Knowledge Discovery from Citation Networks

Zhen Guo, Zhongfei (Mark) Zhang
Computer Science Department
SUNY at Binghamton, Binghamton, NY, 13902
{*zguo,zhongfei*}*@cs.binghamton.edu*

Shenghuo Zhu, Yun Chi, Yihong Gong
NEC Laboratories America, Inc.
10080 N. Wolfe Rd. SW3-350, Cupertino, CA 95014
{*zsh,ychi,ygong*}*@sv.nec-labs.com*

***Abstract*—Knowledge discovery from scientific articles has received increasing attentions recently since huge repositories are made available by the development of the Internet and digital databases. In a corpus of scientific articles such as a digital library, documents are connected by citations and one document plays two different roles in the corpus: *document itself* and *a citation of other documents*. In the existing topic models, little effort is made to differentiate these two roles. We believe that the topic distributions of these two roles are different and related in a certain way. In this paper we propose a *Bernoulli Process Topic* (BPT) model which models the corpus at two levels: *document level* and *citation level*. In the BPT model, each document has two different representations in the latent topic space associated with its roles. Moreover, the multi-level hierarchical structure of the citation network is captured by a generative process involving a Bernoulli process. The distribution parameters of the BPT model are estimated by a variational approximation approach. In addition to conducting the experimental evaluations on the document modeling task, we also apply the BPT model to a well known scientific corpus to discover the latent topics. The comparisons against state-of-the-art methods demonstrate a very promising performance.**

***Keywords*-Unsupervised learning; latent models; text mining**

I. Introduction

One of the learning tasks which play central roles in the data mining field is to understand the content of a corpus such that one can efficiently store, organize, and visualize the documents. Moreover, it is essential in developing the human-machine interface in an information processing system to improve user experiences. This problem has received more and more attentions recently since huge repositories of documents are made available by the development of the Internet and digital databases and analyzing such large-scale corpora is a challenging research area. Among the numerous approaches on the knowledge discovery from documents, the latent topic models play an important role. The topic models extract latent topics from the corpus and the documents have new representations in the new latent semantic space. This new latent semantic space bridges the gap between the documents and the words and thus enables efficient processing of the corpus such as browsing, clustering, and visualization. Probabilistic Latent Semantic Indexing (PLSI) [1] and Latent Dirichlet Allocation (LDA) [2] are two well-known topic models.

A basic assumption underpinning the PLSI and LDA models as well as other topic models is that the documents are independent of each other. However, documents in most of corpora are related to each other in many ways instead of being isolated, which suggests that such information should be considered in analyzing the corpora. For example, research papers are related to each other by citations in the digital libraries. One approach is to treat the citations as the additional features in a similar way to the content features and apply the existing approaches to the new feature space, where Cohn et al. [3] used PLSI model and Erosheva et al. [4] applied LDA model. Zhu et al. [5] formulated a loss function in the new feature space for optimization. The above studies, however, fail to capture two important properties of the citation network. First, one document plays two different roles in the corpus: *document itself* and *a citation of other documents*. We believe that the topic distributions of these two roles are different and are related in a certain way. It should be beneficial to model the corpus at a finer level by differentiating these two roles for each document. For example, in the well-known LDA paper, Blei et al. proposed a graphical model for document modeling and adopted the variational inference approach for parameter estimation. When the LDA paper serves as the citation role, one might be more interested in the graphical model and variational inference approach than other content covered in the LDA paper. This is the case, especially when one is interested in the applications of the LDA model in other contexts, such as the document clustering task. Therefore, the topic distributions of the LDA paper at the two levels (*document level* and *citation level*) are different, as illustrated in Fig. 1. The topic models which simply treat the citations as the features in a peer-level to the content fail to differentiate these two levels.

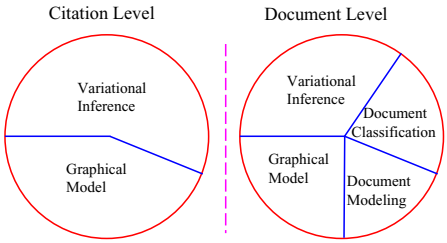

Figure 1. An illustration of the different topic distributions of the LDA paper at the document level and citation level.

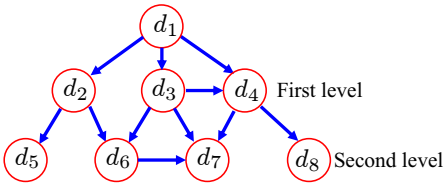

Figure 2. An illustration of the multi-level hierarchial structure of a citation network. Circles represent the papers and arrows represent the citation relationships.

The second property of the citation network that is ignored by the above studies is the multi-level hierarchial structure, which implies that the relations represented by the citations are transitive. A small citation network is illustrated in Fig. 2, where the first level citations of document d_1 are those papers directly cited by d_1 and the second level citations of d_1 are those papers cited by the papers in the reference list of d_1. Although the second level citations are not directly cited by d_1, they are also likely to influence d_1 to a lesser degree than the first level citations. For example, d_5 is not directly cited by d_1; however, d_1 is probably influenced by d_5 indirectly through d_2. A topic model which fails to capture such multi-level structure is flawed.

In this paper we propose a generative model for modeling the documents linked by the citations, called the *Bernoulli Process Topic* (BPT) model, which explicitly exploits the above two properties of the citation network. In our model, the content of each document is a mixture of two sources: (1) the content of the given document, and (2) the content of other documents related to the given document through the multi-level citation structure. This perspective actually reflects the process of writing a scientific article: the authors first learn the knowledge from the literature and then combine their own creative ideas with what they learnt from the literature to form the content of their article. Consequently, the literature they learnt knowledge from forms the citations of their article. Furthermore, the multi-level structure of the citation network is captured by a Bernoulli process which generates the related documents, where the related documents are not necessarily directly cited by the given document. In addition, due to a Bayesian treatment of parameter estimation, BPT can generate a new corpus unavailable in the training stage. We conduct the comprehensive evaluations to investigate the performance of the BPT model. The experimental results on the document modeling task demonstrate that the BPT model achieves a significant improvement over state-of-the-art methods on the generalization performance. Moreover, the BPT model is applied to the well-known Cora corpus to discover the latent topics. The comparisons against state-of-the-art methods demonstrate the promising knowledge discovery capability of the BPT model.

II. Related Work

PLSI [1] is one topic model towards document modeling which treats documents as mixtures of the topics and each topic as a multinomial distribution over the words. However, PLSI cannot generate new documents which are not available in the training stage. To address this limitation, Blei et al. [2] proposed the LDA model by introducing a Dirichlet prior for the topic distributions of the documents. Different from PLSI and LDA, the BPT model in this paper incorporates the link information available in the corpus in the generative process to model the relationships among the documents. BPT is a more general framework in the sense that LDA is a special case of BPT.

PHITS [6] is a probabilistic model for links which assumes a generative process for the citations similar to PLSI, ignores the content of the documents, and characterizes the documents by the citations. Cohn et al. [3] present a probabilistic model which is a weighted sum of PLSI and PHITS (we call it Link-PLSI for the reference purpose). Similarly, Erosheva et al. [4] adopt the LDA model in a similar fashion to consider the citations (we call it Link-LDA). Following this line of research, Nallapati et al. [7] propose the Link-PLSI-LDA model which assumes the Link-PLSI model for the cited documents and the Link-LDA model for the citing documents. The common disadvantage of the above studies is that they fail to explicitly consider the relations of the topic distributions between the cited and citing documents and the transitive property of the citations. Different from the above studies which generate the citations from the documents, the BPT model in this paper considers the citations as the observed information to avoid the unnecessary assumption of generating the citations since we are interested in the latent topics instead of the citations.

Shaparenko et al. [8] consider the influences among non-hyperlinked documents by modeling one document as a mixture of other documents. Similarly, Dietz et al. [9] propose a citation influence model for hyperlinked documents by the citations. The citation influence model, however, fails to capture the multi-level transitive property of the citation network. In addition to the relations represented by the citations, other relations might be also available, for example, the co-author relations among the documents. To model authors' interest, Rosen-Zvi et al. [10] present the author-topic model which extends LDA by including the authors' information. Specifically, the author-topic model considers the topic distribution of a document as a mixture of topic distributions of the authors. Consequently, the author-topic model implicitly considers the relations among the documents through the authors. BPT *explicitly* considers the relations among the documents in a novel way by modeling the topic distributions at the document level as mixtures of the topic distributions at the citation level.

III. Bernoulli Process Topic Model

The *Bernoulli Process Topic* (BPT) model is a generative probabilistic model of a corpus along with the citation

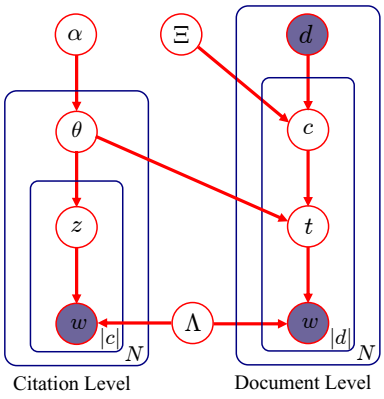

Figure 3. BPT model using the plate notation.

information among the documents. Similar to the existing topic models, each document is represented as a mixture over latent topics. The key differences from the existing topic models are that the topic distributions of the documents are modeled at two levels (*document level* and *citation level*) by differentiating the two different roles and the multi-level hierarchial structure of the citation network is captured by a Bernoulli random process.

Suppose that the corpus consists of N documents $\{d_j\}_{j=1}^N$ in which M distinct words $\{w_i\}_{i=1}^M$ occur. A word is represented by a unit vector that has a single entry equal to 1 and all other entries equal to 0. Thus, the l-th word in the vocabulary is represented by an M-dim vector $\mathbf{w}$ where $\mathbf{w}^l = 1$ and $\mathbf{w}^h = 0$ for $h \neq l$. The s-th document d_s is a sequence of the L_s words denoted by $d_s = (\mathbf{w}_{s1}, \mathbf{w}_{s2}, \cdots, \mathbf{w}_{sL_s})$ where L_s is the length of the document and $\mathbf{w}_{si}$ is the vector representing the i-th word in document d_s. Thus, the corpus is denoted by $\mathcal{D} = (d_1, d_2, \cdots, d_N)$. In addition, each document d might have a set of citations C_d, so that the documents are linked together by these citations.

BPT assumes the following generative process for each document in the corpus at the citation level, where we are interested in the topic distribution of the documents taking the citation role.

- For each document d_j.
  - For the i-th location in document d_j.
    - Choose a topic z_{ji} from the topic distribution of document d_j, $p(z|d_j, \boldsymbol{\theta}_{d_j})$, where the distribution parameter $\boldsymbol{\theta}_{d_j}$ is drawn from a Dirichlet distribution Dir($\boldsymbol{\alpha}$).
    - Choose a word $\mathbf{w}_{ji}$ which follows the multinomial distribution $p(\mathbf{w}|z_{ji}, \boldsymbol{\Lambda})$ conditioned on the topic z_{ji}.

The topic distributions at the citation level reflect the novel ideas instead of those existing approaches. In the illustration example Fig. 1, the topic distribution of the LDA paper at the citation level indicates that "graphical model" and "variational inference" are the two novel ideas in this paper, which are most likely to influence research communities.

Although the topic distributions at the citation level are important in terms of the novel ideas, we are also interested in what the content of the document is. Such information could be obtained from the topic distributions at the document level, which are described in the following generative process.

- For each document d_s.
  - For the i-th location in document d_s.
    - Choose a related document c_{si} from $p(c|d_s, \boldsymbol{\Xi})$, a multinomial distribution conditioned on the document d_s.
    - Choose a topic t_{si} from the topic distribution of the document c_{si} at the citation level, which is described in the previous generative process.
    - Choose a word $\mathbf{w}_{si}$ which follows the multinomial distribution $p(\mathbf{w}|t_{si}, \boldsymbol{\Lambda})$ conditioned on the topic t_{si}.

As shown in the above generative processes, the topic distribution at the document level is a mixture of the topic distributions at the citation level, where $\boldsymbol{\Xi}$ is the mixing coefficient matrix and the composition of $\boldsymbol{\Xi}$ and $\boldsymbol{\theta}$ models the topic distribution at the document level. It is worth noting that $\boldsymbol{\Xi}$ represents how much the content of the given document is from direct or indirect citations. Here for the clarity reason we use t, z to represent the latent topics at the document level and citation level, respectively; but they are both the random variables representing the latent topics. The whole generative processes are shown in Fig. 3.

In this generative model, the number of the latent topics is K and the mixing coefficients are parameterized by an $N \times N$ matrix $\boldsymbol{\Xi}$ where $\Xi_{js} = p(c_{si} = d_j|d_s)$, which we treat as a fixed quantity computed from the citation information of the corpus. The topic distributions at the citation level are parameterized by a $K \times N$ matrix $\boldsymbol{\Theta}$ where $\Theta_{lj} = p(z_{ji} = l|d_j)$, which is to be estimated. Similarly, an $M \times K$ word probability matrix $\boldsymbol{\Lambda}$, where $\Lambda_{hl} = p(\mathbf{w}_{si}^h = 1|t_{si} = l)$, needs to be estimated.

A. Bernoulli Process

Suppose that document d_s has a set of citations Q_{d_s}. A matrix $\mathbf{S}$ is constructed to denote the direct relationships among the documents in this way: $\mathbf{S}_{ls} = \frac{1}{|Q_{d_s}|}$ for $d_l \in Q_{d_s}$ and 0 otherwise, where $|Q_{d_s}|$ denotes the size of the set Q_{d_s}. A simple method to obtain $\boldsymbol{\Xi}$ is to set $\boldsymbol{\Xi} = \mathbf{S}$.

However, this simple strategy is not enough to capture the multi-level structure of the citation network. To model the transitive property of the citations, we assume the following generative process for generating a related document c from the given document d_s.

1) Let $l = s$.
2) Choose $t \sim$ Bernoulli(β).
3) If $t = 1$, choose $h \sim$ Multinomial($\mathbf{S}_{\cdot,l}$), where $\mathbf{S}_{\cdot,l}$ denotes the l-th column; let $l = h$, and return to Step 2.

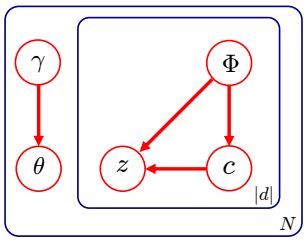

Figure 4. Graphical model representation of the variation distribution.

4) If $t = 0$, let $c = d_l$.

The above generative process combines a Bernoulli process and a random walk on the directed graph together, where the transitive property of the citations is captured. The parameter β of the Bernoulli process determines the probability that the random walk stops at the current node. The parameter β also specifies how much of the content of the given document is influenced from the direct or indirect citations.

As a result of the above generative process, $\mathbf{\Xi}$ can be obtained according to the following theorem which can be proven by the properties of random walk. The proof is omitted due to the space limitation.

Theorem 1. *The probability matrix $\mathbf{\Xi}$ is given as follows*

$$\mathbf{\Xi} = (1-\beta)(\mathbf{I} - \beta\mathbf{S})^{-1} \tag{1}$$

When the probability matrix $\mathbf{\Xi}$ is an identity matrix, the topic distributions at the document level are identical to those at the citation level. Consequently, BPT reduces to LDA [2]. Equivalently, $\beta = 0$ indicates that the relationships among the documents are not considered at all. Thus, LDA is a special case of BPT when $\beta = 0$.

IV. Parameter Estimation and Inference

The above generative processes lead to the joint probability distribution

$$p(\mathbf{c}, \mathbf{z}, \mathcal{D}, \mathbf{\Theta}|\boldsymbol{\alpha}, \mathbf{\Lambda}) = p(\mathbf{\Theta}|\boldsymbol{\alpha}) \prod_{s=1}^{N} p(\mathbf{c}_s|d_s) p(\mathbf{z}_s|\mathbf{c}_s) \prod_{i=1}^{L_s} p(\mathbf{w}_{si}|z_{si}, \mathbf{\Lambda})$$

where $p(\mathbf{\Theta}|\boldsymbol{\alpha}) = \prod_{j=1}^{N} p(\boldsymbol{\theta}_j|\boldsymbol{\alpha})$, $p(\mathbf{c}_s|d_s) = \prod_{i=1}^{L_s} p(c_{si}|d_s)$, and $p(\mathbf{z}_s|\mathbf{c}_s) = \prod_{i=1}^{L_s} p(z_{si}|c_{si}, \boldsymbol{\theta}_{c_{si}})$.

The marginal distribution of the corpus can be obtained by integrating over $\mathbf{\Theta}$ and summing over $\mathbf{c}, \mathbf{z}$

$$\begin{aligned} p(\mathcal{D}) &= \int \sum_{\mathbf{z}} \sum_{\mathbf{c}} p(\mathbf{c}, \mathbf{z}, \mathcal{D}, \mathbf{\Theta}|\boldsymbol{\alpha}, \mathbf{\Lambda}) d\mathbf{\Theta} \\ &= B(\boldsymbol{\alpha})^{-N} \int (\prod_{j=1}^{N} \prod_{i=1}^{K} \Theta_{ij}^{\alpha_i - 1}) \prod_{s=1}^{N} \prod_{i=1}^{L_s} \sum_{l=1}^{K} \sum_{t=1}^{N} \prod_{h=1}^{M} (\Xi_{ts} \Theta_{lt} \Lambda_{hl})^{\mathbf{w}_{si}^h} d\mathbf{\Theta} \end{aligned} \tag{2}$$

where $B(\boldsymbol{\alpha}) = \prod_{i=1}^{K} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{K} \alpha_i)$.

Following the principle of maximum likelihood, one needs to maximize Eq. (2) which is intractable to compute due to the coupling between $\mathbf{\Theta}$ and $\mathbf{\Lambda}$ in the summation. By assuming a particular form of the joint distribution of $\mathbf{c}, \mathbf{z}, \boldsymbol{\theta}$ as shown in Fig. 4, we arrive at the following iterative update rules by the variational approximation approach. The proof is omitted due to the space limitation.

$$\Phi_{sjhl} \propto \Xi_{js} \Lambda_{hl} \exp(\Psi(\gamma_{jl}) - \Psi(\sum_{t=1}^{K} \gamma_{jt})) \tag{3}$$

$$\gamma_{sl} = \alpha_l + \sum_{g=1}^{N} \sum_{h=1}^{M} A_{hg} \Phi_{gshl} \tag{4}$$

$$\Lambda_{hl} \propto \sum_{s=1}^{N} \sum_{j=1}^{N} A_{hs} \Phi_{sjhl} \tag{5}$$

where $A_{hs} = \sum_{i=1}^{L_s} \mathbf{w}_{si}^h$ and $\Psi(\cdot)$ is digamma function. These update rules are performed iteratively in the above order, until convergence. To perform the inference on a new corpus, one only iterates Eqs. (3) and (4) until convergence.

V. Experimental Evaluations

The BPT model is a probabilistic model towards document modeling. In order to demonstrate the performance of the BPT model, the experiments on the document modeling task are conducted. Moreover, the BPT model is applied to the well-known Cora corpus to discover the latent topics.

A. Document Modeling

The goal of the document modeling is to generalize the trained model from the training dataset to a new dataset. Thus, we wish to obtain a high likelihood on a held-out test set. In particular, we compute the perplexity of the held-out test set to evaluate the models. A lower perplexity score indicates a better generalization performance. More formally, the perplexity for a test set of N documents is

$$\text{perplexity}(\mathcal{D}) = \exp(-\sum_{i=1}^{N} \log p(d_i) / \sum_{i=1}^{N} L_i) \tag{6}$$

In this experiment, we use two corpora: Cora [11] and CiteSeer[1], which are the standard datasets with citation information available. These two datasets both contain the papers published in the conferences and journals of different research areas in computer science including artificial intelligence, information retrieval, hardware, etc. We use the subsets of these two datasets, where Cora contains 9998 documents with 3609 unique words and CiteSeer consists of 9135 documents with 889 words. Each dataset is randomly split into two parts (70% and 30%), with the 70% used to train the model and the 30% used as the held-out test set. The BPT model is evaluated against LDA [2] and Link-LDA [4], where Link-LDA incorporates the citation information into the LDA model. Fig. 5 shows the perplexity results on these two corpora where the number of the topics varies from 10 to 200 and the parameter β in the BPT model is simply fixed at 0.99. As can be seen, the BPT model achieves a significant improvement on the generalization performance.

[1] http://citeseer.ist.psu.edu/

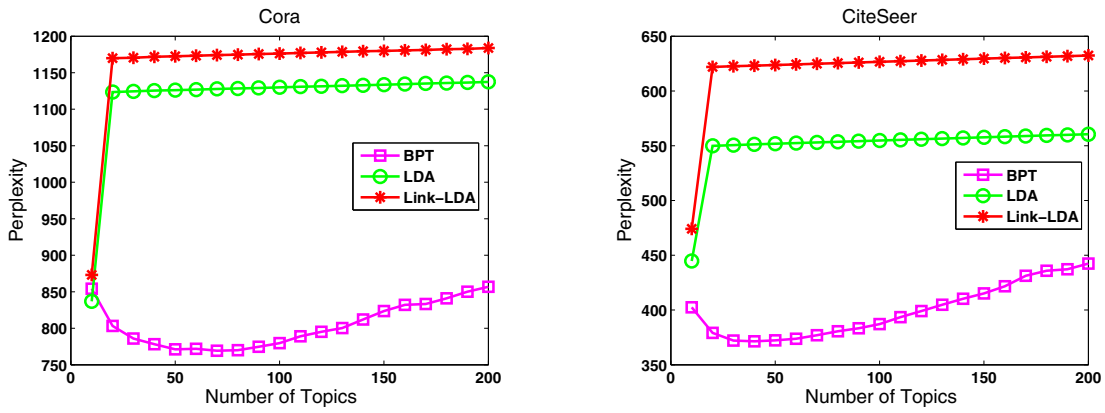

Figure 5. Perplexity comparisons on the Cora and CiteSeer datasets (the lower, the better).

B. BPT model for Cora

To discover the latent topics in details, we apply the BPT model to Cora with the number of the topics fixed at 300. The parameter β is also fixed at 0.99. Lots of applications are possible based on the learned 300 topic model. The LDA [2] and Link-LDA [4] models are also applied to Cora corpus with the same number of the topics for comparison.

1) Topic Distributions at Two Levels: One main advantage of the BPT model is the capacity of differentiating the two roles of the documents. We choose several research topics related to data mining field and investigate the topic probabilities at the document level and citation level. Fig. 6 illustrates the topic probabilities of the paper "Intelligent Query Answering by Knowledge Discovery Techniques" by Jiawei Han et al. in the data mining field, where each topic is denoted by several representative words following the order of the topic. The topic probability conditioned on this paper has a high value on the data mining topic at the document level as expected. However, the topics which this paper has the most influence on are the research topics related to decision tree and information retrieval instead of data mining as indicated at the citation level distribution. In other words, this paper is most likely to be cited by the papers related to decision tree and information retrieval.

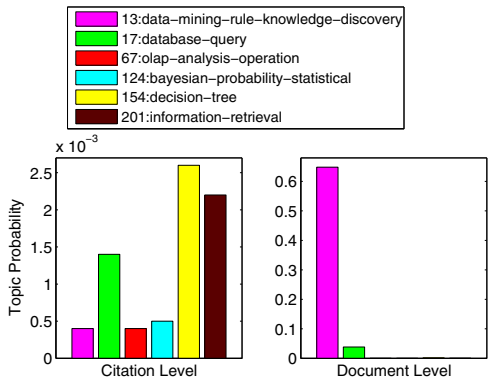

Figure 6. Topic distributions of the paper "Intelligent Query Answering by Knowledge Discovery Techniques".

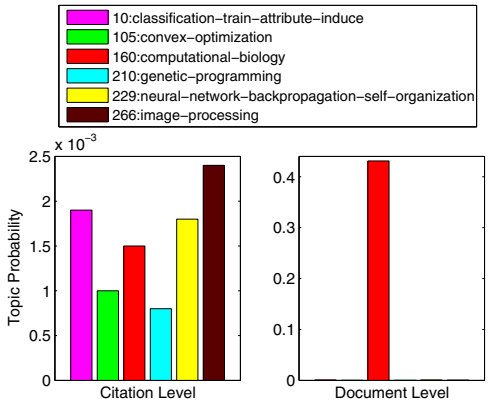

Figure 7. Topic distributions of the paper "The Megaprior Heuristic for Discovering Protein Sequence Patterns".

Another example is from the computational biology field. Since computational biology is an interdisciplinary field where machine learning and image processing techniques play the active roles, the research in the computational biology is very likely to influence these related research areas. Fig. 7 shows the related topic distributions of the paper "The Megaprior Heuristic for Discovering Protein Sequence Patterns" by Timothy L. Bailey et al.. Clearly, the probability of the computational biology topic at the document level is the highest. Yet the research topics related to image processing and classification are more likely to be influenced by this paper as indicated at the citation level distribution.

2) Citation Recommendation: The underlying assumption in the Link-LDA and LDA models is that the documents are independent of each other, which implies the topic distributions of the documents are also independent. This assumption leads to an issue in computing the posterior probability of the documents conditioned on the given topic. According to $p(d|t) \propto p(t|d)p(d)$, one would expect that a longer document (larger $p(d)$) is likely to have a larger posterior probability because the topic distribution of document $p(t|d)$ is assumed to be independent of the document length in the Link-LDA and LDA models. However, intuitively the topic distribution of a document should not be mainly

Table I
TOP 3 CITATIONS RECOMMENDED ACCORDING TO $p(c|z)$, WHERE C-CORA DENOTES THE CITATION COUNT OF THE GIVEN PAPER IN CORA AND C-GS DENOTES THE CITATION COUNT OBTAINED FROM GOOGLE SCHOLAR.

Paper title	$p(c\|z)$	C-Cora	C-GS
Data Mining			
Knowledge Discovery in Databases: An Attribute-Oriented Approach	0.977229	19	354
Bottom-up Induction of Functional Dependencies from Relations	0.005908	2	47
Fast Spatio-Temporal Data Mining of Large Geophysical Datasets	0.001346	2	62
OLAP Analysis			
Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and SubTotals	0.733346	26	1469
Query Evaluation Techniques for Large Databases	0.078250	24	990
The SEQUOIA 2000 storage benchmark	0.036707	2	201
Speech Recognition			
A Telephone Speech Database of Spelled and Spoken Names	0.118541	6	34
ASCII Phonetic Symbols for the World's Languages: Worldbet	0.109741	6	92
Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis&Antidotes	0.095960	5	48
Network QoS Services			
A generalized processor sharing approach to flow control in integrated services networks: The single node	0.957520	75	2370
Comparison of Rate-Based Service Disciplines	0.015441	32	311
A Scheduling Discipline and Admission Control Policy for Xunet 2	0.003878	6	13

determined by its length. The paper "Building Domain-Specific Embedded Languages" is the longest document in Cora corpus. In the evaluations on the Link-LDA and LDA models, this paper has the largest posterior probability for most of the topics, as expected, which does not make a reasonable sense.

The above issue is addressed by the BPT model by explicitly considering the relations among the documents represented by the citations. In the BPT model, the topic distribution of a given document $p(t|d)$ is related to other documents because it is a mixture of the topic distributions of other documents at the citation level. This is also verified by the experiments on the Cora corpus. In BPT model, the documents with a high posterior probability are directly related to the given topic instead of being determined by the document length. Due to space limitation, we do not include the experimental results and make them available online [12].

Since the topic distributions of the documents at the citation level (the matrix $\mathbf{\Theta}$) are directly modeled in the BPT model, it is natural to recommend the most influential citations in the given topic by computing the posterior probabilities $p(c|z)$. Table I shows the citations recommended by the BPT model in several research topics. Since Cora only covers the research papers before 1999, the citation count from Google Scholar is much more than that in Cora. The top 20 citations recommended in all research topics discovered by BPT are also available online [12].

VI. CONCLUSION

A multi-level latent topic model, BPT, is presented in this paper to differentiate the two different roles of each document in a corpus: *document itself* and *a citation of other documents* by modeling the corpus at two levels: *document level* and *citation level*. Moreover, the multi-level hierarchical structure of the citation network is captured by a generative process involving a Bernoulli process. The experimental results on the Cora and CiteSeer corpora demonstrate that the BPT model provides a promising knowledge discovery capability.

ACKNOWLEDGMENT

This work is supported in part by a research internship at NEC Laboratories America, Inc. and the NSF (IIS-0535162, IIS-0812114).

REFERENCES

[1] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999, pp. 50–57.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Journal of Machine Learning Research*, 2003, pp. 993–1022.

[3] D. A. Cohn and T. Hofmann, "The missing link - a probabilistic model of document content and hypertext connectivity," in *NIPS*, 2000, pp. 430–436.

[4] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed membership models of scientific publications," in *Proceedings of the National Academy of Sciences*. press, 2004, p. 2004.

[5] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in *SIGIR*, 2007, pp. 487–494.

[6] D. Cohn and H. Chang, "Learning to probabilistically identify authoritative documents," in *ICML*, 2000, pp. 167–174.

[7] R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *KDD*, 2008, pp. 542–550.

[8] B. Shaparenko and T. Joachims, "Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases," in *KDD*, 2007, pp. 619–628.

[9] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *ICML*, 2007, pp. 233–240.

[10] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI*, 2004, pp. 487–494.

[11] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Inf. Retr.*, vol. 3, no. 2, pp. 127–163, 2000.

[12] http://www.cs.binghamton.edu/~zguo/icdm09.