Inf Retrieval J (2015) 18:1–25
DOI 10.1007/s10791-014-9244-9

Group topic model: organizing topics into groups

Ximing Li · Jihong Ouyang · You Lu · Xiaotang Zhou · Tian Tian

Received: 12 March 2014 / Accepted: 27 August 2014 / Published online: 10 September 2014

Abstract Latent Dirichlet allocation defines hidden topics to capture latent semantics in text documents. However, it assumes that all the documents are represented by the same topics, resulting in the "forced topic" problem. To solve this problem, we developed a group latent Dirichlet allocation (GLDA). GLDA uses two kinds of topics: local topics and global topics. The highly related local topics are organized into groups to describe the local semantics, whereas the global topics are shared by all the documents to describe the background semantics. GLDA uses variational inference algorithms for both offline and online data. We evaluated the proposed model for topic modeling and document clustering. Our experimental results indicated that GLDA can achieve a competitive performance when compared with state-of-the-art approaches.

Keywords Topic modeling · Latent Dirichlet allocation · Group · Variational inference · Online learning · Document clustering

1 Introduction

Large text document collections have recently become readily available online. Systematic analyses of these collections are significantly meaningful to various domains. Consider, for

X. Li · J. Ouyang (✉) · Y. Lu · X. Zhou
College of Computer Science and Technology, Jilin University, Changchun, China
e-mail: ouyj@jlu.edu.cn

X. Li
e-mail: liximing86@gmail.com

X. Li · J. Ouyang · Y. Lu · X. Zhou
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China

T. Tian
College of Software, Jilin University, Changchun, China

example, scientific article archives. (1) We want to organize the articles by subject, and help users explore the archives. This is commonly a substantial multi-label classification task. (2) We want to analyze the article browsing histories of researchers, and build a recommendation system that can list all the interesting and relevant articles. (3) In terms of the submitted journal/conference manuscripts, we want to design a system that can recommend the most professional reviewers.

Although these problems have been well studied, they are still significantly challenging problems in machine learning research. This is because: (1) the text document collections commonly have a high dimensionality, and (2) it is difficult to learn a document's semantics and any correlations between documents. Statistical modeling research has addressed these challenges and developed various approaches for analyzing text documents (Koller and Friedman 2009). In particular, topic modeling approaches (Blei 2012) have provided a realizable avenue for expressing the latent semantics and hidden structures of documents. As a result, these approaches have been widely used in different application domains.

Latent Dirichlet allocation (LDA) (Blei et al. 2003) is acknowledged as one of the most successful topic modeling approaches. In LDA, each document is represented by a distribution over latent topics, and each topic is described by a distribution over words. It defines a Dirichlet prior beyond all the document-topic distributions, so it does not suffer from the same parameter explosion and over-fitting problems as the probabilistic latent semantic index (PLSI) approach (Hofmann 1999). Recently, researchers have proposed many extensions to LDA in terms of various considerations, i.e., relaxing the LDA assumptions (Blei et al. 2010; Blei and Lafferty 2006, 2007; Doyle and Elkan 2009; Wallach 2006; Wang et al. 2009), incorporating meta data (Blei and McAuliffe 2007; Chang and Blei 2010), and applying it to other kinds of data (Li and Perona 2005; Sivic et al. 2008).

In this paper, we focused on a topic modeling approach and investigated relaxing the assumptions of LDA. Intuitively, we know that larger document collections may contain more latent topics. To capture the latent semantics, all documents in LDA are represented by the same K topics. This leads to a "forced topic" problem. For example, consider a large academic paper archive that covers many latent topics such as "artificial intelligence", "data mining", "network", "inorganic chemistry", "organic chemistry", and "high-polymer chemistry". Computer science articles may only contain the three computer-related topics, whereas chemistry articles prefer to cover the three chemistry-related topics. But in LDA, all articles must cover all six topics. That is, chemistry articles do not involve the "network" topic, but they are forced to cover it in LDA. We can organize documents into different groups (i.e., computer science and chemistry) and then assign related topics to the groups (e.g., "network" to computer science and "organic chemistry" to chemistry), as a reasonable method to tackle the "forced topic" problem.

By considering the discussions above, we developed an extension of the LDA model, namely group latent Dirichlet allocation (GLDA). In GLDA, there are two kinds of topics: local topics and global topics. A local topic corresponds to a topic that only occurs in the subset of the corpus, whereas the global topic corresponds one that is ubiquitous to the whole corpus. Closely related local topics are clustered together as latent groups. Each document first selects a group, and then generates the topic distribution over both the local topics with respect to the selected group and the global topics. Finally, it samples words from the corresponding topic-word distributions. Based on the latent groups, GLDA can model the documents using the most related topics, rather than constraining each document to all the topics. In this paper, we used the variational inference algorithm and parameter estimation for

GLDA. Additionally, we developed an online inference algorithm to model large-scale data. We conducted a number of experiments on topic modeling and document clustering to evaluate the proposed model. Our experimental results demonstrate that GLDA can achieve a competitive performance when compared with state-of-the-art approaches.

The rest of this paper is organized as follows. In Sect. 2, we review topic modeling approaches. In Sect. 3, we describe the proposed GLDA model. Our evaluation results are presented in Sect. 4, and our conclusions and some potential future work are discussed in Sect. 5.

2 Topic modeling approach

In this section, we review the history of topic modeling approaches. Table 1 summarizes several important notations used in this paper.

To the best of our knowledge, PLSI (Hofmann 1999) was the first well known topic model for latent semantic analysis (Deerwester et al. 1990). However, it suffers from two intractable problems: parameter explosion and over-fitting. Blei et al. (2003) proposed LDA to tackle these two problems by introducing a Dirichlet prior to the latent topics. They also developed an effective variational inference algorithm to infer the model. As a result, LDA is in widespread use. As shown in Fig. 1a, the generative process of LDA is summarized as follows:

1. For each topic k
   (a) Sample a distribution over words: $\phi_k \sim Dirichlet\,(\beta)$
2. For each document d in the corpus W
   (a) Sample a distribution over topics: $\theta_d \sim Dirichlet\,(\alpha)$
   (b) For each of the N_d words $w_{d,n}$
      i. Sample a topic $z_{d,n} \sim Multinomial\,(\theta_d)$
      ii. Sample a word $w_{d,n} \sim Multinomial\left(\phi_{z_{d,n}}\right)$

In topic modeling research, one active direction is to relax the LDA assumptions to further uncover more sophisticated structures in the documents. Traditionally, the extensions of LDA focus on four fundamental assumptions (Blei 2012): "bag of words", "bag of documents", "fixed topics", and "independent topics".

1. "Bag of words" is an exchangeable assumption that the orders of words in documents do not matter. Although this assumption is reasonable for uncovering a coarse semantic structure and has benefits for computation, it is unrealistic in the sense of human cognition. Wallach proposed the bigram topic model (Wallach 2006), in which the word generation is associated with both its topic and context, i.e., the previous word. Wang et al. (2007) developed the topic N-gram model, which discovers phrases using the word orders and adjacent topics. Boyd-Graber and Blei (2008) considered the syntactic structure and proposed the syntactic topic model. These approaches model words non-exchangeably and improve the language modeling performance.
2. "Bag of documents" is also an exchangeable assumption that the orders of documents in collections do not matter. This assumption is unreasonable for collections that span years. Blei and Lafferty (2006) proposed the dynamic topic model, where the topics

Table 1 Notation description

Notation	Description
D	Number of documents
C	Number of groups
V	Number of words
K	Number of topics in LDA
Kl	Number of local topics for each group
Kg	Number of global topics
Kd	Number of topics for a document, i.e., $Kd = Kl + Kg$
KK	Total number of topics in GLDA, i.e., $KK = C * Kl + Kg$
N_d	Number of unique words w.r.t document d
π	The C-dimension group distribution
θ_d	The K/Kd-dimension topic distribution in LDA/GLDA w.r.t document d
ϕ_k	The V-dimension word distribution w.r.t topic k
α	The topic Dirichlet prior in LDA
$\alpha^{(g)}$	The global topic Dirichlet prior
$\alpha_c^{(l)}$	The local topic Dirichlet prior w.r.t group c
β	The Dirichlet prior for word distributions
z	The topic assignments for all words

change over time. In this method, each topic is a sequence of distributions over words, so it can capture the dynamic latent semantics.

3. The "fixed topics" assumption signifies that the number of topics in LDA is fixed and known. Typically, we must determine the number of topics experimentally. To address this problem, Blei et al. (2010) developed Bayesian nonparametric topic models using the Dirichlet process (Teh et al. 2006). In such topic models, the number of topics is determined by the data itself. Furthermore, they can explore hierarchies of topics such as the tree of topics.
4. "Independent topics" is the assumption that, in LDA, topics are independent from each other. The pachinko allocation model (Li and McCallum 2006) used a topic directed acyclic graph to describe the correlation among topics. With the same goal, the correlated topic model (Blei and Lafferty 2007) used the logistic normal prior for per-document topic proportion, instead of the Dirichlet prior in LDA.

There are other modified topic models that relax various assumptions with respect to LDA, for example, as the spherical topic model (Reisinger et al. 2010) and sparse topic model (sparseTM) (Reisinger et al. 2009). Recently, Wallach (2008) argued that it was unrealistic to force each document to associate with the same K topics. Considering this analysis, they proposed a cluster based topic model (CTM), which organizes topics into different groups and individualizes each group using the group-specific Dirichlet prior of the document-topic distribution. For each document, CTM first generates a group indicator and then samples the local topic distribution from the Dirichlet prior specific to the selected group. Based on CTM, Xie and Xing (2013) further introduced global topics to capture the global semantics, and proposed the multi-grain cluster topic model (MGCTM). As shown in Fig. 1b, for each word, MGCTM must select between local and global topics, and then generate local or global topics using its choice. In our work, we investigated how to relax

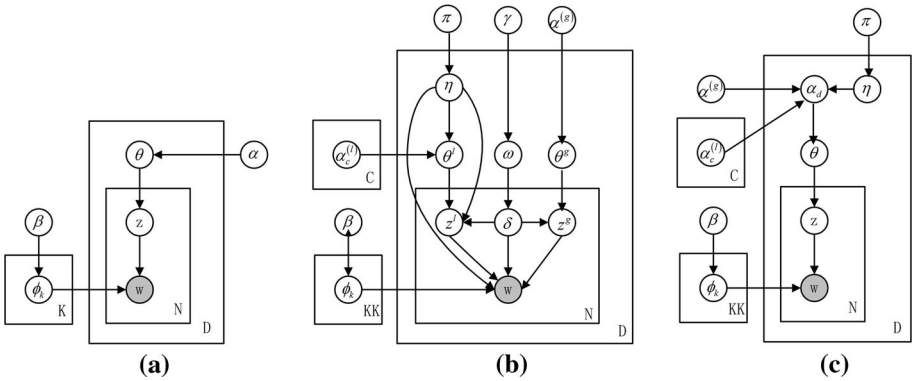

Fig. 1 The three topic models: **a** LDA, **b** MGCTM, and **c** GLDA

the assumptions of LDA, and propose the GLDA model. The proposed GLDA model defines local topics specific to a group as a solution to the "forced topic" problem, and defines global topics to capture the background semantics. It samples the document-topic distributions from a combination of the Dirichlet prior with the selected group's local prior and the global prior. The GLDA representation is less ambiguous than MGCTM. More importantly, GLDA further considers the relationships between local topics and global topics in terms of different groups. There are more detailed discussions in Sect. 3.5.

3 Proposed approach

In this section, we first introduce the GLDA model, and then propose the procedures for inference, parameter estimation and online learning. Finally, we compare MGCTM and GLDA in detail.

3.1 GLDA

In LDA, all documents are represented by the same K topics. This results in the "forced topic" problem, which has two aspects: (1) in practice, the documents belonging to different groups might only involve some topics, but they are forced to cover all topics (an example is shown in Sect. 1); and (2) LDA has no mechanism to cover the background semantics in a corpus, so the semantics must fill in each specific topic. For instance, the words that cover the background semantics are commonly ubiquitous and frequently occur in the corpus. As a result, these meaningless words might dominate topics, e.g., "introduction" to "network" and "organic chemistry". This behavior reduces the expressiveness of the topics.

To address the "forced topic" problem mentioned above, we extended LDA to the GLDA model. In GLDA, we assume that (1) there is a corpus-level multinomial distribution π, which can be used to generate the group indicator for the documents; (2) each group c corresponds to Kl local topics behind the Dirichlet prior $\alpha_c^{(l)}$; and (3) to capture the background semantics, all documents share Kg global topics behind the Dirichlet prior $\alpha^{(g)}$. To formalize the generation process for document d, we first choose a group indicator η_d from the distribution π. We combine the local Dirichlet prior of group η_d with the global

Dirichlet prior, to obtain a merged Kd-dimension Dirichlet prior, that is $\alpha_d = \left[\alpha_{\eta_d}^{(l)}, \alpha^{(g)}\right]$. Then, we sample the document-topic distribution θ_d over the local topics with respect to the selected group η_d, and the global topics (i.e., "selected topics") from the Dirichlet prior α_d. The words are then generated as in LDA.

As shown in Fig. 1c, the generative process of GLDA is as follows:

1. For each topic k
   (a) Sample a distribution over words: $\phi_k \sim Dirichlet\,(\beta)$
2. For each document d in the corpus W
   (a) Sample a group: $\eta_d \sim Multinomial\,(\pi)$
   (b) Sample a distribution over the "selected topics": $\theta_d \sim Dirichlet\,(\alpha_d)$
   (c) For each of the N_d words $w_{d,n}$
      i. Sample a topic $z_{d,n} \sim Multinomial\ (\theta_d)$
      ii. Sample a word $w_{d,n} \sim Multinomial\left(\phi_{z_{d,n}}\right)$

We can summarize model parameters as $U = \left\{\pi,\ \left\{\alpha_c^{(l)}\right\}_{c=1}^{C}, \alpha^{(g)}, \beta\right\}$ and the latent variables as $H = \left\{\left\{z_{d,n}\right\}_{d=1,n=1}^{d=D,n=N_d},\ \{\eta_d\}_{d=1}^{D},\ \{\theta_d\}_{d=1}^{D},\ \{\phi_k\}_{k=1}^{KK}\right\}$.

Reviewing this generation process, we argue that GLDA solves the "forced topics" problem to some extent. On one hand, in contrast to LDA, GLDA gives a two-stage procedure to generate topics. For each document, if a group is chosen then only the topics in this group can be used to describe the document. On other hand, GLDA introduces the concept of global topics to gather the words that describe the background semantics. This helps to purify the specific topics.

3.2 Inference

Given a corpus W, the key inference problem with respect to GLDA is to compute the posterior distribution of the latent variables $p(H|W, U)$. Because this posterior distribution is intractable to estimate, we use the variational inference (Blei et al. 2003) algorithm for approximate estimation.

The basic idea behind variational inference is to use Jensen's inequality to approach the tightest lower bound on the log likelihood. To achieve this, we introduced the variational distribution $q(H|\Omega)$ (see Fig. 2) in terms of the free variational parameters $\Omega = \left\{\{\tilde{\pi}_d\}_{d=1}^{D},\ \{\tilde{\alpha}_d\}_{d=1}^{D},\ \left\{\tilde{\beta}_k\right\}_{k=1}^{KK},\ \left\{\tilde{\theta}_{d,n}\right\}_{d=1,n=1}^{d=D,n=N_d}\right\}$, by removing the coupling edges and nodes in GLDA. That is,

$$q(H|\Omega) = \prod_{k=1}^{KK} q(\phi_k|\tilde{\beta}_k) \prod_{d=1}^{D} \left(q(\eta_d|\tilde{\pi}_d) q(\theta_d|\tilde{\alpha}_d) \prod_{n=1}^{N_d} q(z|\tilde{\theta}_{d,n}) \right) \tag{1}$$

where $\{\tilde{\alpha}_d\}_{d=1}^{D}$ and $\left\{\tilde{\beta}_k\right\}_{k=1}^{KK}$ are Dirichlet parameters; and $\{\tilde{\pi}_d\}_{d=1}^{D}$ and $\left\{\tilde{\theta}_{d,n}\right\}_{d=1,n=1}^{d=D,n=N_d}$ are multinomial distribution parameters.

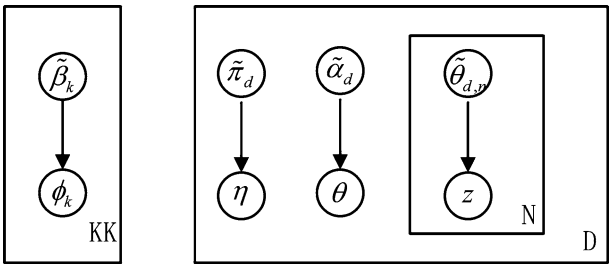

Fig. 2 The graphical model representation of the variational distribution

We transformed the task of finding the tightest lower bound on the log likelihood into the problem of maximizing the following lower bound:

$$\mathscr{L}(\Omega|U) = E_q[\log p(H, W|U)] - E_q[\log q(H|\Omega)] \tag{2}$$

which is described in the Appendix.

We use the fixed point method to maximize this lower bound with respect to the free variational parameters Ω. The derivation of this process is also shown in the Appendix. The updating rules are:

$$\tilde{\pi}_{d,c} \propto \pi_c \times \exp\left(\begin{array}{c} \log\Gamma\left(\sum_{k=1}^{Kd}\alpha_k^{(c)}\right) - \sum_{k=1}^{Kd}\log\Gamma\left(\alpha_k^{(c)}\right) \\ + \sum_{k=1}^{Kd}\left(\alpha_k^{(c)} - 1\right)\left(\Psi\left(\tilde{\alpha}_{d,k}\right) - \Psi\left(\sum_{j=1}^{Kd}\tilde{\alpha}_{d,j}\right)\right) \\ + \sum_{n=1}^{N_d}\sum_{k=1}^{Kl}\tilde{\theta}_{d,n,k}\left(\Psi\left(\tilde{\beta}_{c\cdot k,w_{dn}}\right) - \Psi\left(\sum_{j=1}^{V}\tilde{\beta}_{c\cdot k,j}\right)\right) \end{array}\right) \tag{3}$$

$$\tilde{\theta}_{d,n,k} \propto \exp\left(\begin{array}{c} \left(\Psi\left(\tilde{\alpha}_{d,k}\right) - \Psi\left(\sum_{j=1}^{Kd}\tilde{\alpha}_{d,j}\right)\right) \\ + \sum_{c=1}^{C}\tilde{\pi}_{d,c}\left(\Psi\left(\tilde{\beta}_{c\cdot k,w_{dn}}\right) - \Psi\left(\sum_{j=1}^{V}\tilde{\beta}_{k,j}\right)\right) \end{array}\right) \tag{4}$$

$$\tilde{\alpha}_{d,k} = \sum_{c=1}^{C}\tilde{\pi}_{d,c}\alpha_k^{(c)} + \sum_{n=1}^{N_d}\tilde{\theta}_{d,n,k} \tag{5}$$

where $\alpha^{(c)}$ is the Dirichlet prior that combines the local topics specific to group c with the global topics (i.e., $\alpha^{(c)} = \left[\alpha_c^{(l)}, \alpha^{(g)}\right]$); $\tilde{\beta}_{c\cdot k}$ corresponds to the kth $\tilde{\beta}$ of group c; $\Gamma(\cdot)$ is the gamma function; and $\Psi(\cdot)$ is the digamma function Then,

$$\begin{cases} \tilde{\beta}_{k,v} = \beta_v + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \tilde{\theta}_{d,n,k} w_{d,n}^v & \textit{if } k \textit{ is global} \\ \tilde{\beta}_{c \cdot k,v} = \beta_v + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \tilde{\pi}_{d,c} \tilde{\theta}_{d,n,k} w_{d,n}^v & \textit{otherwise} \end{cases} \tag{6}$$

where

$$w_{d,n}^v = \begin{cases} 1 & \textit{if } w_{d,n} = v \\ 0 & \textit{otherwise} \end{cases}$$

The full variational inference procedure is summarized in *Algorithm* 1.

3.3 Parameter estimation

In this section, we consider the parameter estimation for GLDA. Given a corpus, we wish to optimize the model parameters (U) using a maximum likelihood estimation. Again, the

Algorithm 1 Variational inference for GLDA

```
Initialize latent variables Ω randomly.
While Change in |ℒ(Ω|U)|>0.0001 do.
    E-step:
        For d=1 to D
            Repeat
                Compute π̃_d according to Eq. (3)
                Compute θ̃_d according to Eq. (4)
                Compute α̃_d according to Eq. (5)
            Until Change in |α̃_d|<0.0001
        End for
    M-step:
        Compute β̃_k according to Eq. (6)
End while
```

likelihood function $p(W|U)$ is intractable to compute. So, we use the variational expectation maximization (variational EM) algorithm, which alternatively updates the free variational parameters (Ω) and model parameters (U).

Similar to *Algorithm* 1, the variational EM algorithm is summarized in *Algorithm* 2. In the E-step, we infer $\tilde{\pi}_d, \tilde{\alpha}_d, \tilde{\theta}_d$ using Eqs. (3), (4) and (5) from Sect. 3.2. In the M-step, we estimate $\tilde{\beta}_k$ and U. $\tilde{\beta}_k$ is also updated using Eq. (6). The Dirichlet parameters $(\alpha^{(l)}, \alpha^{(g)}, \beta)$ are optimized by the Newton–Raphson method described in (Blei et al. 2003), and the multinomial parameter π is updated using:

$$\pi_c = \frac{\sum_{d=1}^{D} \tilde{\pi}_{d,c}}{D} \tag{7}$$

Algorithm 2 Variational EM for GLDA

Initialize latent variables Ω randomly.
While Change in $|\mathscr{L}(\Omega|U)|$>0.0001 do.
  E-step:
    For d=1 to D
      Repeat
        Compute $\tilde{\pi}_d$ according to Eq. (3)
        Compute $\tilde{\theta}_d$ according to Eq. (4)
        Compute $\tilde{\alpha}_d$ according to Eq. (5)
      Until Change in $|\tilde{\alpha}_d|$<0.0001
    End for
  M-step:
    Compute $\tilde{\beta}_k$ according to Eq. (6)
    Optimize $\alpha^{(l)}, \alpha^{(g)}, \beta$ using Newton-Raphson algorithm
    Compute π according to Eq. (7)
End while

Comparison with asymmetric LDA GLDA organizes topics into groups. To specialize different groups, we apply asymmetric Dirichlet priors for local topics, so GLDA is in default an asymmetric model. It seems similar with the best version of asymmetric LDA, i.e., AS form (asymmetric topic Dirichlet prior and symmetric word Dirichlet prior), suggested in (Wallach et al. 2009a), so we have stated the relationships between the two models. When inferring a document d, GLDA might equal to AS form LDA with a certain value of $\tilde{\pi}_d$. However, this is infrequent in practice; and more importantly, the values of π_d are totally different for different documents. In other words, we believe that GLDA and asymmetric LDA are two disparate models.

3.4 Online learning

In this section, we extend *Algorithm* 2 to an online inference algorithm (Online GLDA) for modeling large-scale data. This work is based on the spirit of stochastic variational inference (SVI) (Hoffman and Wang 2013; Hoffman and Blei 2010), where each iteration uses only a mini-batch of the documents to generate a stochastic gradient, and a stochastic optimization algorithm is used to learn the global parameters of interest.

In the GLDA context, the local variational parameters are $\tilde{\pi}_d, \tilde{\alpha}_d, \tilde{\theta}_d$ and the global parameters are $\tilde{\beta}_k, \alpha^{(l)}, \alpha^{(g)}, \beta, \pi$. At each iteration t, we first randomly sample M documents and compute their optimal local variational parameters using Eqs. (3), (4) and (5). We then update the global parameters given a learning rate ρ_t as follows:

In terms of $\tilde{\beta}$, we can compute the natural gradient $\nabla_{\tilde{\beta}}\mathscr{L}(\Omega|U)$, and then give the updating rule:

$$\begin{cases} \tilde{\beta}_{k,v} \leftarrow \tilde{\beta}_{k,v} + \rho_t \left(-\tilde{\beta}_{k,v} + \beta_v + \dfrac{D}{M} \displaystyle\sum_{d=1}^{M} \sum_{n=1}^{N_d} \tilde{\theta}_{d,n,k} w_{d,n}^{v} \right) & \textit{if } k \textit{ is global} \\ \tilde{\beta}_{c\bullet k,v} \leftarrow \tilde{\beta}_{c\bullet k,v} + \rho_t \left(-\tilde{\beta}_{c\bullet k,v} + \beta_v + \dfrac{D}{M} \displaystyle\sum_{d=1}^{M} \sum_{n=1}^{N_d} \tilde{\pi}_{d,c} \tilde{\theta}_{d,n,k} w_{d,n}^{v} \right) & \textit{otherwise} \end{cases} \tag{8}$$

In terms of $\alpha^{(l)}, \alpha^{(g)}, \beta$, we extend the Newton–Raphson algorithm to the online case as in (Hoffman and Blei 2010):

$$\begin{cases} \alpha_{c,k}^{(l)} \leftarrow \alpha_{c,k}^{(l)} - \rho_t \hat{\alpha}^{(l)} \\ \alpha_k^{(g)} \leftarrow \alpha_k^{(g)} - \rho_t \hat{\alpha}^{(g)} \\ \beta_v \leftarrow \beta_v - \rho_t \widehat{\beta} \end{cases} \tag{9}$$

where $\hat{\alpha}^{(l)}$ and $\hat{\alpha}^{(g)}$ are the inverse of the Hessian times the gradient $\nabla_{\alpha^{(l)}} \mathscr{L}(\Omega|U)$ and $\nabla_{\alpha^{(g)}} \mathscr{L}(\Omega|U)$; and $\hat{\beta}$ is the inverse of the Hessian times the gradient $\nabla_{\beta}^{\sim} \mathscr{L}(\Omega|U)$.

This is different to the global variational parameters above; it is a constrained maximization of π because $\sum_{c=1}^{C} \pi_c = 1$. So we must subtract a form of projection Z (Zinkevich 2003), when updating π:

$$\pi_c \leftarrow \pi_c + \rho_t \pi_c \left(D \sum_{d=1}^{M} \widetilde{\pi}_{d,c} D \sum_{d=1}^{M} \widetilde{\pi}_{d,c} \Big/ M\pi_c - Z \right), \qquad \text{where } Z = \frac{\left\langle \pi, D \sum_{d=1}^{M} \widetilde{\pi}_d \Big/ M\pi \right\rangle}{C} \tag{10}$$

and $\langle\ ,\ \rangle$ is the inner product.

The Online GLDA is summarized in *Algorithm* 3.

Algorithm 3 Online inference for GLDA

```
Initialize latent variables Ω randomly.
For t = 1, 2, ···, ∞ do
    Sample M documents.
    For d=1 to M do
        Repeat
            Compute π̃_d according to Eq. (3)
            Compute θ̃_d according to Eq. (4)
            Compute α̃_d according to Eq. (5)
        Until Change in |α̃_d|<0.0001
    End for
    Update β̃_k according to Eq. (8)
    Update α^(l), α^(g), β according to Eq. (9)
    Update π according to Eq. (10)
End For
```

3.5 Comparison with MGCTM

MGCTM and GLDA have some similarities, so we have investigated the relationships between the two topic models. In fact, both represent documents using the local topics of the groups and global topics. However, the relationships between the two kind topics are different in MGCTM and GLDA.

A graphical model representation is shown in Fig. 1b. We reviewed the generative process of MGCTM (Xie and Xing 2013) as follows: For each document d, first select a group η_d from the distribution π. Then sample a local topic distribution $\theta_{\eta_d}^{l}$ from the Dirichlet prior $\alpha_{\eta_d}^{(l)}$ with respect to the selected group, and sample a global topic distribution θ^g from the Dirichlet prior $\alpha^{(g)}$. The Beta prior γ samples a Bernoulli distribution ω_d, which is used to make choice between local and global topics. To generate a word $w_{d,n}$, we first choose a topic indicator $\delta_{d,n}$ from the distribution ω_d. If $\delta_{d,n} = 1$, the word $w_{d,n}$ will be assigned a local topic with respect to the group η_d. If $\delta_{d,n} = 0$, the word $w_{d,n}$ will be assigned a global topic. Finally a word is generated as in LDA.

Let $p(t^g = k)$ be the probability of generating the kth global topic. In MGCTM $p(t^g = k) = p\left(\delta_{d,n} = 0|\omega_d\right) \cdot p(t^g = k|\theta^g)$ and its expectation is:

$$E_p[k] = \frac{\gamma_1}{\gamma_1 + \gamma_2} \times \frac{\alpha_k^{(g)}}{\sum_{i=1}^{Kg} \alpha_i^{(g)}} \tag{11}$$

In contrast, GLDA samples the document-topic distribution from the combination Dirichlet prior $\alpha_d = \left[\alpha_{\eta_d}^{(l)}, \alpha^{(g)}\right]$. Therefore, in GLDA equals:

$$E_p[k] = \frac{\alpha_k^{(g)}}{\sum_{i=1}^{Kd} \alpha_{d,i}} \tag{12}$$

We can transform Eq. (12) into:

$$E_p[k] = \frac{\sum_{i=1}^{Kg} \alpha_i^{(g)}}{\sum_{i=1}^{Kg} \alpha_i^{(g)} + \sum_{i=1}^{Kl} \alpha_{\eta_d,i}^{(l)}} \times \frac{\alpha_k^{(g)}}{\sum_{i=1}^{Kg} \alpha_i^{(g)}} \tag{13}$$

Comparing this with Eqs. (11) and (13), we found that the second terms are the same. The first terms are the probabilities of choosing global topics. Because each group c has its specific local topic prior $\alpha_c^{(l)}$, the first term of Eq. (13) should be different for different groups. That is to say, in GLDA the relationships between local and global topics are dependent on the different groups, which is not the case in MGCTM. We argue that the assumption in GLDA is reasonable. For example, computer science articles may be naturally more willing to cover common knowledge topics (i.e., global topics) than chemistry articles. In particular, we argue that this consideration is more significant when modeling collections that contain many latent groups.

4 Experiment

In this section, we present our results when evaluating GLDA on two problem domains, i.e., topic modeling and document clustering.

4.1 Dataset

We considered two widely used offline datasets:[1] 20-NewsGroups (20-NG) and WebKB. 20-NG is a balanced dataset. It contains 18,821 documents, which are equally divided into 20 related categories. We used 11,293 documents as the training data, and the remaining 7,528 documents as the testing data. WebKB contains 4,199 documents, which consists of four categories. In contrast to 20-NG, it is an unbalanced dataset, where the largest category contains 1,641 documents and the smallest category only contains 504 documents. We selected 2,803 documents for training, and used the remaining 1,396 documents for testing.

We also chose an online collection. We randomly downloaded 3M documents from *Wikipedia* (Wiki) using the implementation[2] in (Hoffman and Blei 2010). We then

[1] http://web.ist.utl.pt/~acardoso/datasets/.

[2] http://www.cs.princeton.edu/~mdhoffma/.

processed these documents using a standard vocabulary of 7,700 words. We used 2,000 randomly selected documents from the collection for testing.

4.2 Topic modeling

We evaluated the topic modeling performance of GLDA across the three selected corpora. In terms of the offline datasets, we used three state-of-the-art topic models (LDA Blei et al. 2003, CTM Wallach 2008, and MGCTM Xie and Xing 2013) as performance baselines. We downloaded the public version of LDA[3] and implemented in-house codes for CTM and MGCTM. For fair comparisons, we estimated all of the hyper-parameters of these approaches using the variational EM method, and estimated the GLDA using *Algorithm* 2. In terms of the online collection (Wiki), we used Online LDA[2] (Hoffman and Blei 2010) as the baseline and GLDA was estimated using *Algorithm* 3. All these models used the AS (Wallach et al. 2009a) form (asymmetric topic Dirichlet prior and symmetric word Dirichlet prior). The asymmetric topic Dirichlet priors, including local topic Dirichlet priors and global topic Dirichlet priors, were all initialized[4] as $50/K$ and estimated using the Newton–Raphson algorithm (Blei et al. 2003). The symmetric word Dirichlet prior β was fixed at 0.01.

Naturally, we can consider the topic model as a special probability density function for generating a corpus. So the topic modeling performance can be evaluated by the likelihood on the held-out test data (Wallach et al. 2009b). In our experiments, we trained all the baseline topic models and the GLDA using the training data, and then compared the perplexity scores of the held-out test data. The perplexity, used by convention in language modeling, is equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity represents a higher performance. Given corpora W and W_{test}, the perplexity is defined as:

$$perplexity(W_{test}) = \exp\left\{-\frac{\log p(W_{test}|W)}{\sum_{d=1}^{D} N_d}\right\} \tag{14}$$

4.2.1 Qualitative evaluation

We fit GLDA to two versions of the 20-NG datasets. One is the original 20-NG with stop words, and the other is a filtered 20-NG that has removed stop words[5] (378 in total). For both versions, we set $C = 20, Kl = 5$ and $Kg = 20$.

Table 2 illustrates the 10 most popular words for three global topics learnt by GLDA. The global topics learnt from the original 20-NG are almost filled by stop words. Because the stop words are ubiquitous to all documents, they can be explained as background semantics. In other words, GLDA successfully captured the common semantics. The results are clearer for the filtered 20-NG. We observed that Global Topic 1 is about article writing; Global Topic 2 is about time; Global Topic 3 is about both writing and time. These

[3] http://www.cs.princeton.edu/∼blei/topicmodeling.html.

[4] In topic modeling evaluation, K is the total number of topics, e.g., in CTM, K equals to $C * Kl$; in MGCTM and GLDA, K equals to $C * Kl + Kg$.

[5] http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop.

global topics obviously show the common semantics, and can be generated in all documents.

Table 3 shows the local topics from two estimated groups learnt by GLDA. Obviously, local topics cover the local semantics of each group. In Group 1, the local topics are clearly associated with computers. where Topics 1, 2 and 3 correspond to hardware, operating system and network, respectively. In Group 2, the local topics are about sports, including baseball, hockey and game. Although the results for the original 20-NG were affected by stop words (e.g., "can", "many" and "same"), they effectively captured the local semantics for each group.

Overall, we found that the two-stage generation of GLDA had a positive influence when capturing the semantics. On one hand, the local semantics were first organized at a coarse level (e.g., computer) and then further divided into a fine level (e.g., hardware and network). On the other hand, the common semantics were covered by the global topics. This framework effectively modeled the corpus, even with stop words.

4.2.2 *Quantitative evaluation of offline collections*

We tested the perplexity scores of the offline collections with different numbers of topics. We used the filtered 20-NG and WebKB datasets. The settings for the 20-NG were as follows: For MGCTM and GLDA, we fixed $C = 20$ and $Kg = 20$, and set $Kl = 1, 2, \ldots, 10$. For the CTM, we set the local topics from 2 to 11, for the same number of total topics. For WebKB, in both MGCTM and GLDA, we fixed $C = 4$ and $Kg = 32$, and set $Kl = 8, 12, \ldots, 32$. For CTM, we set $Kl = 16, 20, \ldots, 40$ for the same number of total topics.

The results for 20-NG are shown in Fig. 3. GLDA performed better than LDA and CTM. For LDA, there was a conflict. A larger K (more topics) is required to uncover the complex semantics in large document collections, but many documents naturally only involve some of these topics and the "forced topics" problem is more serious for larger K. Our experimental results confirmed this analysis. The LDA performed better when $K = 40$, and the performance deteriorated for larger K. CTM organizes topics into different groups. Its performance increases with the growth of K. Unfortunately, the CTM lacks the mechanism to distinguish local and global topics. So its peak performance is even worse than LDA for the 20-NG dataset.

Compared with MGCTM, GLDA performed better with respect to the perplexity metric. They both performed worse for $Kl = 1, 2$. Because: (1) a small number of local topics are not enough to adequately cover the local semantics; and (2) relatively few local topics exaggerate the influence of global topics (i.e., one man's loss is another's gain). GLDA outperformed MGCTM for $Kl > 2$, e.g., 3,190 in GLDA and 3,620 in MGCTM for $Kl = 6$, and 3,048 in GLDA and 3,323 in MGCTM for $Kl = 8$. We argue that this is because GLDA considers the relationships between local and global topics in terms of the different groups (see the discussions in Sect. 3.5). Our experimental results further validate this point.

As shown in Fig. 4, GLDA also performed better for WebKB. It performed better than the two simpler models (i.e., LDA and CTM) and slightly outperformed the state-of-the-art MGCTM. For the two simpler topic models, CTM outperformed LDA except when $K = 64$. This is because, for the WebKB dataset, we used a sufficient amount of local topics to capture the local semantics. In particular, we found that the gap between MGCTM and GLDA was smaller than the gap on the 20-NG dataset. This was mainly because GLDA considers the relationships between local and global topics in terms of different

Table 2 The 10 most popular words for several global topics in terms of 20-NG learnt by GLDA

Original 20-NG			Filtered 20-NG		
Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
many	now	talk	body	year	information
we	same	their	information	period	article
probably	sent	gun	section	issue	talk
alt	the	most	talk	time	news
serious	might	many	article	day	situation
should	more	never	reference	volume	state
harvard	talk	take	graphics	talk	groups
section	serious	state	work	future	computer
they	politics	about	collection	output	write
able	meanwhile	computer	manuscript	space	time

groups, in contrast to MGCTM. However, WebKB contains fewer groups ($C = 4$) than 20-NG ($C = 20$). So MGCTM approaches GLDA in this case.

We also investigated how to set the number of local and global topics in GLDA. We used fivefold cross validation, which produced convincing results. Figure 5 shows the averaged perplexity performance for different Kl and Kg using the WebKB dataset. The topic modeling performance was not significantly sensitive to the number of topics, and the variations were not very abrupt. Larger Kl and Kg resulted in a better performance than small values, e.g., the best performance was achieved when $Kl = 32$ and $Kg = 32$, and the worst was when $Kl = 8, 16$ and $Kg = 8$. More importantly, we found that the performance reduced when $Kl > Kg$. We argue that this trend is reasonable, because it intuitively requires more global topics to describe the background semantics that are ubiquitous to all the documents.

4.2.3 Quantitative evaluation on Wiki

Comparion with MGCTM Because of the similarities between MGCTM and (non-online) GLDA, we attempted to further compare the two models using a larger collection. To this end, we randomly selected 50,000 documents from the entire 3M Wiki collection (mini-Wiki) for model training, and evaluated MGCTM and GLDA on the test data that contained the 2,000 documents mentioned above.

Because the true number of groups in Wiki is unknown, we tested the perplexity scores using different numbers of groups. For both models, we fixed $Kl = 10$ and $Kg = 20$, and set $C = 2, 3, \ldots, 10$. The results are shown in Fig. 6. We can see that GLDA performed better than MGCTM in most cases. When there was a small number of groups (e.g., $C = 2$, 3, 4), the gap between the two models was relatively small. As C increased, GLDA rapidly diverges from the other model. As discussed in Sect. 3.5, the main difference between the two models is that GLDA considers the relationships between local and global topics in terms of the different groups, but MGCTM does not. In other words, we argue that GLDA is superior to MGCTM with a relatively large value of C. These empirical results support this view, as expected.

Table 3 The 10 most popular words for several local topics in terms of 20-NG learnt by GLDA

Original 20-NG			Filtered 20-NG		
Group 1			Group 1		
Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
comp	many	network	sys	comp	server
most	server	ftp	hardware	windows	client
hardware	windows	sun	graphics	file	ohio
graphics	file	many	windows	pc	internet
same	version	clients	comp	linux	comp
ibm	might	program	ibm	mac	network
apple	mac	happen	fs	unix	ftp
can	pc	internet	harvard	version	state
mac	follows	server	apple	flash	program
sys	computer	graphics	mac	cis	file
Group 2			Group 2		
Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
sport	off	race	club	sport	rec
rec	game	while	baseball	violence	player
serious	hockey	game	player	body	ohio
baseball	and	rec	game	hockey	money
run	uwm	ohio	serious	race	gov
game	violence	se	food	western	game
they	state	have	space	destroyer	club
club	sport	gov	sdd	health	apr
more	western	player	rec	body	disease
harvard	run	andrew	haven	ece	primate

Online learning we evaluated the performance of Online GLDA on the entire 3M Wiki collection. We set the mini-size $M = 100$ and 500. We fixed $K = 100$ for the Online LDA (Hoffman and Blei 2010), and $C = 8$, Kl $= 10$ and $Kg = 20$ for the online GLDA. The following learning rate is chosen, where the delay τ and forgetting rate κ are set as 1,024 and 1, respectively.

$$\rho_t = (t + \tau)^{-\kappa} \tag{15}$$

The results are shown in Fig. 7. Obviously, Online GLDA outperformed Online LDA. It improved by 150 when $M = 100$ and approximately 180 when $M = 500$. This is because a larger dataset must contain more topics. By organizing the topics into groups, we can cluster the relevant topics together. This experimental result shows that GLDA is useful for large-scale data.

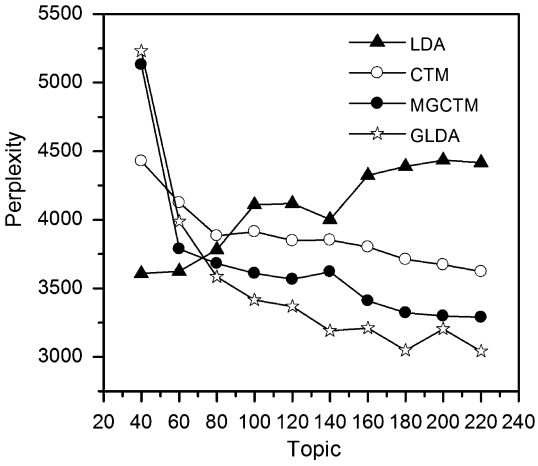

Fig. 3 The perplexity performance on 20-NG dataset

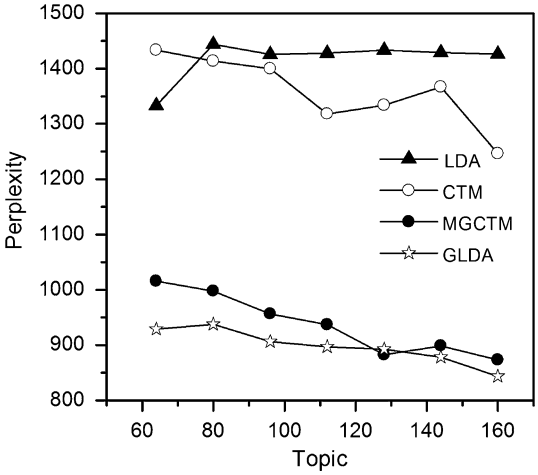

Fig. 4 The perplexity performance on WebKB dataset

4.3 Document clustering

GLDA assumes that documents belong to groups, so it can be naturally be used for clustering. We evaluated the document clustering performance of the proposed GLDA model using the filtered 20-NG and WebKB datasets. For both datasets, we removed the words that have occurred less than 10 times.

4.3.1 Metric

We evaluated the clustering performance by comparing the obtained cluster indices for documents using the clustering algorithm and the true labels. In our experiments, we used two common metrics (Cai et al. 2011; Zhang et al. 2011): clustering accuracy (AC) and normalized mutual information (NMI). For both metrics, a larger score represents a better performance.

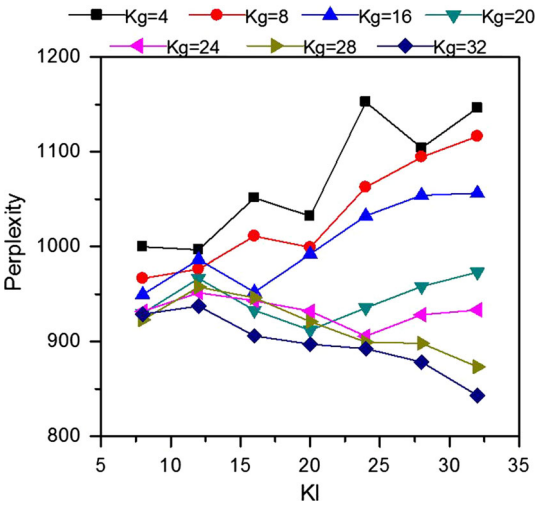

Fig. 5 The perplexity performance with different *Kl* and *Kg* on WebKB dataset

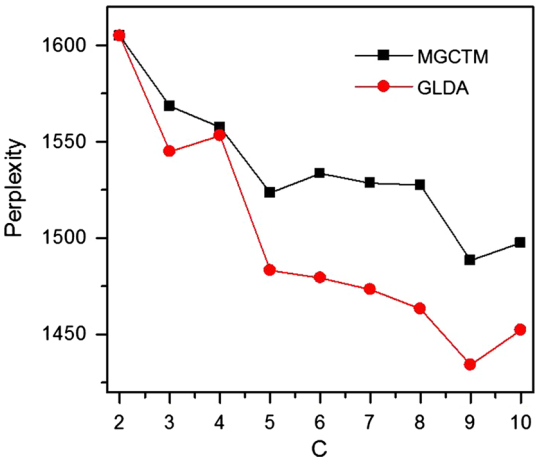

Fig. 6 The perplexity performance across mini-Wiki

The AC is used to evaluate the final clustering performance. Given a document d, let $\widetilde{y_d}$ and y_d denote the cluster index and the true label, respectively. Then the AC can be computed by:

$$AC = \frac{\sum_{d=1}^{D} \delta(y_d, map(\widetilde{y_d}))}{D} \tag{16}$$

where $\delta(x, y)$ is a delta function that is 1 if $x = y$ and 0 otherwise; $map(\cdot)$ is a function that maps each cluster to a label (as defined in the Kuhn–Munkres algorithm Lovasz and Plummer 1986).

NMI is originally used to measure the statistical information shared between two distributions. Let $\widetilde{Y}$ be the set of clusters obtained by the clustering algorithm and Y be the true set of labels. Their mutual information is defined as:

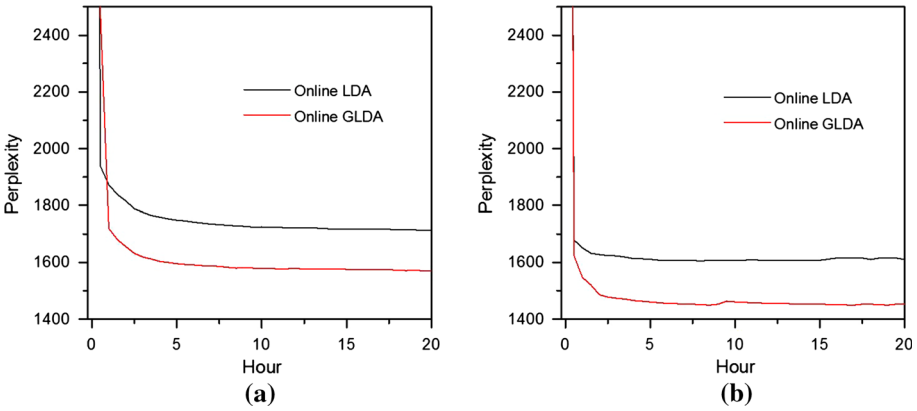

Fig. 7 The perplexity performance across Wiki: **a/b** is the result with $M = 100/500$

$$MI(Y, \widetilde{Y}) = \sum_{y_i \in Y, \widetilde{y_j} \in \widetilde{Y}} p(y_i, \widetilde{y_j}) \log \left(\frac{p(y_i, \widetilde{y_j})}{p(y_i) p(\widetilde{y_j})} \right).$$

where $p(y_i)$ and $p(\widetilde{y_i})$ denote the probabilities that a document belongs to the label y_i and cluster $\widetilde{y_j}$, respectively; $p(y_i, \widetilde{y_j})$ is the joint probability that a document belongs to the label y_i and cluster $\widetilde{y_j}$ at the same time. Here, we normalize $MI(Y, \widetilde{Y})$ using:

$$NMI(Y, \widetilde{Y}) = \frac{MI(Y, \widetilde{Y})}{\max(H(Y), H(\widetilde{Y}))} \tag{17}$$

where $H(Y)$ is the entropy of the true label set Y, and $H(\widetilde{Y})$ is the entropy of the estimated cluster set $\widetilde{Y}$.

4.3.2 *Performance*

We selected several baseline algorithms: non-negative matrix factorization (NMF), entropy weighting K-Means (EWKM) (Jing et al. 2007), LDA (Blei et al. 2003), CTM (Wallach 2008), and MGCTM (Xie and Xing 2013). For the LDA, we followed the experimental studies in (Lu et al. 2011): That is, (1) treat each topic as a cluster (assign a document to cluster x if $x = \arg\max_j \theta_j$); and (2) use symmetric Dirichlet priors α and β, setting $\alpha = 0.1$ and $\beta = 0.01$. For the CTM, we set the number of topics to 120 for the 20-NG dataset and to 40 for the WebKB dataset. In MGCTM and the proposed GLDA model, we used 10 local topics for each group and 20 global topics for the 20-NG dataset, and 32 local topics for each group and 32 global topics for the WebKB dataset. Following the settings in (Xie and Xing 2013), MGCTM initialized the variational document-group distributions with the clustering results of LDA and randomly initialized the other parameters. For GLDA, we initialized the parameters in the same was as MGCTM, and used another version that randomly initialized all the parameters (Ran-GLDA). For all the approaches, we averaged the results over 10 independent runs, and also calculated the pairwise t tests at 5 % significance levels for the GLDA and the baselines.

Table 4 shows the results for the 20-NG dataset. Obviously, the proposed GLDA model achieved the highest scores in both the AC and NMI metrics. GLDA performed much

Table 4 Performance (the average score ± standard deviation) on 20-NG

Algorithm	AC (%)	NMI (%)
NMF	36.42 ± 2.87• (0.0044)	35.42 ± 1.58• (0.0021)
EWKM	38.99 ± 1.23• (0.0062)	37.58 ± 2.17• (0.0043)
LDA	51.32 ± 0.83• (0.0395)	55.81 ± 0.92• (0.0274)
CTM	48.87 ± 1.48• (0.0282)	50.22 ± 3.46• (0.0162)
MGCTM	53.83 ± 0.52 (0.0836)	58.42 ± 1.27• (0.0409)
Ran-GLDA	54.21 ± 2.01• (0.0336)	58.78 ± 2.62• (0.0415)
GLDA	56.42 ± 0.36	61.19 ± 0.68

•/° GLDA is statistically superior/inferior to the compared algorithm. The p value are shown in brackets

better than the traditional approaches (i.e., NMF and EWKM), and performed competitively when compared to the topic modeling approaches. It outperformed LDA by approximately 5 % in AC and 6 % in NMI, and outperformed CTM by approximately 8 % in AC and 11 % in NMI. Ran-GLDA was approximately 0.5 % better in AC and 0.3 % in NMI than the state-of-the-art MGCTM. More importantly, GLDA outperformed MGCTM by approximately 3 %,in both AC and NMI.

Table 5 illustrates the results for the WebKB dataset. As for the 20-NG dataset, GLDA outperformed all the other approaches on the two metrics. For example, GLDA was approximately 4 % better than LDA in AC, and approximately 3.5 % better than CTM in NMI. GLDA outperformed the state-of-the-art MGCTM (approximately 0.3 % better in AC, and 0.6 % better in NMI). Ran-GLDA performed slightly worse than MGCTM (i.e., 0.5 % in AC and NMI), because there are no optimal initial parameters for Ran-GLDA.

Additionally, the p values obtained by the pairwise t tests are reported in Tables 4 and 5. We can clearly see that the proposed GLDA model was statistically superior to the compared algorithms in most cases [i.e., 20-NG (11/12) and WebKB (9/12)]. GLDA was clearly better than NMF, EWKM, LDA and CTM, and was slightly superior to MGCTM. We can also see that the standard deviations of the scores from GLDA were smaller. This further validates the robustness of GLDA.

4.3.3 Study on the number of topics

We investigated the effect of the number of topics on document clustering. Figure 8 illustrates the AC and NMI performance for the WebKB dataset, with different Kl and Kg. We observed that the results were very similar to the results in the topic modeling evaluation. A better performance is achieved when Kl and Kg are both large. In particular, the performance deteriorated when $Kl > Kg$ (e.g., the worst scores were obtained when $Kl = 24, 28, 32$ and $Kg = 20$). This is because, in document clustering, an increase in the number of global topics reduces the discrimination of the local topics for different groups. Therefore, in practice, we suggest the following settings in GLDA: (1) relatively larger Kl and Kg; and (2) $Kl \leq Kg$.

5 Conclusion

In this paper, we developed GLDA as an extension to the LDA model. The highlight of GLDA is that it organizes topics into groups to capture local semantics, and introduces global topics to cover the background semantics. In contrast to existing techniques, GLDA considers the

Table 5 Performance (the average score ± standard deviation) on WebKB

Algorithm	AC (%)	NMI (%)
NMF	46.71 ± 3.93• (0.0086)	33.28 ± 1.26• (0.0135)
EWKM	49.32 ± 1.35• (0.0088)	29.88 ± 1.88• (0.0023)
LDA	53.63 ± 0.54 (0.0793)	34.17 ± 1.26• (0.0374)
CTM	52.97 ± 1.62• (0.0212)	33.96 ± 0.93 (0.0621)
MGCTM	56.68 ± 0.85 (0.1364)	36.54 ± 2.57• (0.0478)
Ran-GLDA	55.11 ± 1.92• (0.0418)	36.07 ± 3.73• (0.0423)
GLDA	56.95 ± 0.36	37.11 ± 0.74

•/° GLDA is statistically superior/inferior to the compared algorithm. The p values are shown in brackets

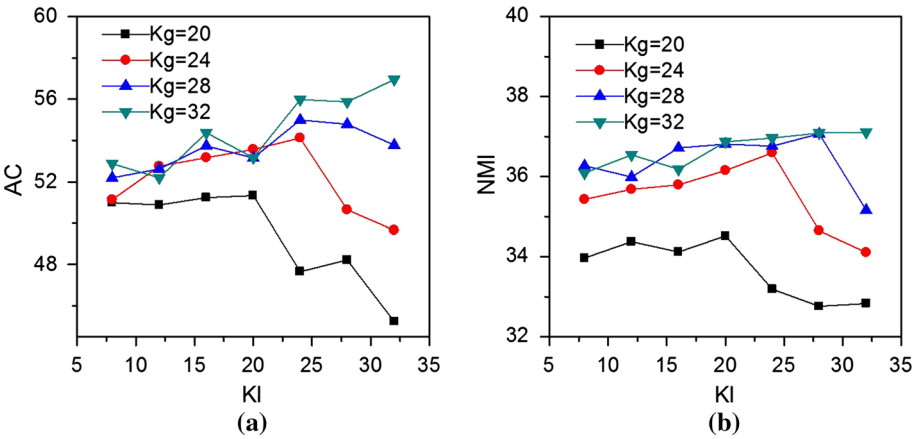

Fig. 8 The clustering performance with different *Kl* and *Kg* on WebKB dataset, i.e., **a** AC and **b** NMI

relationships between local and global topics in terms of the different groups. We developed a variational inference algorithm to model the offline corpora, and further extended an online learning algorithm for GLDA for a large-scale collection and true online data.

We used extensive experiments to evaluate the proposed GLDA model. We compared the topic modeling performance to traditional topic models for both offline and online cases. We also evaluated GLDA for document clustering. Our experimental results demonstrated that GLDA can achieve a state-of-the-art topic modeling performance, and also has a competitive clustering performance when compared with state-of-the-art clustering approaches.

In the future, we hope to develop extensions of GLDA using nonparametric methods, which can adaptively determine the number of groups and topics. It may also be useful to apply GLDA to basic tasks such as classification and sentiment analysis.

Acknowledgments This work was supported by National Nature Science Foundation of China (NSFC) under the Grant Nos. 61170092, 61133011, and 61103091.

Appendix

In this appendix, we derive the variational inference w.r.t GLDA. In terms of the variational distribution defined in Eq. (1), we can bound the log likelihood of the corpus using Jensen's inequality:

$$\begin{aligned}\log P\left(W|\pi,\alpha^{(l)},\alpha^{(g)},\beta\right) \geq E_q\left[\log P\left(W,\eta,z,\theta,\phi|\pi,\alpha^{(l)},\alpha^{(g)},\beta\right)\right] \\ - E_q\left[\log q(\eta,z,\theta,\phi|\tilde{\pi},\tilde{\theta},\tilde{\alpha},\tilde{\beta})\right]\end{aligned} \tag{18}$$

The objective is to maximize this lower bound with respect to $\tilde{\pi},\tilde{\theta},\tilde{\alpha}$ and $\tilde{\beta}$. Let $\mathscr{L}\left(\tilde{\pi},\tilde{\theta},\tilde{\alpha},\tilde{\beta}|\pi,\alpha^{(l)},\alpha^{(g)},\beta\right)$ be the right-hand of Eq. (18). We expand this lower bound as follows:

$$\begin{aligned}&\mathscr{L}\left(\tilde{\pi},\tilde{\theta},\tilde{\alpha},\tilde{\beta}|\pi,\alpha^{(l)},\alpha^{(g)},\beta\right)\\ &= E_q[\log p(\eta|\pi)] + E_q[\log p(\theta|\alpha,\eta)] + E_q[\log p(z|\theta)]\\ &\quad + E_q[\log p(\phi|\beta)] + E_q[\log p(w|\phi,z,\eta)]\\ &\quad - E_q[\log q(\eta|\tilde{\pi})] - E_q[\log q(\theta|\tilde{\alpha})]\\ &\quad - E_q\left[\log q\left(z|\tilde{\theta}\right)\right] - E_q\left[\log q\left(\phi|\tilde{\beta}\right)\right]\end{aligned} \tag{19}$$

We further expand Eq. (19) in terms of the model parameters $\left(\pi,\alpha^{(l)},\alpha^{(g)},\beta\right)$ and the free variational parameters $\left(\tilde{\pi},\tilde{\theta},\tilde{\alpha},\tilde{\beta}\right)$:

$$\begin{aligned}&\mathscr{L}\left(\tilde{\pi},\tilde{\theta},\tilde{\alpha},\tilde{\beta}|\pi,\alpha^{(l)},\alpha^{(g)},\beta\right)\\ &= \sum_{d=1}^{D}\sum_{c=1}^{C}\tilde{\pi}_{d,c}\log\pi_c\\ &\quad + \sum_{d=1}^{D}\sum_{c=1}^{C}\tilde{\pi}_{d,c}\left(\log\Gamma\left(\sum_{k=1}^{Kd}\alpha_k^{(c)}\right) - \sum_{k=1}^{Kd}\log\Gamma\left(\alpha_k^{(c)}\right) + \sum_{k=1}^{Kd}\left(\alpha_k^{(c)}-1\right)E_q\left[\log\theta_{d,k}\right]\right)\\ &\quad + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{Kd}\tilde{\theta}_{d,n,k}E_q\left[\log\theta_{d,k}\right]\\ &\quad + \sum_{k=1}^{KK}\left(\log\Gamma\left(\sum_{v=1}^{V}\beta_v\right) - \sum_{v=1}^{V}\log\Gamma(\beta_v) + \sum_{v=1}^{V}(\beta_v-1)E_q\left[\log\phi_{k,v}\right]\right)\\ &\quad + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{c=1}^{C}\sum_{k=1}^{Kd}\tilde{\pi}_{d,c}\tilde{\theta}_{d,n,k}E_q\left[\log\phi_{c\cdot k,w_{dn}}\right]\\ &\quad - \sum_{d=1}^{D}\sum_{c=1}^{C}\tilde{\pi}_{d,c}\log\tilde{\pi}_{d,c}\\ &\quad - \sum_{d=1}^{D}\left(\log\Gamma\left(\sum_{k=1}^{Kd}\tilde{\alpha}_{d,k}\right) - \sum_{k=1}^{Kd}\log\Gamma\left(\tilde{\alpha}_{d,k}\right) + \sum_{k=1}^{Kd}\left(\tilde{\alpha}_{d,k}-1\right)E_q\left[\log\theta_{d,k}\right]\right)\\ &\quad - \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{Kd}\tilde{\theta}_{d,n,k}\log\tilde{\theta}_{d,n,k}\\ &\quad - \sum_{k=1}^{KK}\left(\log\Gamma\left(\sum_{v=1}^{V}\tilde{\beta}_{k,v}\right) - \sum_{v=1}^{V}\log\Gamma\left(\tilde{\beta}_{k,v}\right) + \sum_{v=1}^{V}\left(\tilde{\beta}_{k,v}-1\right)E_q\left[\log\phi_{k,v}\right]\right)\end{aligned} \tag{20}$$

where $\phi_{c\cdot k}$ corresponds to the kth ϕ of group c and:

$$E_q\left[\log \theta_{d,k}\right] = \Psi\left(\tilde{\alpha}_{d,k}\right) - \Psi\left(\sum_{j=1}^{Kd} \tilde{\alpha}_{d,j}\right) \tag{21}$$

$$E_q\left[\log \phi_{k,v}\right] = \Psi\left(\tilde{\beta}_{k,v}\right) - \Psi\left(\sum_{j=1}^{V} \tilde{\beta}_{k,j}\right) \tag{22}$$

Now we derive the update rules w.r.t the four free variational parameters one by one.

1. **For** $\tilde{\pi}$: we know that $\sum_{c=1}^{C} \tilde{\pi}_{d,c} = 1$. Form the Lagrangian by isolating the terms that contain $\tilde{\pi}_{d,c}$ and adding the Lagrange multipliers λ, we obtain:

$$\begin{aligned}\mathscr{L}\left[\tilde{\pi}_{d,c}\right] = {} & \tilde{\pi}_{d,c} \log \pi_c \\ & + \tilde{\pi}_{d,c}\left(\log \Gamma\left(\sum_{k=1}^{Kd} \alpha_k^{(c)}\right) - \sum_{k=1}^{Kd} \log \Gamma\left(\alpha_k^{(c)}\right) + \sum_{k=1}^{Kd}\left(\alpha_k^{(c)} - 1\right) E_q\left[\log \theta_{d,k}\right]\right) \\ & + \sum_{n=1}^{N_d}\sum_{k=1}^{Kd} \tilde{\pi}_{d,c}\tilde{\theta}_{d,n,k} E_q\left[\log \phi_{c\cdot k,w_{dn}}\right] + \tilde{\pi}_{d,c} \log \tilde{\pi}_{d,c} \\ & + \lambda\left(\sum_{c=1}^{C} \tilde{\pi}_{d,c} - 1\right)\end{aligned} \tag{23}$$

Compute the derivative with respect to $\tilde{\pi}_{d,c}$ as follows:

$$\begin{aligned}\frac{\partial \mathscr{L}\left[\tilde{\pi}_{d,c}\right]}{\partial \tilde{\pi}_{d,c}} = {} & \log \pi_c \\ & + \left(\log \Gamma\left(\sum_{k=1}^{Kd} \alpha_k^{(c)}\right) - \sum_{k=1}^{Kd} \log \Gamma\left(\alpha_k^{(c)}\right) + \sum_{k=1}^{Kd}\left(\alpha_k^{(c)} - 1\right) E_q\left[\log \theta_{d,k}\right]\right) \\ & + \sum_{n=1}^{N_d}\sum_{k=1}^{Kd} \tilde{\theta}_{d,n,k} E_q\left[\log \phi_{c\cdot k,w_{dn}}\right] - \log \tilde{\pi}_{d,c} - 1 + \lambda\end{aligned} \tag{24}$$

Setting Eq. (24) to zero, so:

$$\begin{aligned}\tilde{\pi}_{d,c} \propto {} & \pi_c \\ & \times \exp\left(\begin{array}{c}\log \Gamma\left(\sum_{k=1}^{Kd} \alpha_k^{(c)}\right) - \sum_{k=1}^{Kd} \log \Gamma\left(\alpha_k^{(c)}\right) + \sum_{k=1}^{Kd}\left(\alpha_k^{(c)} - 1\right) E_q\left[\log \theta_{d,k}\right] \\ + \sum_{n=1}^{N_d}\sum_{k=1}^{Kd} \tilde{\theta}_{d,n,k} E_q\left[\log \phi_{c\cdot k,w_{dn}}\right]\end{array}\right)\end{aligned} \tag{25}$$

2. **For** $\tilde{\theta}$: we know that $\sum_{k=1}^{Kd} \tilde{\theta}_{d,n,k} = 1$. Again, form the Lagrangian by isolating the terms that contain $\tilde{\theta}_{d,n,k}$ and adding the Lagrange multipliers λ, we obtain:

$$\begin{aligned}\mathscr{L}\left[\tilde{\theta}_{d,n,k}\right] &= \tilde{\theta}_{d,n,k} E_q\left[\log \theta_{d,k}\right] \\ &+ \sum_{c=1}^{C} \tilde{\pi}_{d,c} \tilde{\theta}_{d,n,k} E_q\left[\log \phi_{c\cdot k, w_{dn}}\right] \\ &- \tilde{\theta}_{d,n,k} \log \tilde{\theta}_{d,n,k} + \lambda\left(\sum_{k=1}^{Kd} \tilde{\theta}_{d,n,k} - 1\right)\end{aligned} \tag{26}$$

Taking the derivative with respect to $\tilde{\theta}_{d,n,k}$, we obtain:

$$\begin{aligned}\frac{\partial \mathscr{L}\left[\tilde{\theta}_{d,n,k}\right]}{\partial \tilde{\theta}_{d,n,k}} &= E_q\left[\log \theta_{d,k}\right] \\ &+ \sum_{c=1}^{C} \tilde{\pi}_{d,c} E_q\left[\log \phi_{c\cdot k, w_{dn}}\right] - \log \tilde{\theta}_{d,n,k} - 1 + \lambda\end{aligned} \tag{27}$$

Setting Eq. (27) to zero, we can obtain:

$$\tilde{\theta}_{d,n,k} \propto \exp\left(E_q\left[\log \theta_{d,k}\right] + \sum_{c=1}^{C} \tilde{\pi}_{d,c} E_q\left[\log \phi_{c\cdot k, w_{dn}}\right]\right) \tag{28}$$

3. **For** $\tilde{\alpha}$: in Eq. (20), the terms that contain $\tilde{\alpha}_{d,k}$ are as follows:

$$\begin{aligned}\mathscr{L}\left[\tilde{\alpha}_{d,k}\right] &= \sum_{c=1}^{C} \tilde{\pi}_{d,c}\left(\sum_{k=1}^{Kd}\left(\alpha_k^{(c)} - 1\right) E_q\left[\log \theta_{d,k}\right]\right) \\ &+ \sum_{n=1}^{N_d} \tilde{\theta}_{d,n,k} E_q\left[\log \theta_{d,k}\right] \\ &- \left(\log \Gamma\left(\sum_{k=1}^{Kd} \tilde{\alpha}_{d,k}\right) - \sum_{k=1}^{Kd} \log \Gamma\left(\tilde{\alpha}_{d,k}\right) + \sum_{k=1}^{Kd}\left(\tilde{\alpha}_{d,k} - 1\right) E_q\left[\log \theta_{d,k}\right]\right)\end{aligned} \tag{29}$$

The corresponding derivative is:

$$\begin{aligned}\frac{\partial \mathscr{L}\left[\tilde{\alpha}_{d,k}\right]}{\partial \tilde{\alpha}_{d,k}} &= \Psi'\left(\tilde{\alpha}_{d,k}\right)\left(\sum_{c=1}^{C} \tilde{\pi}_{d,c} \alpha_k^{(c)} + \sum_{n=1}^{N_d} \tilde{\theta}_{d,n,k} - \tilde{\alpha}_{d,k}\right) \\ &- \Psi'\left(\sum_{j=1}^{Kd} \tilde{\alpha}_{d,j}\right) \sum_{k=1}^{Kd}\left(\sum_{c=1}^{C} \tilde{\pi}_{d,c} \alpha_k^{(c)} + \sum_{n=1}^{N_d} \tilde{\theta}_{d,n,k} - \tilde{\alpha}_{d,k}\right)\end{aligned} \tag{30}$$

where $\Psi'(\cdot)$ is the derivative of $\Psi(\cdot)$ function.

Setting Eq. (30) to zero, we can yield the maximum at:

$$\tilde{\alpha}_{d,k} = \sum_{c=1}^{C} \tilde{\pi}_{d,c} \alpha_k^{(c)} + \sum_{n=1}^{N_d} \tilde{\theta}_{d,n,k} \tag{31}$$

4. **For** $\tilde{\beta}$: in Eq. (20), the terms that contain $\tilde{\beta}_{k,v}$ are as follows:

$$\begin{aligned}\mathscr{L}\left[\tilde{\beta}_{k,v}\right] &= \sum_{v=1}^{V}(\beta_v - 1)E_q\left[\log \phi_{k,v}\right] \\ &+ \sum_{d}^{D}\sum_{c=1}^{C}\sum_{n=1}^{N_d}\tilde{\pi}_{d,c}\tilde{\theta}_{d,n,k}E_q\left[\log \phi_{c\cdot k,w_{dn}}\right] \\ &- \left(\log\Gamma\left(\sum_{v=1}^{V}\tilde{\beta}_{k,v}\right) - \sum_{v=1}^{V}\log\Gamma\left(\tilde{\beta}_{k,v}\right) + \sum_{v=1}^{V}\left(\tilde{\beta}_{k,v} - 1\right)E_q\left[\log \phi_{k,v}\right]\right)\end{aligned} \tag{32}$$

When k is a local topic that belongs to group c, its corresponding derivative is:

$$\begin{aligned}\frac{\partial\mathscr{L}\left[\tilde{\beta}_{c\cdot k,v}\right]}{\partial\tilde{\beta}_{c\cdot k,v}} &= \Psi'\left(\tilde{\beta}_{c\cdot k,v}\right)\left(\beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\tilde{\pi}_{d,c}\tilde{\theta}_{d,n,k}w_{dn}^{v} - \tilde{\beta}_{c\cdot k,v}\right) \\ &- \Psi'\left(\sum_{j=1}^{V}\tilde{\beta}_{c\cdot k,j}\right)\sum_{v=1}^{V}\left(\beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\tilde{\pi}_{d,c}\tilde{\theta}_{d,n,k}w_{dn}^{v} - \tilde{\beta}_{c\cdot k,v}\right)\end{aligned} \tag{33}$$

When k is a global topic, its corresponding derivative is:

$$\begin{aligned}\frac{\partial\mathscr{L}\left[\tilde{\beta}_{k,v}\right]}{\partial\tilde{\beta}_{k,v}} &= \Psi'\left(\tilde{\beta}_{k,v}\right)\left(\beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\tilde{\theta}_{d,n,k}w_{dn}^{v} - \tilde{\beta}_{k,v}\right) \\ &- \Psi'\left(\sum_{j=1}^{V}\tilde{\beta}_{k,j}\right)\sum_{v=1}^{V}\left(\beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\tilde{\theta}_{d,n,k}w_{dn}^{v} - \tilde{\beta}_{k,v}\right)\end{aligned} \tag{34}$$

Setting the Eqs. (33) and (34) to zero, we can yield the maximums at:

$$\begin{cases}\tilde{\beta}_{k,v} = \beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\tilde{\theta}_{d,n,k}w_{d,n}^{v} & \textit{if } k \textit{ is a global topic} \\ \tilde{\beta}_{c\cdot k,v} = \beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\tilde{\pi}_{d,c}\tilde{\theta}_{d,n,k}w_{d,n}^{v} & \textit{otherwise}\end{cases} \tag{35}$$

References

Blei, D., & Lafferty, J. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM.

Blei, D., & McAuliffe, J. (2007). Supervised topic models. In *Proceedings of the neural information processing systems*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D., & Lafferty, J. (2007). A correlated topic model fo science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D., Griffiths, T., & Jordan, M. (2010). The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*(2), 1–30.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Boyd-Graber, J., & Blei, D. (2008). Syntactic topic models. In *Proceedings of neural information processing systems*.

Cai, D., He, X., & Han, J. (2011). Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, *23*(6), 902–913.

Chang, J., & Blei, D. (2010). Hierarchical relational models for document networks. *Annals of Applied Statistics*, *4*(1), 124–150.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th international conference on machine learning* (pp. 281–288). ACM.

Hoffman, M., & Blei, D. (2010). Online learning for latent Dirichlet allocation. In *Advances in neural information processing systems*.

Hoffman, M., Blei, D., & Wang, C. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(1), 1303–1347.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57). ACM.

Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, *19*(8), 1026–1041.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge: MIT Press.

Li, W., & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on machine learning* (pp. 577–584). ACM.

Li, F., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Computer vision and pattern recognition* (Vol. 2, pp. 524–531). IEEE.

Lovasz, L., & Plummer, M. (1986). *Matching theory*. North Holland: Akademiai Kiado.

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*, *14*(2), 178–203.

Reisinger, J., Waters, A., Silverthorn, B., & Mooney, R. (2009). Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Proceedings of neural information processing systems* (pp. 1982–1989). (2009).

Reisinger, J., Waters, A., Silverthorn, B., & Mooney, R. (2010). Spherical topic models. In *Proceedings of the 27th international conference on machine learning*. ACM.

Sivic, J., Russell, B., Zisserman, A., Freeman, W., & Efros, A. (2008). Unsupervised discovery of visual object class hierarchies. In *Proceedings of the computer vision and pattern recognition* (pp. 1–8). IEEE.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd international conference on machine learning* (pp. 977–984). ACM.

Wallach, H. M. (2008). *Structured topic models for language*. Ph.D. thesis. Newnham College, University of Cambridge.

Wallach, H., Mimno, D., & McCallum, A. (2009a). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimn, D. (2009b). Evaluation methods for topic models. In *Proceedings of the 26th conference on uncertainty in artificial intelligence* (pp. 1105–111). ACM.

Wang, X., Mccallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE international conference on data mining*. IEEE.

Wang, C., Thiesson, B., Meek, C., & Blei, D. (2009). Markov topic models. In *Proceedings of the 12th international conference on artificial intelligence and statistics* (pp. 583–590). Journal of Machine Learning Research.

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. In *Proceedings of the 20th conference on uncertainty in artificial intelligence* (pp. 694–703).

Zhang, D., Wang, J., & Si, L. (2011). Document clustering with universum. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 873–882). ACM.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *20th international conference on machine learning*. ACM.