Generative Models for Authors’ Influence in Document Networks

Saurabh Kataria, Prasenjit Mitra and C. Lee Giles
College of Information Science & Technology
Pennsylvania State University, University Park, PA 16802
Email: {skataria, pmitra, giles}@ist.psu.edu

Abstract

In a document network such as citation network of scientific documents, web-logs, etc., the content produced by authors exhibit their interest in certain topics whereas some authors tend to 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 a 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 remain an open problem. We propose a novel document generation schemes that incorporates the context while modeling the interests of citing authors and influence of the cited authors simultaneously. We apply the proposed models to two text corpora: CiteSeer and CiteULike dataset. The experiments based upon log-likelihood fit on the test documents suggest significant improvements over the baseline models.

1 Introduction

Modeling the interest of authors given a corpus of documents has been studied to answer important queries about the authors such as who produces similar work [19], who belongs to the same research community [13], etc. These queries form the basis of several information retrieval and machine learning tasks such as expert search, community detection, etc. Recently, several generative models of document corpus have begun exploring latent structures such as topics, present in the documents. Probabilistic approaches such as Latent Dirichlet Allocation(LDA) [3] and Probabilistic Latent Semantic Analysis(PLSA) [10] model the co-occurrence patterns present in text and identify a probabilistic membership of words and documents in a lower dimensional space. Rosen-Zvi, et al., [19] extended these approaches to answer queries related to authors’ topic of interests. However, these approaches are unable to answer another fundamental question about the attribution of topics to authors: 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.

The citations present in any given document corpus, e.g., scholarly articles, weblogs etc. establishes an authority of the document over the content. The quantitative estimates of this authority has successfully been utilized in web search and information retrieval tasks [12, 4]. Researchers have recently begun to explore the probabilistic modeling of content and links to quantify the influence of a document across latent topics in a given corpus. Dietz, et al., [6], Nallapati, et al., [16] and Cheng, et al., [5] model content and citations of a document together to obtain the influences of the citations over the new content corresponding to a given topic. However these statistical methods for parameterizing the influence of a document cannot be easily extended to identify the influence of the authors because one document often has multiple authors.

In this work, we present generative models for a linked corpus of documents that simultaneously model the content of the document, interests of the authors and influence of the authors on topics. In
Author Topic Model (ATM) [19], the documents are modeled as a mixture of topics, as in LDA [3], and the weights of the mixture are determined by the authors of the document. In order to capture the influence of cited authors, we extend the ATM to let the cited set of authors in the document be represented as a mixture of topics and again the weight of the topics are determined by the authors of the document. Moreover, we hypothesize that the context in which a cited document appears in the citing document provides extra information about the topic of the cited authors. The citation context does not necessarily portray the entire content of the cited document, but, provides a description of some attribute of the cited document from the citing author’s perspective in relation to the citing document’s topic. Therefore, the topics of the words in the citation context help distinguish the topics related to which an authors’ work is cited and consequently a better quantitative estimate of the authors’ influence on the topics can be obtained.

The contributions are as follows: (1) We propose generative models for author-author linkage in an interlinked corpus of documents conditioned on the topic of interest to authors. The proposed models are able to distinguish between authors’ interest and authors’ influence on the topics. (2) 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.

2 Related work

Since topic models such as LDA [3] have achieved considerable success in discovering underlying latent structure in the textual documents, there is a growing interest in developing better topic models by utilizing additional information present along with the text. One such stream of efforts focus on author-centric topic models [19] where the authors and the content are simultaneously modeled with coupling of hyper-parameters for interest of authors and topics of text with latent topic variable. Bhattacharya, et al., [2] extended the ATM with a noise model to disambiguate incomplete or unresolved references to authors. Another stream of author-centric modeling deals with expert finding [7, 1] where an expert is defined as a person knowledgeable in the field. This corresponds to “interest” factor in our modeling, however our main goal is to model the influence 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. Mixed membership model [21], also referred as linked-LDA [16], 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 [16, 9, 5], influence propagation [6], community of links detection [13], context-sensitive citation and text modeling [11]. Another stream of topic modeling focuses on social networks of entities where entity-entity relationships conditioned upon topics are explored. McCallum, et al., [15] 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., [17, 20] proposed another entity-entity relationship model for named-entities in news articles where documents are modeled as mixture of topics over both entities and words.

3 Models

Notations: Let \( V \), \( D \), \( A \), \( a_d \), \( d_c \) and \( N_d \) denote the size of the word vocabulary, the number of documents, the number of authors, set of authors for a document \( d \), a document cited by \( d \) and number of words in document \( d \) respectively. Let \( T \) denote the number of topics and suppose there exist a \( T \times V \) topic-word distribution matrix \( \phi \) that indexes a probabilistic distribution over words given the topic and a \( T \times A \) topic-author distribution matrix \( \theta \) that indexes the probability with which an author shows interest in a topic. Our definition of citation context for a cited document is similar to the one adopted in Kataria et al., [11]. We define citation context for a cited document as a bag of words that contains a fixed number of words appearing before and after the citation’s mention in a citing document. In case a cited document is mentioned multiple times in citing document, we assimilate all the corresponding context words.
Citations among documents exhibit the biases of citing authors towards certain influential authors who has key contributions in the topic of discourse. We quantify the influence of an author given a topic by the probability that the author’s work gets cited when there is a mention of the topic in a citing document. Since the Author Topic Model 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 the ATM, a set of authors $a_d$ decides to write a document and to generate each word, an author $x$ is chosen uniformly randomly from $a_d$, and a topic is sampled from the chosen author specific distribution, and the word is generated from the 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 $a_d$ would choose to reference those authors’s work that are influential in that topic. Fig. 1(a) shows the plate diagram for the model for Author Link Topic Model. Formally, given the model parameters $\alpha$, $\phi$, and $\omega$, the joint distribution of an author $x$, 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 documents in the document $d$.

$$p(x,c,w|a_d, \alpha, \phi, \omega) = \int \int \prod_{n=1}^{N_d} p(x|a_d)p(z_n|x, \theta_x)p(w_n|z, \phi_{z_n})p(\theta_x|\alpha_x)p(\phi_{z_n}|\alpha_{\phi}) \times$$

$$\prod_{l=1}^{L_d} p(x|a_d)p(z|x, \theta_x)p(c|z_l, \varphi_{z_l})p(\theta_x|\alpha_x)p(\varphi_{z_l}|\alpha_{\varphi})d\theta_x d\phi_x d\varphi$$

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 with ALT model, 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(b)). With this independence assumption, the topic sampled for words in the citation context window generates both word and a reference to the cited author. Note that the variable $c$ is hidden for the ACT model since, for each citation, we can observe only one author from the set of co-authors of the cited document. The parameters of the ACT model remain same as that of ALT model, however the complete data log-likelihood function is different due to a difference in generation process. The log-likelihood function to optimize can be written as follows:

Figure 1: Plate diagrams for (a) Author Link Topic Model (ALT), (b) Author Cite Topic Model (ACT) and (c) Switch Author Cite Topic Model (SACT)
\[
p(x, c, z, w | a_d, \alpha_x, \alpha_y, \alpha_z) = \int \int \prod_{n=1}^{N_d-C_d} p(x | a_d)p(z_n | x, \theta_x)p(w_n | z, \phi_{z_n})p(\theta_{z_n} | \alpha_y)p(\phi_{z_n} | \alpha_y) \times (1)
\]

\[
\prod_{n=1}^{C_d} (p(z_n | x, \theta_x)p(w_n | z, \phi_{z_n})p(\phi_{z_n} | \alpha_y)p(\phi_{z_n} | \alpha_y))d\theta d\phi \quad 2
\]

\(C_d\) is the total length (number of words) of all citations contexts in the document \(d\). Intuitively, Eq. 1 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 \(p(z_n | x, \theta_x), p(w_n | z, \phi_{z_n})\) 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.

### 3.2.1 Relaxing the Independence Assumption with Model Switching

The independence assumption in Author Cite Topic Model can be restrictive in cases where certain words are attributed to the authors of the cited document and these words appear in the context only because of these documents. In other words, if there exists a strong correlation between the word in the citation context and the citation the independence assumption break down. For example, the words such as “HITS” and “Pagerank” are attributed to the authors of corresponding documents, i.e., Jon Kleinberg [12] and Sergey Brin and Lawrence Page [4] respectively. To relax the independence assumption, we propose Switch Cite Author Topic Model.

**The Switch Author Cite Topic Model:** Switch Author Cite Topic Model (SACT) relaxes the restrictive independence assumption by providing the model with the capability of switching between parameters. The intuition behind this model is that if certain words appear too often around the citation mention, then with high probability these words belong to the topic of cited authors. Therefore, we choose to throw a coin in the context window to decide whether the topic be sampled from cited author’s influence vector or the cited author’s interest vector. Eq. 2 shows the complete log-likelihood of the model. Here, \(a_c\) denotes the cited set of authors for the cited document \(d_c\).

\[
P(x, c, z_c, w | \lambda, a_c, \alpha_x, \alpha_y, \alpha_z, \alpha_{\lambda}) = \prod_{n=1}^{N_d-C_d} p(x | a_d)p(z_n | x, \theta_x)p(w_n | z, \phi_{z_n})p(\theta_{z_n} | \alpha_y)p(\phi_{z_n} | \alpha_y) \times
\]

\[
\prod_{n=1}^{C_d} (p(z_n | x, \theta_x)p(w_n | z, \phi_{z_n})p(\phi_{z_n} | \alpha_y)p(\phi_{z_n} | \alpha_y))
\]

### 3.3 Inference using Gibbs Sampling

As in basic LDA [3], the computation of the posterior distribution of the hidden variables \(\theta\) and \(z\) is intractable. Therefore, we need to utilize approximate methods e.g. variational methods [16] or sampling techniques [8] 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 [18], we utilize Gibbs sampling as a tool to approximate the posterior distribution for both the models.

Gibbs sampler for all the three models, i.e., ALT, ACT and SACT need to sample \(p(z_i = k, x_i = x | z_{-i}, w, x_{-i})\) for words in the document that are not in citation context. Since ALT is context insensitive, all the words in the document are sampled by Eq.(i) in Table 1. ALT model samples topic for each cited author in the reference list of document using Eq.(ii) in Table 1. ACT model needs to sample \(p(z_i = k, x_i = x, c_i = c | z_{-i}, x_{-i}, c_{-i}, w)\) for the words that are in any citation context in the document and algebraic form of the conditional probability is shown in the Eq. (iii) in Table 1. For SACT, we first need to draw a coin for each context word and then, based upon the outcome of the coin, we draw topic for the word in the citation context. Eq. (iv & v) show the conditional distribution for the coin’s draw and Eq. (vi & vii) show the conditional distribution for the topic draw for each word. We omit the derivation of the sampling equation due to lack of space.
following values:  
\[ \alpha \]

however, we choose to fix the parameter and evaluate different models. We set the hyper-parameters

\[ 4 \] Experiments

entity and list only the algebraic form of these equations. Here, in generalization, \( n_{i}(b) \) stands for counts of

entity \( b \) with entity \( a \).

4 Experiments

We describe our data set and experimental setting below and, in § 4.2, we provide the details of evaluations

task 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 \(^1\) labeled 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. For each citation, we add 20 words in the radius of 10 words originating at the citation mention in the document. We fix the radius for the experiments reported in this paper.

We employ CiteSeer-DS1 because of various previous studies \(^{[16], [5]}\) have used the dataset for link prediction task, however CiteSeer-DS1 is a hand-picked dataset prepared for document classification purposes \(^{[14]}\). For both qualitative and quantitative evaluations on a user selected scientific citation network of the sample. The resultant CiteSeer-DS2 contains 18354 documents in which 1511 authors are cited. After standard preprocessing of removing stop words, we supplement the data set with the context information for each citation. For each citation, we add 20 words in the radius of 10 words originating at the citation mention in the document. We fix the radius for the experiments reported in this paper.

\(^1\)http://www.cs.umd.edu/ sen/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

\(^5\)mapping is obtained from http://citeulike.org

\[ p(z_{i} = k, x_{i} = x|z_{-i}, x_{-i}, w) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k}}{\sum_{k=1}^{K} n_{i}(k)} \prod_{c} \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k}}{\sum_{k=1}^{K} n_{i}(k)} \]

(i)

\[ p(z_{i} = k, x_{i} = x, c_{i} = c|z_{-i}, x_{-i}, c_{-i}) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k}}{\sum_{k=1}^{K} n_{i}(k)} \prod_{c} \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k}}{\sum_{k=1}^{K} n_{i}(k)} \]

(ii)

\[ p(z_{i} = k, x_{i} = x, c_{i} = c|z_{-i}, x_{-i}, c_{-i}, w) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k}}{\sum_{k=1}^{K} n_{i}(k)} \prod_{c} \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k}}{\sum_{k=1}^{K} n_{i}(k)} \]

(iii)

\[ p(s_{i} = 0|c, s_{-i}) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1}{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1} \]

(iv)

\[ p(s_{i} = 1|c, s_{-i}) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1}{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1} \]

(v)

\[ p(z_{i} = k|z_{-i}, x, c, w, s_{i} = 0) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1}{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1} \]

(vi)

\[ p(z_{i} = k|z_{-i}, x, c, w, s_{i} = 1) \propto \frac{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1}{\sum_{k=1}^{K} n_{i}(k) + \alpha_{k} + \alpha_{\lambda_{k}} - 1} \]

(vii)

Table 1: Gibbs updates for ALT(i & ii), ACT(i & iii) and SACT (i, iv, v, vi & vii)
model are calculated by taking expectations of the corresponding counts from 10 samples collected during test iterations.

4.2 Log-likelihood 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 [19] where the inference algorithm is run exclusively on the unseen words in test set of 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 test set only.

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:

\[
p(w|\Pi) = \sum_{n=1}^{N_d} \prod_{k=1}^{K} \left( \frac{1}{|\{d\}|} \sum_{x \in a_d} p(w_n|z_n = k).p(z_n = k|d = x) \right) = \prod_{n=1}^{N_d} \prod_{k=1}^{K} \phi_{k,n} \theta_{k,x,n} \tag{3}
\]

We treat author topic model as a baseline for this task. Fig. 2 shows the comparison of log-likelihood on test set of CiteSeer-DS1 and CiteSeer-DS2 respectively. ACT outperforms the context insensitive ALT which, in turn, outperforms ATM. We believe that the better performance of ALT than ATM is due to the fact that the author-author relationship present in the citation helps better reveal the interest of authors whereas the topic assignment to contextual words helps further in improving the interest attribution to authors.

Table 2 shows the most likely words, interested and influential authors in 6 topics from the CiteSeer-DS2 dataset obtained using ACT model (e.g. griffith, beal, etc., as interested authors whereas mackay, ghahramani and hinton as influential authors in Bayesian learning).

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 the topics. We also propose context sensitive extensions of the proposed model that incorporates the context of the cited document to infer topic of both cited and citing authors with better quality.

References


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

<table>
<thead>
<tr>
<th>Topic-97</th>
<th>Topic-98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Words</td>
<td>Top interested authors</td>
</tr>
<tr>
<td>scale</td>
<td>0.01434</td>
</tr>
<tr>
<td>object</td>
<td>0.01385</td>
</tr>
<tr>
<td>images</td>
<td>0.01065</td>
</tr>
<tr>
<td>matching</td>
<td>0.01055</td>
</tr>
<tr>
<td>recognition</td>
<td>0.00846</td>
</tr>
<tr>
<td>features</td>
<td>0.00772</td>
</tr>
<tr>
<td>local</td>
<td>0.00751</td>
</tr>
</tbody>
</table>