A Two-Level Topic Model Towards Knowledge Discovery from Citation Networks

Zhen Guo, Zhongfei (Mark) Zhang, Shenghuo Zhu, Yun Chi, and Yihong Gong

Abstract—Knowledge discovery from scientific articles has received increasing attention 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 considers 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 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. An efficient computation algorithm is proposed to overcome the difficulty of matrix inverse operation. In addition to conducting the experimental evaluations on the document modeling and document clustering tasks, we also apply the BPT model to well known corpora to discover the latent topics, recommend important citations, detect the trends of various research areas in computer science between 1991 and 1998, and to investigate the interactions among the research areas. The comparisons against state-of-the-art methods demonstrate a very promising performance. The implementations and the data sets are available online [1].

Index Terms—Unsupervised learning, latent models, text mining

1 Introduction

One of the learning tasks which play central roles in the data mining and pattern discovery 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 retrieval system to improve user experiences. This problem has received more and more attentions recently since huge repositories of documents are made available with 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 a 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) [15] and latent Dirichlet allocation (LDA) [6] are two well-known topic models.

A fundamental assumption underpinning 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 independent, 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. [9] used the PLSI model and Erosheva et al. [11] applied the LDA model. Zhu et al. [29] 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 a corpus: *document itself* and *a citation of other documents*. We observe that the topic distributions of these two roles are different and are related in a certain way. Consequently, it is 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, at the document level Blei et al. [6] proposed a graphical model for document modeling and adopted the variational inference approach for parameter estimation; they also discussed the topic on the application of the model to document classification at the same level. When the LDA paper serves as the citation role, one might be more interested in the graphical model and the 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

- *Z. Guo is with Yahoo Labs, Santa Clara, CA 95054. E-mail: zguo@cs.binghamton.edu.*
- *Z.M. Zhang is with the Department of Computer Science, State University of New York at Binghamton, Binghamton, NY 13902. E-mail: zhongfei@cs.binghamton.edu.*
- *S. Zhu and Y. Chi are with NEC Laboratories America, Inc., 10080 N. Wolfe Rd. SW3-350, Cupertino, CA 95014. E-mail: {zsh, ychi}@nec-labs.com.*
- *Y. Gong is with the School of Electronic and Information Engineering, Xi'an Jiaotong University, China 710054. E-mail: ygong@mail.xjtu.edu.cn.*

Manuscript received 3 Aug. 2011; revised 11 Apr. 2012; accepted 22 Mar. 2013; date of publication 4 Apr. 2013; date of current version 18 Mar. 2014. Recommended for acceptance by B.C. Ooi.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2013.56

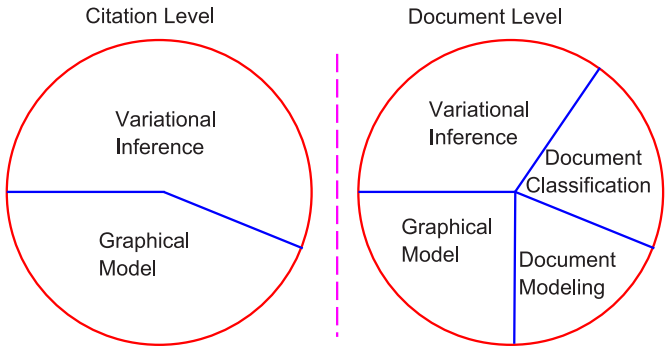


Fig. 1. An illustration of the different topic distributions of the LDA paper at the document level and citation level.

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.

The second property of the citation network that is ignored by the above studies is the multi-level hierarchical structure, which implies that the relations represented by the citations are transitive. An exemplary 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 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 have learnt from the literature to form the content of their article. Consequently, the literature they have 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 the parameter estimation, BPT can generate a new corpus unavailable in the training stage. In summary, we highlight the contributions of this work as follows.

- We propose a novel generative model for modeling documents linked by citations at a finer level by differentiating the two roles of each document. More specifically, the topic distribution of each document at the document level is a mixture of the topic distributions of the related documents at the citation level.

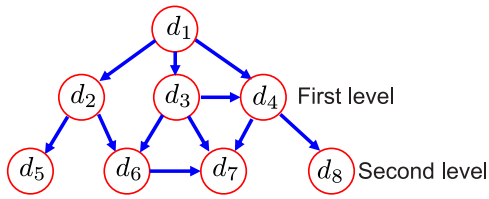


Fig. 2. An illustration of the multi-level hierarchical structure of a citation network. Circles represent the papers and arrows represent the citation relationships.

- The multi-level hierarchical structure of the citation network is captured by a random Bernoulli process which involves a random walk on a directed graph. In addition to the direct relations among documents, the indirect relations are also fully considered.
- 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 the state-of-the-art methods on the generalization performance. Moreover, the BPT model is applied to well-known corpora to discover the latent topics, cluster the documents, detect the trends of various research areas in computer science between 1991 and 1998, and investigate the interaction among the research areas. The comparisons against the state-of-the-art methods demonstrate the promising knowledge discovery capability of the BPT model.

Compared with our previous publication on this research work [12], this paper provides thorough theoretical analysis for the BPT model. Moreover, the complexity analysis of parameter estimation is conducted and an efficient algorithm is proposed to significantly reduce the computation time. The comprehensive experimental evaluations are reported to demonstrate the promising knowledge discovery capability of the BPT model.

2 Related Work

PLSI [15] is a topic model towards document modeling which considers documents as mixtures of the topics and each topic as a multinomial distribution over the words. Then the principle of maximum likelihood leads to a parameter estimation algorithm derived in the EM framework. However, PLSI cannot generate new documents which are not available in the training stage. To address this limitation, Blei et al. [6] 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 a 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 [8] 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. [9] 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. [11] adopt the LDA model in a similar fashion to consider the citations (we call it Link-LDA for the reference purpose). Following this line of research, Nallapati et al. [22] 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 the 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. [24] consider the influences among non-hyperlinked documents with the time stamps available. In [24], Shaparenko et al. assume that each document is generated by a unigram language model. To characterize the relationship among documents, they assume that the language model of a document d_i with the time stamp t_i can be approximately expressed as a mixture distribution over the language models of all the previous documents with the time stamp before t_i. The influences among documents are determined by the likelihood ratio test approach for each pair of documents. Since [24] does not consider the latent space representation of documents, it is impossible to perform other tasks on corpus such as clustering. The BPT model proposed in this paper, however, is able to perform various tasks on large data sets with the efficient computation algorithm, including document modeling, document clustering, topic detection, citation recommendation, etc.

Dietz et al. [10] propose a copycat model and a citation influence model for the documents hyperlinked by the citations. The copycat model extends the LDA model by considering each document as a weighted sum of documents it cites. Each topic in a document is drawn from the topic mixtures of its citations, and the distribution of draws from citations is modeled by a multinomial distribution over citations. Therefore, the copycat model enforces each word in a document to be associated with one of its citations, and it is unable to model the new or evolving topics. To overcome this limitation, Dietz et al. propose the citation influence model, where a document may choose to draw the topic of each word from a topic mixture of its citations or from its own topic distribution that models innovation topics. This choice is modeled by a flip of an unfair coin. He et al. [14] propose the inheritance topic model (ITM) model to detect topic evolution in scientific literature. Although the ITM model has different plate notation from the citation influence model and it does not differentiate citing documents and cited documents in the plate notation, it is essentially the same as the citation influence model. [14] divides the document corpus into exclusive temporal subsets according to the time stamps of the documents, and then applies the ITM model to each subset to learn the latent topics for different time period. The similarities between the topics from different time period are used to detect topic evolution. These models consider the topic distribution of each document as a mixture of the topic distributions of its citations. They, however, fail to capture the multi-level transitive property of the citation network. The BPT model in this paper considers the multi-level transitive property of the citation network by a random Bernoulli process; thus the indirect relations are fully considered and the superior performance is expected. We will compare some of these models with our BPT model in experimental evaluations.

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. [23] 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. More recently, Kataria et al. [17] propose a framework in which the *context* of a citation, defined as the words near the occurrence of the citation in the citing paper, is incorporated into the topic model. In [17], two variants of the proposed model are the extensions of Link-LDA [11] and Link-PLSI-LDA [22], and one additional generative process is added into Link-LDA and Link-PLSI-LDA, where each word inside the context is associated with one citation through a topic-citation distribution. In this paper, we do not assume the context information to be available, although incorporating context information into BPT is one of our future directions.

3 Bernoulli Process Topic Model

The *Bernoulli process topic* model is a generative probabilistic model of a corpus along with the citation 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 that the multi-level hierarchical 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 lth 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 sth document d_s is a sequence of the L_s words denoted by $d_s = (\mathbf{w}_{s1}, \mathbf{w}_{s2}, \ldots, \mathbf{w}_{sL_s})$ where L_s is the length of the document and $\mathbf{w}_{si}$ is the vector representing the ith word in the document d_s. Thus, the corpus is denoted by $\mathcal{D} = (d_1, d_2, \ldots, d_N)$. In addition, each document d might have a set of citations C_d, so that the documents are linked together by these citations.

3.1 The Generative Process of BPT

BPT assumes the following generative process for each document in the corpus at the citation level, where we are

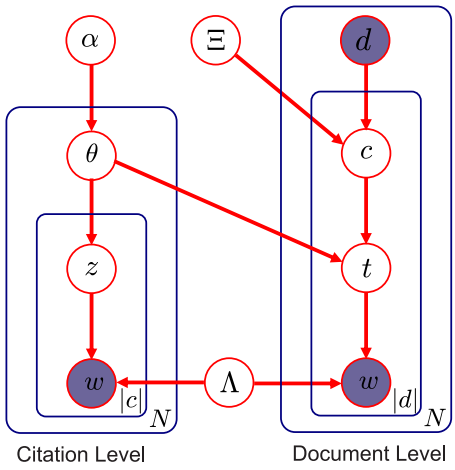


Fig. 3. The BPT model using the plate notation.

interested in the topic distribution of the documents taking the citation role.

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

The topic distributions at the citation level reflect the novel ideas instead of those existing approaches. In the illustrative 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 the 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 ith location in the document d_s.
    a. Choose a related document c_{si} from $p(c|d_s, \Xi)$, a multinomial distribution conditioned on the document d_s.
    b. 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.
    c. Choose a word $\mathbf{w}_{si}$ which follows the multinomial distribution $p(\mathbf{w}|t_{si}, \Lambda)$ conditioned on the topic t_{si}.

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. Note that $|c| = N$, i.e., a document may potentially cite all the documents (including itself) in the data set. 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 Ξ is the mixing coefficient matrix and the composition of Ξ and θ models the topic distribution at the document level.

It is worth noting that Ξ represents how much the content of the given document is from direct or indirect citations. In this paper, Ξ is constructed in the way such that the diagonal elements are not zeros. Therefore, the related document c_{si} could be the document d_s itself. In other words, the topics of each document include two parts: the "novel" topics from the document itself and the topics from all the direct/indirect citations. Furthermore, when one document has lots of direct/indirect citations (e.g., survey papers), the topic mixture of this document is influenced by all the direct/indirect citations. Thus, for the very dense citation graph, the topic distribution of large number of documents at the citation level are mixed together to form the topic distributions of documents at the document level. However, if there are some documents which only have a few direct/indirect citations, the mixture can occur from only these direct/indirect citations even if the whole citation graph is very dense.

In this generative model, the number of the latent topics is K and the mixing coefficients are parameterized by an $N \times N$ matrix Ξ where $\Xi_{js} = p(c_{si} = d_j|d_s)$, which we consider as a fixed quantity computed from the citation information of the corpus. In the next section, we will describe in detail how Ξ is derived in the BPT model. The topic distributions at the citation level are parameterized by a $K \times N$ matrix Θ where $\Theta_{lj} = p(z_{ji} = l|d_j)$, which is to be estimated. Similarly, an $M \times K$ word probability matrix Λ, where $\Lambda_{hl} = p(\mathbf{w}^h_{si} = 1|t_{si} = l)$, needs to be estimated.

3.2 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: $\mathbf{S}_{cs} = \frac{1}{|Q_{d_s}|}$ for $d_c \in Q_{d_s}$ and 0 otherwise, where $|Q_{d_s}|$ denotes the size of the set Q_{d_s}. A simple method to obtain Ξ is to set $\Xi = \mathbf{S}$.

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 $t \sim \mathrm{Bernoulli}(\beta)$.
3. If $t = 1$, choose $h \sim \mathrm{Multinomial}(\mathbf{S}_{\cdot,l})$, where $\mathbf{S}_{\cdot,l}$ denotes the l-th column; let $l = h$, and return to Step 2.
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 specifies 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 by the direct or indirect citations.

As a result of the above generative process, Ξ can be obtained according to the following theorem. The proof is provided in the appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org 10.1109/TKDE.2013.56.

Theorem 1. *The probability matrix Ξ is given as follows:*

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

When the probability matrix Ξ is an identity matrix, the topic distributions at the document level are identical to those at the citation level. Consequently, BPT reduces to LDA [6]. Equivalently, $\beta = 0$ indicates that the relationships among the documents are not considered at all. Thus, LDA is a special case of BPT when $\beta = 0$.

Note that $\Xi = 0$ when $\beta = 1$, which means that all the topics of the document come from the citations and the documents do not have any additional new content. This will not happen in the reality since each document must have some additional content to be eligible to be published, even for the survey papers. The optimal value of β depends on the data and the citation graph. When all the documents are closely related to each other by the direct/indirect citations, large value of β will achieve a good performance. In other words, the optimal β is a good indicator of how the documents are related to each other.

4 Parameter Estimation and Inference

The above generative processes lead to a joint probability distribution

$$\begin{aligned} &p(\mathbf{C},\mathbf{Z},\mathcal{D},\Theta \mid \alpha,\Lambda,\Xi) \\ &\quad = p(\Theta|\alpha)\prod_d p(\mathbf{c}_d \mid \Xi_d)p(\mathbf{z}_d \mid \mathbf{c}_d,\Theta)p(\mathbf{w}_d \mid \mathbf{z}_d,\Lambda), \end{aligned} \tag{2}$$

where $p(\Theta|\alpha) = \prod_{c=1}^{N} p(\Theta_c \mid \alpha)$, $p(\mathbf{c}_d \mid \Xi_d) = \prod_{i=1}^{L_d} p(c_{di})$, $p(\mathbf{z}_d|\mathbf{c}_d,\Theta) = \prod_{i=1}^{L_d} p(z_{di} \mid c_{di},\Theta)$, and $p(\mathbf{w}_d \mid \mathbf{z}_d,\Lambda) = \prod_{i=1}^{L_d} p(w_{di}|z_{di},\Lambda)$. By marginalizing Θ, $\mathbf{C}$, and $\mathbf{Z}$ in Eq. (2), we obtain the likelihood

$$\begin{aligned} L(\alpha,\Lambda;\mathcal{D},\Xi) &= \int \sum_{\mathbf{Z},\mathbf{C}} p(\mathbf{C},\mathbf{Z},\mathcal{D},\Theta|\alpha,\Lambda,\Xi)d\Theta \\ &= B(\alpha)^{-N}\int\left(\prod_{c,z}\Theta_{zc}^{\alpha_z-1}\right)\prod_{d,w}[\Lambda\Theta\Xi]_{wd}^{A_{wd}}d\Theta, \end{aligned}$$

where c and d enumerate over $\mathcal{D}$, z over $\mathcal{Z}$, w over $\mathcal{W}$, Beta function $B(\alpha) = \prod_z \Gamma(\alpha_z)/\Gamma(\sum_{z=1}^{K}\alpha_z)$, and the number of term occurrence $A_{wd} = \sum_{i=1}^{L_d}[w_{di} = w]$ ($[\bullet]$ is the indicator function).

Following the principle of the maximum likelihood, one needs to maximize Eq. (3) which is intractable to compute due to the integration of Θ. Similar to LDA, an approximate solution, however, can be obtained by introducing the variational parameters.

Proposition 1. *Function $f(\alpha,\Lambda,\Omega)$ is defined as*

$$\sum_{d,c,z,w} A_{wd}\Phi_{wzcd}\log\left(\frac{\Lambda_{wz}\Xi_{cd}}{\Phi_{wzcd}}\right) + \sum_c \log\frac{B(\gamma_c)}{B(\alpha)}, \tag{4}$$

where d and c enumerate over $\mathcal{D}$, z over $\mathcal{Z}$, and w over $\mathcal{W}$; Ω is a nonnegative matrix of size $K \times N$, $\Phi = \varphi(\Lambda,\Omega)$ defined as $\Phi_{wzcd} = (\Lambda_{wz}\Omega_{zc}\Xi_{cd})/[\Lambda\Omega\Xi]_{wd}$, and $\gamma_c = \{\gamma_{zc} : \gamma_{zc} = \alpha_z + \sum_{d,w}\Phi_{wzcd}A_{wd}\}$. Then the inequality

$$\log L(\alpha,\Lambda;\mathcal{D},\Xi) \geq \sup_{\Omega} f(\alpha,\Lambda,\Omega)$$

holds true.

The proof for Proposition 1 is provided in the appendix, available in the online supplemental material. This proposition gives a variational lower bound of the likelihood. The approximate solution to Eq. (3) can be obtained by maximizing the lower bound $f(\alpha,\Lambda,\Omega)$, which, however, is not a convex function. Thus, the global optimum solution is not realistic and we aim at obtaining a local maximum.

4.1 Update Rules

In order to achieve the lower bound, the parameters can be estimated by an alternative descend algorithm similar to the NMF algorithm [18]. To facilitate the derivation, we define a mixture projection from vector $\mathbf{x}$ onto a simplex as vector $\mathbf{y}$ ($y_k = x_k/\sum_l x_l$), denoted as $\mathbf{y} = \mathcal{P}_{\mathcal{M}}(\mathbf{x})$.[1] Similarly, a *Dirichlet adjustment* is defined as follows.

Definition 1. *A K-dimensional vector $\mathbf{y}$ is the Dirichlet adjustment of a K-dimensional vector $\mathbf{x}$ with respect to Dirichlet distribution $\mathrm{Dir_K}(\alpha)$ if*

$$y_k = \exp(\Psi(\alpha_k + x_k) - \Psi\left(\sum_l(\alpha_l + x_l)\right)), \quad \forall k$$

where $\Psi(\cdot)$ is digamma function. It is denoted by $\mathbf{y} = \mathcal{P}_{\mathcal{D}}(\mathbf{x},\alpha)$.

The above operations can be extended to a matrix by applying the operations on each column of the matrix, which can be denoted by $\mathbf{Y} \overset{\mathcal{P}_{\mathcal{M}}}{\longleftarrow} \mathbf{X}$ and $\mathbf{Y} \overset{\mathcal{P}_{\mathcal{D}}(\cdot,\alpha)}{\longleftarrow} \mathbf{X}$, respectively, where $\mathbf{X},\mathbf{Y}$ are matrices. The parameters in the BPT model can be estimated by these operations, as shown in the following proposition, where $\mathbf{X}\circ\mathbf{Y}$ is the element-wise product of matrices $\mathbf{X}$, $\mathbf{Y}$ and $\frac{\mathbf{X}}{\mathbf{Y}}$ is the element-wise division.

Proposition 2. *The local maximum of $f(\alpha,\Lambda,\Omega)$ is obtained by iteratively sequentially applying the following update rules:*

$$\Lambda \overset{\mathcal{P}_{\mathcal{M}}}{\longleftarrow} \left[\frac{\mathrm{A}}{\Lambda(\Omega\Xi)}(\Omega\Xi)\right]\circ\Lambda \tag{5}$$

$$\Omega \overset{\mathcal{P}_{\mathcal{D}}(\cdot,\alpha)}{\longleftarrow} \left[\Lambda\frac{\mathrm{A}}{\Lambda(\Omega\Xi)}\Xi\right]\circ\Omega \tag{6}$$

1. It is known as m-projection onto simplex in information geometry.

$$\alpha_z \leftarrow \alpha_z \frac{\sum_c \{\Psi(\gamma_{zc}) - \Psi(\alpha_z)\}}{\sum_c \{\Psi(\sum_z \gamma_{zc}) - \Psi(\sum_z \alpha_z)\}}, \quad 1 \le z \le K. \quad (7)$$

The proof for Proposition 2 is provided in the appendix, available in the online supplemental material. Inference on a new corpus can be obtained by computing the variational bound of $L(\alpha, \Lambda; \mathcal{D}, \Xi)$ for the given α, Λ. In other words, we can fix α, Λ and iteratively apply Eq. (6) to find the maximum of $f(\alpha, \Lambda, \Omega)$.

4.2 Complexity Analysis

The computation complexity of the BPT model depends on the size of corpus and the density of the probability matrix Ξ. According to the Eqs. (5)-(7), the computation complexity is $O((K+M)N^2 + KMN)$, where M is the number of unique words, N is the number of documents, and K is the number of topics. In addition, the complexity $O(N^3)$ is required to compute matrix inverse in Eq. (1). Therefore, the overall computation complexity is $O(N^3 + (K+M)N^2 + KMN)$. Compared to the LDA model which has the complexity $O(KMN + (K+N)M)$, the BPT model requires more expensive computation. The learning algorithm of the BPT model, however, can be optimized to significantly reduce the computation complexity, as described below.

A matrix inverse operation is necessary to compute the relation matrix Ξ by Eq. (1), which could be extremely slow, especially for the large corpus. Also, the inversion loses the sparseness of the link structure; thus it requires a large space to store Ξ and a long time to compute the matrix multiplication operations. One important observation from Eqs. (5) and (6) is that we only need to compute the product $\mathbf{X}\Xi$ or $\mathbf{X}\Xi^\top$ without explicitly computing Ξ, where $\mathbf{X}$ is an $K \times N$ matrix. Thus, we can take advantage of the sparse LU factorization to speed up the computation.

Let $\mathbf{B} = (1-\beta)^{-1}(\mathbf{I} - \beta\mathbf{W}) = \Xi^{-1}$. The matrix $\mathbf{B}$ can be factored as $\mathbf{PBQ} = \mathbf{LU}$, where $\mathbf{P}$ and $\mathbf{Q}$ are the permutation matrices, $\mathbf{L}$ is a sparse unit lower triangular matrix, and $\mathbf{U}$ is a sparse upper triangular matrix. Thus, we have

$$\mathbf{X}\Xi = \mathbf{XQU}^{-1}\mathbf{L}^{-1}\mathbf{P}. \quad (8)$$

The operations involving $\mathbf{U}^{-1}$ and $\mathbf{L}^{-1}$ can be efficiently computed by the forward and backward substitution. Because of the sparseness of the matrices, we can efficiently compute $\mathbf{X}\Xi$. $\mathbf{X}\Xi^\top$ can be computed in a similar way.

The matrix inverse operation takes $\frac{8}{3}N^3$ flops (floating-point operation) [7] and the operation $\mathbf{X}\Xi$ takes $2N^2K$ flops as Ξ is usually dense. On the other hand, sparse LU factorization takes $\frac{2}{3}N^3$ flops or less by taking advantage of sparse $\mathbf{B}$. The operations in Eq. (8) takes $4TNK$ flops, where T is the maximum number of nonzero entries per row in $\mathbf{L}$ and $\mathbf{U}$. Therefore, the computation complexity is significantly reduced by taking the sparse LU factorization without explicitly computing the inverse.

5 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 and document clustering tasks are conducted. Moreover, the BPT model is applied to well known corpora to discover the latent topics, to recommend important citations, and to detect the trends of various research areas in computer science between 1991 and 1998, as well as to investigate the interaction among the research areas. The complexity analysis in Section 4.2 is also verified by the experimental results.

5.1 Document Modeling

The goal of the document modeling is to generalize the trained model from the training data set to a new data set. 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

$$\text{perplexity}(\mathcal{D}) = \exp\Big\{ -\sum_{i=1}^{N} \log p(d_i) \Big/ \sum_{i=1}^{N} L_i \Big\}, \quad (9)$$

where L_i is the length of the ith document.

In this experiment, we use two corpora: Cora [3], [20] and CiteSeer [2], [28], which are the standard data sets with citation information available. These two data sets both contain the papers published in the conferences and journals of different research areas in computer science including artificial intelligence, information retrieval, and hardware. We conduct further pre-processing by removing stop words and infrequent words (the processes are partially described in [28]). As a result, our experimental evaluations are based on the subsets of these two data sets, where Cora contains 9,998 documents with 3,609 unique words and CiteSeer consists of 9135 documents with 889 words. Each data set is randomly split into two parts (70 and 30 percent), with the 70 percent used to train the model and the 30 percent used as the held-out test set. To make sure the repeatability of our experimental results, we put the source code of BPT and the used data sets online [1].

The BPT model is compared against LDA [6], Link-LDA [11], Copycat [10], and Citation-Influence model [10]. The Link-LDA model incorporates the citation information into the LDA model. The Copycat and Citation-Influence model are two probabilistic models which only consider the direct citations. Fig. 4 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.

To investigate the impact of the parameter β, we perform the experiments with different β values while the number of the topics is fixed at 300, as shown in Fig. 5. As expected, when β approaches to zero, the BPT model performs similar to the LDA model. As β becomes larger, BPT provides a better document modeling than other comparing methods. It is interesting to notice that BPT achieves the best performance when $\beta = 0.99$ on both Cora and CiteSeer data sets.

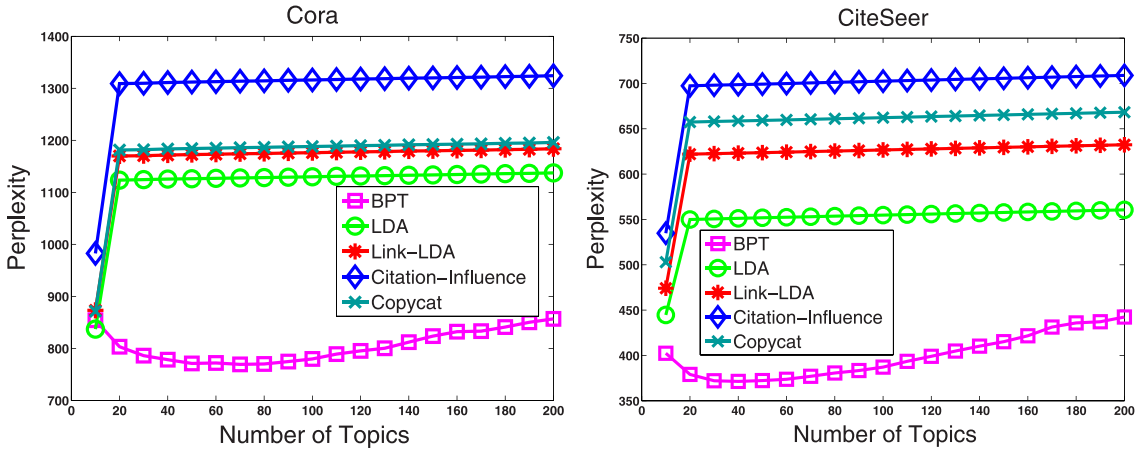


Fig. 4. Perplexity comparisons on the Cora and CiteSeer data sets (the lower, the better).

5.2 Document Clustering

Document clustering is performed on Cora which is categorized into 10 classes. For each paper, a unique label is assigned to indicate the research area it belongs to.

5.2.1 Evaluation Metrics

The two widely used metrics to measure the clustering performance are accuracy (AC) and normalized mutual information (NMI). Suppose that $\mathbf{t}$ and $\mathbf{g}$ are the cluster labels (obtained by a certain clustering algorithm) and the ground truth labels, where $\mathbf{t}_i$ and $\mathbf{g}_i$ are the labels for document d_i. The best mapping function π from $\mathbf{t}$ to $\mathbf{g}$ can be found by Hungarian algorithm [19]. The accuracy is defined as $\mathrm{AC} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta(\mathbf{g}_\mathrm{i}, \pi(\mathbf{t}_\mathbf{i}))$, where $\delta(x, y) = 1$ if $x = y$ and 0 otherwise.

The following normalized mutual information which takes a value between zero and one measures the clustering performance from the viewpoint of information theory:

$$\mathrm{NMI} = \mathrm{MI}(t, g)/\max(\mathrm{H}(t), \mathrm{H}(g)), \tag{10}$$

where t and g are the random variables corresponding to the cluster distributions of $\mathbf{t}$ and $\mathbf{g}$, respectively; $\mathrm{MI}(\mathrm{t}, \mathrm{g})$ is the mutual information between random variables t and g; $\mathrm{H}(\mathrm{t})$ is the entropy of the random variable t.

One disadvantage of NMI is that Eq. (10) only considers the maximum of the entropies and the smaller one does not contribute at all. A more reasonable metric should take into account both entropies. Inspired by the F1 score measure used to measure the classification performance, we propose the information F1 score (IF1) which is the harmonic mean of information recall (IR) and information precision (IP)

$$\mathrm{IR} = \frac{\mathrm{MI}(t, g)}{\mathrm{H}(g)} \qquad \mathrm{IP} = \frac{\mathrm{MI}(t, g)}{\mathrm{H}(t)} \qquad \mathrm{IF1} = \frac{2 * \mathrm{IR} * \mathrm{IP}}{\mathrm{IR} + \mathrm{IP}}.$$

Note that IF1 is identical to the symmetric uncertainty [26].

5.2.2 Performance Comparisons

By representing the documents in terms of latent topic space, the topic models can assign each document to the most probable latent topic according to the topic distributions of the documents. To demonstrate how our method improves the clustering performance over the state-of-the-art clustering methods, we compare the BPT model with the following representative clustering methods.

1. Traditional K-means.
2. Spectral clustering with normalized cuts (Ncut) [25].

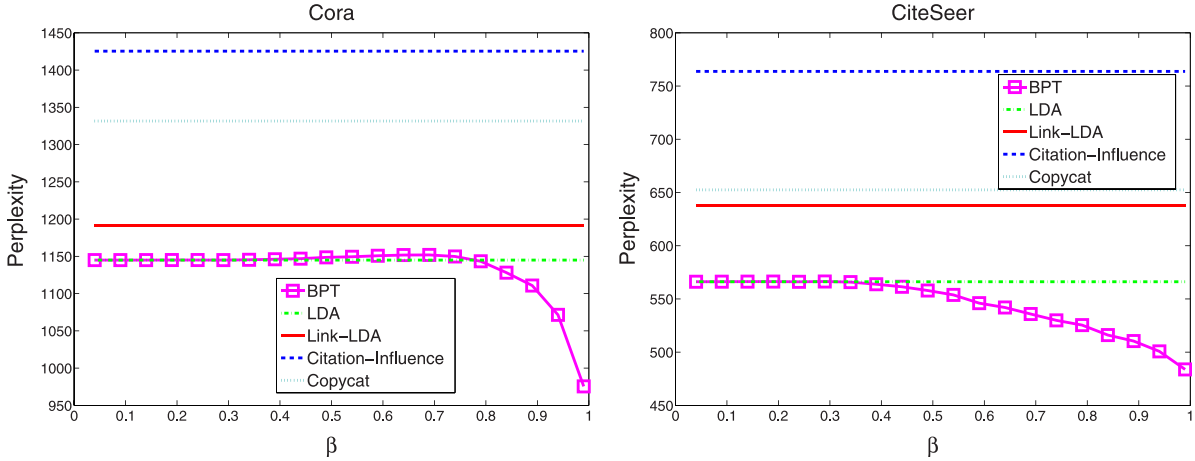


Fig. 5. Perplexity comparisons with different β values on the Cora and CiteSeer data sets.

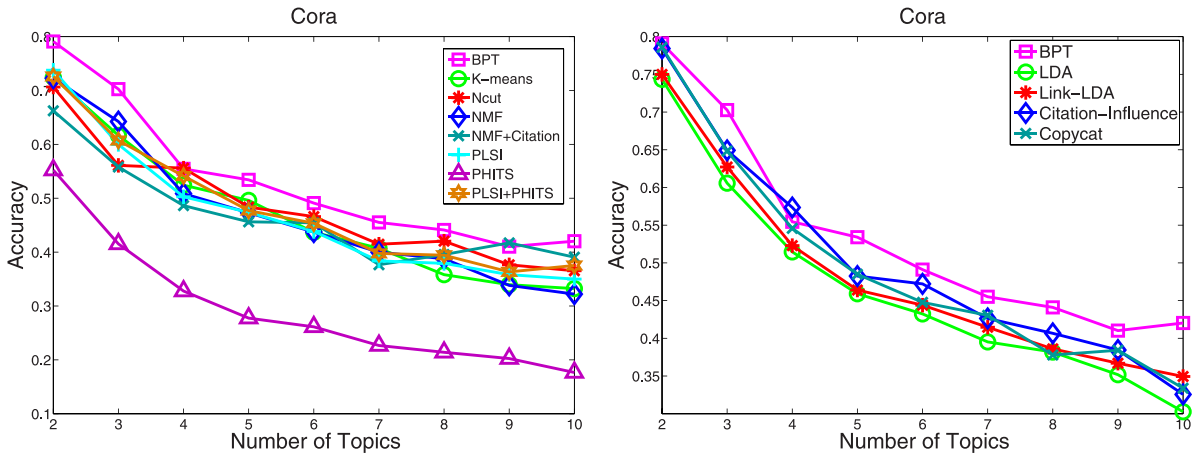


Fig. 6. Accuracy comparisons on the Cora data set.

3. Nonnegative matrix factorization (NMF) [27] which only factorizes the document-term matrix.
4. NMF [27] which factorizes the matrix which is formed by concatenating the document-term matrix and document-citation matrix (denoted by NMF+Citation model for reference purpose).
5. Probabilistic latent semantic indexing [15].
6. Latent Dirichlet allocation [6].
7. PHITS [8].
8. PLSI+PHITS, which corresponds to $\alpha = 0.5$ in [9].
9. Link-LDA [11].
10. Citation-Influence [10].
11. Copycat [10].

For the probabilistic models (BPT, PLSI, LDA, PHITS, PLSI+PHITS, Link-LDA, Citation-Influence, Copycat), the original term-document matrix is used for clustering. For all other non-probabilistic models, we take the standard tf-idf scheme, followed by the normalization step to make each column of the data matrix to be unit Euclidean length.

We adopt the evaluation strategy in [27] for the clustering performance. The test data used for evaluating the clustering methods are constructed by mixing the documents from multiple clusters randomly selected from the corpus. The evaluations are conducted for different numbers of clusters K. At each run of the test, the documents from a selected number K of clusters are mixed, and the mixed document set, along with the cluster number K, are provided to the clustering methods. For each given cluster number K, 20 test runs are conducted on different randomly chosen clusters, and the final performance scores are obtained by averaging the scores over the 20 test runs. Since all the evaluated clustering methods except Ncut are not guaranteed to find the global optimum, the standard approach is to perform the clustering several times with different initial values and choose the best one in terms of the criteria they attempt to optimize. In practice, a few number of trials are enough to find a satisfactory solution. In all of our evaluations, 10 trials are performed in each test run.

Figs. 6 and 7 report the comparisons on the Cora data set with the number of clusters ranging from 2 to 10, where the comparison methods are split into two groups for better illustration. The comparison shows that BPT has the best performance in terms of accuracy and achieves significant improvements in terms of information F1 score. The evaluations on the Cora also show that the relations among the documents do help in the document clustering. On the other hand, some comparison methods only have a good performance in terms of a certain metric. For example, Ncut which is a representative spectral clustering method gives a good accuracy, but does not perform well in terms of information

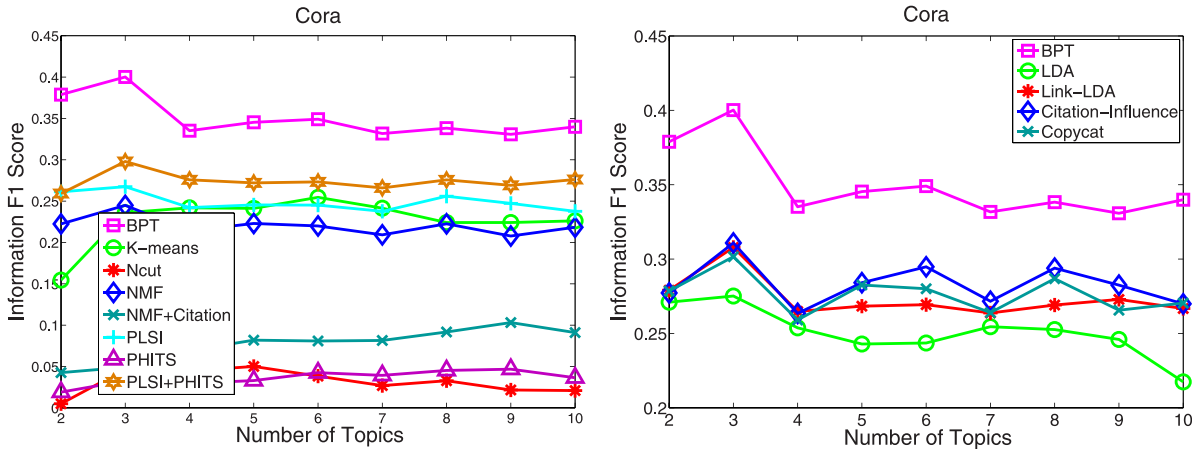


Fig. 7. Information F1 score comparisons on the Cora data set.

TABLE 1
p-Value with the Significance Level 0.05

Methods	paired t-test		signed-rank test	
	AC	IF1	AC	IF1
K-means	1.05e-5	1.58e-5	3.91e-3	3.91e-3
Ncut	3.59e-3	2.72e-10	7.81e-3	3.91e-3
NMF	8.75e-7	2.94e-9	3.91e-3	3.91e-3
NMF+Citation	1.49e-3	2.47e-8	7.81e-3	3.91e-3
PLSI	1.27e-6	3.20e-8	3.91e-3	3.91e-3
PHITS	5.65e-10	5.17e-10	3.91e-3	3.91e-3
PLSI+PHITS	5.59e-5	2.27e-6	3.91e-3	3.91e-3
LDA	1.68e-5	8.14e-8	3.91e-3	3.91e-3
Link-LDA	4.27e-6	5.06e-8	3.91e-3	3.91e-3
Citation-Influence	7.68e-3	2.65e-6	1.95e-2	3.91e-3
Copycat	9.42e-4	4.07e-7	3.91e-3	3.91e-3

F1 score. By examining the Cora corpus in details, we find that the Cora data set is very unbalanced, which means that Ncut can obtain a good accuracy by assigning most of the documents to the clusters of large sizes, but the information F1 score is very low.

To investigate whether BPT improves the clustering performance over the comparing methods or not from the viewpoint of statistics, we perform the paired hypothesis tests based on the results in Figs. 6 and 7 for the pairs of BPT and each comparing method. Two hypothesis tests are performed: paired right-tail t-test and paired two-sided Wilcoxon signed-rank test, where the null hypothesis is that the difference between the results of the two methods comes from a distribution with zero mean and the alternative hypothesis is that the mean is greater than zero (right-tail t-test) or is not zero (signed-rank test). According to the p-value shown in Table 1, the null hypotheses for all pairs are rejected, which indicates that BPT statistically improves the clustering performance by modeling the relations among the documents represented by the explicit link information.

TABLE 3
Four Topics Found by the Link-LDA Model

Topic 9		Topic 42		Topic 129		Topic 291	
system	0.013	problem	0.011	energy	0.066	compile	0.345
network	0.012	compute	0.011	circuit	0.065	routing	0.057
perform	0.012	perform	0.010	supply	0.024	persist	0.045
problem	0.010	application	0.010	conserve	0.021	program	0.030
present	0.010	provide	0.009	charge	0.017	grasp	0.027
approach	0.009	describe	0.008	resist	0.012	profit	0.022
process	0.008	learn	0.008	cmo	0.011	trial	0.020
method	0.008	system	0.007	dissipate	0.010	haskel	0.020
generate	0.007	language	0.007	capacity	0.010	inline	0.019
technique	0.007	method	0.007	vlsi	0.008	benefit	0.017

Topics are represented by their 10 most probable words, i.e., the words are ordered according to $p(w|z)$. The number is the probability of each word within the topics.

5.3 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 Link-LDA [11] model is also applied to Cora corpus with the same number of the topics for the comparison.

5.3.1 Topic Detection from Cora

Topic detection and tracking aims at understanding the corpus and discovering topically related documents from the corpus. It is a challenging task to identify the research areas from the scientific corpus for the research communities. To show a few examples, eight topics are illustrated in Table 2. The first four topics (125, 154, 160, 266) are the examples of the specific topics on Internet security, decision tree, computational biology, and image processing. The last four topics (13, 67, 124, 201) are examples of topics related to data mining field—data mining itself, database OLAP analysis, Bayesian learning, and information retrieval. Each topic is hand-labeled according to the top ten words within

TABLE 2
Eight Topics Found by the BPT Model

Topic 125		Topic 154		Topic 160		Topic 266		Topic 13		Topic 67		Topic 124		Topic 201	
security	0.117	rule	0.377	sequence	0.223	image	0.286	mine	0.042	cube	0.053	bayesian	0.050	document	0.197
protocol	0.073	tree	0.191	dna	0.029	shape	0.067	data	0.033	table	0.040	probability	0.020	retrieve	0.117
authenticate	0.037	decision	0.159	align	0.024	deform	0.028	rule	0.032	multidimension	0.039	statistic	0.020	filter	0.061
attack	0.026	expert	0.033	site	0.022	region	0.028	database	0.032	aggregate	0.038	causal	0.020	cluster	0.024
cryptograph	0.025	list	0.014	protein	0.021	segment	0.024	knowledge	0.021	olap	0.029	variable	0.018	category	0.022
signature	0.023	produce	0.013	motif	0.018	line	0.019	discovery	0.014	dimension	0.016	model	0.018	relevant	0.018
encrypt	0.022	rule-based	0.012	fragment	0.017	invariant	0.018	association	0.014	group	0.013	infer	0.015	link	0.014
hash	0.016	comprehension	0.010	hmm	0.014	surface	0.017	large	0.013	propose	0.012	condition	0.013	word	0.013
message	0.016	except	0.010	assemble	0.014	contour	0.016	algorithm	0.009	on-line	0.011	posterior	0.012	automat	0.013
scheme	0.015	accuracy	0.010	tree	0.014	histogram	0.015	efficient	0.008	spatial	0.011	graphical	0.012	index	0.013

Topics are represented by their 10 most probable words, i.e., the words are ordered according to $p(w|z)$ The number is the probability of each word within the topics.

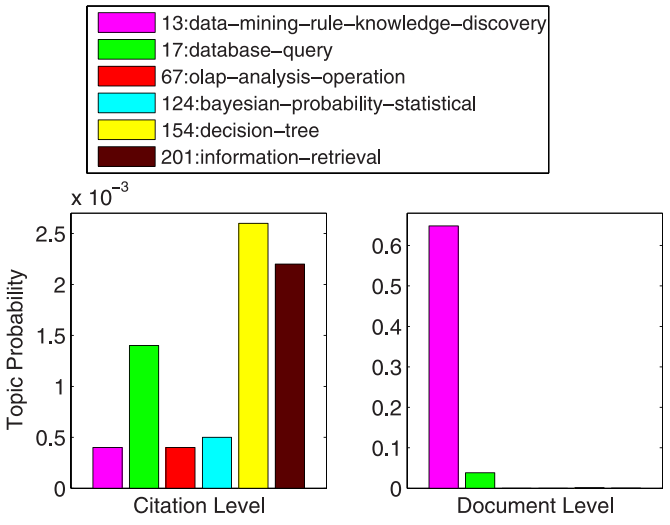


Fig. 8. Topic distributions of the paper "intelligent query answering by knowledge discovery techniques".

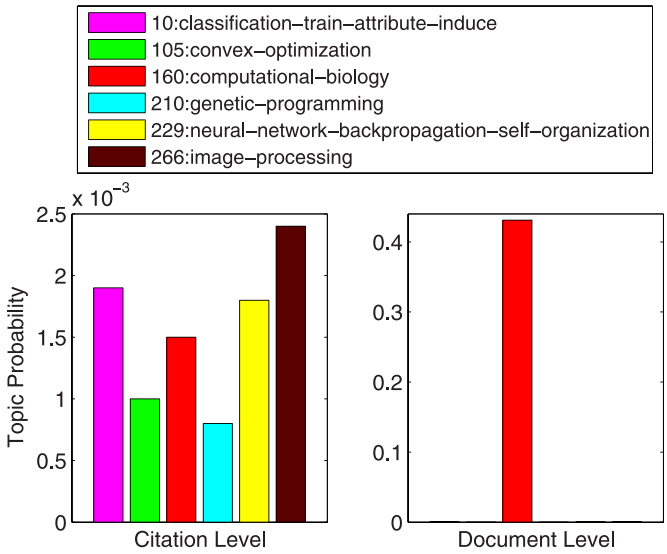


Fig. 9. Topic distributions of the paper "the megaprior heuristic for discovering protein sequence patterns".

the topic. In addition to the topics shown here, the BPT model discovers several other topics related to data mining, such as neural networks and classification, as well as topics that span the full range of research areas covered by the documents in Cora.

For comparison, four topics are chosen from the 300 topics learned by Link-LDA as illustrated in Table 3. The last two topics (129, 291) are the specific research areas on circuit design and compiler. The first two topics (9, 42) contain the most popular words in the corpus such as "system" and "problem". In fact the last two topics are only two topics whose top ten words are related to a specific research area. For all other topics, only those frequently used words in the research papers appear in the top ten words list. We believe that the reason is that the relations among the documents are not considered in Link-LDA which attempts to group the whole corpus into one topic. To investigate this issue in details, we apply Link-LDA with different numbers of the topics. Specially we are interested in the performance when the number of the topics is the ground truth (10 research areas and 62 sub-areas). When the number of the topics is fixed at 10, the various research areas are discovered; however, more specific sub-areas cannot be detected because of the small number of the topics specified. When the number of the topics is fixed at 62, most of the topics discovered are represented by the frequently used words in the corpus, the same situation in the 300 topics. Other various numbers of the topics are also attempted for Link-LDA. As the number of the topics gets larger, more topics are represented by the frequently used words in the top ten words, such as "system", "problem", and "model". Therefore, the specific research topics which the researchers are interested in cannot be discovered by Link-LDA. The same situation is observed for the LDA model. The detailed topics are not included because of the space limitation and they are available online [1].

5.3.2 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 the citation level. Fig. 8 illustrates the topic probabilities of the paper "Intelligent Query Answering by Knowledge Discovery Techniques" by Han et al. [13] 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. In other words, this paper is most likely to be cited by the papers related to decision tree and information retrieval.

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. 9 shows the related topic distributions of the paper "The Megaprior Heuristic for Discovering Protein Sequence Patterns" by Bailey et al. [5]. On the one hand, the probability of the computational biology topic at the document level is the highest. On the other hand, in addition to the computational biology topic, the research topics related to the image processing and classification are also very likely to be influenced by this paper.

5.3.3 Citation Recommendation

The underlying assumption in Link-LDA and LDA models is that the documents are independent of each other, which implies that 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 expects that a longer document (larger $p(d)$) is likely to have a larger posterior probability because the topic distribution of the document $p(t|d)$ is

TABLE 4
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

independent of the document length. This is similar to the problem discussed in Section 5.3.1 where most of the topics discovered are represented by the frequently used words in the corpus. The paper "Building Domain-Specific Embedded Languages" [16] is the longest document in Cora corpus. In the evaluations on the Link-LDA and LDA model, this paper has the largest posterior probability for most of the topics, as expected.

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 the 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 the BPT model, the documents with a high posterior

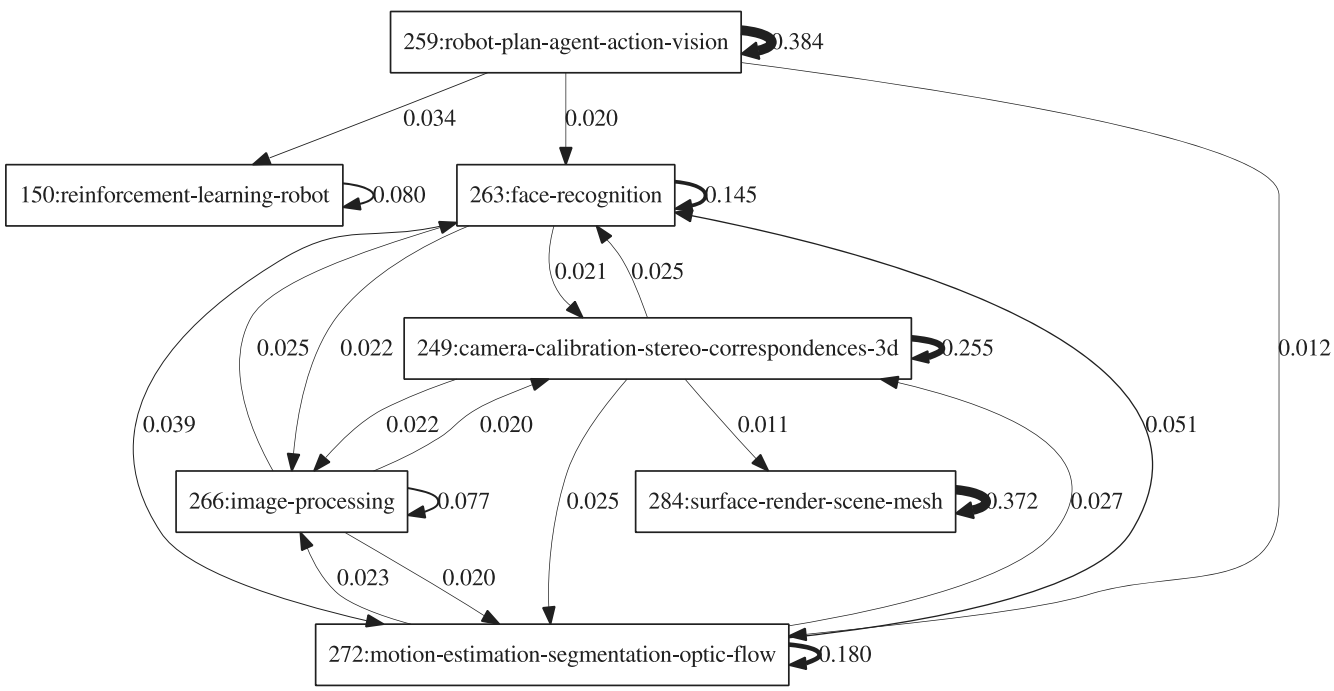


Fig. 10. Influence graph among the topics related to computer vision. The directional arrows show the directions of the influence and the number on an edge is the probability $p(z|t)$ which is proportional to the thickness of the edge.

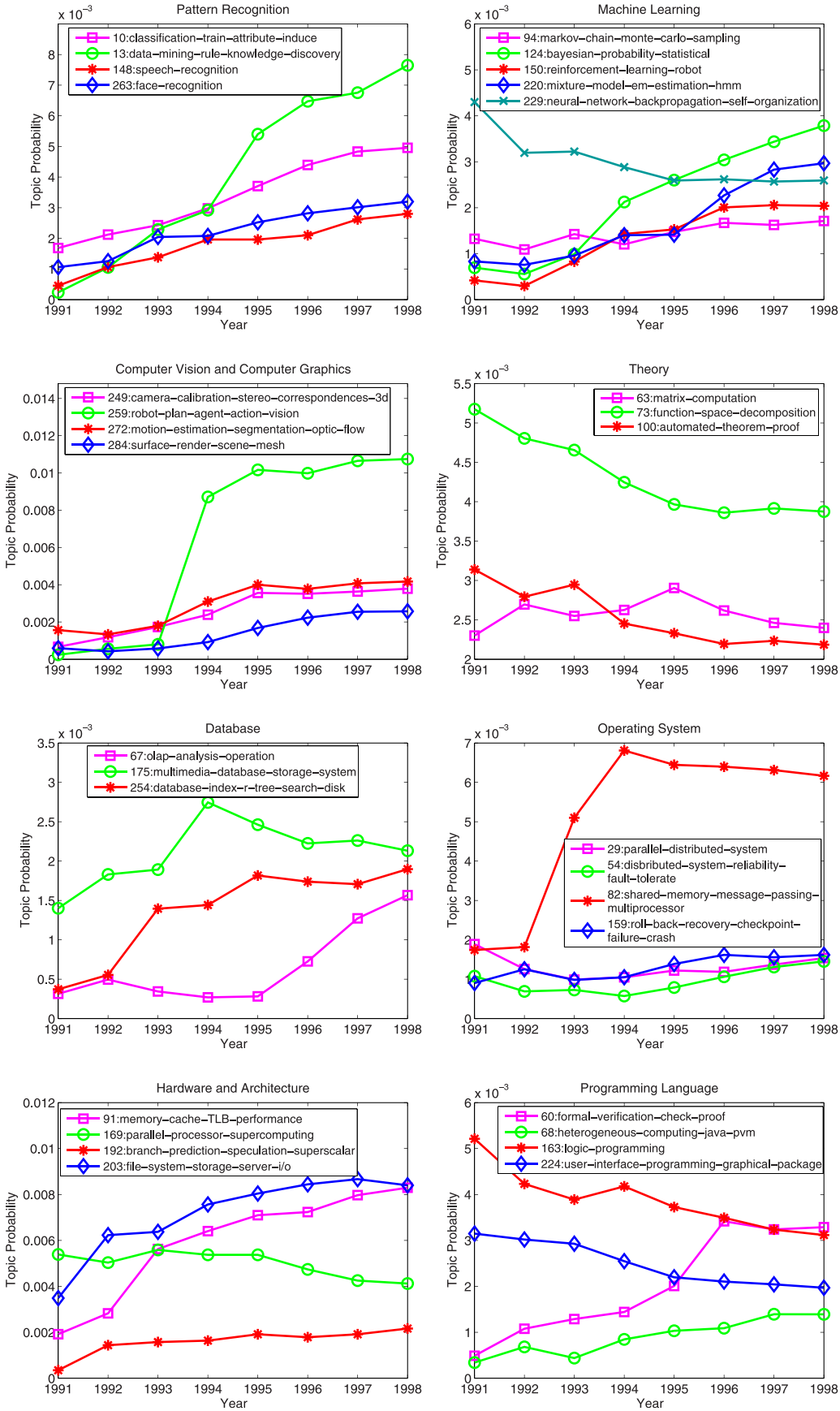


Fig. 11. Topic trend discovery over time.

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 they are available online [1].

Since the topic distributions of the documents at the citation level (the matrix Θ) are directly modeled in the BPT model, it is natural to recommend the most influential citations in a given topic by computing the posterior probabilities $p(c|z)$. Table 4 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 [1].

5.3.4 Influence among the Topics

The BPT model obtains the ability of modeling the relations among the topics by differentiating the two roles of the documents in a corpus. More formally, given two topics z, t, we are interested in the citation probability of the topic z for the topic t which is simply the conditional probability $p(z|t)$. It can be computed as $p(z|t) = \sum_{d,c} p(d|t)p(c|d)p(z|c)$. This citation probability can also be interpreted as the influence of the topic z on the topic t. We choose several research topics related to compute vision area and investigate the influence among these topics, as shown in Fig. 10. The insignificant influences are not shown in the graph. As expected, the within-topic citation probability is very high (indicated by the self-pointing arrows). Some interesting patterns are observed in the graph. For example, the influence of the motion estimation on the face recognition topic is stronger than that in the reverse direction. Also, the influence between reinforcement learning and image processing is very low (ignored in the graph) because there is not much overlap between these two topics.

5.3.5 Topic Trend Discovery over Time

The research papers in Cora corpus span from 1991 to 1998. We investigate the topic trends over time using the learned 300 topics model. Since the papers are published no earlier than their citations, all the research papers before/on the given year should be considered in order to obtain the topic distributions for that year. Therefore, we inference on all the research papers published before/on the given year to obtain the topic probabilities of that year using the learned 300 topics model. The topic probability provides a useful indicator of the topic popularity in the research literature. Fig. 11 shows the topic trends according to the topic probabilities in different research areas.

On the first row, on the left we observe a steady increase in several research topics related to pattern recognition area, and in particular the data mining area began to undergo a substantial emergence to become an independent research area around 1995, which is reflected well with the fact. On the right, however, the neural network topic decreases steadily while other machine learning topics grow, which also agrees with the fact.

On the second row, the left graph illustrates trends in computer vision and computer graphics area: a substantial increase in the robot vision topic around 1994 and slow but steady increase in other research topics. The matrix computation topic holds steady over time but other two theoretical topics decrease over time in the graph on the right.

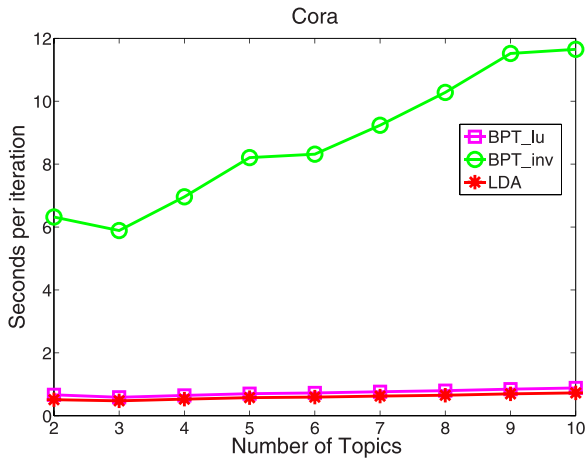


Fig. 12. Complexity comparisons on the Cora data set with different numbers of topics.

As shown in the left plot in the third row, the multimedia database topic reaches to the peak around 1994 and other two topics within database research area grow steadily. In the right plot, the inter-process communication topic in the multiprocessor research area began to emerge as a hot topic around 1994 while other topics hold steady over time.

In the hardware research area in the last row, the parallel processor topic holds steady over time, which is consistent with the trend of the parallel system in operating system area, and other three topics grow over time. The trends within programming language area are reported in the right plot.

5.4 Complexity Evaluations

To verify the complexity analysis in Section 4.2, we perform the experiments on the Cora data set using two different strategies Eq. (1) (we call it BPT_inv for the reference purpose) and Eq. (8) (we call it BPT_lu). In addition, we also investigate the complexity of LDA. We use the whole data set while the number of the topics varies from 2 to 10. Fig. 12 reports the comparisons in terms of the number of the seconds one iteration takes. The evaluations are performed by Matlab running on a Linux machine with quad-core Intel Xeon CPU 2.66 GHz and 8 Gb memory. As is shown LU factorization reduces the computation complexity significantly. It is worth noting that although LU-factorization-based approach is seemingly more complex than LDA (LDA is a special case of BPT), its complexity is comparable to that of LDA.

6 Conclusion

A multi-level latent topic model, BPT, is presented in this paper to explicitly differentiate the two different roles of each document in a corpus based on an important observation: *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. To overcome the prohibitive computation of the matrix inverse, LU factorization is used to speed up the computation with the update rules. The experimental results on the Cora and CiteSeer corpora demonstrate that the BPT model provides a better document modeling than the existing, peer methods in the literature. Furthermore, the investigation of the various applications suggests the promising knowledge discovery capability of the BPT model from a citation network.

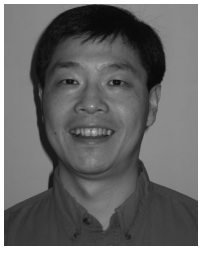
ACKNOWLEDGMENTS

The first and second authors of this work are supported in part by US National Science Foundation through grants IIS-0535162, IIS-0812114, and CCF-1017828. This work was completed during the PhD program of Zhen Guo at the Computer Science Department, SUNY at Binghamton. This work was also completed when the first author was supported in part by a summer internship at NEC Laboratories America. The second author was also supported in part by National Basic Research Program of China (2012CB316400) and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis. The authors thank Prof. C. Lee Giles and Dr. Ding Zhou for providing us the CiteSeer data.

REFERENCES

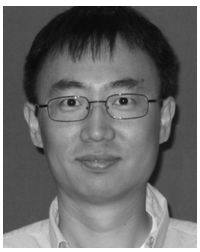
[1] http://www.cs.binghamton.edu/%7Ezguo/icdm09, 2014.
[2] http://citeseer.ist.psu.edu/, 2014.
[3] http://people.cs.umass.edu/%7Emccallum/data.html, 2014.
[4] H. Alzer, "Inequalities for the Beta Function of n Variables," *ANZIAM J.*, vol. 44, pp. 609-623, 2003.
[5] T.L. Bailey and M. Gribskov, "The Megaprior Heuristic for Discovering Protein Sequence Patterns," *Proc. Int'l Conf. Intelligent System Moleculer Biology*, 1996.
[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
[7] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
[8] D. Cohn and H. Chang, "Learning to Probabilistically Identify Authoritative Documents," *Proc. 17th Int'l Conf. Machine Learning (ICML)*, pp. 167-174, 2000.
[9] D.A. Cohn and T. Hofmann, "The Missing Link—A Probabilistic Model of Document Content and Hypertext Connectivity," *Proc. Advances in Neural Information Processing Systems (NIPS '00)*, pp. 430-436, 2000.
[10] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised Prediction of Citation Influences," *Proc. 17th Int'l Conf. Machine Learning (ICML)*, pp. 233-240, 2007.
[11] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed Membership Models of Scientific Publications," *Proc. Nat'l Academy of Sciences USA*, p. 2004, 2004.
[12] Z. Guo, Z. Zhang, S. Zhu, Y. Chi, and Y. Gong, "Knowledge Discovery from Citation Networks," *Proc. IEEE Int'l Conf. Data Mining (ICDM)*, 2009.
[13] J. Han, Y. Huang, N. Cercone, and Y. Fu, "Intelligent Query Answering by Knowledge Discovery Techniques," *IEEE Trans. Knowledge Data Eng.*, vol. 8, no. 3, pp. 373-390, June. 1996.
[14] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting Topic Evolution in Scientific Literature: How Can Citations Help?" *Proc. 18th ACM Conf. Information and Knowledge Management (CIKM)*, pp. 957-966, 2009.
[15] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, pp. 50-57, 1999.
[16] P. Hudak, "Building Domain-Specific Embedded Languages," *ACM Computing Surveys*, vol. 28, article 196, 1996.
[17] S. Kataria, P. Mitra, and S. Bhatia, "Utilizing Context in Generative Bayesian Models for Linked Corpus," *Proc. 24th AAAI Conf. Artificial Intelligence (AAAI)*, 2010.
[18] D.D. Lee and H.S. Seung, "Algorithms for Non-Negative Matrix Factorization," *Proc. Advances in Neural Information Processing Systems*, pp. 556-562, 2000.
[19] L. Lovasz and M.D. Plummer, *Matching Theory (North-Holland Mathematics Studies)*. Elsevier, 1986.
[20] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the Construction of Internet Portals with Machine Learning," *Information Retrieval*, vol. 3, no. 2, pp. 127-163, 2000.
[21] T. Minka and J. Lafferty, "Expectation-Propagation for the Generative Aspect Model," *Proc. 18th Conf. Uncertainty in Artificial Intelligence*, pp. 352-359, 2002.
[22] R. Nallapati, A. Ahmed, E.P. Xing, and W.W. Cohen, "Joint Latent Topic Models for Text and Citations," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08)*, pp. 542-550, 2008.
[23] M. Rosen-Zvi, T.L. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," *Proc. 20th Conf. Uncertainty in Artificial Intelligence (UAI '04)*, pp. 487-494, 2004.
[24] B. Shaparenko and T. Joachims, "Information Genealogy: Uncovering the Flow of Ideas in Non-Hyperlinked Document Databases," *Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '07)*, pp. 619-628, 2007.
[25] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 22, no. 8, pp. 888-905, Aug. 2000.
[26] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.
[27] W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non-Negative Matrix Factorization," *Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, pp. 267-273, 2003.
[28] D. Zhou, S. Zhu, K. Yu, X. Song, B.L. Tseng, H. Zha, and C.L. Giles, "Learning Multiple Graphs for Document Recommendations," *Proc. 17th Int'l Conf. World Wide Web (WWW)*, pp. 141-150, 2008.
[29] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining Content and Link for Classification Using Matrix Factorization," *Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, pp. 487-494, 2007.

Zhen Guo received the BS and MS degrees in electrical engineering from Xi'an Jiaotong University, China, and the PhD degree in computer science from the State University of New York at Binghamton. He is currently a researcher at Yahoo! Labs USA and his research interests include machine learning, information retrieval, and data mining.

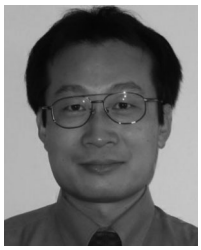


Zhongfei (Mark) Zhang received the BS (cum laude) degree in electronics engineering, the MS degree in information science, both from Zhejiang University, Hangzhou, China, and the PhD degree in computer science from the University of Massachusetts at Amherst. He is currently a professor of computer science at the State University of New York (SUNY) at Binghamton. He directs the Multimedia Research Laboratory at Binghamton. He has published more than 100 peer-reviewed academic papers in leading international journals and conferences and several invited papers and book chapters, has edited or co-edited two books. He is associate editors and guest editors for several international journals.

Shenghuo Zhu received the PhD degree in computer science from the University of Rochester, Rochester, New York, in 2003. He is currently a senior research staff member with NEC Laboratories America, Cupertino, CA. His primary research interests include information retrieval, machine learning, and data mining.



Yun Chi received the MS degree in electrical engineering from the University of Notre Dame in 2000, the MS and PhD degrees in computer science from the University of California, Los Angeles, in 2001 and 2005, respectively. He has been a research staff member in NEC Laboratories America, Inc. (Cupertino, CA) since 2005. His primary research interests include data mining, machine learning, information retrieval, and databases.



Yihong Gong received the BS, MS, and PhD degrees in Electronic Engineering from the University of Tokyo in 1987, 1989, and 1992, respectively. He then joined the Nanyang Technological University of Singapore, where he worked as an assistant professor in the School of Electrical and Electronic Engineering for four years. From 1996 to 1998, he worked for the Robotics Institute, Carnegie Mellon University, as a project scientist. In 1999, he joined NEC Laboratories America, where he worked as a senior research staff, department head, and site manager. In 2012, he joined Xi'an Jiaotong University in China, and became a professor of the "Thousand Talent Program". His research interests include video content analysis and summarization, and machine learning.