An Improved LDA Algorithm for Text Classification

Dexin Zhao[1] , Jinqun He[1], Jin Liu[2]
1.Tianjin Key Laboratory of Intelligent Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China
2.Tianjin Keyilong Decoration Engineering Co., Ltd., Tianjin 300202, China
qiqiharxin@163.com

***Abstract*—Latent Dirichlet Allocation is a classic topic model which can extract latent topic from large data corpus. This model assumes that if a document is relevant to a topic, then all tokens in the document are relevant to that topic. In this paper, we present an algorithm called gLDA for topic text classification by adding topic-category distribution parameter to LDA, which can make the document generated from the most relevant category. Gibbs sampling is employed to conduct approximate inference, and experiment results in two datasets show the effectiveness of this method.**

Keywords—topic model; LDA; text classification

I. INTRODUCTION

Latent Dirichlet Allocation (LDA) [1] is a classic topic model proposed by Blei in 2003, which offers a strong theoretical framework and tests to be effective in many text classification applications, this model has been used to discover the hidden thematic structure in large archives of documents [5,8]. The model assumes that document is composed by a set of topics, which can group words into "topics" using vector dimension reduction [6,11]. This also helps put words and documents map into a lower dimension space, so as to make a good performance in latent topic extract process[4,9]. As we know, words in the LDA model are assumed to occur independently, and documents in the LDA model are represented as "bag of words". The words that didn't relevant or less relevant also regarded as the document topic [10,12]. Therefore, how to get the topic from the documents more precisely, is a problem worthy of study. In this paper, based on LDA model, we propose an improved LDA topic model gLDA, documents in the model are generated into different categories, and each category has a special set of "topics". Through defining the category each document most relevant, each document is generated in the category that they most probably belong to. Through limiting the generation scope by topic-category distribution parameter, we can largely avoid wrong assignment of topic-words.

The paper is organized as follows. In Section 2, we introduce the basic notation and terminology of gLDA, and present its main idea. In Section 3, we discuss the inference and parameter estimation for gLDA. In section 4, the experiment results are presented. And specific discussion of the results is given in section 5.

978-1-4799-3197-2/14/$31.00 ©2014 IEEE

II. GLDA

To address the problem stated above, we add the category-topic distribution in gLDA model on the foundation of LDA model. The category-topic distribution indicates which topics are important for that category. This model divides documents in the corpus into different categories, especially, these documents within each category share the specific set of semantic "topic", and represented as a multinomial distribution over those *T* topics.

As the graphical model Fig.1 shows, the following conditional probability specifies the i*th* word token in a document:

$$p(\theta, z, w|\alpha, \beta, \gamma) = p(\theta|\alpha, \gamma)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (1)$$

θ is the document-topic distribution, inference from the joint estimate of hyperparameter α and category variable parameter γ . Z is the random variable sampled from $\theta^{(d,c)}$, which represents the topic indices. $p(w_n|z_n, \beta)$ is the probability of word w_n under topic z_n . $p(z_n|\theta)$ is the probability that topic z_n sampled from the document-topic distribution $\theta^{(d,c)}$ for the i*th* word token w_i .

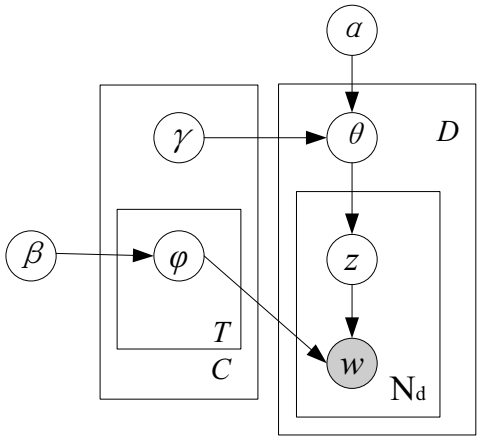

Fig. 1. The graphical model of gLDA

The word vocabulary set V is comprised by the set of W possible word token, $V = \{word_1, word_2, ..., word_N\}$. Each document d is comprised by a sequence of N_d tokens.

The generate process of gLDA can be formalized as follows:

1. For each topic $t \in \{1,...,T\}$, select a word distribution $\phi^{(t)} \sim$ Dirichlet(β)
2. For each document $d \in \{1,...,D\}$
   a. Select a category $c_i \sim$Dirichlet(γ)
   b. Select a distribution over topics $\theta^{(d,c_i)} \sim$ Dirichlet(α)
   c. For each word position i in document d
      (i) Select a topic $z_i \sim$ Dirichlet（ $\theta^{(d,c_i)}$ ）
      (ii) Generate a word token from topic z_i , $w_i \sim$ Dirichlet（ $\phi^{(z_i)}$ ）

The category variable parameter γ indicates the category that the document belongs to. In these particular categories, documents are generated with a certain set of topics. Topic t is represented by a multinomial distribution over the V word types in the corpus. $\theta^{(d,c)}$ is the document-topic distribution which determined by both hyperparameter α and category variable parameter γ : $\theta^{(d,c)} = [\theta_1^{(d,c)}, ..., \theta_T^{(d,c)}]$, $\theta^{(d,c)} = p(t|d,c)$ and $1 \le t \le T$. When the word token is generated for a document d, the model will compare the document-topic distribution $\theta^{(d,c)}$ and the category-topic distribution γ, and corresponding to the document-topic distribution $\theta^{(d,c)}$, words are generated from the topic-word distribution $\varphi^{(t)}$. Variable $\varphi^{(t)}$ indicates which words are important for one topic, both $\theta^{(d,c)}$ and $\varphi^{(t)}$ are used to smooth the word-topic and topic-document distributions respectively. $\phi^{(t)} = [\varphi_1^{(t)}, ..., \varphi_v^{(t)}]$, where $\varphi_w^{(t)} = p(w|t)$, $\phi^{(t)} = p(w_n|z_n, \beta)$ and $1 \le w \le V$.

III. PARAMETER ESTIMATION

For each word token i, documents are presented by a set of category indices ci, word indices wi and document indices di. The Gibbs sampling procedure [2,3,7] turns to consider each word token in the text collection. The estimated parameters α , is the assumption that the current topic has been assigned to other words, considering the text collection probability assigned to each word of other topics, so as to estimate the topic-word model parameter α.

The conditional distribution $p(z_i = j | z_{-i}, w_i, d_i, c_i, .)$, which can be calculated as:

$$p(z_i = j | z_{-i}, w_i, d_i, c_i, .) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{w_i j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i j}^{DT} + T\alpha} \frac{C_{c_i j}^{CT} + \gamma}{\sum_{l=1}^{L} C_{c_i j}^{CT} + C\gamma} \quad (2)$$

In the above conditional distribution, the left part is the probability of word w under topic j, the middle part is the probability that topic j under the current topic distribution for document d, whereas the right part is the probability of topic j under category group c. If category c has been used multiple times in one document, it will increase the probability that the topics relevant to the category assigned to category c.

C^{CT} is the matrix of document-category with dimension C x T, $C_{w_i j}^{WT}$ denotes the word-topic matrix and $C_{d_i j}^{DT}$ is the document-topic matrix with dimension D x T. From this conditional distribution, topic is sampled as the new topic assignment for each word token, and category is stored as the new category assignment for each topic. $z_i = j$ represents the probability of topic assignment of token i to topic j, z_{-i} refers to the topic assignments of all other word tokens, and "$\cdot$" refers to all other known or observed information such as all other word and document indices w_{-i} ,d_{-i} , c_{-i} , hyperparameters α, β, γ.

IV. EXPERIMENTS

A. The Experiment of Text Classification on Reuters-21578 dataset

To verify the validity of the model, we assess the performance of gLDA by evaluating its capability to classify the new documents that the model has not been trained on. In the models that trained from a certain genre, documents should be generalized from the same genre.

One formal way to assess generalization performance is through perplexity [1]. Perplexity is a quantitative measure for comparing topic models and is widely used to compare the predictive performance of topic models. Although perplexity does not directly measure aspects of a model, such as interpretability or coverage, it is nonetheless a useful general predictive metric for assessing the quality of a topic model [13,14].

Perplexity is equivalent to the inverse of the geometric mean of the likelihood of holdout data. The perplexity of a collection of test documents given the training set is defined as:

$$\text{Perp}(w_{test} | D_{train}) = \exp\left(-\frac{\sum_{d=1}^{Dtest} \log p(w_d | D_{train})}{\sum_{d=1}^{Dtest} N_d}\right) \quad (3)$$

Where w_{test} is the set of word tokens in the test documents, w_d is the set of word tokens in document d of the test set, D_{train} is the training set, and N_d is the number of words tokens in document d. The lower perplexity scores indicate that the model's predicted distribution is much closer to the true distribution.

The experiment in this section is based on the Reuters-21578 dataset. The dataset contains 8000 documents and

15,818 words. We trained the models on a random subset of 90% of documents, which are classified as *science*, *social studies*, *snack* and *tour*. By training the models, we obtain the estimating for the word-topic distributions φ, topic-document distributions θ, the assignments of word tokens to topics, and the hyperparameter γ on the topic-category distribution.

We evaluate generalization performance on the remaining documents in the *science*, *snack* and *tour* genres and also on a subset of documents classified as *social studies*. By testing on these classified documents, we first evaluate the models' ability to generalize within the same genre, and then evaluate the models' ability through the genres in different category. For each test document, we use a random 50% of the document's words to estimate document-topic and topic-category distributions. And using the estimated distributions measure perplexity on other remaining 50% of words.

Through varying the amount of training data for each model, we can observe the results of gLDA. A and C in Fig. 2 show the results when the model is trained and tested on *science*, *snack* and *tour* documents. B and D in Fig. 2 show the results when the model is trained on the documents of *science*, *snack* and *tour* genres, and tested on *social studies* documents. As the learned topics are entirely data driven, on the basis of just a few hundred documents, there is not enough statistical information to build accurate representations in LDA. Through providing a category prior to the learning algorithm, gLDA matches the category which is the document belong to. Through this matching algorithm, gLDA also appropriate for small training data. When there is a small training data (e.g., up to 500 documents), gLDA model's performance is better.

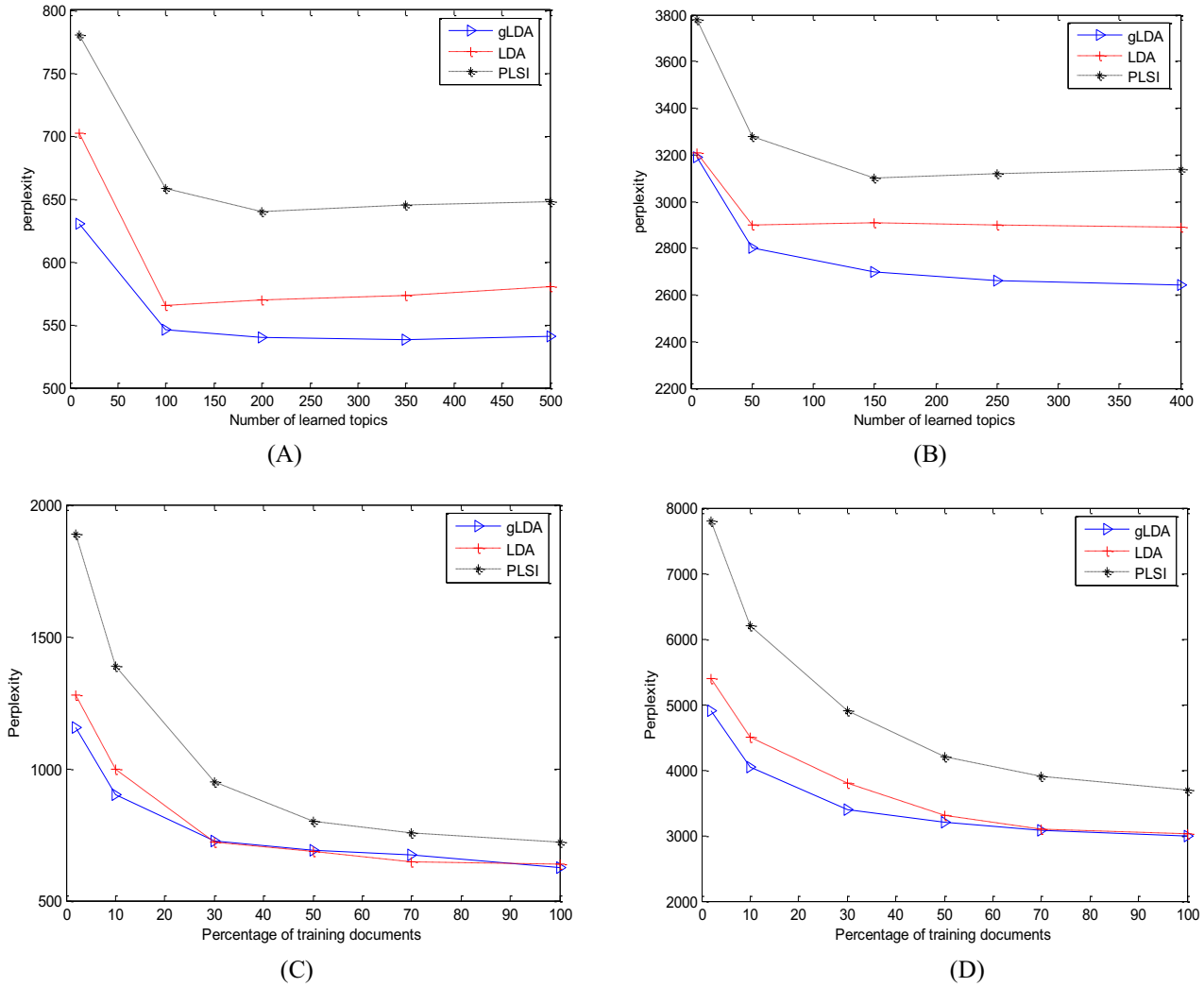

Fig. 2. Perplexity results on Reuters-21578 dataset for gLDA, the topic model LDA, and pLSI. As a function of number of topics (A–B) and percentage of training documents (C–D). Panels (A) and (C) show the results when the model is trained and tested on documents from the *science*, *snack* and *tour* genres. Panels (B) and (D) shows the results when the model is trained on documents from the *science*, *snack* and *tour* genres, but tested on documents from the social studies genre.

As the figure shown above, gLDA imposes hierarchical constraints to the topics. When small training data are available (See A and B in Fig. 2.) or documents are generated from a different genre (See B and D in Fig. 2.), gLDA outperforms LDA. Through the documents training process of gLDA, documents in the corpus are grouped into different categories. For example, training gLDA to learn the concepts about *apple* brand and its related topics, *apple* computer and *apple* mobile phone can be predicted with high probability. It is a good way to avoid generating the word *apple* from a document(such as the document about fruits) that has little relevant to the topic word.

B. *The Experiment of Text Classification on Fudan Chinese Text Classification Corpus*

Another formal way to assess performance is by accuracy. Given the training set, the accuracy of a collection testing documents can be defined as:

$$Accu(w_{test} \mid D_{train}) = \frac{1}{k}\sum_{i=1}^{k}\frac{C_i}{T_i} \qquad (4)$$

K is the genre number, C_i is the number of correct classified documents in the i*th* genre, T_i is the total number of documents which are tested in the i*th* genre. We have a further experiment on the Chinese text classification corpus of Fudan University, which consist of D = 20000 documents, passages are excerpted from educational texts. The documents are divided into twenty different educational genres. According to the size of the documents, they can be roughly divided into two levels: one is the eleven small educational genres, the document number of each genre less than 100, and another is the nine educational genres with large number of documents. The proportion of training set and testing set is divided with 1：1.

In this part, we compare the algorithm performance of gLDA with the traditional topic model. Using the parameters settings that obtained from the training process of Reuters-21578 dataset experiment, Fig.3 shows the experimental results. We can see that, when the topic number of each category is 10, compared with the traditional LDA, the category performance of gLDA have a 3.4% increase (increased from 85.7% to 89.1%). While, when having a comparison between the traditional LDA model and PLSI, the category performance between them is just about 1%. This experiment on this corpus also shows the satisfaction results on topic classification.

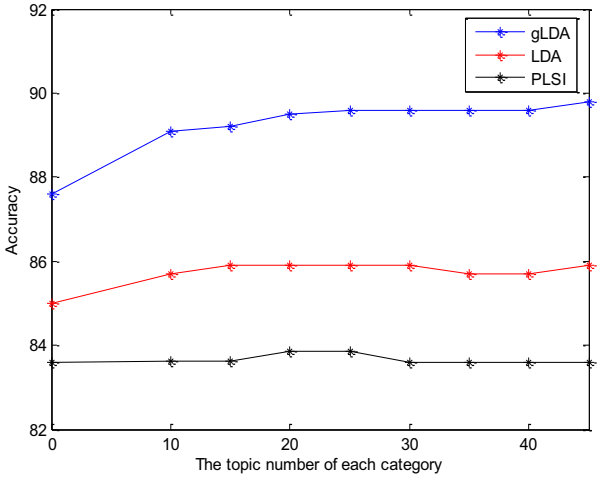

Fig. 3. The accuracy comparison

V. DISCUSSION AND CONCLUSION

We have embedded topic-category distribution parameter in the framework of LDA, a flexible generative probabilistic model for collections of discrete data. Based on a simple exchangeability assumption for the words and topics in a document, gLDA added categories for documents. The topics are built by aggregating documents for selected categories, and probability distributions are obtained by normalizing the word counts of the associated documents. Documents in this model can be grouped into different categories automatically. Instead of using the EM algorithm, Gibbs sampling was adopted for parameter eatimation in our model, which tested to have a lower perplexity consumption on parameter estimation.

For the purpose of word-sense disambiguation, by using a Dirichlet forest prior over topics in a category, gLDA has a more effective modeling process compared with [4]. Our framework is somewhat more general. By specify which words should have high probability in a topic, we improve the quality of making predictions on text data, model specifies the topics which should have high probability in a category. By adding the multinomial distribution over topics in a category, documents in a same field knowledge can be taken into account in the iteration of the topic modeling process, this is much different from an iterative topic modeling process which mainly depend on human inspection introduced in [8]. In addition, our topic model doesn't require labeled data. And there are several potentially useful directions, which the improved hierarchical topic model can be extended. The application of gLDA in information retrieval is a significance direction that deserve to address.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No.61202169&No.61170027).

REFERENCES

[1] D. Blei, M. Ng, A. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research. pp. 993-1022, March 2003.

[2] M. Homan, D. Blei, F. Bach, "On-line learning for latent Dirichlet allocation", In Neural Information Processing Systems, 2010.

[3] D. Blei, T. Griths, M. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies", Journal of the ACM. Vol.57, pp. 1-30, 2010.

[4] D. Boyd-Graber, D. Blei, X. Zhu, "A topic model for word sense disambiguation", Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Computational Natural Language Learning, New York: ACM, pp. 1024-1033, 2007.

[5] C. Kemp, J. B. Tenenbaum, "The discovery of structural form", Proceedings of the National Academy of Sciences. Vol.105, pp. 10687-10692, 2008.

[6] T. Hofmann, "Probabilistic latent semantic indexing", Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, New York: ACM Press, pp. 50-57, 1999.

[7] Y. W. Teh, M. Jordan, M. Beal, D. Blei, "Hierarchical Dirichlet processes", Journal of the American Statistical Association. Vol. 101, pp. 1566-1581, 2006.

[8] D. Andrzejewski, X. Zhu, M. Craven, "Incorporating domain knowledge into topic modeling via Dirichlet forest priors", The 26th International Conference on Machine Learning (ICML), New York: ACM, 2009.

[9] M. N. Jones, D. J. K. Mewhort, "Representing word meaning and order information in a composite holographic lexicon", Psychological Review. Vol.114 , pp. 1-37, 2007.

[10] T. L. Griffiths, M. Steyvers, J. B. T. Tenenbaum, "Topics in semantic representation", Psychological Review. Vol.114, pp. 211-244,2007.

[11] J. Boyd-Graber, D. Blei, "Syntactic topic models", In Neural Information Processing Systems, 2009.

[12] A. Asuncion, M. Welling, P. Smyth, Y. Teh, "On smoothing and inference for topic models", The 25th Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 2009.

[13] J. Reisinger, A. Waters, B. Silverthorn, R. Mooney, "Spherical topic models", the 27 th International Conference on Machine Learning (ICML-10), Haifa, Israel, 2010.

[14] J. Chang, D. Blei, "Hierarchical relational models for document networks", Annals of Applied Statistic. Vol. 4, 2010.