a topic. In contrast, Author link topic model (ALT) generates the references to cited authors along with the words from a mixture of topics. As in ATM, a set of authors $\mathbf{a_d}$ decides to write a document. To generate each word, an author x is chosen uniformly at random from $\mathbf{a_d}$, and a topic is sampled from the chosen author's specific distribution. Then the corresponding word is generated from the chosen topic. For each author in the referenced set of authors in the document d, again an author x is chosen to generate a topic, and based upon the topic, an author c is selected from the topic specific distribution over authors. ALT model captures the intuition that given a topic and a list of relevant authors to be cited, authors from $\mathbf{a_d}$ would choose to reference those authors's work that are influential in that topic. Fig. 1(b) shows the plate diagram for the ALT model.

In the following subsections, we will use $\mathbf{w}$ and $\mathbf{c}$ to denote the words and observed cited authors in a document and $\mathbf{z}$ to denote the vector of topic assignments in the document.



(a) Author Topic Model (b) Author Link Topic Model (c) Author Cite Topic Model

Figure 1: Plate diagram for: (a) Author Topic Model; (b) Author Link Topic Model; and (c) Author Cite Topic Model.

With the model hyper-parameters α_θ, α_ϕ and α_φ, the joint distribution of authors **x**, the topic variables **z**, the document **w** and the cited authors **c** can be written as below. Here, L_d stands for the number of cited authors in the document d.

$$p(\mathbf{x,c,z,w}|\mathbf{a_d}, \alpha_\theta, \alpha_\phi, \alpha_\varphi) = \tag{1}$$
$$\int\int\int \prod_{n=1}^{N_d} p(x|\mathbf{a_d})p(z_n|x,\theta_x)p(w_n|z_n,\phi_{z_n})p(\theta_x|\alpha_\theta)p(\phi_{z_n}|\alpha_\phi)$$
$$\prod_{l=1}^{L_d} p(x|\mathbf{a_d})p(z_l|x,\theta_x)p(c_l|z_l,\varphi_{z_l})p(\theta_x|\alpha_\theta)p(\varphi_{z_l}|\alpha_\varphi)d\theta d\phi d\varphi$$

3.2 Context sensitive modeling of Author-linkage : Author Cite Topic Model

ALT model does not utilize the context in which a document cites an author. Although ALT models the cited authors in the citing document, yet, because of the bag of words assumption, the topic assignment to the authors does not explicitly depend upon the topics assigned to the content in that document. To enforce this dependence, we model the cited authors along with the context of the citation. In contrast to ALT, the Author Cite Topic (ACT) model associates cited authors and the words in the citation context of the cited authors with topic assignments to the context words. This association is based upon the assumption that given a topic, the choice of words and the authors to be cited are independent (see the plate diagram in Fig 1(c). With this independence assumption, the topic sampled for words in the citation context window generates both a word and a reference to the cited author. Since we observe a set of authors for a cited document, we treat c as hidden similar to x. The parameters of the ACT model remain the same as those of the ALT model, however the complete data log-likelihood function is different due to a difference in the generation process. The log-likelihood function to optimize can be written as below. Here, C_d is the total length (number of words) of all citation contexts in the document d.

$$p(\mathbf{x,c,z,w}|\mathbf{a_d}, \alpha_\theta, \alpha_\phi, \alpha_\varphi) =$$
$$\int\int\int \prod_{n=1}^{N_d - C_d} \Big(p(x|\mathbf{a_d})p(z_n|x,\theta_x)p(w_n|z_n,\phi_{z_n})p(\theta_x|\alpha_\theta)$$
$$p(\phi_{z_n}|\alpha_\phi)\Big) \prod_{n=1}^{C_d} \Big(p(x|\mathbf{a_d})p(z_n|x,\theta_x)p(\theta_x|\alpha_\theta)p(w_n|z_n,\phi_{z_n})$$
$$p(\phi_{z_n}|\alpha_\phi)p(c_n|z_n,\varphi_{z_n})p(\varphi_{z_n}|\alpha_\varphi)\Big)d\theta d\phi d\varphi \tag{2}$$

Intuitively, Eq. 2 implies that the author first picks the words from the topic and then chooses to cite an author's work or vice versa. The product $p(z_n|x,\theta_x).p(w_n|z_n,\phi_{z_n})$ acts as the mixing proportions for the author "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, and in turn authors, with documents than without utilizing the citation context explicitly.

3.3 Dynamic Selection of Length of Context Window

Since the ACT model imposes independence assumption in the context window surrounding the citation mention, it becomes important to identify the context that refers to the cited article. Previous work on context utilization in topic models, either assumes a fixed window of 10 words radius surrounding the citation mention [Kataria *et al*., 2010] or the whole document as the context for any citation mention [Tu *et al*., 2010]. However, the amount of relevant context in the vicinity of the citation anchor depends upon various factors such as the strength of the influence of cited article over the citing article, the location of the citation mention in the citing article, etc. Therefore, we propose to identify a dynamic window surrounding the citation anchor with the following method.

Let $\overleftarrow{d}$ represent the cited document for a given citation anchor c_i^d, where i ranges over all citation mentions in the citing document d. Let $S(c_i^d)$ (or simply S_i) represent the bag of words in the citation context surrounding c_i^d. The objective function that we choose to maximize is $f(\overleftarrow{d}|S_i)$ which is defined as:

$$f(\overleftarrow{d}|S_i) = \sigma(Z_{\overleftarrow{d}}.Z_{S_i}) \tag{3}$$

Here, Z_p is the topic vector defined as $\frac{1}{N_p}\sum_n z_{p,n}$ where n ranges over all the tokens in the bag p and N_p denotes the

cardinality of p. σ represents the sigmoid function and $'.'$ represents the dot product between two vectors. Intuitively, $f(\overleftarrow{d}|S_i)$ represents the topical similarity between cited document and its corresponding context.

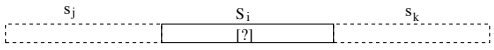


Figure 2: An illustrative citation context window

Next we describe our dynamic context selection procedure. We allow our window to grow over sentences beginning with the sentence that has the citation mention, although the method proposed is general enough to be applicable to any building block such as words or paragraphs. We choose to begin with the sentence that contains the citation mention as the sentence carries most of the information about the cited document. We let S_i denote the current context window and s_j and s_k are the next left and right candidates to either include in the window or to let the growth stop in either direction. We update the window as defined below and continue to grow in the direction which maximizes the objective function 3.

$$S_i = max\{f(\overleftarrow{d}|S_i), f(\overleftarrow{d}|\{S_i, s_j\}), f(\overleftarrow{d}|\{S_i, s_k\}), f(\overleftarrow{d}|\{S_i, s_j, s_k\})\} \quad (4)$$

3.4 Inference using Gibbs Sampling

We utilize Gibbs sampling as a tool to approximate the posterior distribution for both the models. Specifically, we want to estimate θ, ϕ and φ parameters of the multinomial distributions $Multi(.|\theta)$, $Multi(.|\phi)$ and $Multi(.|\varphi)$, respectively, in Fig. 1(b) and 1(c).

According to Eq. 2, the joint probability distribution of the latent and the observed variables can be factorized as follows:

$$\begin{aligned} &p(\mathbf{x}, \mathbf{c}, \mathbf{z}, \mathbf{w}|\mathbf{a_d}, \alpha_\theta, \alpha_\phi, \alpha_\varphi) \\ &= p(\mathbf{w}|\mathbf{z}, \alpha_\phi)p(\mathbf{c}|\mathbf{z}, \alpha_\varphi)p(\mathbf{z}|\mathbf{x}, \mathbf{a_d}, \alpha_\theta)p(\mathbf{x}|\mathbf{a_d}) \end{aligned} \quad (5)$$

To generalize the notations, let $n_a^{(b)}$ denote the number of times entity b is observed with entity a. Particularly, if an observation of topic z is made with author x, then $n_x^{(z)}$ denotes the number of times this observation is made in the whole corpus. Similarly, we define $n_z^{(t)}$, $n_z^{(c)}$ where t and c stand for term and cited author respectively. Here, we derive $p(\mathbf{c}|\mathbf{z}, \alpha_\varphi)$. Other factors can be obtained in a similar fashion. The target posterior distribution for cited author generation, i.e., $p(\mathbf{c}|\mathbf{z}, \alpha_\varphi)$, can be obtained by integrating over all possible values of φ:

$$\begin{aligned} &p(\mathbf{c}|\mathbf{z}, \alpha_\varphi) \\ &= \int \prod_{k=1}^{K} \frac{1}{\Delta(\alpha_\varphi)} \prod_{c=1}^{A} \varphi_{z,c}^{n_z^{(c)}+\alpha_\varphi^c-1} d\varphi_z = \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}\varphi} + \alpha_\varphi)}{\Delta(\alpha_\varphi)} \end{aligned} \quad (6)$$

where $\Delta(\alpha_\varphi) = \frac{\prod_{i=1}^{dim(\alpha_\varphi)} \Gamma(\alpha_\varphi^i)}{\Gamma(\sum_{i=1}^{dim(\alpha_\varphi)} \alpha_\varphi^i)}$ and $\mathbf{n}_{\mathbf{z}\varphi} = \{n_z^{(c)}\}_{c=1}^{A}$

With the likely treatment to other factors, the joint distribution can be written as:

$$\begin{aligned} &p(\mathbf{x}, \mathbf{w}, \mathbf{c}, \mathbf{z}|\alpha_\theta, \alpha_\phi, \alpha_\varphi) \\ &= \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}\phi} + \alpha_\phi)}{\Delta(\alpha_\phi)} \prod_{z=1}^{K} \frac{\Delta(\mathbf{n}_{\mathbf{z}\varphi} + \alpha_\varphi)}{\Delta(\alpha_\varphi)} \prod_{x=1}^{A} \frac{\Delta(\mathbf{n}_{\mathbf{x}} + \alpha_\theta)}{\Delta(\alpha_\theta)} \end{aligned} \quad (7)$$

where $\mathbf{n}_{\mathbf{z}\phi} = \{n_z^{(t)}\}_{t=1}^{V}$ and $\mathbf{n}_{\mathbf{x}} = \{n_x^{(z)}\}_{z=1}^{K}$

Starting with a random assignment of topics $\mathbf{z}$ and authors $\mathbf{x}$ from the list of co-authors in a document, Gibbs sampler iterates through each word and cited authors in a document, for all the documents in the corpus. For the ALT model, we need to sample topic assignment for each word variable and cited author variable. Since we have two unobserved random variables x and z for both types of assignments, our Gibbs sampler performs blocked sampling on these two random variables. We draw a sample from $p(z_i = k, x_i = x|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w})$ for the word variable and from $p(z_i = k, x_i = x|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{c})$ for the cited author variable. The subscript $-i$ indicates that we leave the i^{th} token out from the otherwise complete assignment. After algebraic manipulation to Eq. 7, we arrive at the sampling equations as given in Eq. (i & ii) in Table 1.

Unlike the ALT model, Author Cite Topic (ACT) model has one additional unobserved random variable c that appears inside the citation context of a given citation in any document. We initialize c from the co-authors of the cited documents by uniformly selecting one author. The remaining initializations remains the same as above. We block x, z and c while sampling and for each word in the citation context, we sample from the conditional distribution, i.e. $p(z_i = k, x_i = x, c_i = c|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{c}_{-i}, \mathbf{w})$. The algebraic form of the conditional distribution is given in Eq.(iii) in Table 1.

4 Experiments

We describe our data set and experimental settings below and, in § 4.2 and § 4.3, we provide the details of evaluation tasks with corresponding results.

4.1 Data Sets and Experimental Settings

We use two different subsets of scientific documents for our evaluation purpose. For the first dataset (referred as *CiteSeer-DS1*), we use publicly available [1] subset of the CiteSeer [2] 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. The dataset contains 4699 unique authors[3] where 1511 authors are cited. After standard preprocessing of removing stop words, we supplement the data set with the context information for each citation.

We employ CiteSeer-DS1 because various previous studies [Nallapati *et al*., 2008], [Chang and Blei, 2009] have used the dataset for link prediction task, however CiteSeer-DS1 is a hand-picked dataset prepared for document classification purposes [Lu and Getoor, 2003]. For both qualitative and quantitative evaluations on a user selected scientific documents dataset in a collaborative setting, we also acquired dataset from CiteULike [4] for over 2 years from November 2005 to January 2008 (referred as *CiteSeer-DS2*). The dataset is available at http://citeulike.org. Overall, there are 33,456 distinct

[1] http://www.cs.umd.edu/šen/lbc-proj/LBC.html

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

[3] we use disambiguated authors for each documents available at http://CiteSeerx.ist.psu.edu/about/metadata

[4] http://citeulike.org

$$p(z_i = k, x_i = x|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}) \propto \frac{n^{(t)}_{k,-i}+\alpha^t_\phi}{\sum_{t=1}^{V} n^{(t)}_{k,-i}+V.\alpha^t_\phi} \cdot \frac{n^{(k)}_{x,-i}+\alpha^k_\theta}{\sum_{k=1}^{K} n^{(k)}_{x,-i}+K.\alpha^k_\theta} \quad \text{(i)}$$

$$p(z_i = k, x_i = x|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{c}) \propto \frac{n^{(c)}_{k,-i}+\alpha^c_\varphi}{\sum_{c=1}^{C} n^{(c)}_{k,-i}+C.\alpha^c_\varphi} \cdot \frac{n^{(k)}_{x,-i}+\alpha^k_\theta}{\sum_{k=1}^{K} n^{(k)}_{x,-i}+K.\alpha^k_\theta} \quad \text{(ii)}$$

$$p(z_i = k, x_i = x, c_i = c|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{c}_{-i}, \mathbf{w}) \propto \frac{n^{(t)}_{k,-i}+\alpha^t_\phi}{\sum_{t=1}^{V} n^{(t)}_{k,-i}+V.\alpha^t_\phi} \cdot \frac{n^{(c)}_{k,-i}+\alpha^c_\varphi}{\sum_{c=1}^{C} n^{(c)}_{k,-i}+C.\alpha^c_\varphi} \cdot \frac{n^{(k)}_{x,-i}+\alpha^k_\theta}{\sum_{k=1}^{K} n^{(k)}_{x,-i}+K.\alpha^k_\theta} \quad \text{(iii)}$$

Table 1: Gibbs updates for ALT(i,ii), ACT(i,iii)

papers in CiteULike sample. We map the document ids of CiteULike documents to document ids of CiteSeer documents [5] to gain access to citation network of the sample. The resultant CiteSeer-DS2 contains 18354 documents in which 9571 documents are cited. There are a total of 29645 unique authors in CiteSeer-DS2 out of which 15967 authors are cited at least once. We follow the same preprocessing step as the CiteSeer-DS1 dataset.

Experimental Set-up: We choose to fix the hyper-parameters and evaluate different models with the same setting. We set the hyper-parameters to the following values [Rosen-Zvi *et al*., 2004]: $\alpha_\theta = 50/T$, $\alpha_\phi = 0.01$, $\alpha_\varphi = 0.01$. We run 1000 iterations of Gibbs sampling for training and extend the chain with 100 iterations over test set. For dynamic window selection, we collect 10 samples from the chain after every 10 iterations starting from 1000 iterations, and compute the new window with the average of the samples using Eq. 4. After the window update, we let the chain converge and start to update the window again. Starting with the sentence that contains the citation mention, we allow our window to grow up to a maximum of 5 sentences in either direction. The multinomial parameters of the model are calculated by taking expectations of the corresponding counts from 10 samples collected during test iterations.

4.2 Model Evaluation on Unseen Content

This task quantitatively estimates the generalization capabilities of a given model on unseen data. In particular, we compute the *perplexity* on the held-out test set. We run the inference algorithm exclusively on the unseen words in the test set of documents, same as [Rosen-Zvi *et al*., 2004], to obtain the log-likelihood of test documents. Before extending the Gibbs sampling chain and *sweeping* the test set, we first initialize the topic assignment to authors and unseen words randomly and run the Gibbs iteration on the test set with following Gibbs updates:

$$p(z^u_i, x^u_i|w^u_i = t, \mathbf{z^u_{-i}}, \mathbf{w^u_{-i}}, \mathbf{x^u_{-i}}) = \frac{n^{(t)}_{k,-i}+\beta}{\sum_{t=1}^{V} n^{(t)}_{k,-i}+V.\beta} \cdot \frac{n^{(k)}_{x^u,-i}+\alpha^k_\theta}{\sum_{k=1}^{K} n^{(k)}_{x^u,-i}+K.\alpha^k_\theta} \quad (8)$$

Superscript $(.^u)$ stands for any unseen element. The sampling updates in Eq. 8 can be used to calculate the model parameters, $\Pi = (\theta, \phi, \varphi)$ for unseen documents as:

$$\theta_{x^u,k} = \frac{n^{(k)}_{x^u}+\alpha^k_\theta}{\sum_{k=1}^{K} n^{(k)}_{x^u}+K.\alpha^k_\theta}; \phi_{k,t} = \frac{n^{(t)}_k+\beta_t}{\sum_{t=1}^{V} n^{(t)}_k+\beta_t} \quad (9)$$

The predictive log-likelihood of a text document in the test set, given the model $\Pi = (\theta, \phi, \varphi)$, can be directly expressed as a function of the multinomial parameters:

$$p(\mathbf{w}|\mathbf{\Pi}) = \prod_{n=1}^{N_x} \sum_{k=1}^{T} \Big(\frac{1}{|a_d|} \sum_{x \in a_d} p(w_n|z_n = k).p(z_n = k|d = x) \Big)$$
$$= \prod_{n=1}^{N_x} \Big(\frac{1}{|a_d|} \sum_{k,x \in a_d} \phi_{k,t}\theta_{x,k} \Big) \quad (10)$$

Next, we compute the perplexity as defined below. Here, N_w is the total number of word occurrences in the test set.

$$Perplexity(\mathbf{w}) = exp\Big(\frac{-log\ p(\mathbf{w})}{N_w}\Big) \quad (11)$$

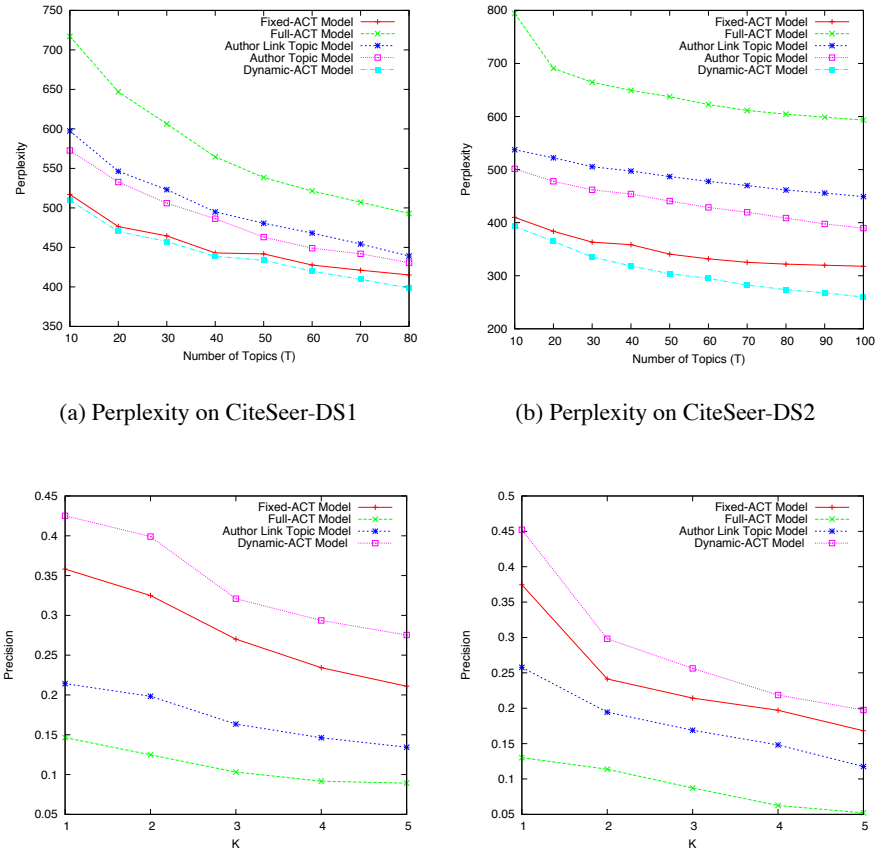
Baselines: We use following two baselines from [Rosen-Zvi *et al*., 2004] and [Tu *et al*., 2010], namely Author Topic Model (ATM) and Citation Author Topic Model (CAT) respectively. Since ATM does not learn from links among documents, comparison with ATM signifies the importance of learning from links along with the text of the documents. CAT model treats all the content of a citing document as context for any cited document within, therefore, comparison with CAT highlights the importance of choosing a context window surrounding the citation mention. We compare these baselines against the proposed Author Link model (ALT), fixed length window Author Cite Topic Model (Fixed-ACT) [6] and dynamically selected window based ACT model (Dynamic-ACT). For our experiments, the training data consists of 4 splits with 75% documents (training docs) along with the 25% words of the remaining 25% of the documents (test docs). The rest 75% words in test documents are used to calculate log-likelihood. The average value over the 4 splits are reported in the experiments.

Fig. 3 (a)&(b) show the comparison of perplexity on test set of CiteSeer-DS1 and CiteSeer-DS2, respectively. The ATM model performs slightly better than the ALT model. We believe that this is because the links considered separately from the content actually deteriorate the prediction capability of the models over words. In contrast, while training, links along with the content help to learn the topics better. However, when all the content is treated as context for every cited article [Tu *et al*., 2010] in a given citing document, the performance deteriorates significantly. Therefore, we argue that a wise selection of context window is essential when a context sensitive topic modeling approach is considered.

Dynamic-ACT outperforms all the other approaches (see Fig. 3 (a)&(b)). During our experiments, we observed that the length of a relatively large fraction of citation contexts

[5] mapping is obtained from http://citeulike.org

[6] we set the radius to be 10 words from the citation mention after stop word removal, i.e., 20 words window



(a) Perplexity on CiteSeer-DS1

(b) Perplexity on CiteSeer-DS2

(c) P@K on CiteSeer-DS1

(d) P@K on CiteSeer-DS2

Figure 3: Experimental results for (a) Perplexity on CiteSeer datasets DS1, (b) Perplexity on CiteSeer datasets DS2, (c) Precision @ K for cited author prediction on CiteSeer datasets DS1 and (d) Precision @ K for cited author prediction on CiteSeer datasets DS2

was limited to a single sentence that contains the citation mention. The fraction decreases as we increase the number of topics. Specifically, for CiteSeer-DS1, 78% of the total citation contexts were composed of only one sentence when we set the number of topics to 10. This number drops to 65% with 100 topics. Also, we found the average window length on CiteSeer-DS1 to be 1.4 with 10 topics and 1.6 with 100 number of topics. We observe the similar trend with CiteSeer-DS1 where 81% of the total total citation contexts were composed of only one sentence with topic count 10 whereas the number decreases to 62% with 100 topics. Considering that the topic assignment to words is fine grained with a large number of topics, the growth outside the window is more likely to explain the finer details mentioned in the cited document.

4.3 Cited Author Prediction

In this task, we evaluate the capability of the models to predict the authors that this document links to. That is, given the text of a test document, which authors' work should this document cite to? The experimental design for this task is very similar to the one in the previous subsection. We again perform the Gibbs update following the sampling from conditional distribution in Eq. 8 and calculate the model parameters. With the model parameters for the ALT and ACT models, the probability $p(c|\mathbf{w_d})$, where c is the author to be cited given a document $\mathbf{w_d}$ is:

$$p(c|\mathbf{w_d}) = \sum_z p(c|z) \int_{x \in a_d} \frac{1}{|a_d|} p(z|\theta_x) d\theta_x \propto \sum_k \frac{1}{|a_d|} \varphi_{c,k}.\theta_{d,k} \tag{12}$$

Baselines: Because the ATM does not model the links, it is not possible to treat ATM as a baseline for this task. We keep all the other four comparisons intact for this task. The training data consists of 4 splits with 75% documents and their outgoing links to cited authors (training docs) and the 25% outgoing links of the remaining 25% of the documents (test docs). The rest of 75% outgoing links in the test documents are used for this task. We set the number of topic to be 100 for this task. We use Precision@K as the evaluation metric. The average value over the 4 splits are reported in the experiments.

Topic-45						Topic-71					
Top Words		Top interested authors		Top influential authors		Top Words		Top interested authors		Top influential authors	
scale	0.01663	k. mikolajczyk	0.95714	c. schmid	0.08392	retrieval	0.03634	s.-fu chang	0.04901	j. r. smith	0.05583
shape	0.01434	j. ponce	0.95641	j. malik	0.07407	images	0.01635	s. mehrotra	0.04856	t. s. huang	0.04141
object	0.01385	t. lindeberg	0.95619	d. g. lowe	0.06075	texture	0.01572	r. paget	0.04451	y. rui	0.03214
images	0.01069	s. lazebnik	0.95412	s. belongie	0.04564	color	0.01184	j. z. wang	0.04399	r. jain	0.03025
matching	0.01000	r. fergus	0.86715	k. mikolajczyk	0.03996	features	0.01016	m. ortega	0.04214	a. efros	0.02561
recognition	0.00846	a. c. berg	0.86306	j. puzicha	0.03863	content	0.00958	p. harrison	0.04118	t. leung	0.02385
features	0.00772	g. loy	0.85624	j. shi	0.02436	search	0.00763	g. wiederhold	0.04098	w.-ying	0.01933
local	0.00751	e. rosten	0.04269	d. p. huttenlocher	0.01704	visual	0.00754	r. peteri	0.04058	j. malik	0.01732
Topic-6						Topic-97					
Top Words		Top interested authors		Top influential authors		Top Words		Top interested authors		Top influential authors	
learning	0.02424	xiaoli li	0.04352	t. mitchell	0.09649	model	0.01829	t. l. griffiths	0.04683	d. j.c. mackay	0.07512
classification	0.02189	k. nigam	0.04203	k. nigam	0.08227	data	0.01164	m. j. beal	0.04588	z. ghahramani	0.06245
text	0.01635	t. mitchell	0.04146	a. mccallum	0.05819	learning	0.00817	z. ghahramani	0.04376	g. e. hinton	0.04727
training	0.01420	a. mccallum	0.04076	a. blum	0.05808	bayesian	0.00791	b. j. frey	0.04345	l. r. rabiner	0.03903
unlabeled	0.01351	yang dai	0.04031	d. d. lewis	0.04469	mixture	0.00773	d. m. blei	0.04263	t. hofmann	0.03840
examples	0.01150	andrew ng	0.03843	s. thrun	0.03260	inference	0.00689	d. j.c. mackay	0.04158	c. e. rasmussen	0.03226
set	0.00913	r. gilleron	0.03619	ken lang	0.02693	distribution	0.00657	r. m. neal	0.04147	r. m. neal	0.02999
Topic-46						Topic-61					
Top Words		Top interested authors		Top influential authors		Top Words		Top interested authors		Top influential authors	
algorithms	0.01376	e. zitzler	0.04734	d. e. goldberg	0.06921	matrix	0.01443	luh yen	0.04252	yair weiss	0.08598
quantum	0.01087	k. deb	0.04655	k. deb	0.06606	algorithms	0.01137	heinrich voss	0.04194	j. malik	0.07566
genetic	0.01043	k. sastry	0.04552	p. j. fleming	0.05680	spectral	0.01003	w. freeman	0.04068	andrew y. ng	0.06003
optimization	0.00847	t. goel	0.04523	c. m. fonseca	0.04943	graph	0.00970	d. verma	0.03988	m. i. jordan	0.03339
objective	0.00792	l. thiele	0.04520	n. srinivas	0.04930	segmentation	0.00670	s. t. roweis	0.03973	m. belkin	0.02860
pareto	0.00713	l. barbulescu	0.04503	k. l. clarkson	0.03059	embedding	0.00667	m. saerens	0.03915	p. niyogi	0.02394
population	0.00708	d. aharonov	0.04448	l. k. grover	0.02802	eigenvectors	0.00647	b. d. packer	0.03896	s. vempala	0.02352
evolutionary	0.00658	k. svozil	0.04429	j. horn	0.02654	cut	0.00614	a. goldberg	0.03886	r. kannan	0.02279

Table 2: Top words, interested authors and influential authors for 6 topics in CiteSeer-DS2

To evaluate the prediction accuracy for the proposed models, we first label the actual authors that are cited by a test document as its relevant result set. We rank the authors in the train corpus against each test document using $p(c|\mathbf{w_d})$ and compare the models based upon the precision of the retrieved results. Fig. 3(c) & (d) shows the results for the three methods on CiteSeer-DS1 and CiteSeer-DS2, respectively.

4.4 Anecdotal Evidences

Table 2 shows the most likely words, interested and influential authors in 6 topics from the CiteSeer-DS2 dataset obtained using the ACT model (e.g. Griffith, Beal, etc., as interested authors and Mackay, Ghahramani and Hinton as influential authors in Bayesian learning). For each topic shown in the table, most influential authors are well known in their respective areas and their authored papers gets cited in the respective fields.

5 Conclusion

We propose novel models for author-author linkage conditioned on topics latent in the content of the documents. We exploit the citations between documents to infer influence of certain authors over topics. We also propose context sensitive extensions of the proposed model that incorporates the context of the cited document and how it infers the topic of both cited and citing authors with better quality.

6 Acknowledgement

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

References

[Balog *et al.*, 2009] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Inf. Process. Manage.*, 45(1):1–19, 2009.

[Bhattacharya and Getoor, 2006] I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. In *SDM*, 2006.

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

[Chang and Blei, 2009] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, 2009.

[Dietz *et al.*, 2007] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML*, pages 233–240, 2007.

[Erosheva *et al.*, 2004] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. In *Proceedings of the National Academy of Sciences*, 2004.

[Fang and Zhai, 2007] H. Fang and C. Zhai. Probabilistic models for expert finding. In *ECIR*, pages 418–430, 2007.

[Gruber *et al.*, 2008] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Latent topic models for hypertext. In *UAI*, pages 230–239, 2008.

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

[Kataria *et al.*, 2010] S. Kataria, P. Mitra, and S. Bhatia. Utilizing context in generative bayesian models for linked corpus. In *AAAI*, 2010.

[Liu *et al.*, 2009] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: joint models of topic and author community. In *ICML*, pages 665–672, 2009.

[Lu and Getoor, 2003] Q. Lu and L. Getoor. Link-based classification. In *ICML*, pages 496–503, 2003.

[McCallum *et al.*, 2007] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res. (JAIR)*, 30:249–272, 2007.

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

[Newman *et al.*, 2006] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD*, pages 680–686, 2006.

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

[Shiozaki *et al.*, 2008] H. Shiozaki, K. Eguchi, and T. Ohkawa. Entity network prediction using multitype topic models. *IEICE - Trans. Inf. Syst.*, E91-D(11):2589–2598, 2008.

[Tu *et al.*, 2010] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier. Citation author topic model in expert search. In *COLING*, 2010.

[Wang *et al.*, 2005] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and their attributes. In *NIPS*, 2005.