A Topic Model for Linked Documents and Update Rules for its Estimation

Zhen Guo[†], Shenghuo Zhu[‡], Zhongfei (Mark) Zhang[†], Yun Chi[‡], Yihong Gong[‡]
†Computer Science Department, SUNY at Binghamton, Binghamton, NY 13905
‡ NEC Laboratories America, Inc., 10080 N. Wolfe Rd. SW3-350, Cupertino, CA 95014
†{zguo,zhongfei}@cs.binghamton.edu, ‡{zsh,ychi,ygong}@sv.nec-labs.com

Abstract

The latent topic model plays an important role in the unsupervised learning from a corpus, which provides a probabilistic interpretation of the corpus in terms of the latent topic space. An underpinning assumption which most of the topic models are based on is that the documents are assumed to be independent of each other. However, this assumption does not hold true in reality and the relations among the documents are available in different ways, such as the citation relations among the research papers. To address this limitation, in this paper we present a *Bernoulli Process Topic* (BPT) model, where the interdependence among the documents is modeled by a random Bernoulli process. In the BPT model a document is modeled as a distribution over topics that is a mixture of the distributions associated with the related documents. Although BPT aims at obtaining a better document modeling by incorporating the relations among the documents, it could also be applied to many applications including detecting the topics from corpora and clustering the documents. We apply the BPT model to several document collections and the experimental comparisons against several state-of-the-art approaches demonstrate the promising performance.

Introduction

Unsupervised learning from documents is a fundamental problem in machine learning, which aims at modeling the documents and providing a meaningful description of the documents while preserving the basic statistical information about the corpus. Many learning tasks, such as organizing, clustering, classifying, or searching a collection of the documents, fall into this category. This problem becomes even more important with the existing huge repositories of text data, especially with the rapid development of Internet and digital databases, and thus receives an increasing attention recently.

There has been comprehensive research on the unsupervised learning from a corpus and the latent topic models play a central role among the existing methods. The topic models extract the latent topics from the corpus and therefore represent the documents in the new latent semantic space. This new latent semantic space bridges the gap between the documents and words and thus enables the efficient processing of the corpus such as browsing, clustering, and visualization. PLSI (Hofmann 1999) and LDA (Blei, Ng, and Jordan 2003) are two well known topic models toward document modeling by treating each document as a mixture of a set of topics. In these and other existing probabilistic models, a basic assumption underpinning the generative process is that the documents are independent of each other. More specifically, they assume that the topic distributions of the documents are independent of each other. However, this assumption does not hold true in practice and the documents in a corpus are actually related to each other in certain ways; for example, research papers are related to each other by citations. The existing approaches treat the citations as the additional features similar to the content. For example, Cohn et al. (2000) applies the PLSI model to a new feature space which contains both content and citations. The LDA model is also exploited in a similar way (Erosheva, Fienberg, and Lafferty 2004). As another example, Zhu et al. (2007) combine the content and citations to form an objective function for optimization.

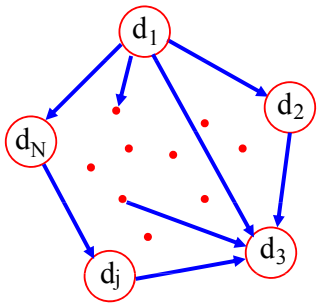

Figure 1: An example of a paper corpus in which the papers are related to each other by the citations.

The above studies, however, fail to fully capture the relations represented by the citations by simply treating the citations in the same way as the content. An example in Fig. 1 illustrates the relations represented by the citations. Paper d_1 cites paper d_2 and thus the topic distributions of d_1 and d_2 depend on each other. If the content of paper d_2 focuses on the "document clustering" problem, one should expect that paper d_1 is more or less related to the "document clustering" problem as well. Such relations cannot be captured by simply treating the citations as the additional features. Another disadvantage of the above studies is that they ignore the transitive property of the relations. In Fig. 1, although d_3 is not

cited directly by d_N, they are related indirectly through d_j. In other words, d_3 still has influence on d_N.

To address the above limitations in the existing approaches, in this paper we propose a *Bernoulli Process Topic* (BPT) model which explicitly considers the relations among the documents when extracting the latent topics from the corpus. In order to model the dependence among the documents, the content of each document could be considered to be from two sources: *the related documents* and *the document alone*. Therefore, in the BPT model each document is modeled as a distribution over the topics that is a mixture of the distributions associated with the related documents. In order to model transitive property of the relations a Bernoulli random process is proposed to generate the related documents from the given document. The variational approximation approach is adopted to estimate the parameters due to the intractability of the computation of the posterior probabilities.

The BPT model could be applied to lots of applications including detecting the topics from corpora and clustering the documents. We apply the BPT model to several document collections and the experimental comparisons against state-of-the-art approaches demonstrate the promising performance.

Related Work

PLSI (Hofmann 1999) is one well known topic model towards document modeling which treats each document as a mixture of the topics and each topic as a multinomial distribution over the words. Based on PLSI, LDA model (Blei, Ng, and Jordan 2003) is a parametric empirical Bayes model introducing a Dirichlet prior for the topic distributions of the documents, which makes it possible to generate new documents not available in the training stage. Different from PLSI and LDA, the BPT model in this paper incorporates the relations available in the corpus in the generative process to model the interdependence among the documents. BPT is a more general framework in the sense that LDA is a special case of BPT.

PHITS (Cohn and Chang 2000) is a probabilistic model for the citations similar to PLSI, where the content of the documents is ignored. Cohn et al. (2000) combine PLSI and PHITS in a linear fashion (we call it PLSI+PHITS for reference purpose). Similarly, Erosheva et al. (2004) consider the citations in LDA model (we call it Link-LDA). Following this line of research, Nallapati et al. (2008) propose a Link-PLSI-LDA model which assumes a PLSI+PHITS model for the cited documents and a Link-LDA model for the citing documents. The above studies, however, fail to fully capture the relations represented by the citations by simply treating the citations as the additional features. 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.

The relations within the corpus have received attentions recently. Dietz et al. (2007) propose a citation influence model for the hyperlinked documents by the citations. Similarly, Shaparenko et al. (2007) consider the relations among the non-hyperlinked documents by modeling one document as a mixture of other documents. To model the authors' interest, Rosen-Zvi et al. (2004) 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. The transitive property of the relations, however, is ignored in the above studies.

Bernoulli Process Topic Model

Bernoulli Process Topic (BPT) 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 feature that distinguishes the BPT model from the existing topic models is that the relationships among the documents are modeled by a Bernoulli process such that the topic distribution of each document is a mixture of the distributions associated with the related documents.

Suppose that the corpus $\mathcal{D}$ consists of N documents in which M distinct words form the vocabulary set $\mathcal{W}$. A document d is a sequence of L_d words denoted by $\mathbf{w}_d = (w_{d1}, w_{d2}, \cdots, w_{dL_d})$ where L_d is the length of the document and $w_{di} \in \mathcal{W}$ is the word in the i-th position of the document. In addition, each document d may have a set of citations C_d, so that the documents are linked together by these citations. Therefore, the corpus can be represented by a directed graph as shown in Fig. 1. Other types of relationships among the documents are also possible such as hyperlinks among the webpages and they also lead to a directed graph. Consequently, BPT model is applicable to the general scenario where the linked documents can be represented by a directed graph. For simplicity, we focus on the situation where citations among the documents are available. The extension to other scenarios is straightforward.

The BPT model assumes the following generative process for each word w_{di} in each document d in the corpus:

1. Choose a related document c_{di} from $p(c|\mathbf{\Xi}_d)$, a multinomial probability conditioned on the document d.
2. Choose a topic z_{di} from the topic distribution of the document c_{di}, $p(z|c_{di}, \mathbf{\Theta})$.
3. Choose a word w_{di} which follows the multinomial distribution $p(w|z_{di}, \mathbf{\Lambda})$ conditioned on the topic z_{di}.

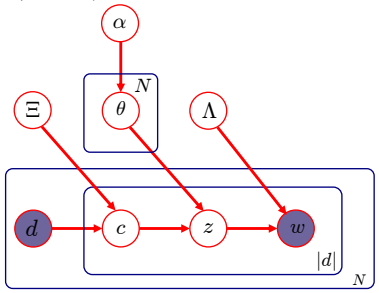

Figure 2: The BPT model using the plate notation.

The corpus is obtained once every document in the corpus is generated by this process, as shown in Fig. 2. In this gen-

erative process, the latent topic set is $\mathcal{Z}$ where $|\mathcal{Z}| = K$. The relations among the documents are represented by an $N \times N$ matrix $\mathbf{\Xi}$ with the entry $\Xi_{cd} = p(c_{di} = c)$, which we treat as a fixed quantity computed from the citation information of the corpus. We denote by an $K \times N$ matrix $\mathbf{\Theta}$ the topic distributions of the documents, where $\Theta_{zc} = p(z_{di} = z|c_{di} = c), \forall d, i$. Each column of $\mathbf{\Theta}$ denotes the topic distribution of the corresponding document (θ in Fig. 2). The c-th column of $\mathbf{\Theta}$ is denoted by $\mathbf{\Theta}_c$ and similarly, the d-th column of $\mathbf{\Xi}$ by $\mathbf{\Xi}_d$. We further assume that for each c, $\mathbf{\Theta}_c$ is drawn from a Dirichlet distribution, $\mathrm{Dir}_K(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a K-dimensional vector. The word probabilities are parameterized by an $M \times K$ matrix $\mathbf{\Lambda}$, where $\Lambda_{wz} = p(w_{di} = w|z_{di} = z), \forall d, i$.

This generative process leads to a joint probability distribution

$$\begin{aligned} &p(\mathbf{C}, \mathbf{Z}, \mathcal{D}, \mathbf{\Theta}|\boldsymbol{\alpha}, \mathbf{\Lambda}, \mathbf{\Xi}) \\ &= p(\mathbf{\Theta}|\boldsymbol{\alpha}) \prod_d p(\mathbf{c}_d|\mathbf{\Xi}_d) p(\mathbf{z}_d|\mathbf{c}_d, \mathbf{\Theta}) p(\mathbf{w}_d|\mathbf{z}_d, \mathbf{\Lambda}) \end{aligned} \quad (1)$$

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

$$\begin{aligned} L(\boldsymbol{\alpha}, \mathbf{\Lambda}; \mathcal{D}, \mathbf{\Xi}) &= \int \sum_{\mathbf{Z},\mathbf{C}} p(\mathbf{C}, \mathbf{Z}, \mathcal{D}, \mathbf{\Theta}|\boldsymbol{\alpha}, \mathbf{\Lambda}, \mathbf{\Xi}) d\mathbf{\Theta} \\ &= B(\boldsymbol{\alpha})^{-N} \int (\prod_{c,z} \Theta_{zc}^{\alpha_z - 1}) \prod_{d,w} [\mathbf{\Lambda}\mathbf{\Theta}\mathbf{\Xi}]_{wd}^{A_{wd}} d\mathbf{\Theta} \end{aligned} \quad (2)$$

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

Bernoulli Process

The document relation matrix $\mathbf{\Xi}$ is computed from the citation information of the corpus. Suppose that document d has a set of citations Q_d. A matrix $\mathbf{S}$ is constructed to denote the direct relationships among the documents in this way: $S_{cd} = 1/|Q_d|$ for $c \in Q_d$ and 0 otherwise, where $|Q_d|$ denotes the cardinality of set Q_d. A simple method to compute $\mathbf{\Xi}$ is to set $\mathbf{\Xi} = \mathbf{S}$. However, this strategy is not enough to capture the relationships among the documents. In the example in Fig. 1, d_N does not cite d_3 directly and $\Xi_{3,N} = 0$ according to the above strategy. But d_N is related to d_3 indirectly through d_j. Therefore, $\Xi_{3,N}$ should not be equal to 0.

To incorporate the indirect relations among the documents, we assume the following generative process for generating a related document c from the given document d.

1. Let $l = d$.
2. Draw $t \sim$ Bernoulli(β).
3. If $t = 1$, draw $h \sim$ Multinomial($\mathbf{S}_l$), where $\mathbf{S}_l$ denotes the l-th column of $\mathbf{S}$; let $l = h$, and go to Step 2.
4. Let $c = l$.

The above generative process combines the Bernoulli process and the random walk on the directed graph together, where the transitive property of the relations is captured. The parameter β of the Bernoulli process determines the probability that the random walk stops at the current node. As a result of the above generative process, $\mathbf{\Xi}$ can be obtained according to the following proposition which can be proven by the properties of random walk. The proof is omitted due to space limitation.

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

$$\mathbf{\Xi} = (1 - \beta)(\mathbf{I} - \beta\mathbf{S})^{-1} \quad (3)$$

When $\mathbf{\Xi}$ is an identity matrix (equivalently, $\beta = 0$), the relations among the documents are not considered at all and BPT reduces to LDA (Blei, Ng, and Jordan 2003). Thus, LDA is a special case of BPT when $\beta = 0$.

Variational Parameter Estimation

Following the principle of the maximum likelihood, one needs to maximize Eq. (2) which is intractable to compute due to the integration of $\mathbf{\Theta}$. Similar to LDA, an approximate solution, however, can be obtained by introducing the variational parameters.

Proposition 2. *Function $f(\boldsymbol{\alpha}, \mathbf{\Lambda}, \mathbf{\Omega})$ is defined as*

$$\sum_{d,c,z,w} A_{wd} \Phi_{wzcd} \log(\frac{\Lambda_{wz}\Xi_{cd}}{\Phi_{wzcd}}) + \sum_c \log \frac{B(\boldsymbol{\gamma}_c)}{B(\boldsymbol{\alpha})} \quad (4)$$

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

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

holds true.

The proof is provided in the appendix. This proposition gives a variational lower bound of the likelihood. The approximate solution to Eq. (2) can be obtained by maximizing the lower bound $f(\boldsymbol{\alpha}, \mathbf{\Lambda}, \mathbf{\Omega})$, which, however, is not a convex function. Thus, the global optimum solution is not realistic and we aim at obtaining a local maximum.

Update Rules

In order to achieve the lower bound, the parameters can be estimated by an alternative descend algorithm similar to NMF algorithm (Lee and Seung 2000). 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 by $\mathbf{y} = \mathcal{P}_M(\mathbf{x})$[1]. Similarly, a *Dirichlet adjustment* is defined as the following.

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

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

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

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

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

Proposition 3. *The local maximum of $f(\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Omega})$ is obtained by iteratively sequentially applying the following update rules*

$$\boldsymbol{\Lambda} \overset{\mathcal{P}_M}{\longleftarrow} \left[\frac{\mathbf{A}}{\boldsymbol{\Lambda}(\boldsymbol{\Omega}\boldsymbol{\Xi})}(\boldsymbol{\Omega}\boldsymbol{\Xi})^\top\right] \circ \boldsymbol{\Lambda} \tag{5}$$

$$\boldsymbol{\Omega} \overset{\mathcal{P}_D(\cdot,\boldsymbol{\alpha})}{\longleftarrow} \left[\boldsymbol{\Lambda}^\top \frac{\mathbf{A}}{\boldsymbol{\Lambda}(\boldsymbol{\Omega}\boldsymbol{\Xi})}\boldsymbol{\Xi}^\top\right] \circ \boldsymbol{\Omega} \tag{6}$$

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

This proposition is proven in the appendix. Inference on a new corpus can be obtained by computing the variational bound of $L(\boldsymbol{\alpha}, \boldsymbol{\Lambda}; \mathcal{D}, \boldsymbol{\Xi})$ for given $\boldsymbol{\alpha}$, $\boldsymbol{\Lambda}$. In other words, we can fix $\boldsymbol{\alpha}$, $\boldsymbol{\Lambda}$ and iteratively apply Eq. (6) to find the maximum of $f(\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Omega})$.

Experimental Evaluations

BPT is a probabilistic model towards unsupervised learning from linked documents. Thus, it can be applied to lots of applications such as organizing, classifying, clustering, or searching a collection of documents. In this section, we investigate two important applications: document modeling and document clustering. In all the experiments, the parameter β in the BPT model is simply fixed at 0.99.

Document Modeling

The goal of document modeling is to generalize the trained model from the training dataset to a new dataset. Thus, we wish to obtain 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 defined as $\text{perplexity}(\mathcal{D}) = \exp\left(-\sum_{i=1}^N \log p(d_i) / \sum_{i=1}^N L_i\right)$. We conduct the experiment on a subset of the CiteSeer[2] corpus which is a standard dataset with citation information available. CiteSeer contains papers published in the conferences and journals of different research areas in computer science including artificial intelligence, information retrieval, hardware, etc. There are 9135 papers with 889 unique words in the subset used in the evaluations. The whole corpus 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. We compare BPT against other two state-of-the-art topic models LDA (Blei, Ng, and Jordan 2003) and Link-LDA (Erosheva, Fienberg, and Lafferty 2004). Fig. 3 shows the perplexity results where the number of the topics varies from 10 to 200. As can be seen, the BPT model achieves a significant improvement on the generalization performance, which substantiates that the relations among the documents do offer help in the document modeling. Note that Link-LDA has higher perplexity than LDA since the additional citation features are introduced in Link-LDA.

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

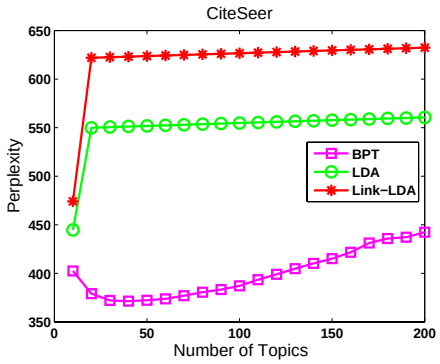

Figure 3: Perplexity comparisons on the CiteSeer dataset.

Document Clustering

Document clustering is performed on another standard dataset with the citation information available: Cora (McCallum et al. 2000). There are 9998 papers with 3609 unique words in Cora, which is categorized into 10 classes. For each paper, a unique label is assigned to indicate the research area it belongs to.

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 (Lovasz and Plummer 1986). The accuracy is defined by $\text{AC} = \frac{1}{N}\sum_{i=1}^N \delta(\mathbf{g}_i, \pi(\mathbf{t_i}))$, 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.

$$\text{NMI} = \text{MI}(t, g)/\max(\text{H}(t), \text{H}(g)) \tag{8}$$

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

One disadvantage of NMI is that Eq. (8) 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).

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

Note that IF1 is identical to the symmetric uncertainty (Witten and Frank 2005).

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) (Shi and Malik 2000).
3. Nonnegative Matrix Factorization (NMF) (Xu, Liu, and Gong 2003).
4. Probabilistic Latent Semantic Indexing (PLSI) (Hofmann 1999).
5. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003).
6. PHITS (Cohn and Chang 2000).
7. PLSI+PHITS, which corresponds to $\alpha = 0.5$ in (Cohn and Hofmann 2000).
8. Link-LDA (Erosheva, Fienberg, and Lafferty 2004).

For the probabilistic models (BPT, PLSI, LDA, PHITS, PLSI+PHITS, Link-LDA), 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 (Xu, Liu, and Gong 2003) 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 number 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.

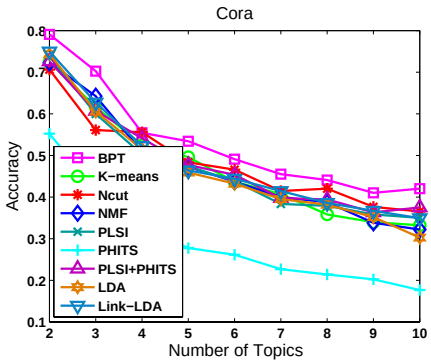

Figure 4: Accuracy comparisons on the Cora dataset.

Figs. 4 and 5 report the comparisons on the Cora dataset with the number of clusters ranging from 2 to 10, which show 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

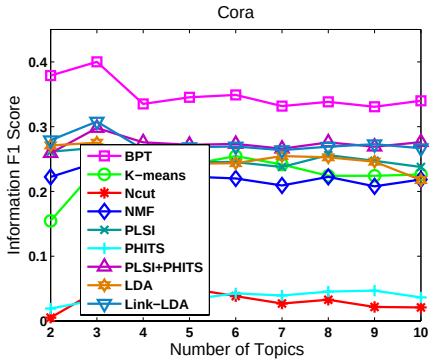

Figure 5: Information F1 score comparisons on the Cora dataset.

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 F1 score. By examining the Cora corpus in details, we find that the Cora dataset 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.

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
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

To investigate that whether BPT improves the clustering performance over the comparison methods or not from the viewpoint of statistics, we perform the paired hypothesis tests based on the results in Figs. 4 and 5 for the pairs of BPT and each comparison 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 the 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.

Conclusion

A probabilistic generative model BPT is presented in this paper to incorporate the relations among the documents into the topic model. We apply the BPT model to several document collections for document modeling and document clustering, and the experimental comparisons against state-of-the-art approaches demonstrate the promising performance.

Acknowledgements

This work is supported in part by an internship at NEC Laboratories America, Inc. and NSF (IIS-0535162, IIS-0812114, IIS-0956924).

References

Alzer, H. 2003. Inequalities for the beta function of n variables. *ANZIAM Journal* 44:609–623.

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

Cohn, D., and Chang, H. 2000. Learning to probabilistically identify authoritative documents. In *ICML*, 167–174.

Cohn, D. A., and Hofmann, T. 2000. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, 430–436.

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

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

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *NIPS*, 556–562.

Lovasz, L., and Plummer, M. D. 1986. *Matching Theory (North-Holland mathematics studies)*. Elsevier Science Ltd.

McCallum, A.; Nigam, K.; Rennie, J.; and Seymore, K. 2000. Automating the construction of internet portals with machine learning. *Inf. Retr.* 3(2):127–163.

Minka, T., and Lafferty, J. 2002. Expectation-propagation for the generative aspect model. In *In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 352–359. Morgan Kaufmann.

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

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

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

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(8):888–905.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *SIGIR*, 267–273.

Zhu, S.; Yu, K.; Chi, Y.; and Gong, Y. 2007. Combining content and link for classification using matrix factorization. In *SIGIR*, 487–494.

Appendix

Proof of Proposition 2. By Jensen's inequality, we have $[\mathbf{\Lambda\Theta\Xi}]_{wd} \geq \prod_{z,c}(\frac{\Lambda_{wz}\Theta_{zc}\Xi_{cd}}{\Phi_{wzcd}})^{\Phi_{wzcd}}$ because $\Phi_{wzcd} > 0$ and $\sum_{zc}\Phi_{wzcd} = 1$. Substituting this inequality into Eq. (2), we obtain

$$L(\boldsymbol{\alpha},\mathbf{\Lambda};\mathcal{D},\mathbf{\Xi}) \geq B(\boldsymbol{\alpha})^{-N}\prod_{w,z,c,d}\left(\frac{\Lambda_{wz}\Xi_{cd}}{\Phi_{wzcd}}\right)^{\Phi_{wzcd}A_{wd}} \times \int(\prod_{z,c}\Theta_{zc}^{\gamma_{zc}+\alpha_z-1})d\mathbf{\Theta}$$

$$= \left(\prod_c \frac{B(\boldsymbol{\gamma}_c)}{B(\boldsymbol{\alpha})}\right)\prod_{w,z,c,d}\left(\frac{\Lambda_{wz}\Xi_{cd}}{\Phi_{wzcd}}\right)^{\Phi_{wzcd}A_{wd}}$$

Taking the logarithm on both sides of the above inequality completes the proof. □

Lemma 1. *Let*

$$g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}}) = \sum_{w,z,c,d}\Phi_{wzcd}A_{wd}\log(\frac{\Lambda_{wz}\Xi_{cd}}{\Phi_{wzcd}}) + \sum_{z,c}(\Psi(\tilde{\gamma}_{zc}) - \Psi(\sum_{t=1}^{K}\tilde{\gamma}_{tc}))(\gamma_{zc}-\tilde{\gamma}_{zc}) + \sum_c \log\frac{B(\tilde{\boldsymbol{\gamma}}_c)}{B(\boldsymbol{\alpha})},$$

where $\tilde{\mathbf{\Phi}} = \varphi(\tilde{\mathbf{\Lambda}},\tilde{\mathbf{\Omega}})$ *and*

$$\tilde{\boldsymbol{\gamma}}_c = \{\tilde{\gamma}_{zc} : \tilde{\gamma}_{zc} = \alpha_z + \sum_{d,w}\tilde{\Phi}_{wzcd}A_{wd}\}$$

$$\boldsymbol{\gamma}_c = \{\gamma_{zc} : \gamma_{zc} = \alpha_z + \sum_{d,w}\Phi_{wzcd}A_{wd}\}.$$

Then

$$f(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Omega}) \geq g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}}), \quad (9)$$

$$f(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Omega}) = g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\mathbf{\Phi}). \quad (10)$$

Proof. Due to the logarithm convexity of Beta function (Alzer 2003), we have

$$\ln B(\boldsymbol{\gamma}_j) \geq \ln B(\tilde{\boldsymbol{\gamma}}_j) + \sum_{i=1}^{K}(\Psi(\tilde{\gamma}_{ij}) - \Psi(\sum_{i=1}^{K}\tilde{\gamma}_{ij}))(\gamma_{ij}-\tilde{\gamma}_{ij})$$

Substituting the above inequality into Eq. (4) leads to $f(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Omega}) \geq g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}})$. It is easy to verify the equality. □

Proof of Proposition 3. According to Lemma 1, $g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}})$ is an auxiliary function of $f(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Omega})$ (Lee and Seung 2000). To maximize $g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}})$ over $\mathbf{\Phi}$ with the constraints $\sum_{z,c}\Phi_{wzcd} = 1$, we have the Lagrangian

$$\mathcal{L}(\mathbf{\Phi}) = g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}}) + \sum_{w,d}\lambda_{wd}(\sum_{zc}\Phi_{wzcd} - 1)$$

Taking the derivative with respect to Φ_{wzcd}, one obtains

$$\frac{\partial\mathcal{L}(\mathbf{\Phi})}{\partial\Phi_{wzcd}} = A_{wd}\left(\log(\frac{\Lambda_{wd}\Xi_{cd}}{\Phi_{wzcd}}) - 1 + \Psi(\tilde{\gamma}_{zc}) - \Psi(\sum_{t=1}^{K}\tilde{\gamma}_{tc})\right) + \lambda_{wd}$$

Setting this derivative to zero leads to

$$\Phi_{wzcd} \propto \Lambda_{wz}\Xi_{cd}\exp(\Psi(\tilde{\gamma}_{zc}) - \Psi(\sum_{t=1}^{K}\tilde{\gamma}_{tc}))$$

So we have $\Omega_{zc} = \exp(\Psi(\tilde{\gamma}_{zc}) - \Psi(\sum_{t=1}^{K}\tilde{\gamma}_{tc}))$. To maximize $g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}})$ over $\mathbf{\Lambda}$ with the constraints $\sum_{w=1}^{M}\Lambda_{wz} = 1$, we have the Lagrangian

$$\mathcal{L}(\mathbf{\Lambda}) = g(\boldsymbol{\alpha},\mathbf{\Lambda},\mathbf{\Phi},\tilde{\mathbf{\Phi}}) + \sum_{z=1}^{K}\lambda_l(\sum_{w=1}^{M}\Lambda_{wz} - 1)$$

Setting the derivative with respect to Λ_{wz} to zero, one obtains $\Lambda_{wz} \propto \sum_{c,d=1}^{N}A_{wd}\Phi_{wzcd}$. Eq. (7) follows the Dirichlet estimation in (Minka and Lafferty 2002). Rearranging the above results leads to the update rules. □