International Journal of Hybrid Information Technology
Vol.8, No.12 (2015), pp. 269-282
http://dx.doi.org/10.14257/ijhit.2015.8.12.20

# On-Line labeled Topic Model Based on Global and Local Topic

YongHeng Chen, Yaojin Lin and Hao Yue

*College of Computer Science, Minnan Normal University, zhangzhou 363000, China*
*Response author: Yongheng Chen email: cyh771@163.com*

## *Abstract*

*A large number of electronic documents are labeled using human-interpretable annotations. High-efficiency text mining on such data set requires generative model that can flexibly comprehend the significant of observed labels while simultaneously uncovering topics within unlabeled documents. This paper presents a novel and generalized on-line labeled topic model based on global and local topic (GL-OLT) tracking the time evolution of topics in a sequentially organized multi-labeled corpus. GL-OLT topic model has an incrementally update principle based on time slices by an on-line fashion, and each label has not only a set of local topics, but also has several global topics. Empirical results are presented to demonstrate significant improvements accuracy of label predictive, and lower perplexity and high performance of our proposed model when compared with other models.*

***Keywords:*** *Text Information Processing, Latent Dirichlet Allocation (LDA), Topic Modeling, Natural Language Processing*

## 1 Introduction

As peoples began memorizing their documents in digital, Information Retrieval as a domain were rose in computer science. Information Retrieval increased with the wide extends application of web technology. Peoples need find relevant web pages that are satisfying their information need from millions of web pages on the internet through a convenient and efficient way. Describing the feathers of documents' content is a typical problem emphasized in Information Retrieval. Researchers usually employ the characteristic of the content of document to search, form, or classify the corpus.

Recently, generative models for corpus have been used to detect topic-based content presentations; each document is modeled as a mixture of probabilistic topics. Probabilistic Latent Semantic Indexing (PLSI), a statistical generative model, is proposed by Hofmann, which is one of topic model and employs topics represented by latent variables to connect documents and words [1]. A document is regarded as a mixture of topics. The content, the words in a document, can be produced presented the small set of topics (or latent variables). Reversing this process, *i.e.*, matching the generative model to the words in training set, equivalent to deducing the latent variables and, therefore, inferring the potential topics' distributions. Blei, *et al.*, proposed Latent Dirichlet Allocation (LDA), which develops the generative model to accomplish the ability of concluding generalizing the topic distributions so that research can also employ LDA model to create unseen document [2]. The achieving success of Latent Dirichlet Allocation in the research field far exceeds the domain of Information Retrieval. There has been a wide range of applications employed LDA model in relevant area, for instance, multimedia classification and data mining.

Though LDA is enough to model multi-topics per document, this model is not suitable for labeled corpora since, as an unsupervised model, it gives unconspicuous pattern of

ISSN: 1738-9968 IJHIT

integrating a supervised label collection into its learning process. In order to incorporate supervised labels, some alterations of Latent Dirichlet Allocation are been put forward in existing literature. If employing LDA model to get this purpose, it is generally to improve the capability of a collection of latent topics, such as in [3, 4], instead of modeling the supervised labels set into model's learning analysis. Latent Semantic Indexing [5] and related methods [6] are also popular unsupervised approaches. While unsupervised label topics are suited to acquire wider patterns in corpus, the trained topics do not usually align with human provided labels. On the contrary, the supervised Latent Dirichlet Allocation model emphasizes the prediction issue through deducing the most predictive potential topics of document paired with a response [7]. The Dirichlet-multinomial regression (DMR) topic model is put forward by Mimno, *et al.*, which includes a log-linear prior on the document-topic distributions, where the prior is a function of the observed document features [8]. The essential difference between Dirichlet-multinomial regression and supervised Latent Dirichlet Allocation is that, while supervised Latent Dirichlet Allocation model regard observed characteristic as generated variables, Dirichlet-multinomial regression treats the observed characteristic as a set of conditioned variables. However, these models are single label document supervision and learning algorithm and cannot be applied to multi-labeled corpus. Recently multi-Labeled generative model referred as LLDA is proposed by Daniel Ramag, *et al.*, in 2009, which bounds LDA model by demonstrating a one-to-one communication between this model's topics and labels [9]. Labeled LDA considers every document is tagged using a collection of provided labels, and that these labels play a direct role in generating the document's words from per-label distributions over terms [9]. This can enable Labeled LDA to directly learn word-tag correspondences. However, the presence of any potential topics is not assumed by Labeled LDA. Besides, labeled LDA ignores the difference between the topics of computer recognition and artificial labels, which leads to model's insufficient fitness with document's data and poor generalization ability.

Moreover, big data analysis with above-mentioned topic models can be computationally difficult. A primary study challenge for topic modeling is to efficiently serve models to big corpora. These motivated researchers to look for an optimization model, and ultimately several online topic models have been proposed. Alsumait, *et al.*, (2008) proposes Online Topic Model (OLDA) that copes with documents in an on-line way through resampling topic distributions for documents from the new stream updating parameters. OLDA uses collapsed Gibbs sampling for approximate inference [10]. Hoffman, *et al.*, (2010) adopted an online LDA variational Bayes as the approximate posterior inference algorithm that can analyze massive collections of documents for Latent Dirichlet Allocation (LDA) [19]. However, in order to identify novel topics and analyze the evolution of them, OLDA model need measure between the word distribution of each topic before and after an update using Kullback Leibler or Jensen-Shannon divergence [19]. With the increasing of identified topics, the performance of topic will be greatly affected. In addition, OLDA model only consider static vocabulary across time.

These motivated us to look for an optimization model, and ultimately the novel and generalized on-line labeled topic model based on global and local topic, namely GL-OLT, has been proposed. GL-OLT model implements multi-labeled learning of document through mapping label to the combination of multiple topics, similar to LLDA. However, in order to eliminate the similarity of topics classified differently label, in GL-OLT model each label has not only a set of local topics, but also has several global topics. For example, a global topic that belong simultaneously to "machine learning" and "data mining" label added into "machine learning" and "data mining" label. So GL-OLT model make the generated topics far more independence and distinguish, and is feasible and valid for the similarity and dependency of topics. In order to further improve performance of corpus with time information, the time information is considered and incorporated into GL-OLT model.

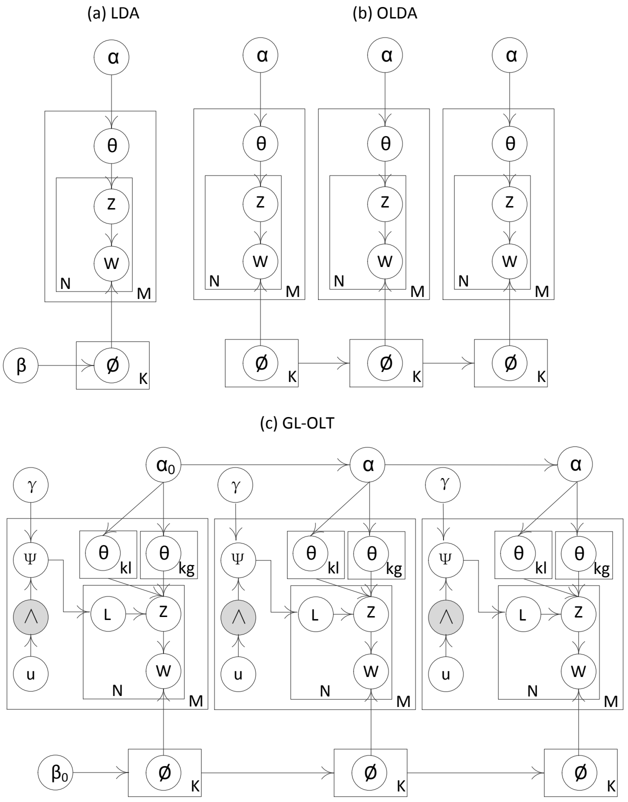


**Figure 1. GL-OLT is shown with Two Inspiration Models: (a) LDA Model (b) OLDA Model (c) GL-OLT Model**

**Table 1. Notation used in the Paper**

| SYMBOL | DESCRIPTION |
|---|---|
| $K_L$ | number of local topics |
| $K_g$ | number of global topics |
| M | Total number of documents |
| N | Total number of unique words |
| $N_d$ | number of word tokens in document d |
| $\gamma$ | dirichlet priors for Ψ |
| $\alpha_d$ | K-vector of dirichlet priors for ducument d |
| $\beta_k$ | V-vector of dirichlet priors for topic k |
| $\theta_d$ | the multinomial distribution of topics specific to the document d |
| $\emptyset_k$ | the multinomial distribution of words specific to the topic k |
| $\Psi_d$ | the multinomial distribution of labels specific to document d |
| $\wedge$ | a sparse binary vector of usable labels |
| L | a space of label |
| $Z_{di}$ | the topic associated with the ith token in the document d |
| $W_{di}$ | the ith token in document d |

To my best of our knowledge, we are the first to deal with a sequentially organized corpus of documents associated with labels by an on-line fashion. The outline of this paper is organized as follows. In the next section, our on-line labeled topic model is introduced. In Section 3, approximate variation inference is given. We present the experiments we carry out on the basis of Xinhua News corpus and the results we obtained in Section 4. Our final conclusions and suggestions for future work are discussed.

# 2. Methodology

## 2.1. Modeling Documents with Topics

Before presenting on-line labeled topic model based on global and local topic (GL-OLT) model, let us review the basic Latent Dirichlet Allocation model. A glossary of notations used in the paper is summarized in Table 1, and the graphical model representations of our GL-OLT model is shown in Figure 1.

LDA has acquired prevalence among theoreticians and practitioners serving as a way for corpus summarization and visualization. LDA model, a completely unsupervised algorithm that models each document as a mixture of topics, generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics [18]. Most importantly, LDA makes the explicit assumption that each word is generated from one underlying topic [18]. Latent Dirichlet Allocation model, Figure 1 (a), is a hierarchical Bayesian network that product a document making use of a mixture of topics. The parameters θ and Ø, the topic and document distributions are conditionally separation because of the words that can be observed. Moreover, the direct connection between documents and words is interrupted [11]. Rather, by introducing additional potential variable z, which represents the responsibility of an especial topic in employing that word in the document, *i.e.,* the topic(s) that the document is centralized on, the connection is managed [12]. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics. The document and topic distributions adopt respectively α and β as the Dirichlet priors. To deal with unseen documents the generative model of LDA is generalized. For LDA model the words in a document is exchangeable, this is the same for the documents in a corpus. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics [13]. LDA produced documents through selecting a distribution on topics θ from a Dirichlet distribution. P (z) is determined by θ for words in document. The words in the document are then generated by picking a topic j from this distribution and then picking a word from that topic according to P(w|z=k), which is determined by a fixed Øk [14]. The estimation problem becomes one of maximizing P (w|Ø, α) =∫P(w|θ, Ø)P(θ|a)dθ,where P(θ) is a Dirichlet(a) distribution. θ and Ø are usually estimated by using sophisticated approximations, either Gibbs sampling or Variational Expectation-Maximization [15].

Latent Dirichlet Allocation model regards words as interchangeable is a simplification that it is compliance with the target of recognizing the semantic topics for each document. For a large number of sets of interest, but nonetheless, the supposition of interchangeable documents is improper [16]. For many data sets, such as news articles, research papers, the sequence of the documents implies a trending collection of topics. OLDA model, Figure 1(b), treats the temporal sequencing information and regards that the data sets are distributed by time slices, and documents as interchangeable within each time slice. OLDA model extends LDA model to operate in an on-line fashion. OLDA model the newly arrived data sets within each slice by a k components topic model, where the generated model associated with slice t-1 is employed as a prior for LDA at slice t, as a newly arrived data stream is available for disposing. OLDA model regards hyper-parameters β as the prior observation counts on the number of times words are sampled from a topic before any word from the corpus is observed [10]. Therefore, model runs LDA on data set at current time, and then constructs the count of words in topics that can be employed as the priors for the successive data stream.

### 2.2. GL-OLT Model

While OLDA is expressional plenty to reveal model's topics of documents related with a sequential distribution over time slice, it is not adequate for modeling data set paired with labels, since, as an unsupervised model, it do not has the capabilities to provide mean of incorporating meta-data into training procedure. This simulated us to optimize and extend OLDA model to combine and handle labeled document collection by on-line style. We propose on-line labeled topic model based on global and local topic (GL-OLT) model (see Figure1(c)), which is a generative model for sequentially organized data set of labeled-documents. GL-OLT model has the ability to update itself incrementally based on time slices by an on-line fashion, on the application side, find the latent low-dimensional structure of labeled-document and analyze dynamic evolution for the classified topics by labels.

GL-OLT model is separated with time slice, a disconnection period, *e.g.*, a day, or a year and documents are distributed for different time slices according to their time. The received sequence of documents within a time slice T are expressed as $S_T$= $\{d_1,\ldots,d_N\}$, where N is the number of received documents within T. In the section below, there are two ways of describing its generative process: firstly, the static portion within GL-OLT model that is considering a single time slice only, and for the second, the dynamic portion within GL-OLT model that is the migration of model over time. In our formalization of GL-OLT model, a document d contains a multi-set of words $w_d$ from a dynamic vocabulary V and a set of labels $\wedge_d$ from a space of labels L (indexed by 1…L), each of label has been related with topics $K_l$ (indexed by $1\ldots K_L$), $Z_l$ = $\{z_{l1}, z_{l2}, \cdots ,z_{lkl}\}$, and where every topic $Ø_{l,k}$ is expressed by a multinomial distribution over $w_d$ constructed by a symmetric Dirichlet prior β. Here the topics included in $Z_l$ denote local topic. In addition to local topics, there is global topics kb, Z' = $\{z_1', z_2', \cdots ,z_{kb}'\}$. The topics included in Z' are global topics, which can been related to all labels. Local and global topics distributions adopt α as the Dirichlet priors. Supposing the being of a general latent label L that can be used to total documents within the cluster, latent topics that have not been stuck any label are optionally modeled. ɣ denotes the ∧d is constructed by a prior ɣ. But because each document's label-set ∧d is observed, its sparse vector prior ɣ is unused. In GL-OLT model it is included for completeness.

GL-OLT model supposes that each local topic can only participate in one label, and each global topic can participate all labels. In GL-OLT model, we presume that a document d is constructed as following. An explicit sub-collection of usable labels is expressed by a sparse binary vector $\wedge_d$. A document-explicit mix θ over local and global topics is constructed by a symmetric Dirichlet prior with hyper parameters α for every label included in $\wedge_d$, and then ψ is constructed by Dirichlet ( ɣ ). Each word w within a document d is constructed by some label's topic's word distribution, *i.e.*, a label l from $\psi_d$ is selected firstly and then a local $Z_l$ or global topic Z' is chosen from $\theta_{d,l}$ that have correlation with that label l. Here supposes that the selection probability of $Z_l$ is pz, and z' is 1-pz. A variable identifier t∈{0, 1} is added. When t is equal one, the word w is constructed by local $Z_l$, otherwise by global Z'. In this paper we realize the value of t through Bernoulli distribution. Finally, word w is selected from $Ø_{l,k}$, where k is from $Z_l$ if t is one, otherwise Z'. The static portion of GL-OLT is shown as following.

```
For each topic k ∈ {1, . . . ,K}:
    Generate $Ø_k=(Ø_{k,1},...,Ø_{k,v})^T$ ~Dir(.|β)
For each document d:
    Select labels of documents ∧d
    Generate $\Psi^{(d)}=(\Psi_{d,1},...,\Psi_{d,L})^T$~Dir(.|∧d)
    Generate topic distribution $\theta_{(d)}=(\theta_{d,1},...,\theta_{d,Kl})^T$ ~Dir(.|α, ∧d), where $\wedge_{(l)}$=1
    Generate topic distribution $\theta'_{(d)}=(\theta_{d,1},...,\theta_{d,Kb})^T$ ~Dir(.|α, ∧d)
```

```
  For each word i in $\{1,\ldots,N_d\}$
     Generate $l_i$~Multinomial($\Psi_{d,1},...,\Psi_{d,L}$)
Generate ti~Bernoulli(pz)
If(ti==1)
  Generate $z_i$~Multinomial($\theta_{(d)}$)
  Generate $w_i$~Multinomial($Ø_{k,zi,1},...,Ø_{k,zi,V}$)
Else
  Generate $z'_i$~Multinomial($\theta'_{(d)}$)
  Generate$w_i$~Multinomial($Ø_{k,}z'_{i,1}, \ldots ,Ø_{k,}z'_{i,v}$)
```

For the dynamic portion, GL-OLT model treats the temporal sequencing information and regards that the data sets are distributed by time slices, and documents as interchangeable within each time slice. GL-OLT model extends LLDA model to cope with documents in an on-line way through resampling topic distributions for documents from the new stream updating parameters. GL-OLT model the newly arrived data sets within each slice by a k components topic model, where the generated model associated with slice t-1 is employed as a prior for LDA at slice t, as a newly arrived data stream is available for disposing. GL-OLT model regards hyper-parameters $\beta$ as the prior observation counts on the number of times words are sampled from a topic before any word from the corpus is observed. Therefore, model runs the static portion of GL-LDA on data set at current time, and then constructs the count of words in topics that can be employed as the priors for the successive data stream.

In the dynamic portion $\kappa$ is represented as a sliding window that comprises a fixed number of time slices and preserves a dynamic vocabulary, where each word relates to the total number in documents contained in current sliding window [10]. In order to keep $\kappa$ constant, the documents partitioned into first old time slice are deleted when documents within a new time slice T appear. In this process, vocabulary will be updated by decreasing the total number related with words in documents contained within time slice T-1 by one. If the total number equals zero, the corresponding word will be deleted. Further update of vocabulary is performed on words in documents within time slices T by the reverse operation. This serves two purposes. The first cause is that the consecutive model will grow indefinitely over time and become less sensitive to modified topic if documents partitioned into different time slices are all saved. So we use sliding window to control of excessive growth over time. The second cause is optimizing LLDA model through considering dynamic vocabulary across time. Because a fixed vocabulary is considered by LLDA, this assumption is not rational for a practical online topic model, where it is impractical to pre-compute the vocabulary in advance. Therefore, GL-OLT model further optimizes LLDA by re-generating vocabulary when adding documents in new time slice to sliding window. The dynamic portion of GL-OLT is shown as following.

```
set initial values $\alpha_0$, $\beta_0$; topic number K; contribution factor $\rho$; and window size $|k|$;
Iterative step for new time slice $T_{new}$:
append $T_{new}$ to sliding window;
if the number of time slices sliding window contains less than $|\kappa|$
  update vocabulary in sliding window by adding different words within $T_{new}$
else
  remove time slice that is first old in sliding window
  update vocabulary in sliding window by
   (a) removing words in documents within the removed time slice
   (b) adding new words within $T_{new}$
Calculate priors $\alpha'$ and $\beta'$
Resample z using $\alpha'$ and $\beta'$ for documents in sliding windows
```

When documents for time slice T+1 arrive, the current GL-OLT model is incrementally updated by representing topic distributions z for all documents within time slices included in sliding window κ, employing ψ, θ and Ø inferred according to previous model in time slice T to construct hyper-parameters α' and β', the Dirichlet priors, for current GL-OLT model. Let ρ be a contribution which determines its degree of contribution of previously known parameters in computing the priors of the successive new mode [13]. The Dirichlet hyper-parameters of a topic t in current time slice T+1 can be obtained as follows:

$$\alpha_t^{T+1} = \alpha_0 \times \frac{D_T}{N_T} \times n_{.,t}^{(ki)} \qquad \beta_t^{T+1} = \beta_0 \times (1 + \rho \times (\frac{V_{T+1}}{N_T} \times n_{t,.}^{(vi)} - 1))$$

where $\alpha_t^{T+!}$ and $\beta_t^{T+!}$ are natural parameter for topic t in time slice T+1; $N_T$ and $D_T$ are the number of documents included into sliding {T-|κ|,...,T} and the number of tokens contained those documents, respectively; $V_T$ is the number of vocabulary in sliding {T-|κ|,...,T}; $n_{.,t}^{(ki)}$ and $n_{t,.}^{(vi)}$ are the total number of topic t related with documents in slice T and the total number of words in those documents related with topic t.

Because the topic is the distribution of words, the label of topic can be calculated by the label of words, and furthermore a unlabeled-document can be labeled by its topics' label. A time slice can be predicted given the words within the document. In general, tracking the evolution of captured topics only need taking into consideration topics associated with same label. For example, if we want to measure the trending of football topic within someone time slice, we only calculate the distance of the captured topics associated with label of sport within that time slice using KL or JS divergence. So in order to further improve performance of tracking topics' evaluation, GL-OLT model can assign the detected topics according to associated label and adopt multithread technology to implement parallel computing.

## 3. Approximate Variation Inference

We have represented the intention behind GL-OLT model and clarified its conceptual benefits over other models. In this section, assuming there is a sliding window κ that comprises a fixed number of time slices, {T-|κ|-1,...,T-1}, after a new time slice T+1 arrives, we shift our attention to process for inference and parameter estimation under GL-OLT model, *i.e.,* find parameters within time slice T+1.

Gibbs sampling is a Markov chain Monte Carlo(MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariateprobability distribution (*i.e.,* from the joint probability distribution of two or more random variables), when direct sampling is difficult. Complexity reduction to be done through the Gibbs sampling algorithm let us transform parameter calculation question into a not complicated counting and sampling course. In this section, we employ Gibbs sampling to perform approximate inference. Through this process, we will obtain parameters θ, the multinomial distribution of topics specific to the document's labels, Ø, the multinomial distribution of words specific to label's topic, and ψ, the multinomial distribution of labels specific to document.

We are interested in finding an efficient way to compute the joint likelihood of the observed word w with the unobserved label and topic assignments, *i.e.,* find the special distribution of parameters θ, Ø and ψ that maximize (marginal) likelihood of p(z, w,l,t|ɣ , α, β,Λ , pz). This Likelihood probability can been decomposed:

p(z, w,l,t|ɣ , α, β,Λ , pz)= p(w|z,β)P(z,l,t|α,Λ , pz, ɣ )

This joint likelihood can be used to derive efficient updates for parameters θ, Ø and ψ. First, the probability p(w|z,β) can be decomposed into∫p(w,Ø|z,β)dØ. The detailed derivation process of it can be computed as following:

$$
\begin{aligned}
&\int p(w,\varphi \mid z,\beta) d\varphi \\
&= \int p(w \mid z,\beta,\varphi)\, p(\varphi \mid \beta) d\varphi \\
&= \int \prod_{k=1}^{K} \Big(\prod_{v=1}^{V} \varphi_{k,v}^{n_{k,v}}\Big) \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \Big(\prod_{v=1}^{V} \varphi_{k,v}^{\beta_v - 1}\Big) d\varphi \\
&= \prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \frac{\prod_{v=1}^{V} \Gamma(\beta_v + n_{k,v}^{(vi)})}{\Gamma(\sum_{v=1}^{V} \beta_v + n_{k,v}^{(vi)})} d\varphi
\end{aligned}
\tag{1}
$$

Using the model's independence assumptions, we expect the further expansion of remaining joint probability P(z,l,t|α, Λ, pz, ɣ ) is showed as P(z,l,t|α, Λ, pz, ɣ )= P(l| ɣ ,Λ)P(z|l, t, α)P(t|pz). Firstly, let's consider to the probability P (l|ɣ ,Λ), which, on account of Λ is supposed known, could be decomposed into:

$$
\begin{aligned}
&P(l \mid \gamma, \wedge) \\
&= P(l \mid \wedge, \psi) P(\psi \mid \gamma, \wedge) \\
&= \int_{\Phi} \prod_{d=1}^{M} P(\psi_d \mid \gamma, \wedge_d) \prod_{i=1}^{W_d} P(\mathrm{l}_{d,i} \mid \wedge_d, \psi_d) d\Phi \\
&= \prod_{d=1}^{M} \prod_{j \in \wedge_d} \frac{\Delta(\mathrm{n}_{d,:,.,.} + \gamma)}{\Delta(\gamma)}
\end{aligned}
\tag{2}
$$

Here we introduce the notation in [17] where ▽(x) is defined as following:

$$
\Delta(\mathrm{x}) = \frac{\prod_{k=1}^{\dim x} \Gamma(\mathrm{x}_k)}{\Gamma(\sum_{k=1}^{\dim x} \mathrm{x}_k)}
$$

Now let's return to the computation of P(z|l, α) and p(t| pz). These can be computed as equation (3) and (4).

$$
\begin{aligned}
&P(z \mid l, t, \alpha) \\
&= \int_{\theta} P(\mathrm{z} \mid \mathrm{l}, t, \theta)\, \mathrm{P}(\theta \mid t, \alpha)\, \mathrm{d}\theta \\
&= \int_{\theta} \prod_{d=1}^{M} \prod_{i=1}^{M_d} P(\mathrm{z}'_{d,i} \mid \mathrm{l}_{d,i}, \theta'_{d,i})^{1-t_{d,i}} P(\mathrm{z}_{d,i} \mid \mathrm{l}_{d,i}, \theta_{d,l_{d,i}})^{t_{d,i}}\, \mathrm{P}(\theta \mid t, \alpha)\, \mathrm{d}\theta
\end{aligned}
\tag{3}
$$

$$
= \int_{\theta} \prod_{d=1}^{M} \prod_{i=1}^{M_d} (\theta_{d,l_{d,i},z_{d,i}})^{t_{d,i}} (\theta'_{d,i})^{1-t_{d,i}}\, \mathrm{P}(\theta \mid t, \alpha)\, \mathrm{d}\theta
$$

$$
P(t \mid pz) = (pz)^{nt=1} (pz)^{nt=0} \tag{4}
$$

In Gibbs sampling a MCMC is organized to have a particular stationary distribution. In this paper, we want to acquire a MCMC that can converge to the posterior distribution over z given training set M, α and β. Obtaining a sampling from the distribution p (l,z,t|M, α, β,), using Gibbs sampling by (a) sampling a topic distribution $z_{d,i}$ and label distribution $l_{d,i}$ for an individual word $w_{d,i}$, and (b) iterating this procedure for every word. In equation (5) we show how to derive the basic equation though equations (1-4) needed for the Gibbs sampler when t is equal to one.

$$
\begin{aligned}
&p(l_{d,i} = j, z_{d,i} = k, t_{d,i} = 1 \mid l_{\neg d,i}, z_{\neg d,i}, w_{d,i} = v, \gamma, \alpha, \beta) \\
&\propto I[j \in \wedge_d, k \in 1..K_j] \left( \frac{\alpha + n^{(\neg d,i)}_{d,j,k,.}}{k_j \alpha + n^{(\neg d,i)}_{d,j,.,.}} \right) \\
&\left( \frac{\beta + n^{(\neg d,i)}_{.,j,k,t}}{V\beta + n^{(\neg d,i)}_{.,j,k,.}} \right) \left( \frac{n^{(\neg d,i)}_{d,j,.,.} + \gamma}{n^{(\neg d,i)}_{d,.,.,.} + \sum_{j' \in \wedge d} \gamma} \right) pz
\end{aligned}
\tag{5}
$$

When t equal zero, the basic equation though equations (1-4) needed for the Gibbs sampler is shown as (6).

$$
\begin{aligned}
&p(l_{d,i} = j, z_{d,i} = k, t_{d,i} = 0 \mid l_{\neg d,i}, z_{\neg d,i}, w_{d,i} = v, \gamma, \alpha, \beta) \\
&\propto I[j \in \wedge_d, k \in 1..K_g] \left( \frac{\alpha + n^{(\neg d,i)}_{d,j,k,.}}{k_j \alpha + n^{(\neg d,i)}_{d,j,.,.}} \right) \\
&\left( \frac{\beta + n^{(\neg d,i)}_{.,j,k,t}}{V\beta + n^{(\neg d,i)}_{.,j,k,.}} \right) \left( \frac{n^{(\neg d,i)}_{d,j,.,.} + \gamma}{n^{(\neg d,i)}_{d,.,.,.} + \sum_{j' \in \wedge d} \gamma} \right) pz
\end{aligned}
\tag{6}
$$

## 4. Experiments

In this paper we use data sets from Xinhua News Agency for evaluation. An insufficient expression of the data sets adopted in our experiments is presented. Xinhua News corpus includes of news articles and ordered by time. For our experiments, we use the data set contains articles from Xinhua News between 2013, 6, 1 to 2013, 11, 30. We only extract eight classifications consist of arts, business, and health and so on. These data sets were preprocessed for down-casing, deleting extremely common words and numbers, and removing the words that frequency less than six times in the data set. Each document associated with classification labels has a time stamp that is determined by the day. In our experiments, the time slice and sliding windows are set to1 day and 3 days. The hyper-parameter $\alpha_0$ is set to 0.002, a small initial value of α is employed to construct a sparse topic distribution over documents. 0.02 is adopted for $\beta_0$.

### 4.1. Perplexity for Different Models

The density measurement, expressing the potential configuration of data, is the intention of document modeling. Measuring the model's universal performance on formerly unobserved document is general method to estimate that. Perplexity is a canonical measure of goodness that is used in language modeling to measure the likelihood of a held-out test data to be generated from the potential distributions of the model [4]. Better universal performance is manifested by a lower perplexity and the higher the likelihood on a held-out document. Formally, for a test set of M documents, the perplexity is [14]:

$$perplexity(M^{text}) = \exp\left\{ - \frac{\sum_{d=1}^{M} \sum_{v=1}^{V} n_m^{(t)} \log(\sum_{k=1}^{K} \beta_{k,v} \theta_{m,k})}{\sum_{d=1}^{M} N_d} \right\}$$

GL-OLT model is the extended model which is based on improved labeled LDA and modified online topic model. In experiments, we compared Online Topic Model (OLDA), Labeled Latent Dirichlet Allocation (LLDA), Latent Dirichlet Allocation (LDA), and our presented on-line label topic model (GL-OLT). The averaged perplexity as a function of the number of topics trained on XinHua news is presented in Figure 2.

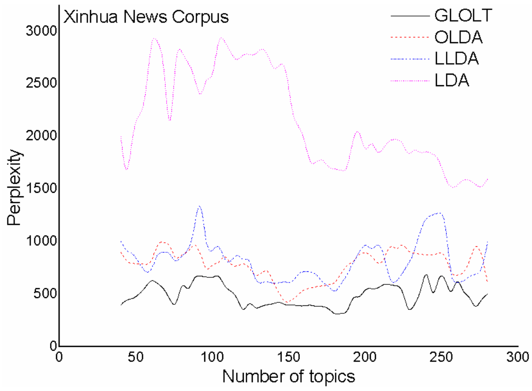


**Figure 2. Comparisons of Perplexity**

OLDA and GL-OLT model all explore new ways of incorporating metadata of the time stamp information into their models. However, we proposed model using knowledge of the labels associated with document to extend a better prior for the data set based on OLDA. So GL-OLT further develops the generalization performance of OLDA model to document associated with labels. One can see shown in Figure 1 that that making use of the label information significantly improves the predictive log-likelihood.

In addition, the documents clustering of the standard LDA model can be seen as the clustering operation based on topics settled by user. Since LDA model do not make use of the label or time information for Labeled corpus ordered by time, which limits its generalization performance, LDA is distinctly not better than either of topic-based models, as demonstrated by its high perplexity. As shown in Figure 1, the ability of document clustering of GL-OLT model and LLDA always performs better than LDA model. So in most cases, the effective use of documents' labels or time information can improve the effect of document clustering.

During the process of document clustering, GL-OLT and LLDA model, viewed as partially supervised clustering model, all adopt new way, which is incorporating label information within corpus into their models, to promote the generation of document's topics. So the effect of document clustering is obviously improved with respect to LDA model. However, GL-OLT model further develops the generalization performance of document clustering of LLDA model through making use of the time information within corpus. One can see shown in Figure 1 that that making use of the time information significantly improves the predictive log-likelihood.

### 4.2. Classification Accuracy

The distribution of a document over topics can be considered a reduced description of the document in a new space spanned by the small set of latent variables [2]. So, the performance of the topic model can be evaluated by investigating the amount of discriminative information that is preserved in the document distributions. One way to do this is through solving a classification problem [10]. For evaluation, classification accuracy is common measure. This means that we need definitely know the classification label of topic. Considering the topic is the distribution of words, the determination of label for topic can be done by the label of words associated for topic. But every word has different weight for label, according the discrimination of the word to document in tfidf, we take into account tfidf during calculating the discrimination of word to label, combine the word frequency to obtain the relationship matrix of word to response variable as following:

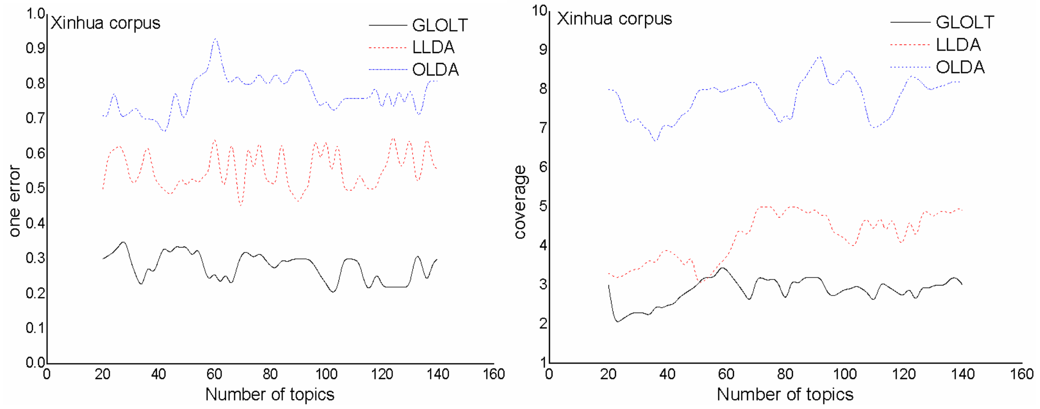
$$C = \sum_{v} \{ \overline{\sum_{d \in C} \sum_{v \in d} n_v \times tfidf_v} \} \times \varphi_v$$

C denotes the labels associated with document; v is the word in that document. For every word with this label we obtain the weight of word to label by the product of word frequency and tfidf, and take the average. So the topic's label can been obtained. As a document d is related with a multinomial distribution θ over topics, we can further predict the unlabeled-document's labels by $\theta_d$.

In order to verify the GL-OLT model for the accuracy of the classification task, employing labeled XinHua corpus, GL-OLT model will been compared with LLDA and OLDA model according to one-error, coverage and hamming loss as the accuracy criteria that are commonly used by Multi-label classification. As GL-OLT model is label ranking algorithm, the value of hamming loss can not been directly calculated. In this paper hamming loss is made some transformations, the variant calculation of hamming loss is as follows.

$$hamming - loss = \frac{1}{m} \sum_{i=1}^{m} \frac{| Y_i \Delta Z_i |}{| Y_i |}$$

The reference to m refers to the number of testing documents. $Y_i$ and $| Y_i |$ refer to the labels contained in the document i and the number of labels, respectively. The testing documents i will be sorted, and then, the first $Y_i$ labels, $Z_i$, are used as the result of model's prediction. Δ refers the set of symmetry with unequal between the set $Y_i$ and $Z_i$. $|Y_i \Delta Z_i|$ indicates the number of two sets of unequal elements.



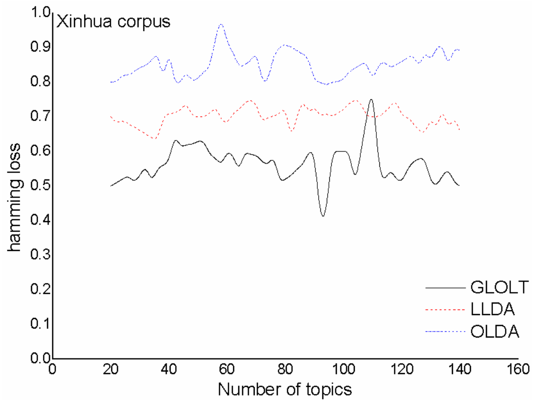


**Figure 3. Comparisons of Predictive Ability**

Figure 3 illustrates the test results. As shown in Figure 3, GL-OLT model has stronger discrimination ability of labels relative to LLDA and OLDA model. The error rate of GL-OLT model is significantly lower than LLDA and OLDA model, especially based on evaluation criterion one-error and coverage.

### 4.3. Algorithmic Efficiency

The comparison of the average iteration time is shown in Figure 4. Because there is linear relationship between the complexity of the model and the number of topics, the average iteration time grows with the number of topics.

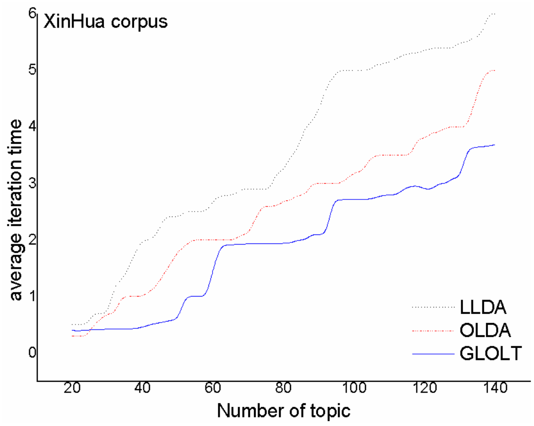


**Figure 4. Comparisons of Running Time**

GL-OLT and OLDA are able to detect topics without making use of the overall data. However LLDA model need entire data to be stored for future processing, the average iteration time of LLDA is remarkably higher than GL-OLT and OLDA model. Furthermore, GL-OLT model adopts the dynamic maintenance of sliding windows before and after an update to control of excessive growth and only sample topics of labels of document relative to OLDA. After several enhancements for GL-OLT model, as it shown the running time of GL-OLT model is the shortest

## 5. Summary

In this paper, the proposed GL-OLT model offers a relatively simple probabilistic model for exploring the relationships among documents, labels, topics, time slices and words. Firstly, this model prefers advanced constraints on latent topics that cause them to align with human-provided labels. Secondly, in order to eliminate the similarity of topics classified differently label, in GL-OLT model each label has not only a set of local topics, but also has several global topics. Finally, for purpose of further improving performance of corpus with time information, the time information is considered and incorporated into GL-OLT model. Empirical results prove that GL-OLT model provides expressively optimized performance in the fields of perplexity, classification accuracy and run time.

From an ordinary perspective, our model can also be applied to other labeled data sets. However, during practical testing, we do not test GL-OLT model on other corpus. In addition, this paper only incorporates labels information by an on-line fashion. So possible future directions for this work include test model on other corpus, integrates other side information, such as user's label, to perform detection and analysis of topics.

## Acknowledgments

This work is supported by science and technology project of Fujian provincial education department: No. JA13196; the National Natural Science Foundation of China No.61303131; the National Natural Science Foundation of China No. 61402216.

## References

[1] "T. H. Probabilistic latent semantic indexing", In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, **(1999)**, pp. 50-57.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, **(2003)**, pp. 993-1022.

[3] H. DaumeI, "Markov Random Topic Fields", In ACL, **(2009)**.

[4] T. Iwata, T. Yamada and N. Ueda, "Modeling Social Annotation Data with Content Relevance using aTopic Model", In NIPS, **(2009)**.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, vol. 41, no. 6, **(1999)**, pp. 391–407.

[6] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations", In International conference on Machine learning, **(2006)**, pp. 577-584.

[7] D. M. Blei and J. M. Supervised Topic Models, In NIPS, **(2007)**, pp. 21-46.

[8] "D. M and A. M. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression", In Uncertainty in Artificial Intelligence, **(2008)**, pp. 411-418.

[9] D. Ramage, D. Hall, R. Nallapati and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-label corpora", In EMNLP, **(2009)**, pp. 248-256.

[10] L. AlSumait, D. Barbará and C. Domeniconi, "On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking", In Proceedings of the Eighth IEEE International Conference on Data Mining, **(2008)**, pp. 3–12.

[11] X. R. W. "Structured Topic Models: Jointly Modeling Words and Their Accompanying Modalities", University of Massachusetts Amherst. Computer Science, **(2009)**.

[12] X. W, A. M. Topics over Time: A Non-Markov continuous time model of topical trends, In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **(2006)**, pp. 424–433.

[13] L. A, D. B. On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking, In ICDM Data Mining, **(2008)**.

[14] M. S and T. L. G. Probabilistic Topic Models, Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, **(2005)**.

[15] D. M. Blei, "Probabilistic Topic Models", Communications of the ACM, vol. 55, no. 4, **(2012)**, pp. 77-84.

[16] P. F. Hu and W. Liu, "Latent topic model for audio retrieval", Pattern Recognition, vol. 3, no. 47, **(2014)**.

[17] G. H. Parameter estimation for text analysis, Arbylon publications, **(2007)**.

[18] R. Daniel, S. T. Dumais and J. Daniel, "Characterizing Microblogs with Topic Models", The AAAI Press, **(2010)**.

[19] M. Hoffman, D. M. Blei and F. Bach, "Online learning for latent dirichlet allocation", In Advances in Neural Information Processing Systems, **(2010)**, pp. 856–864.