Bayesian Latent Topic Clustering Model

Meng-Sung Wu and Jen-Tzung Chien
Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan 70101, ROC
{mswu,chien}@chien.csie.ncku.edu.tw

Abstract

Document modeling is important for document retrieval and categorization. The probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) are popular paradigms of document models where word/document correlations are inferred by latent topics. In PLSA and LDA, the unseen words and documents are not explicitly represented at the same time. Model generalization is constrained. This paper presents the Bayesian latent topic clustering (BLTC) model for document representation. The posterior distributions combined by Dirichlet priors and multinomial distributions are not only calculated in document level but also in word level. The modeling of unseen words and documents is tackled. An efficient variational inference method based on Gibbs sampling is presented to calculate the posterior probability of complex variables. In the experiments on TREC and Reuters-21578, the proposed BLTC performs better than PLSA and LDA in model perplexity and classification accuracy.

Index terms- Bayesian latent topic clustering, document model, Gibbs sampling, document categorization

1. Introduction

Language model captures the statistical regularities of language generation, which have been successfully applied in speech recognition and many other systems. It is well-known that the data clustering scheme is critical in language model and document retrieval. Brown et al. [3] proposed the word clustering and the class-based language modeling from the training data. Xu and Croft [12] grouped the documents into clusters and viewed each of the clusters as a topic representation. Also, the generative models of documents are developed for topic-based content representation [2][6][9]. The latent semantic analysis (LSA) [6] was presented and applied in n-gram language model [1][5]. The idea of LSA is to represent the words or documents in a low dimensional vector space consisting of the common semantic factors. All words and documents are mapped into the common semantic space, which is constructed via the singular value decomposition of a word-by-document matrix. More attractively, Hofmann [9] presented the probabilistic latent semantic analysis (PLSA) to model the aspects in documents in a probabilistic way that the document-word joint distributions are expressed conditionally on latent mixtures or topics. PLSA was applied to language modeling in [7]. The problem of PLSA is that the number of parameters grows linearly with the size of the document collection. Additionally, Blei et al. exploited the latent Dirichlet allocation (LDA) [2] method to deal with the problem of PLSA that the aspect models were only estimated for seen documents in training set. LDA was viewed as a Bayesian extension of unigram language model and extended for language model adaptation in [11]. The parameter estimation and learning were performed by the variational EM algorithm, since the exact solutions were intractable.

This paper presents a novel Bayesian latent topic clustering (BLTC) model where the word, topic and document variables and their hyperparameters are merged in document representation. A mixture model is used to represent the word clusters and the topic clusters so as to build the document model. Each mixture component is modeled by a multinomial distribution over the latent topics. Rather than LDA model generating topics solely based on the word distribution, the proposed BLTC model represents the clusters of documents by the low-dimensional features and identifies the clusters of words as the topic information. Also, BLTC is more general than PLSA owing to its capability in predicting new documents. Furthermore, our method uses the cluster as the computed groups and smoothes the parameters of unseen data. Instead of PLSA and LDA modeling in word space, BLTC method performs the document clustering in topic space. In this study, the non-parametric Bayesian modeling is employed in finding the approximate empirical posterior distribution of cluster variables based on the Gibbs sampling method [8]. This approach is efficient in implementation with small memory requirement. In the experiments, we conduct evaluation of perplexity and document classification using different latent topic models.

2. Probabilistic Topic Models

The generative document models have been a recent trend in machine learning, speech recognition and information retrieval.

2.1. Probabilistic latent semantic analysis

First of all, the probabilistic latent semantic analysis (PLSA) adopts the aspect model to represent the co-occurrence data associated with a topic or hidden variable and uses the maximum likelihood (ML) method to estimate model parameters. The graphical model of PLSA is shown in Figure 1(a). Let the text corpus D consist of document-word pairs \( \{d_m, w_n\}\) collected from \( M \) documents \( d_m \in \{d_1, \ldots, d_M\} \) with a vocabulary of \( N \) words \( w_n \in \{w_1, \ldots, w_N\} \). The joint probability of an observed pair \( (d_m, w_n) \) is represented by [9]

\[
P(d_m, w_n) = P(d_m) \sum_{z_k} P(w_n | z_k) P(z_k | d_m),
\]

assuming that \( d_m \) and \( w_n \) are independent conditionally on the associated topic \( z_k \). We accumulate log likelihood of overall training data \( \{d_m, w_n\} \) by
\[
\log P(d_m, w_n | \Theta) = \sum_{m=1}^{M} \sum_{n=1}^{N} n(d_m, w_n) \log P(d_m, w_n),
\]
where \(n(d_m, w_n)\) is the count of word \(w_n\) occurring in document \(d_m\) and \(\Theta\) is the PLSA parameter set. The new ML PLSA parameters \(\hat{\Theta}\) are obtained by [9]
\[
\hat{P}_{\text{PLSA}}(w_n | z_k) = \frac{\sum_{m=1}^{M} n(d_m, w_n) P(z_k | d_m, w_n)}{\sum_{n=1}^{N} \sum_{m=1}^{M} n(d_m, w_n) P(z_k | d_m, w_n)},
\]
where the posterior probability is calculated by
\[
P(z_k | d_m, w_n) = \frac{P(w_n | z_k) P(z_k | d_m)}{\sum_{l=1}^{L} P(w_n | z_l) P(z_l | d_m)},
\]
using the current estimates \(\hat{\Theta} = \{P(w_n | z_k), P(z_k | d_m)\}\). In [4], a Bayesian PLSA was presented to compensate the domain mismatch between training and test documents.

### 2.2. Latent Dirichlet allocation

Next, the latent Dirichlet allocation (LDA) method is used to model the latent topic of a text corpus. The basic idea is to view each document as a random mixture over latent topics, where the topic is characterized by a distribution over words. The graphical model representation of LDA is given in Figure 1(b). The major difference between PLSA and LDA is that the latent variables in PLSA are dependent on each document, while the topic mixture in LDA is drawn from a Dirichlet prior, which remains the same for all documents. The choice of conjugate prior is attractive for Bayesian inference. Using LDA [2], the words in a document \(w = \{w_n\}\) are assumed to be sampled from a random mixture over latent topics \(z\), where each topic is characterized by a distribution over words. The joint probability over the observed random variables is calculated by
\[
P_{\text{LDA}}(w, z, \alpha, \beta) = P(\alpha) \prod_{n=1}^{N} P(w_n | z_n, \beta) P(z_n | \alpha),
\]
where \(\beta\) is used in word probabilities conditioned on the topic and \(\alpha\) is topic mixture weight that drawn from a priori Dirichlet distribution with parameter \(\alpha\). The marginal distribution of a document is represented by summing out the hidden variable \(z\) and integrating over \(\alpha\) by
\[
P_{\text{LDA}}(w | \alpha, \beta) \propto \int P(\alpha) \prod_{n=1}^{N} P(w_n | z_n, \beta) P(z_n | \alpha) d\alpha.
\]
The inference problem is to compute the posterior distribution of hidden variables given a document \(w\) with parameters \(\alpha\) and \(\beta\)
\[
P_{\text{LDA}}(z, \theta | w, \alpha, \beta) \propto P_{\text{LDA}}(w, z, \theta | \alpha, \beta) / P(w | \alpha, \beta).
\]
However, the exact inference using this model is not available. Blei et al. [2] presented a variational method for approximate inference. The basic idea is to use a variational distribution for latent variables \((z, \theta)\)
\[
q_{\text{LDA}}(z, \theta | \eta, \phi) = q(\theta | \eta) \prod_{n=1}^{N} q(z_n | \phi_n),
\]
where \(\eta\) and \(\phi\) are variational Dirichlet parameter and multinomial parameter, respectively. The variational parameters were found by minimizing the Kullback-Leibler divergence between variational distribution and true posterior. The variational parameters were determined by
\[
\phi_{n,i} = \beta_{n,i} \exp\{\xi(\eta_i) - \zeta_i(\sum_{i} \eta_i)\},
\]
\[
\eta_i = \alpha_i + \sum_{i} \phi_{n,i},
\]
where \(\xi(\cdot)\) were the digamma function. Equations (10)-(11) were applied iteratively until convergence condition was met. In M-step, the conditional multinomial parameters \(\beta\) were updated by
\[
\beta_{i,j} \propto \sum_{m} \sum_{n} w_{mn} \phi_{m,n,j} w_m.
\]
Parameters of Dirichlet prior \(\alpha\) were estimated by the Newton-Raphson or gradient descent procedure. Nevertheless, the new document is likely to contain words that did not appear in any of the documents in a training corpus. The standard solution to this problem is to perform parameter smoothing for multinomial distributions. Alternatively, if the topic clustering is available, we are able to make predictions for new documents by using the seen documents, which correspond to the same cluster of new documents.

![Figure 1: Graphical representations of PLSA, LDA and BLTC](image)

3. Bayesian Latent Topic Clustering Model

This paper presents the Bayesian latent topic clustering (BLTC) model and draws upon the strengths of topic modeling and clustering.

### 3.1. New probabilistic topic model

BLTC is designed as a generative probabilistic model for document collection \(D\). The generative process of BLTC model for each document in a corpus is displayed in Figure 1(c). The distributions of underlying variables are defined by

1. Choose \(\psi_c \sim \text{Dirichlet}(\delta)\)
2. Choose \(\theta \sim \text{Dirichlet}(\alpha)\)
3. Choose \(\beta \sim \text{Dirichlet}(\gamma)\)
4. For each document \(d\)
   (a) Choose a cluster \(c\) and assume \(\psi_{\theta,c} = \psi_c\)
   (b) For each word \(w_n\)

2163
Given the parameters $\alpha, \gamma$ and $\delta$, the joint probability of all observed and hidden variables is expressed by

$$P_{\text{BLTC}}(w, z, \theta, \psi, \beta | \alpha, \gamma, \delta) = P(\theta | \alpha)P(\psi | \delta)P(\beta | \gamma) \prod_{n=1}^{N} P(w_n | z_n, \beta)P(z_n | \psi_n, \delta).$$

Integrating over $\theta, \beta$ and $\psi$ and summing up over $z$, the marginal probability of observing document $w$ is given by

$$P_{\text{BLTC}}(w | \alpha, \gamma, \delta) = \sum_{z} \prod_{n=1}^{N} P(w_n | z_n, \beta)P(z_n | \psi_n, \delta).$$

Finally, the likelihood of the complete corpus is calculated by taking the product of the marginal probabilities over all individual documents

$$P_{\text{BLTC}}(D | \alpha, \gamma, \delta) = \prod_{w \in D} P_{\text{BLTC}}(w | \alpha, \gamma, \delta).$$

### 3.2. Model inference

The inference problem in BLTC model is to compute a posteriori distribution of hidden variables $z, \theta, \psi, \beta$ given the observations $w$ and hyperparameters $\alpha, \gamma, \delta$:

$$P_{\text{BLTC}}(z, \psi, \theta, \beta | w, \alpha, \gamma, \delta) = \frac{P(w, z, \theta, \psi, \beta | \alpha, \gamma, \delta)}{P(w | \alpha, \gamma, \delta)}.$$ (16)

However, the exact inference on this model is intractable. Alternative methods have been used to estimate the parameters of topic model, including variational inference [2], expectation propagation [10], and Gibbs sampling [8]. EM algorithm tends to face local optimal problems [2]. Gibbs sampling is a relatively efficient algorithm for approximate inference in high-dimensional model. In this paper, we use Gibbs sampling, which estimates the joint probability distribution of complex variables from the observed data. Instead of estimating model parameters directly, we evaluate the posterior distribution on $z$ and use the result to infer $\theta, \psi$ and $\beta$. In each run of Gibbs sampling, a subset of variables is sequentially sampled from their distribution conditioned on the current values of all other variables. This operation proceeds until the sampled values approximate the target distribution. We begin with the joint distribution $P(w, z, \theta, \psi | \alpha, \gamma, \delta)$ and calculate the conditional probability $P(z_i | \psi_i, w, z_j, \psi_j, \alpha, \gamma, \delta)$ using the chain rule by

$$P(z_i | \psi_i, w, z_j, \psi_j, \alpha, \gamma, \delta) \propto P(w_i | \psi_i, z_i, \psi_j)P(z_i | \psi_i, \psi_j)P(\psi_i | \psi_j)P(w_j | \psi_j, z_j).$$ (17)

The multinomial distribution is assumed to have a Dirichlet prior with hyperparameters $\alpha, \gamma$ and $\delta$. Here, $z_i, z_j$ denotes the topic assignments $z_i$ for all words not including the current word $i$ and $\psi_i, \psi_j$ represent the cluster for all words except current word $i$. In Gibbs sampling procedure, we draw iteratively a topic assignment $z_i$ and cluster assignment $\psi_i$ for each word $w_i$ in the corpus according to the conditional probability distribution

$$P(z_i | w_i, z_j, \psi, \psi_j, \alpha, \gamma, \delta) \propto P(w_i | \psi_i, z_i, \psi_j)P(z_i | \psi_i, \psi_j)P(\psi_i | \psi_j)P(w_j | \psi_j, z_j).$$ (18)

The relationship of PLSA, LDA and BLTC models can be demonstrated by Figure 1. Using PLSA, each word is assumed to be a sample from a mixture model and a document is assumed to be generated from different topics. However, PLSA is not a generative model for documents so that it does not provide the way to assign probabilities to documents, which are unseen in training corpus. LDA is an extension of PLSA, where the topic mixture parameters are drawn from a Dirichlet conjugate prior. The topic mixture in LDA is only sampled for current document as given in equation (6). LDA is not used to model similar documents with unseen words. This study proposes the clustering of documents in topic space instead of word space. With the knowledge from a group of related topics, the unseen information in new document can be predicted. The clustering of latent topics provides some useful model information. As reported in Table 1, in terms of space complexity, PLSA spends more parameters than BLTC. Given the same $K$ latent variables, PLSA model has $K$ multinomial distributions of size $V$ and $M$ mixtures over the $K$ latent topics. Therefore, PLSA requires the storage $KM + KV$ parameters, which is linearly growth by $M$. LDA with parameters $(\alpha, \beta)$ needs the storage $K + KV$ parameters and does not grow with the size of training corpus. BLTC model needs the storage of $C + K + KV$ parameters, which is moderate in these models.

### 4. Experiments

<table>
<thead>
<tr>
<th>No. of Parameters</th>
<th>PLSA</th>
<th>LDA</th>
<th>BLTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM+KV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K+KV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C+K+KV</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We evaluated the performance of PLSA, LDA and BLTC models by using two public-domain document collections: Associated Press Newswire (1988) dataset and Reuters-21578. The model perplexity and the document categorization were evaluated. We performed preprocessing stages of stemming and stop word removal for all documents. We picked up all the words that occurred at least in five documents in Reuters-21578. In the experiments, we fixed the Dirichlet priors $\gamma = 0.01$, $\delta = 0.1$ and $\alpha = 50/K$ with 500 Gibbs sampling iterations.

4.1. Evaluation of model perplexity

When evaluating the model perplexity, we used AP88 dataset containing 79,919 documents with the size of vocabulary being 26,208 words in model estimation. The dataset were held out 10% for test purpose. The models were trained by the remaining 90%. Perplexity was used to measure the average word branching factor of a document model. The lower the perplexity is, the smaller the source uncertainty that is obtained to achieve better modeling. In this set of experiments, the number of clusters is fixed to be 10 and the number of latent topics $k$ is set to be 2, 16, 32, 64 and 128. Figure 2 displays the perplexity results for different numbers of latent topics using unigram language model (denoted by LM), PLSA, LDA and BLTC methods. We can see that the baseline language model has the perplexity of 6594.61. The latent topic models perform better than LM without merging topic information. BLTC obtains lower perplexity compared to the other models especially when the amount of latent topics increases. BLTC reduces perplexity relative to LM by a factor of 2.11 (6594.61/3121.46), while PLSA and LDA achieve a reduction factor of 1.82 and 1.69, respectively.

![Figure 2: Comparison of perplexity using PLSA, LDA and BLTC with different numbers of latent topics](image)

4.2. Evaluation on document categorization

In evaluation of document categorization, we used Reuters-21578 dataset consisting of 8,113 documents and 12,886 words in the vocabulary. Samples of five most populous categories were divided into two subsets which three-fourth documents were used for model training and the other one-fourth documents were used for test. The largest category is “earn” having documents related to the domains of earnings. In this set of experiments, the numbers of clusters and topics are fixed to be 10 and 64, respectively. Precision, recall, and $F$-measure are used as the effectiveness measurements for text categorization. The results are shown in Table 2. The precision of BLTC is lower than other models. However, BLTC shows a good improvement in terms of recall and $F$-measure.

<table>
<thead>
<tr>
<th></th>
<th>PLSA</th>
<th>LDA</th>
<th>BLTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>88.59</td>
<td>88.87</td>
<td><strong>90.62</strong></td>
</tr>
<tr>
<td>Precision</td>
<td>99.59</td>
<td><strong>99.69</strong></td>
<td>98.30</td>
</tr>
<tr>
<td>$F$-measure</td>
<td>93.77</td>
<td>93.97</td>
<td><strong>94.30</strong></td>
</tr>
</tbody>
</table>

5. Conclusions

This paper presented a generative model that used low-dimensional features for document representation and grouped the words into topics. The Bayesian latent topic clustering model was proposed and illustrated to be better than PLSA and LDA methods in evaluation of model perplexity and document categorization. Efficient variational inference based on Gibbs sampling algorithm was developed for empirical Bayesian parameter estimation. Experimental results showed the promising performance by using the proposed model. The ongoing work is the extension of proposed BLTC method to language model adaptation and spoken document retrieval.

6. References