Data Min Knowl Disc (2018) 32:885–912
https://doi.org/10.1007/s10618-018-0555-0

Labeled Phrase Latent Dirichlet Allocation and its online learning algorithm

Yi-Kun Tang[1,2] · Xian-Ling Mao[1] · Heyan Huang[1]

Received: 3 March 2017 / Accepted: 8 February 2018 / Published online: 27 February 2018

Abstract There is a mass of user-marked text data on the Internet, such as web pages with categories, papers with corresponding keywords, and tweets with hashtags. In recent years, supervised topic models, such as Labeled Latent Dirichlet Allocation, have been widely used to discover the abstract topics in labeled text corpora. However, none of these topic models have taken into consideration word order under the *bag-of-words* assumption, which will obviously lose a lot of semantic information. In this paper, in order to synchronously model semantical label information and word order, we propose a novel topic model, called Labeled Phrase Latent Dirichlet Allocation (LPLDA), which regards each document as a mixture of phrases and partly considers the word order. In order to obtain the parameter estimation for the proposed LPLDA model, we develop a batch inference algorithm based on Gibbs sampling technique. Moreover, to accelerate the LPLDA's processing speed for large-scale stream data, we further propose an online inference algorithm for LPLDA. Extensive experiments were conducted among LPLDA and four state-of-the-art baselines. The results show

Responsible editor: Pauli Miettinen.

✉ Xian-Ling Mao
maoxl@bit.edu.cn

Yi-Kun Tang
tangyk@bit.edu.cn

Heyan Huang
hhy63@bit.edu.cn

1 Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

2 Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China

(1) batch LPLDA significantly outperforms baselines in terms of case study, perplexity and scalability, and the third party task in most cases; (2) the online algorithm for LPLDA is obviously more efficient than batch method under the premise of good results.

Keywords Topic model · Labeled Phrase LDA · Batch Labeled Phrase LDA · Online Labeled Phrase LDA

1 Introduction

There is a substantial amount of labeled data in the world, such as web pages with corresponding categories, papers with their keywords, and tweets with their hashtags. Compared with the unlabeled data, labeled data contains more useful semantic label information. The labels are usually summaries of documents and can help people get which domain the documents belong to or the main meaning of documents in a short time. Thus, it is usual to obtain better performance in many tasks by using label information.

Supervised topic modeling is a kind of widely used text analyzing technology for labeled data, which can take advantage of label information. It can discover the abstract topics and identify the latent semantic components in labeled text corpora. Thus, lots of supervised topic models have been proposed, such as Labeled Latent Dirichlet Allocation (LLDA) (Ramage et al. 2009a) and Partially Labeled Dirichlet Allocation (PLDA) (Ramage et al. 2011). Compared with unsupervised topic model, topics learnt by many supervised topic models like LLDA and PLDA can be more easily interpreted, since users can use a concise semantic label to understand its corresponding latent topic.

However, existing supervised topic models do not consider word order under the *bag-of-words* assumption. As we know, a single word, especially some proper nouns, prepositions or verbs, sometimes fails to express the meaning completely or exactly. For example, the phrase *white sox* is related to baseball, but neither of the single word *white* or *sox* is related to baseball. Thus, the existing supervised topic models are not the best solution to capture the latent semantic structure of labeled documents.

In this paper, we first propose a novel topic model, called Labeled Phrase LDA (LPLDA), to synchronously consider the supervised label information and the word order. The proposed model takes each document as a mixture of phrases under the *bag-of-phrases* assumption and considers the impact of word collocation on semantics; meanwhile, it can establish a one-to-one correspondence between latent topics and human-provided tags, which restricts the target topics of each labeled document to the corresponding label set of the document. Then, we develop a batch inference algorithm to estimate parameters in LPLDA.

Moreover, although the proposed LPLDA performs better than some of the existing state-of-the-art topic models as demonstrated in our extensive experiments, it is difficult to apply batch LPLDA to large-scale data or stream data due to its batch learning algorithm. As we know, there's more data created than ever before. For example, Twitter users tweet 277,000 times every minute, and Yelp users post 26,380 reviews every

minute.[1] A batch learning algorithm cannot efficiently process these large-scale data or stream data. Compared with batch methods, an online learning method updates the weight parameters after processing one unit of the training data, which means that its storage cost is very low and its efficiency is very high. Thus, in this paper, in order to efficiently handle large-scale data or stream data, we further propose an online algorithm for LPLDA based on particle filter, called online Labeled Phrase LDA (online LPLDA).

The rest of the paper is organized as follows. In Sect. 2, we review the related work. We introduce our proposed topic model (LPLDA) and its batch learning algorithm in Sect. 3. In Sect. 4, we further propose an online algorithm for LPLDA to improve the inference efficiency. In Sect. 5, we present extensive experiments on three datasets: paper dataset, Twitter dataset and Yahoo! Answers dataset. Finally, our conclusions and future work are presented in Sect. 6.

2 Related work

2.1 Supervised topic modeling

In order to process labeled data, lots of supervised topic models have been proposed. The existing supervised topic models can be divided into two categories: supervised non-hierarchical topic models and supervised hierarchical topic models. Supervised non-hierarchical topic models are widely studied. Two such models, Supervised LDA (Mcauliffe and Blei 2008) and Disc LDA (Lacoste-Julien et al. 2009), are first proposed to model documents associated with only a single label. In order to deal with multi-labeled documents, many supervised topic models, such as the MM-LDA (Ramage et al. 2009b), Author TM (Rosen-Zvi et al. 2004), Flat-LDA (Rubin et al. 2012), Prior-LDA (Rubin et al. 2012), Dependency-LDA (Rubin et al. 2012), Gibbs MedLDA (Zhu et al. 2013), Diagonal Orthant Latent Dirichlet Allocation (DOLDA) (Magnusson et al. 2016), Partially Observed Topic (POT) (Xiao et al. 2009), rPLSA (Zhao et al. 2015), weakly-supervised nPLSA model (Tang et al. 2014), Labeled LDA (LLDA) (Ramage et al. 2009a), Partially LDA (PLDA) (Ramage et al. 2011) and Conceptualization Labeled LDA (Tang et al. 2018) etc., are not constrained to one label per document because they model each document as a bag of words with a bag of labels. Among all these models, only few models, such as LLDA (Ramage et al. 2009a), PLDA (Ramage et al. 2011) and Conceptualization Labeled LDA (Tang et al. 2018), can obtain topics that correspond directly with the labels, which will facilitate the interpretation of the learned topics. In this paper, we will follow the idea to propose a novel topic model. None of these non-hierarchical supervised models, however, leverage on dependency structure, such as parent–child relation, in the label space. Thus, for hierarchical labeled data, several models, such as hLLDA (Petinot et al. 2011), HSLDA (Perotte et al. 2011), SSHLDA (Mao et al. 2012), Tree Labeled LDA (tLLDA) (Slutsky et al. 2013) and Labelset Topic Model (LsTM) (Li et al. 2016), are proposed to handle the label relations in data.

[1] https://www.domo.com/learn/data-never-sleeps-2 (Accessed date: March 1, 2017).

However, the existing supervised topic model are based on *bag-of-words*, and they all neglect the importance of word order.

2.2 Word order issue for topic modeling

Most topic models are under the *bag-of-words* assumption, however, in fact, the meaning of a text has much to do with the order of words. There have been some existing topic models taking into consideration word order of text data.

A part of them are based on Hidden Markov Model (Eddy 1996), and they make use of the dependencies between the sentiment as well as the syntactic ordering of the words. HMM-LDA (Griffiths et al. 2005) is a generative model that uses both kinds of dependencies, and is capable of simultaneously finding syntactic classes and semantic topics despite having no knowledge of syntax or semantics beyond statistical dependency. Based on HMM-LDA, the CFACTS model (Lakkaraju et al. 2011) combines syntax and semantics and introduces dependencies between the sentiment and facet topic variables for neighboring words. Joint Author Sentiment Topic Model (JAST) (Mukherjee et al. 2014) uses a Hidden Markov Model to capture short range syntactic and long range semantic dependencies in reviews coherence in author writing style.

Some of them apply bigram model to topic model. Bigram Topic Model (Wallach 2006) integrates bigram model and topic model beyond *bag-of-words* assumption. Topical n-grams (Wang et al. 2007) discovers topics as well as topical phrases, and uses additional latent variables and word-specific multinomials to model bi-grams.

There are some other methods devoted to applying the phrase mining algorithm to the topic model. Phrase Discovering LDA (Lindsey et al. 2012) is a hierarchical generative probabilistic model of topical phrases. There are other topic models processing documents as a mixture of phrases, such as Topic Similarity Model (Xiao et al. 2012) and Constructing a Topical Hierarchy (Wang et al. 2013) etc. PhraseLDA (Elkishky et al. 2014) combines phrase mining and LDA, which overcomes the high-complexity of Topical n-grams. The top phrases associated with each topic learned by PhraseLDA have more intact meaning than words learned by tradition LDA. Also, there is a model combining topic model with lexical argument structure (Spagnola and Lagoze 2011), which learns word order by examining syntactic dependency parse trees from Wikipedia article.

However, all these models considering the word order are unsupervised topic models and they cannot take advantage of label information.

2.3 Online learning for topic modeling

Since exact posterior inference is intractable for topic modeling, both variational (Blei et al. 2003) and Monte Carlo (Griffiths and Steyvers 2004) methods have been widely developed for approximate inference, which are normally able to deal with medium-sized datasets. In order to deal with large-scale data, lots of online inference algorithms based on the variational and Monte Carlo methods have been proposed, which can not only deal with massive datasets but also deal with dynamic streaming data.

Much progress has been made for developing online variational methods for topic modeling (AlSumait et al. 2008; Foulds et al. 2013; Hoffman and Blei 2015; Hoffman et al. 2010, 2013; Kingma and Welling 2014; Lakkaraju et al. 2011; Wang et al. 2012). One representative work is the online variational inference method for LDA (Hoffman et al. 2010), which considers LDA as a probabilistic factorization of the matrix of word counts into a matrix of topic weights and a dictionary of topics. Most of them have adopted stochastic approximation of posterior distribution by sub-sampling a given finite data set, whose drawback is that the data size needs to know in advance. To overcome the drawback, some researchers made streaming updates to the estimated posterior (Broderick et al. 2013; Ghahramani and Attias 2000). In addition, McInerney et al. (2015) introduced the population Variational Bayes (PVB) method which combines traditional Bayesian inference with the frequentist idea of the population distribution for streaming inference. Tweet Propagation Model (TPM) (Ren et al. 2013) can infer dynamic probabilistic distributions over user's interests and topics. Dynamic user clustering topic model (dynamic UCT) (Liang et al. 2017; Zhao et al. 2016) can capture the users' dynamic topic distributions in sparse data settings for short text streams. Social collaborative viewpoint regression (sCVR) can predict item ratings based on user opinions and social relations (Ren et al. 2017). Online sparse topical coding (OSTC) was proposed to improve efficiency by using an online learning algorithm to learn the dictionary (Zhang et al. 2013). Shi and Zhu (2014) proposed the Online Bayesian Passive-Aggressive (BayesPA) method for max-margin Bayesian inference of online streaming data.

In contrast, the attempts towards developing streaming Monte Carlo methods for LDA also have some success. One online MCMC approach adapts collapsed Gibbs sampling by sampling topic assignments based on the topic assignments and data for all previously analyzed words, instead of all other words in the corpus (Song et al. 2005). This algorithm is fast and has constant memory requirements, but is not guaranteed to converge to the posterior. Online LDA (Banerjee and Basu 2007) trains a part of the training data using the batch LDA. And then for each new word, it incrementally samples a latent topic according to the prior latent topics and words. But it never resamples old topic variables. Two alternative online MCMC approaches, incremental LDA and particle filtering LDA, are proposed by Canini et al. (2009). Incremental LDA periodically resamples the topic assignments for previously analyzed words while particle filtering approach uses particle filter instead of collapsed Gibbs sampling. Online Labeled LDA (Zhou et al. 2015) is also an online method of Labeled LDA based on particle filter. To further improve the performance, streaming Gibbs sampling (SGS) algorithm (Gao et al. 2016) is proposed to naturally extend the collapsed Gibbs sampling to the online learning setting, which only stores the current mini-batch.

In this paper, we will also develop an online algorithm for our proposed topic model besides a batch inference method.

3 Labeled Phrase LDA

In this paper, we will model the word order by considering phrases in the documents. Thus, it is indispensable to find human-interpretable phrases. Borrowing the idea in

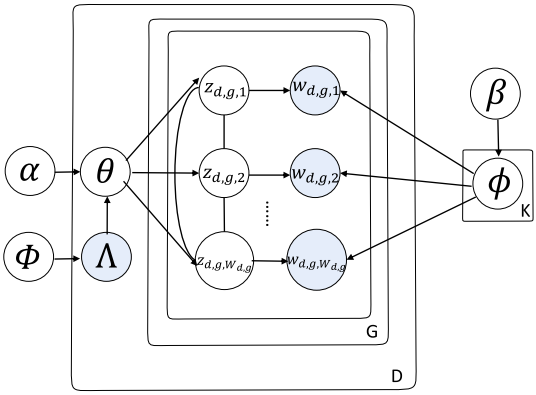

Fig. 1 Graphical model for LPLDA. In LPLDA, we segment each document into phrases, and we assume that each word in a phrase shares the same latent topic. The target topics of a document are chosen from its labels

PhraseLDA (Elkishky et al. 2014), we record the frequency of appearance for contiguous words in the corpus. We gradually increase the size of sliding window over the corpus to generate candidate phrases and obtain aggregate counts. We follow two principles in our algorithm to find candidate phrases. The first one is the downward closure lemma (Agrawal et al. 1994): any super-phrase of an infrequent phrase is also an infrequent phrase. And another one is data-antimonotonicity (Han et al. 2011): the longest frequent phrase of a document containing no frequent phrases of length n is no more than n in length. The two principles can help increase the efficiency of candidate phrases mining. And then at each iteration, we merge neighboring candidate phrases and consider them as a new phrase. We judge whether the new phrase is of high quality via frequency record. Finally, we segment documents into phrases.

The proposed Labeled Phrase Latent Dirichlet Allocation (LPLDA) is a supervised topic model processing multi-labeled corpora, and its graphical model is presented in Fig. 1. It is restricted that the topics of each document are in the domain of the labels in the document. The topic distribution in each document draws from the Dirichlet distribution with an M dimension prior parameters, where M changes with the number of labels of each document. For example, suppose the total number of topics in the datasets is five, denoted as T_a, T_b, T_c, T_d and T_e, that is $K = 5$. If the dth document has labels T_b, T_c and T_e, then $\theta^{(d)}$ is drawn from a Dirichlet distribution with parameters $\alpha^{(d)} = (\alpha^2, \alpha^3, \alpha^5)^T$, where α is the topic prior. The above example explains that the target topics of a document depend on its labels.

LPLDA segments the documents into phrases based on *bag-of-phrases* assumption, and all words in the gth phrase of the dth document have a latent topic, random in $\{z_{d,g,1}, \ldots, z_{d,g,W_{d,g}}\}$. Notations used in LPLDA are listed in Table 1.

The generative process for our LPLDA model can be found in Algorithm 1, where we define an indicator function $I^{(d)}(k)$ as below:

$$I^{(d)}(k) = \begin{cases} 1 & \text{if the } k\text{th topic is in the set of} \\ & \text{labels of the } d\text{th document.} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Table 1 Notations used in LPLDA

Variable	Description
D, K, G, V	Number of documents, topics, phrases, size of vocabulary
$W_{d,g}$	Number of tokens in gth phrase of dth document
β_k	Parameter of the multinomial distribution of the kth topic
α	Parameter of the Dirichlet distribution of the topic prior
ϕ	Parameter of the Dirichlet distribution of the word prior
Φ_k	Labeled prior for topic k
$z_{d,g,i}$	Latent topic for ith token in gth phrase of dth document (latent topic for $w_{d,g,i}$)
$C_{d,g}$	The collection of words in phrase g in a document d
$w_{d,g,i}$	ith token in gth phrase of document d
$N_{d,k}$	Number of tokens assigned to topic k in document d
$N_{w_{d,g,j},k}$	Number of tokens with value $w_{d,g,j}$ and topic k

Algorithm 1 Generative process for Labeled Phrase LDA.

1: For each topic $k \in \{1, \ldots, K\}$:
2: Generate $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,V})^T \sim Dir(\cdot|\phi)$
3: For each document d:
4: For each topic $k \in \{1, \ldots, K\}$:
5: Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim Bernoulli(\cdot|\Phi_k)$
6: Generate $\alpha^{(d)} = I^{(d)} \times \alpha$
7: Generate $\theta^{(d)} = (\theta_{l_1}, \ldots, \theta_{l_{M_d}})^T \sim Dir(\cdot|\alpha^{(d)})$
8: For each g in phrases $\{1,\ldots,G_d\}$:
9: For each i in words $\{1,\ldots,W_{d,g}\}$:
10: Generate $z_{d,g,i} \in \{\lambda_1^{(d)}, \ldots, \lambda_{M_d}^{(d)}\} \sim Mult(\cdot|\theta^{(d)})$
11: Generate $w_{d,g,i} \in \{1, \ldots, V\} \sim Mult(\cdot|\beta_{z_{d,g,i}})$

The generative process can be divided into three parts. First, drawing the vocabulary distribution from Dirichlet distribution for each topic in the datasets (lines 1–2). Then, drawing the topic distribution for each document from Dirichlet distribution only in the domain of the labels of each document (lines 3–7). Finally, generating latent topic for each word in each phrase (lines 8–11).

3.1 Learning and inference

In LPLDA, we first segment each document into single or multi-word phrases like PhraseLDA. We assume that all the words in the same phrase share the same latent topic. Therefore, the Gibbs sampling probability of $C_{d,g}$ (phrase g in a document d), for a topic in LPLDA can be:

$$P(C_{d,g} = k|W, Z_{\setminus C_{d,g}}) \propto I^{(d)}(k) \cdot \prod_{j=1}^{W_{d,g}} \left(\alpha_k^{(d)} + N_{d,k\setminus C_{d,g}} + j - 1\right) \frac{\left(\beta_{k,w_{d,g,j}} + N_{w_{d,g,j},k\setminus C_{d,g}}\right)}{\left(\sum_{x=1}^{V} \beta_{k,x} + N_{k\setminus C_{d,g}} + j - 1\right)} \tag{2}$$

where $N_{d,k\setminus C_{d,g}}$ is the number of tokens assigned to topic k in document d without the gth phrase, $N_{w_{d,g,j},k\setminus C_{d,g}}$ is the number of tokens with value $w_{d,g,j}$ and topic k without the gth phrase in document d, and $I^{(d)}(k)$ is the indicator function.

The parameter estimation for any single Gibbs Sampling is as follows:

$$\hat{\phi}_{w,k} = \frac{N_{w,k} + \beta_{k,w}}{N_{(\cdot),k} + \sum_{x=1}^{V} \beta_{k,x}} \tag{3}$$

$$\hat{\theta}_k^{(d)} = \frac{N_{d,k} + \alpha_k^{(d)}}{N_{d,(\cdot)} + \sum_{i=1}^{K} \alpha_i^{(d)}} \tag{4}$$

where $N_{w,k}$ is the number of tokens with value w and topic k, $\beta_{k,w}$ is the multinomial distribution over words in topic k and word w, and $N_{(\cdot),k}$ is the number of tokens with topic k in Eq. (3). In Eq. (4), $N_{d,k}$ is the number of tokens assigned to topic k in document d. The topic-specific distribution ϕ can be used to obtain topical abstracts for topics; meanwhile the topic distribution θ for each document can be used to discover the most relevant topics for a document and find documents with similar topics.

4 Online Labeled Phrase LDA

The algorithm we introduced in the previous section is the batch algorithm for LPLDA. In this section, we introduce our proposed online algorithm for LPLDA based on particle filters, called online Labeled Phrase LDA (online LPLDA). As demonstrated in Algorithm 2, our online LPLDA algorithm contains two main phases: the initialization phase and the online learning phase.

In the initialization phase, particles are created. We assume the total number of particles is P. We use a small part of the dataset to do the initialization based on batch LPLDA, which draws the initial topic distribution over vocabulary stored in each particle.

In the online learning phase, we first initialize each particle's weight to the same value P^{-1}. G_d is the number of phrases in the document d. While documents are coming successively, new labels or words are added to the label set and vocabulary set. As for the new phrases, we record the number of adjacent words' occurrences and use the phrase mining method used in LPLDA to estimate whether to form new phrases. For each phrase in each particle, we use Eq. (5) to calculate the weights.

$$\omega_i^{(p)} = \omega_{i-1}^{(p)} P\left(G_i|\mathbf{z}_{i-1}^{(p)}, \mathbf{w}_{i-1}\right) \tag{5}$$

Algorithm 2 Online Labeled Phrase LDA.

Initialization Phase:

for $p = 1, \ldots, P$ **do**
  while the times of Gibbs Sampling do not reach the set point **do**
    draw topic for each word in initial corpus, using Eq. (2)

Online learning Phase:

initialize weights $\omega_0^{(p)} = P^{-1}$ for $p = 1, \ldots, P$
for $d = 1, \ldots, D$ **do**
  for $i = 1, \ldots, G_d$ **do**
    for $p = 1, \ldots, P$ **do**
      calculate the weights using Equation (5)
      sample $z_i^{(p)}$ using Equation (6)
    normalize weights ω_i to sum to 1
    if $||\omega_i||^{-2} \leq ESS$ threshold **then**
      resample particles
      for j in $\Re(i)$ **do**
        for $p = 1, \ldots, P$ **do**
          resample $z_i^{(p)}$ using Equation (9)
      set $\omega_i^{(p)} = P^{-1}$ for $p = 1, \ldots, P$

$P(G_i|\mathbf{z}_{i-1}^{(p)}, \mathbf{w}_{i-1})$ is the probability of the phrase G_i assigned to latent topic $z_i^{(p)}$. We use Eq. (6) to sample latent topic in online LPLDA.

$$
\begin{aligned}
&P\left(C_{d,g}^{(p)} = k|\mathbf{W}_{d,g}, Z_{\backslash C_{d,g}}^{(p)}\right) \\
&\propto I^{(d)}(k) \cdot \prod_{j=1}^{W_{d,g}} \left(\alpha_k^{(d)} + N_{d,k\backslash C_{d,g}}^{(p)} + j - 1\right) \frac{\left(\beta_{k,w_{d,g,j}} + N_{w_{d,g,j},k\backslash C_{d,g}}^{(p)} + n_{w_{d,g,j},k\backslash C_{d,g}}^{(p)}\right)}{\left(\sum_{x=1}^{V} \beta_{k,x} + N_{k\backslash C_{d,g}}^{(p)} + n_{k\backslash C_{d,g}}^{(p)} + j - 1\right)}
\end{aligned}
\tag{6}
$$

Same as batch LPLDA, the first part of the equation is an indicator function which restricts the latent topic of phrase in the label set of the document. The second part of the equation expresses the distribution of documents over topics currently. The last part is the distribution of topics over vocabulary currently. $\mathbf{W}_{d,g}$ is all the words came before the gth phrase in the dth document, that is words before the current phrase. The superscript p represents the variable in particle p. $Z_{\backslash C_{d,g}}^{(p)}$ is latent topics without the topic of $C_{d,g}$ before the current phrase in particle p. $N_{d,k\backslash C_{d,g}}^{(p)}$ is the number of tokens assigned to topic k without the gth phrase in the document d of particle p. $N_{w_{d,g,j},k\backslash C_{d,g}}^{(p)}$ is the number of tokens with value $w_{d,g,j}$ and topic k without the gth phrase in the document d of particle p. And $n_{w_{d,g,j},k\backslash C_{d,g}}^{(p)}$ is the number of tokens with value $w_{d,g,j}$ and topic k without the gth phrase in the document d of the initial corpus.

Then we normalize weights of all particles to sum to 1. In particle filter algorithm, posterior probability of latent topics can be calculated by Eq. (7).

$$P(\mathbf{z}_i|\mathbf{w}_i) \approx \sum_{p=1}^{P} \omega_i^{(p)} \cdot 1_{z_i}\left(\mathbf{z}_i^{(p)}\right) \tag{7}$$

$1_{z_i}(\mathbf{z}_i^{(p)})$ is an indicator function of $\mathbf{z}_i$:

$$1_{z_i}(\mathbf{z}_i^{(p)}) = \begin{cases} 1 & \text{if } \mathbf{z}_i \text{ equals } \mathbf{z}_i^{(p)}. \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

This indicator function shows that particles with the same latent topic vector $\mathbf{z}$ have the same posterior probability of latent topics $P(\mathbf{z}|\mathbf{w})$, which is sum of these particles' weights. However, the topic vector $\mathbf{z}$ is often high-dimensional in reality, which makes it cost too much to check whether particles have the same topic vector. And there isn't so many particles with the same topic vector. Therefore, we modify this procedure to choose the particle with highest weight for estimation.

However, weights assigned to each particle may change a lot during processing, and better analyzation result of the observed data may be provided by only a few particles. To solve this problem, we use resampling method. In this method, new sets of particles with high weight are produced whenever the weights become large. To measure weight variance, we use an approximation to the sample size, $||\omega_i||^{-2} \leq ESS$. The threshold ESS is associated with particle number. We resample phrases in the rejuvenation sequence $\Re(i)$, where $\Re(i)$ is a sequence of phrase with no more than i phrases. The number of phrases in the sequence, $|\Re(i)|$, is a function of i. The sequence is selected from $\{1, \ldots, i\}$ randomly, which means resampling phrases from phrases before the ith phrase.

The posterior probability of latent topics in resampling can be calculated by Eq. (9).

$$\begin{aligned} &P\left(C_{d_r,g_r}^{(p)} = k|\mathbf{W}_{d,g}, Z_{\backslash C_{d_r,g_r}}^{(p)}\right) \\ &\propto I^{(d)}(k) \cdot \prod_{j=1}^{W_{d_r,g_r}} \left(\alpha_k^{(d)} + N_{d,k\backslash C_{d_r,g_r}}^{(p)} + j - 1\right) \frac{\left(\beta_{k,W_{d_r,g_r,j}} + N_{W_{d_r,g_r,j},k\backslash C_{d_r,g_r}}^{(p)} + n_{W_{d_r,g_r,j},k\backslash C_{d_r,g_r}}^{(p)}\right)}{\left(\sum_{x=1}^{V} \beta_{k,x} + N_{k\backslash C_{d_r,g_r}}^{(p)} + n_{k\backslash C_{d_r,g_r}}^{(p)} + j - 1\right)} \end{aligned} \tag{9}$$

$C_{d_r,g_r}^{(p)}$ is the latent topic of the phrase resampled in particle p. And each variable is under current state, for example, $N_{W_{d_r,g_r,j},k\backslash C_{d_r,g_r}}^{(p)}$ is the number of times word $w_{d_r,g_r,j}$ assigned to topic k without C_{d_r,g_r} before the gth phrase in the document d of particle p currently.

After resampling, particle's weights are all reset to P^{-1}, for each particle is drawn from the same distribution now and the previous weights only reflect their resampling frequencies.

Generally, it costs too much time to resample in particle filter for the reason that if a particle is resampled more than once, we have to copy its topic assignment to every of its child particle. So in reality, in child particle we only store topic assignment different from its parent particle, and record whether the particle is active. If a particle

is inactive and none of its child particle is resampled in the future, its child particles will be merged with it, maintaining the bound on the tree depth.

5 Experiment

5.1 Experiment for Labeled Phrase LDA

5.1.1 Experiment setting

In these experiments, we use four state-of-the-art topic models as baselines: Labeled LDA (LLDA) (Ramage et al. 2009a), HMM-LDA (Griffiths et al. 2005), Topical N-Grams (TNG) (Wang et al. 2007) and PhraseLDA (Elkishky et al. 2014). LLDA is a generative model for multiple labeled corpora. It is a supervised topic model, in which the topics of each document are restricted to its labels. HMM-LDA (Griffiths et al. 2005) is a generative model that uses both kinds of dependencies, and is capable of simultaneously finding syntactic classes and semantic topics despite having no knowledge of syntax or semantics beyond statistical dependency. TNG is an n-gram topic model that discovers topics as well as topical phrases. It uses additional latent variables and word-specific multinomials to model bi-grams. PhraseLDA based on *bag-of-phrases* assumption can partly consider word order of documents.

We conducted the experiments on three datasets [found at http://pan.baidu.com/s/1geLoPLh (Accessed date: November 30, 2017)].

One of our datasets, we call it **Conf** in the rest of the paper, contains full papers of four conferences (CIKM, SIGIR, SIGKDD and WWW) from the year 2011 to the year 2013. We use the keywords of each paper as labels. And we filtrate the raw datasets by removing the documents with only one label. After filtrating, it remains 1169 documents, and all documents in our datasets are multi-labeled.

The second is a corpus of tweets downloaded from Twitter,[2] a website where people can post short messages about their current activities. We call it **Twitter** later. It contains tweets on Twitter from August 2009 to September 2009. We use the hashtags of each tweet as labels. And we filtrate the raw datasets by removing labels tagged < 2000 documents, and then removing tweets that contain only one label, since the LDA family is under the multi-topic assumption. After filtrating, the **Twitter** remains 883,799 tweets.

As for the last datasets, we crawled the questions and associated answer pairs (QA pairs) of one of a top category on Yahoo! Answers,[3] *Health*. This produced twenty-three subcategories from 2005.11 to 2008.11, and we selected 330,000 QA documents from it randomly. We use the category or subcategory of each question or answer as labels of each document. And all documents have more than one label. We refer the Yahoo! Answer data as **Yahoo! Answers**.

The statistics of all datasets are summarized in Table 2. All documents in our three datasets have more than one label. And we conduct all the experiments on a server with

[2] http://twitter.com/ (Accessed date: March 1, 2017).

[3] https://answers.yahoo.com/ (Accessed date: March 1, 2017).

Table 2 The statistics of the datasets

Datasets	**Conf**	**Twitter**	**Yahoo! Answers**
Size of documents	1169	883,799	330,000
Size of labels	855	209	24
Size of vocabulary	27,312	27,350	40,307

an Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00 GHz and 125 GB memory. In the rest subsections we will compare the proposed model Labeled PhraseLDA (LPLDA) with Labeled LDA, HMM-LDA, TNG and PhraseLDA in terms of case study, perplexity, scalability and a multi-labeled text classification task.

5.1.2 Case studies

We ran the proposed Labeled Phrased LDA (LPLDA) and the baselines on the three corpus described above. All models are based on the standard collapsed Gibbs sampling, and are set the same initial hyperparameters, where the values of α and β both equal to 0.01.

The top associated words or phrases of topics can be used as good descriptors for topics (Blei et al. 2003; Mcauliffe and Blei 2008). Tables 3, 4 and 5 show top ten words or phrases from five topics learned on the **Conf**, **Twitter** and **Yahoo! Answers** in our proposed model LPLDA and the baseline models LLDA, HMM-LDA, TNG and PhraseLDA, respectively. All topics are directly named with the labels of each document. And the topics learned by unsupervised topic model, PhraseLDA, HMM-LDA and TNG, were matched to a topic in LPLDA by using Kullback–Leibler divergence. Since the generation step of phrases in HMM-LDA is after the inference of the model, and each phrase contains both syntactic class words and semantic topic words, there does not exist top phrases in each single topic. In LPLDA, we consider words in each phrase as a whole in the inference step. Therefore, phrases in these two models are different, and in this experiment, for HMM-LDA, we only list its top content words.

We have two major observations from the results: (1) Compared to Labeled LDA, phrases learned by LPLDA show more particular information and contain more distinct meaning than single words learned by Labeled LDA. For example, the phrase *breast cancer* learned by LPLDA obviously contains more clear information than either the word *cancer* or *breast* learned by LLDA for the topic *Cancer* on the **Yahoo! Answers**. (2) Comparing LPLDA, HMM-LDA, TNG and PhraseLDA, we can see that as a supervised topic model, each topic learned by LPLDA can be a one-to-one correspondence with its label, named by the semantic meaning of the label. Therefore, it is more interpretable and understandable for topics in LPLDA than the unsupervised baselines, HMM-LDA, TNG and PhraseLDA. So LPLDA performs better in finding interpretable and understandable topics data and the most appropriate phrases for topics.

Consequently, LPLDA performs better than the baseline models in term of case study.

Table 3 Top ten phrases and words from five topics learned on the **Conf** in our proposed model LPLDA and the baseline models, LLDA, HMM-LDA, TNG and PhraseLDA, respectively

Label	Model				
	LPLDA	LLDA	PhraseLDA	HMM-LDA	TNG
Language models	language mode, retrieval models, translation model, negative feedback, difficult queries, ACM SIGIR, general negative language model, statistically significant, information retrieval, query expansion	language, documents, negative, model, document, retrieval, query, feedback, based, models	method outperforms, performance of our approach, significantly outperforms, compare the performance, baseline methods, compare our approach, outperforms the baseline, outperforms significantly the RPF methods, compare our method, compare our algorithm	vector, data, feedback, cost, prior, nuclear, trade, well, yahoo, access	eduan evaluation, ecology states, language modeling similarities, deliberately divert, igk jg, car cd, stank dard measurement, question groupingin, qt sj, sql procedure
Topic model	topic models, topic distribution, latent topics, topic proportions, multinomial distribution, Dirichlet prior, Gibbs sampling, topic document, generative process, topic assignments	topic, model, topics, words, models, document, distribution, word, lda, number	ACT model, ACTC model, influential authors, author topic, topic models, conferences such as SIGMOD, topical authority, Author Author, topic model for authors, author s most influential	topic, did, document, context, d, until, drop, refer, post, size	significant topic, word levels, topic stream, modeled topics, identifies topics, topic modeling systems, lag relationships, random initiation, prior generated, include vsparql
Book search	book search, digital libraries, internal and external, Open Library, faceted filtering, advanced search interface, test collections, external search, search behavior, external sessions	internal, book, external, library, filtering, digital, books, interaction, operators, faceted	digital libraries, internal and external, book search, Open Library, Google Books, external search, external and internal, external sessions, searches over books, book recommending	textbook, interesting, focused, current, method, multimedia, extract, xt, only, way	external queries, search operators, field usage, moby dick, external wses, search options, search features, biased implicit, informationseeking process, book searchers

Table 3 continued

Label	Model				
	LPLDA	LLDA	PhraseLDA	HMM-LDA	TNG
Recommendation system	interest patterns, recommender systems, implicit feedback, user interests, exploratory search, recommendation performance, search history, search results, recommends books, book recommending	recommendation, patterns, brek, cluster, book, exploratoriness, long-lastingness, searches, similarity, exploratory	method outperforms, performance of our approach, significantly outperforms, compare the performance, baseline methods, compare our approach, outperforms the baseline, outperforms significantly the RPF methods, compare our method, compare our algorithm	spatial, same, topic, idea, recommend, cluster, than, outperform, has, hot	latent user factor, diversified recommendation, lfp pmf, nwn wma, lfp puresvd, user profile level, text clueslin, spainjun wanguniversity college, iptv service, instantly modify recommendations
Collaborative filtering	collaborative filtering, user item, recommender systems, user ratings binary codes, matrix factorization, data set, rated items, rating of user, web pages	user, users, items, data, ratings, set, recommendation, item, collaborative, filtering	method outperforms, performance of our approach, significantly outperforms, compare the performance, baseline methods, compare our approach, outperforms the baseline, outperforms significantly the RPF methods, compare our method, compare our algorithm	will, user-item, asker, web, context, strong, way, need, protocol, basic	collaborative filtering variants, visiting behaviors, sampling strategy ensures reasonable cluster, significantly outnumber, boundary expansionoftentimes, twitter aretopmentione-drandomfollowees, pression rates, classifier rises fairly, southern califorinia, fewer incorrect

Table 4 Top ten phrases and words from five topics learned on the **Twitter** in our proposed model LPLDA and the baseline models, LLDA, HMM-LDA, TNG and PhraseLDA, respectively

Label	Model				
	LPLDA	LLDA	PhraseLDA	HMM-LDA	TNG
Web	web design, web seo internetmarketing, web domain hosting business seo, tags web jobs hiring, web app, web hosting, hosting web domain seo, design web, web tech, tech web	web, design, seo, article, internetmarketing, hosting, domain, business, marketing, tags	seo jobs, web seo internetmarketing, sem jobs, seo marketing, seo sem, link building, search engine, web site, freelance seo job seo, freelance seo job	age, weekend, trend, wasnt, webdesign, mike, lead, viral, haha, tom	web design, iranelection iran, tcot orca, logo design, freelance seo job, freelance web design job, hiring mortgage, iran iranelection, graphic design, freelance seo job seo
Media	media noisemachine, japan tech media, fox news, social media news, business media television news markets media, fox media, reuters japan tech media, news media, glenn beck, social media	media, news, fox, japan, noisemachine, social, tech, business, obama, glenn	fox news, rt rt, media noisemachine, rt fox news, mainstream media, news politics, news media, media matters, abc news, rt cnn	dl, haven, accept, spirit, vid, multimedia, custom, actualiteit, type, cost	social media, hiring director, tcot tlot, nsfw porn, socialmedia marketing, rt top, iran iranelection, rt ways, rt twitter, twitter facebook
UN	security council, human rights, iranelection amnesty, tcot gop, amnesty iranelection, general assembly, violence against women un news, global warming, iranelection iran, whn minor	news, iranelection, execution, amnesty, tcot, rights, iran, sentenced, minor, women	rt indico vip, indico vip rt, rt rt indico vip, rt rt, rt rt rt, rt rt rt indico vip, indico vip, rt rt rt rt, rt rt rt rt indico vip, rt rt rt rt rt	christian, camera, overhaul, white, tom, crash, england, corn, humanitarian, hire	human rights, hiring director, tcot tlot, rt iranian, iranelection humanrights, iran iranelection humanrights, porn nsfw, conceited person, rt humanrights, health healthcare

Table 4 continued

Label	Model				
	LPLDA	LLDA	PhraseLDA	HMM-LDA	TNG
Showbiz	nieuws showbiz gossip, nieuws showbiz entertainment, reuters nieuws showbiz gossip, ap nieuws showbiz gossip, lady gaga, michael jackson, pictures nieuws showbiz gossip, sex and the city, mad men, chris brown	showbiz, nieuws, gossip, reuters, entertainment, film, video, back, jackson, michael	nieuws sport, nieuws showbiz gossip, nieuws actualiteit, nieuws showbiz entertainment, voor nieuws sport, op nieuws sport, van nieuws sport, nieuws internet ict, bij nieuws sport, met nieuws sport	film, subject, mass, martin, laser, hero, kingdom, hier, pr, yall	tcot ocra, rt rt, nieuws showbiz entertainment, nieuws sport, hiring finance, voor nieuws, hiring finance representative, op nieuws, iran iranelection, bij nieuws
Baseball	mlb baseball, baseball mlb, win mlb baseball, red sox, game mlb baseball, white sox, sports baseball, free agent, sports baseball mlb, season mlb baseball	baseball, mlb, win, giants, dodgers, sox, game, angels, tigers, reds	mlb baseball, yankees angels, yankees mlb, yankees phillies, phillies yankees, angels yankees, support supernatural add twibbon to your avatar, support philadelphia phillies add twibbon to your avatar, world series, dodgers phillies	fuck, happen, thought, vanilla, document, redsox, numb, doing, patriot, face	world series, tcot tlot, yankees mlb, yankees phillies, angels yankees, phillies yankees, red sox, mlb baseball, redsox angels, iranelection iran

Table 5 Top ten phrases and words from five topics learned on the **Yahoo! Answers** in our proposed model LPLDA and the baseline models, LLDA, HMM-LDA, TNG and PhraseLDA, respectively

Label	Model				
	LPLDA	LLDA	PhraseLDA	HMM-LDA	TNG
Optical	eye doctor, contact lenses, wear contacts, wear glasses, contact lens, eye drops, dry eyes, eye exam, left eye, colored contacts	eye, eyes, contacts, glasses, lenses, wear, vision, contact, doctor, don't	eye doctor, contact lenses, wear glasses, wear contacts, pink eye, eye drops, close your eyes, contact lens, dry eyes, left eye	the, to, sofa, cambia, sternum, micro, cdcs, basin, steak, eyebrow	contact lenses, pink eye, eye drops, contact lens, eye doctor, left eye, good luck, blood vessels, wear glasses, eye exam
Heart diseases	heart attack, blood pressure, high blood pressure, chest pain, heart disease, heart problems, heart rate, heart beat, heart failure, heart murmur	heart, blood, pressure, doctor, high, attack, normal, chest, pain, rate	heart attack, heart disease, panic attacks, heart problems, heart rate, heart failure, heart beat, kidney stones, gall bladder, bowel movement	the, and, to, seamen, healthcareforless, sane, rudolph, tenia, shrink, neglect	blood pressure, heart attack, heart rate, good luck, chest pain, high blood pressure, blood flow, high blood, heart failure, chest pains
Infectious diseases	chicken pox, sore throat, strep throat, staph infection, wash your hands, immune system, flu shot, HIV AIDS, high fever, bacterial infection	infection, flu, virus, fever, throat, doctor, people, cold, antibiotics, don't	sore throat, yeast infection, cold sores sinus infection, ear infection, strep throat, bacterial infection, staph infection, cold or flu, bladder infection	the, to, dormant, matador, disappear, visual, mederma, groin, spoke, bluetooth	sore throat, genital herpes, immune system, cold sores, cold sore, sinus infection, strep throat, chicken pox, runny nose, bacterial infection

Table 5 continued

Label	Model				
	LPLDA	LLDA	PhraseLDA	HMM-LDA	TNG
Respiratory diseases	sleep apnea, shortness of breath, asthma attack, sore throat, chest pain, sinus infection, cystic fibrosis, carbon monoxide, deep breath, lung disease	asthma, doctor, cough, lungs, breathing, smoking, smoke, sleep, chest, throat	sore throat, yeast infection, cold sores sinus infection, ear infection, strep throat, bacterial infection, staph infection, cold or flu, bladder infection	the, to, discover, nosey, damp, duller, sip, custom, trachea, stride	good luck, fall asleep, quit smoking, stop smoking, sleep apnea, cold turkey, blood pressure, hand smoke, sleeping pills, feel tired
Cancer	breast cancer, lung cancer, type of cancer, skin cancer, cancer cells, lymph nodes, prostate cancer, cure for cancer, colon cancer, cervical cancer	cancer, breast, doctor, treatment, chemo, years, time, good, people, don't	breast cancer, lymph nodes, cervical cancer, lung cancer, type of cancer, CT scan, cancer cells, prostate cancer, skin cancer, brain tumor	the, complexion, rs, martial, billion, squash, rhumatoid, atenolol, fibromyalgia, iu	breast cancer, side effects, immune system, lung cancer, heart disease, years ago, united states, lymph nodes, long term, ct scan

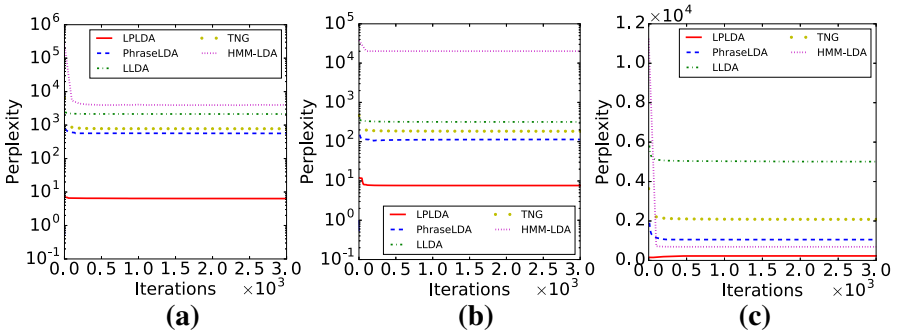

Fig. 2 A comparison of the perplexity of LPLDA versus the baseline models, LLDA, HMM-LDA, TNG and PhraseLDA, during Gibbs sampling inference on three datasets. **a Conf**. **b Twitter**. **c Yahoo! Answers**

Table 6 A comparison of the run-time of LPLDA versuss the baselines

Method	Datasets		
	Conf (h)	**Twitter** (h)	**Yahoo! Answers** (h)
LPLDA	**0.444**	**0.839**	**0.175**
PhraseLDA	13.865	5.464	0.728
HMM-LDA	17.608	6.035	2.509
TNG	78.783	39.204	6.391

The bold values are the results of our model

5.1.3 Perplexity

Besides comparing the case studies of the five models on the above corpus, we also evaluated the perplexity, which can measure the quality of different models quantificationally. The equation to evaluate perplexity for test dataset is shown in Eq. (10).

$$\text{Perplexity}(D) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{10}$$

Figure 2 shows the results. It is clear that the value of perplexity of LPLDA is much smaller than that of all the other four baselines. Therefore, it obviously indicates that LPLDA is much better than the baselines on all the three datasets in this experiment.

5.1.4 Scalability

To evaluate the scalability of our method, we compute our model's runtime on different datasets and compare them to HMM-LDA's, TNG's and PhraseLDA's. Since LLDA is based on bag-of-words assumption and does not contain the phrase mining step, it is unfair to compare the runtime of LPLDA and LLDA. The experimental results are shown in Table 6. It is obvious that LPLDA is much more efficient than all the baselines, especially when the datasets has a large number of labels. For example, in the **Conf**, the runtime of LPLDA is 0.444 h, however, in HMM-LDA, TNG and PhraseLDA the runtime increases to 17.608, 78.783 and 13.865 h respectively.

5.1.5 Multi-labeled text classification task

In this section, to evaluate the performance of LPLDA, we compare LPLDA and LLDA on a third party task, multi-labeled classifier for text data. Average precision and one-error (Schapire and Singer 2000) are widely used to evaluate the performance of models in multi-labeled text classification task, which are both rank-based evaluation metrics.

Non-interpolated average precision (Schapire and Singer 2000) can assess the label ranking of a multiclass system as a whole, and is used to evaluate the effectiveness of the label rankings. The higher the value of average precision, the better. For a ranking H with respect to a training set S, the average precision can be computed by Eq. (11).

$$avgprec_S(H) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{\ell\in Y_i}\frac{|\{\ell^{'} \in Y_i | rank_f(x_i, \ell^{'}) \leq rank_f(x_i, \ell)\}|}{rank_f(x, \ell)} \tag{11}$$

One-error (Schapire and Singer 2000) is a simple generalization of classification error for multiclass multi-label problems. It evaluates how many times the top-ranked label was not in the set of possible labels. Thus, if the goal of a multiclass system is to assign a single label to a document, the one-error measures how many times the predicted label was not in Y. The lower the value of one-error, the better. The equation to compute one-error is shown in Eq. (12).

$$one\text{-}err_S(H) = \frac{1}{m}\sum_{i=1}^{m}[[H(x_i) \notin Y_i]] \tag{12}$$

$H: X \rightarrow Y$ is a classifier that assigns a single label for a document x by setting $H(x) = arg\ max_{l\in Y} f(x, y)$, and m is the number of documents. $[[H(x_i) \notin Y_i]] = 1$ if $H(x_i) \notin Y_i$ and 0 otherwise.

We use 20% of each dataset mentioned above as test set in this experiment. In the training phase, words are assigned to topics within the label set of each document. When testing, we ignore the labels of each document, and allow each word in the document to be assigned to any topic in the dataset. We rank each dimension of θ^d in the test set, and consider the ranked θ^d as the feature of each document. We compute the above two evaluation metrics of LPLDA and LLDA on the three datasets, and the experimental results are shown in Table 7. On the **Yahoo! Answers**, our proposed LPLDA perform significantly better than LLDA, while on the **Conf** and **Twitter**, LPLDA fails. We analyzed the difference among these three datasets, and the reason may be the difference in the number of labels that leads to the different experimental results, i.e the number of **Yahoo! Answers** is 24, while the number of the **Conf** and **Twitter** are both more than 200. In order to confirm our guess, we further reduced the number of labels of the **Conf** and **Twitter** to 33 and 30 respectively, and conducted the experiment (seen in the last two lines of Table 7). The experimental result on the **Conf** is the same as the results we expected, however, result on the **Twitter** goes the opposite. The reason may be that text in **Twitter** are short text with colloquialism, and contain very few phrases that are not enough to provide enough features for text

Table 7 A comparison of the average precision and one-error of LPLDA versus LLDA

Datasets	Method			
	Average precision (%)		One-error (%)	
	LPLDA	LLDA	LPLDA	LLDA
Conf	26.31	28.47	66.24	58.97
Twitter	37.01	45.67	58.86	34.04
Yahoo! Answers	**70.17**	54.09	**18.99**	42.62
Conf$_{label\ number=33}$	**49.96**	49.27	**52.63**	55.26
Twitter$_{label\ number=30}$	56.63	62.46	28.39	17.58

The bold values are the results of our model

Table 8 Some example of phrases and the original documents learned by LPLDA on the **Twitter**

Original documents	Phrases
you might like this xo lol celebrity	xo, lol celebrity
rt tehran is tell of iranelection after to to	rt, tehran iranelection
swimming fail lol fail it looks like it she was boy	swim, fail lol fail, boy
some of the best are too long to rt tcot	long, rt, tcot
well be hiring for more than new jobs on october	hiring jobs, october

classification. We then examined the phrases in each document learned by LPLDA on the **Twitter**. Some examples of phrases and the original documents are shown in Table 8, and the different phrases are separated by commas. From the examples, we can see that when a tweet is too short, phrases learned by LPLDA cannot express the full meaning of the document.

Therefore, LPLDA is better than LLDA when the number of labels in the dataset is small and the average length of the documents is not too short, however, when the number of labels is large, LPLDA does not perform well with LLDA.

5.2 Experiment for online Labeled Phrase LDA

5.2.1 Experiment setting

In this experiment, we simulated the initial hyperparameters and set them the same as in the former subsection. For online LPLDA, we simulate the situation where documents are coming continuously. And we use 5% of each dataset mentioned above for initialization. Since LPLDA performs beter than any of the baselines (seen in Sect. 5.1), we only compare the performance of online LPLDA and batch LPLDA in this subsection. To control the variable of the input dataset for the batch and online algorithm for LPLDA, we excluded documents sampled in the initialization phase when computing perplexity and runtime. We ran online LPLDA and batch LPLDA on the three corpus. Both models are based on the standard collapsed Gibbs sampling. In online LPLDA, the number of particles is set to 10, and the threshold ESS is set to 1.5.

Table 9 Top ten phrases from five topics learned on the **Conf** in our proposed model batch LPLDA and online LPLDA, respectively

Label	Model	
	LPLDA	Online LPLDA
Multidomain search	search engines, search queries, Query Expansion, shown in Figure, World Wide Web, refining queries , query refinement, search for more information, Information Search and Retrieval, Information Storage and Retrieval	result set, data sources, search engines, search queries, Query Expansion, shown in Figure, World Wide Web, query refinement, set and the result, refining queries
Twitter	social network, social media, data mining, sentiment analysis, March April, WWW DemoMarch April, World Wide Web, machine learning, large number, web pages	social network, social media, data mining, sentiment analysis, March April, WWW DemoMarch April, World Wide Web, machine learning, web pages, Mining ModuleThe Data
Clustering	large scale, March April, WWW PosterMarch April, held by the author owner s WWW, social network, proposed method, clustering algorithms, recommender systems, large number, Categories and Subject DescriptorsH Information	large scale, WWW PosterMarch April, held by the author owner s WWW, March April, proposed method, social network, Computer Science, clustering algorithms, recommender systems, large number
Summarization	knowledge base, March April, WWW DemoMarch April, World Wide Web, web pages, Information Search, web search, search engines, Categories and Subject, clustering algorithms	March April, WWW DemoMarch April, World Wide Web, web pages, web search, social network, held by the author owner s WWW, shows the results, result set, search engines
Clickthrough data	Query Expansion, query similarity, similar queries, web search, search engines, search queries, results show, Information Search, Storage and Retrieval, Categories and Subject DescriptorsH Information	Query Expansion, query similarity, similar queries, web search, search engines, search queries, results show, WWW PosterMarch April, Information Search, held by the author owner s WWW

5.2.2 Case studies

We ran both batch LPLDA and online LPLDA over the three corpus and then picked up the top ten phrases of each topic learned from each algorithm to represent a topic. Each topic is associated with its corresponding label. Tables 9, 10 and 11 show the performance of the two algorithms. From the results, we can see most topics generated by online LPLDA are as interpretable as the batch algorithm, which indicates online LPLDA performs as well as LPLDA in case study, and can process large-scale data.

5.2.3 Perplexity

We then compared the perplexity of online LPLDA and batch LPLDA according to Eq. (10) with each iterating for 1000 times and converging to a certain value. Figure

Table 10 Top ten phrases from five topics learned on the **Twitter** in our proposed model batch LPLDA and online LPLDA, respectively

Label	Model	
	LPLDA	Online LPLDA
Photography	postrank photography, photography art, photo photography, travel photography, art photography, photography photo, photography travel, photography photog, blog post, photog photography	postrank photography, photography art, photo photography, travel photography, photography photo, art photography, photography travel, blog post, flickr photography, photography photog
PHP	php job follow, php job, job php freelance, jobs php, php developer, web developer, php freelance, job follow, software developer, job data entry	php job follow, php job, job php freelance, jobs php, php developer, web developer, php freelance, job follow, software developer, job data entry
Politics	postrank politics, news politics, politics news, fox news, tcot politics sgp, politics tcot, van jones, tcot tlot, tcot obama politics, tcot politics	news politics, politics news, postrank politics, fox news, tcot politics sgp, politics tcot, van jones, glenn beck, tcot tlot, tcot politics
Finance	finance money, professional jobs finance, jobs finance, professional jobs finance atlanta, money finance, senior accountant, economy finance, business finance, account manager, staff accountant	professional jobs finance, finance money, jobs finance, professional jobs finance atlanta, business finance, senior accountant, money finance, economy finance, account manager, staff accountant
iPhone	iphone app, iphone apple, apple iphone, app iphone, iphone ipod apps, blackberry or iphone, mac iphone, iphone ipod, iphone gs, iphone and ipod touch	iphone app, iphone apple, apple iphone, app iphone, blackberry or iphone, mac iphone, iphone ipod apps, iphone ipod, iphone and ipod touch, apple mac iphone

3 shows that the perplexity curves have a downward trend in general. The perplexity of online LPLDA is much smaller than that of batch LPLDA, which indicates that our online LPLDA is better than batch LPLDA. Moreover, in Figure 3a, the perplexity curve of batch LPLDA is not always downward, while that of online LPLDA is still downward. The reason may be that **Conf** contains much less documents compared with other datasets. We use 5% of each dataset as test set resulting test set of **Conf** containing only 60 documents. So it may be the small size of **Conf** that influences the difference in the perplexity curve of batch LPLDA in Fig.3a. But for online LPLDA, the size of the dataset does not have too much effect on its perplexity. Therefore, it is another merit of online LPLDA.

5.2.4 Efficiency

We further compared the run-time of online LPLDA and that of batch LPLDA, and the experimental results are shown in Fig. 4. From Fig. 4, we can see that online LPLDA is much more efficient than batch LPLDA. From the run-time increasing speed of view, the red curve is exponential growth, and the other one is linear growth. This means

Table 11 Top ten phrases from five topics learned on the **Yahoo! Answers** in our proposed model batch LPLDA and online LPLDA, respectively

Label	Model	
	LPLDA	Online LPLDA
Diabetes	ear tubes, genital herpes, ear infection, HPV types, herpes simplex, nerve damage, optic nerve, personal experience, reduce the risk, HPV infection	ear tubes, genital herpes, heart disease, HPV types, ear infection, immune system, nerve damage, cancer cells, herpes simplex, optic nerve
General Health Care	blood sugar, heart rate, weight loss, lymph nodes, blood test, lose weight, blood pressure, high blood pressure, immune system, blood flow	http www, blood sugar, heart rate, lymph nodes, weight loss, http www youandaids org, high blood pressure, healthcare professional, blood clot, blood test
Injuries	colorectal cancer, cervical cancer, colon cancer, risk of developing, Pulmonary Fibrosis, increased risk, pap smear, HIV infection, risk factors, genital warts	colorectal cancer, cervical cancer, colon cancer, risk of developing, increased risk, pap smear, HIV infection, Pulmonary Fibrosis, risk factors, genital warts
Other—General Health Care	genital warts, HPV virus, people living, cervical cancer, optic nerve, rib cage, HPV types, oral sex, beta blockers, HIV AIDS	genital warts, HPV virus, people living, cervical cancer, optic nerve, rib cage, oral sex, beta blockers, HIV AIDS, HPV types
Diseases and conditions	blood pressure, blood sugar, immune system, lose weight, high blood pressure, heart attack, heart disease, type diabetes, weight loss, low carb	http www, lose weight, blood sugar, blood pressure, http www youandaids org, http en wikipedia org wiki, high blood pressure, heart attack, type diabetes, web site

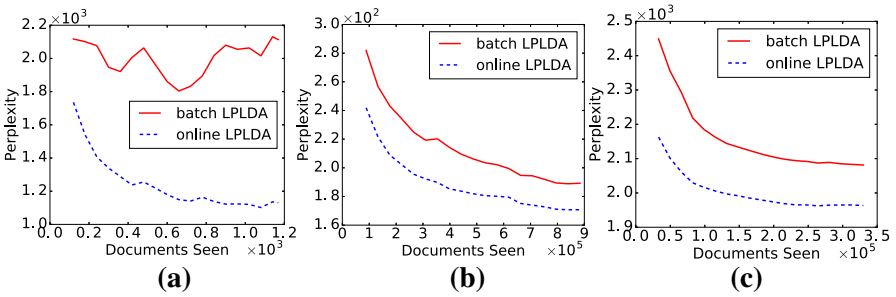

Fig. 3 A comparison of the perplexity of online LPLDA versus batch LPLDA during documents' arriving on three datasets. **a Conf**. **b Twitter**. **c Yahoo! Answers**

that for large-scale data processing, the online learning algorithm is significantly more efficient than batch processing. Therefore, the online LPLDA proposed in this paper can effectively process large-scale data efficiently.

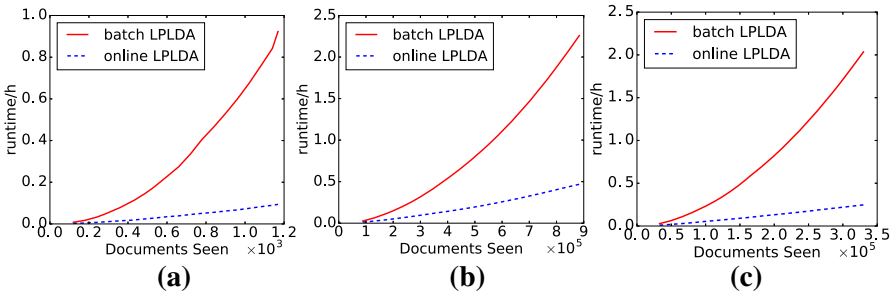

Fig. 4 A comparison of the run-time of online LPLDA versus batch LPLDA. **a Conf**. **b Twitter**. **c Yahoo! Answers**

6 Conclusion and future work

In this paper, we proposed a supervised topic model, called Labeled Phrase LDA (LPLDA). We also developed the batch and online inference algorithms for the proposed model. Extensive experiments have been conducted on three different datasets in different domain by comparing LPLDA with four state-of-the-art topic models, Labeled LDA, HMM-LDA, TNG and PhraseLDA. And we also compared the performance of the batch and online algorithms for LPLDA. The results of the experiments shows that our LPLDA model performs better than the baselines in terms of case study, perplexity and scalability, and the third party task in most cases. The online LPLDA is obviously more efficient than batch method under the premise of good results.

Further work may focus on methods to manage the following two challenges: (1) Filtrate similar phrases, especially in spoken corpus. (2) Consider the use of word order except phrase.

Acknowledgements This work was supported by National Key Research and Development Program of China (2016YFB1000902), China National Science Foundation (61402036, 61772076), Beijing Advanced Innovation Center for Imaging Technology (BAICIT-2016007), Open Fund Project from Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDD201701) and Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (MJUKF201738). A preliminary version of this work appears in Tang et al. (2016).

References

Agrawal R, Srikant R et al (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th very large data bases (VLDB) conference, vol 1215, pp 487–499

AlSumait L, Barbará D, Domeniconi C (2008) On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: 2008 Eighth IEEE international conference on data mining. IEEE, pp 3–12

Banerjee A, Basu S (2007) Topic models over text streams: a study of batch and online unsupervised learning. In: Proceedings of the 2007 SIAM international conference on data mining. SIAM, pp 431–436

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Broderick T, Boyd N, Wibisono A, Wilson AC, Jordan MI (2013) Streaming variational Bayes. In: Advances in neural information processing systems. Curran Associates Inc., pp 1727–1735

Canini KR, Shi L, Griffiths TL (2009) Online inference of topics with latent Dirichlet allocation. In: Proceedings of the twelfth international conference on artificial intelligence and statistics, vol 5. PMLR, pp 65–72

Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6(3):361–365

Elkishky A, Song Y, Wang C, Voss CR, Han J (2014) Scalable topical phrase mining from text corpora. In: Proceedings of The VLDB endowment, vol 8, no 3, pp 305–316

Foulds J, Boyles L, DuBois C, Smyth P, Welling M (2013) Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 446–454

Gao Y, Chen J, Zhu J (2016) Streaming Gibbs sampling for LDA model. ArXiv preprint arXiv:1601.01142

Ghahramani Z, Attias H (2000) Online variational Bayesian learning. Slides from talk presented at neural information processing systems workshop on online learning

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235

Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB (2005) Integrating topics and syntax. In: Advances in neural information processing systems, vol 17. MIT Press, pp 537–544

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam

Hoffman M, Blei D (2015) Stochastic structured variational inference. In: Proceedings of the eighteenth international conference on artificial intelligence and statistics, vol 38. PMLR, pp 361–369

Hoffman M, Bach FR, Blei DM (2010) Online learning for latent Dirichlet allocation. In: Advances in neural information processing systems, vol 23. Curran Associates Inc., pp 856–864

Hoffman MD, Blei DM, Wang C, Paisley JW (2013) Stochastic variational inference. J Mach Learn Res 14(1):1303–1347

Kingma DP, Welling M (2014) Auto-encoding variational Bayes. In: 2nd international conference on learning representations (ICLR2014), Ithaca, NY.

Lacoste-Julien S, Sha F, Jordan MI (2009) DiscLDA: discriminative learning for dimensionality reduction and classification. In: Advances in neural information processing systems, vol 21. Curran Associates Inc., pp 897–904

Lakkaraju H, Bhattacharyya C, Bhattacharya I, Merugu S (2011) Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: Proceedings of the 2011 SIAM international conference on data mining. SIAM, pp 498–509

Li X, Ouyang J, Zhou X (2016) Labelset topic model for multi-label document classification. J Intell Inf Syst 46(1):83–97

Liang S, Ren Z, Zhao Y, Ma J, Yilmaz E, Rijke MD (2017) Inferring dynamic user interests in short text streams for user clustering. ACM Trans Inf Syst (TOIS) 36(1):10:1–10:37

Lindsey RV, Headden III WP, Stipicevic MJ (2012) A phrase-discovering topic model using hierarchical Pitman–Yor processes. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, association for computational linguistics, pp 214–222

Magnusson M, Jonsson L, Villani M (2016) DOLDA—a regularized supervised topic model for high-dimensional multi-class regression. ArXiv preprint arXiv:1602.00260

Mao XL, Ming ZY, Chua TS, Li S, Yan H, Li X (2012) SSHLDA: a semi-supervised hierarchical topic model. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, pp 800–809

Mcauliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, vol 20. Curran Associates Inc., pp 121–128

McInerney J, Ranganath R, Blei D (2015) The population posterior and Bayesian modeling on streams. In: Advances in neural information processing systems, vol 28. Curran Associates Inc., pp 1153–1161

Mukherjee S, Basu G, Joshi S (2014) Joint author sentiment topic model. In: Proceedings of the 2014 SIAM international conference on data mining. SIAM, pp 370–378

Perotte AJ, Wood F, Elhadad N, Bartlett N (2011) Hierarchically supervised latent Dirichlet allocation. In: Advances in neural information processing systems, vol 24. Curran Associates Inc., pp 2609–2617

Petinot Y, McKeown K, Thadani K (2011) A hierarchical model of web summaries. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol 2. Association for Computational Linguistics, pp 670–675

Ramage D, Hall D, Nallapati R, Manning CD (2009a) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol 1. Association for Computational Linguistics, pp 248–256

Ramage D, Heymann P, Manning CD, Garcia-Molina H (2009b) Clustering the tagged web. In: Proceedings of the second ACM international conference on web search and data mining. ACM, pp 54–63

Ramage D, Manning CD, Dumais S (2011) Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 457–465

Ren Z, Liang S, Meij E, de Rijke M (2013) Personalized time-aware tweets summarization. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 513–522

Ren Z, Liang S, Li P, Wang S, de Rijke M (2017) Social collaborative viewpoint regression with explainable recommendations. In: Proceedings of the tenth ACM international conference on web search and data mining. ACM, pp 485–494

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence. AUAI Press, pp 487–494

Rubin TN, Chambers A, Smyth P, Steyvers M (2012) Statistical topic models for multi-label document classification. Mach Learn 88:157–208

Schapire RE, Singer Y (2000) BoosTexter: a boosting-based systemfor text categorization. Mach Learn 39:135–168

Shi T, Zhu J (2014) Online Bayesian passive-aggressive learning. In: Proceedings of the 31st international conference on international conference on machine learning, vol 32. JMLR.org, pp I-378–I-386

Slutsky A, Hu X, An Y (2013) Tree labeled LDA: a hierarchical model for web summaries. In: IEEE international conference on big data. IEEE, pp 134–140

Song X, Lin CY, Tseng BL, Sun MT (2005) Modeling and predicting personal information dissemination behavior. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. ACM, pp 479–488

Spagnola S, Lagoze C (2011) Word order matters: measuring topic coherence with lexical argument structure. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. ACM, pp 21–24

Tang J, Zhang M, Mei Q (2014) "Look ma, no hands!" A parameter-free topic model. ArXiv preprint arXiv:1409.2993

Tang YK, Mao XL, Huang H (2016) Labeled phrase latent Dirichlet allocation. In: International conference on web information systems engineering. Springer, pp 525–536

Tang YK, Mao XL, Huang H, Shi X, Wen G (2018) Conceptualization topic modeling. Multimedia Tools Appl 77(3):3455–3471

Wallach HM (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp 977–984

Wang C, Danilevsky M, Desai N, Zhang Y, Nguyen P, Taula T, Han J (2013) A phrase mining framework for recursive construction of a topical hierarchy. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 437–445

Wang X, McCallum A, Wei X (2007) Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: Seventh IEEE international conference on data mining (ICDM 2007). IEEE, pp 697–702

Wang Y, Agichtein E, Benzi M (2012) TM-LDA: efficient online modeling of latent topic transitions in social media. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 123–131

Xiao H, Wang X, Du C (2009) Injecting structured data to generative topic model in enterprise settings. In: Advances in machine learning: first Asian conference on machine learning, ACML 2009. Springer, Berlin, pp 382–395

Xiao X, Xiong D, Zhang M, Liu Q, Lin S (2012) A topic similarity model for hierarchical phrase-based translation. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers, vol 1. Association for Computational Linguistics, pp 750–758

Zhang A, Zhu J, Zhang B (2013) Sparse online topic models. In: Proceedings of the 22nd international conference on world wide web. ACM, pp 1489–1500

Zhao WX, Wang J, He Y, Nie J, Wen J, Li X (2015) Incorporating social role theory into topic models for social media content analysis. IEEE Trans Knowl Data Eng 27(4):1032–1044

Zhao Y, Liang S, Ren Z, Ma J, Yilmaz E, de Rijke M (2016) Explainable user clustering in short text streams. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 155–164

Zhou Q, Huang H, Mao XL (2015) An online inference algorithm for labeled latent Dirichlet allocation. In: Proceedings on web technologies and applications: 17th Asia-Pacific web conference, APWeb 2015, Guangzhou, China, 18–20 Sept 2015. Springer, pp 17–28

Zhu J, Chen N, Perkins H, Zhang B (2013) Gibbs max-margin topic models with fast sampling algorithms. In: Proceedings of the 30th international conference on machine learning, vol 28. PMLR, pp 124–132