Topic-weak-correlated Latent Dirichlet Allocation[1]

Yimin TAN, Zhijian OU
Department of Electronic Engineering, Tsinghua University, Beijing
Corresponding email: ozj@tsinghua.edu.cn

Abstract—**Latent Dirichlet allocation (LDA) has been widely used for analyzing large text corpora. In this paper we propose the topic-weak-correlated LDA (TWC-LDA) for topic modeling, which constrains different topics to be weak-correlated. This is technically achieved by placing a special prior over the topic-word distributions. Reducing the overlapping between the topic-word distributions makes the learned topics more interpretable in the sense that each topic word-distribution can be clearly associated to a distinctive semantic meaning. Experimental results on both synthetic and real-world corpus show the superiority of the TWC-LDA over the basic LDA for semantically meaningful topic discovery and document classification.**

Keywords - topic modeling, weak-correlated topics

I. INTRODUCTION

Pioneered by latent Dirichlet allocation (LDA) [1], probabilistic topic modeling is becoming a popular tool for analyzing large unstructured discrete data such as text corpora. The basic idea is that the words of each document are assumed to be independently drawn from a mixture of multinomials. Each multinomial component is a word distribution over the vocabulary, which we call the topic-word distribution. The topic-word distributions, or topics, are shared by all documents. Each document has its own mixing proportion, which we call the document-topic proportion. Learning with topic models allows us to discover the latent topics from unstructured text corpora. Posterior inference for the document-topic proportion is useful for dimensionality reduction, classification, and information retrieval.

Since the introduction of the basic LDA, there are a lot of works developing new topic models and their applications. Among them, there are many works related to exploring new priors for the LDA, namely, the priors over the topic proportion and over the topic-word distribution respectively. In the basic LDA, the two priors are both assumed to be dirichlet. In contrast to the researches on exploring different priors over the topic proportion such as using the logistic normal prior [2][3] or the Dirichlet tree prior [4] to develop correlated topic models, there have been relatively few works on exploring new priors over the topic-word distributions, which is the main issue addressed in this paper.

First, it should be pointed out that the priors over the topic-word distributions is not merely for smoothing in estimating the topic-word probabilities, as introduced in the original paper [1]. They have practical effects. For example, using nested Chinese restaurant process as the priors can learn topic hierarchies from data [5]. Using Gaussian Markov random fields as the priors can capture the relationships between topics across multiple corpora [6]. Second, note that the *topic* term in the LDA is more a metaphor, with no epistemological claims [1]. The learned topics are usually named after we inspect the top words from the learned topic-word distributions. Topics are expected to be distinct in order to convey information [7]. We think, the distinction between different topics can be quantified by the weak-correlation between different topic-word distributions. So if we can reduce the overlapping between the topic-word distributions, it will make the learned topics more interpretable in the sense that each topic-word distribution can be clearly associated to a distinct semantic meaning.

The above two considerations motivate us to propose the topic-weak-correlated LDA (TWC-LDA) for topic modeling, which constrains different topics (i.e. topic-word distributions) to be weak-correlated. This is technically achieved by placing a special prior over the topic-word distributions, which exponentially decreases as the correlation between different topics increases. Variational inference procedure is derived for the new model. Experimental results on both synthetic and real-world corpus show that the TWC-LDA can successfully discover the weak-correlated topics which have clearer and more distinctive semantic meanings than topics learned by the basic LDA. A direct consequence of this is that TWC-LDA eliminates the need to manually remove stop-words. In the document classification task on Reuters-21578 dataset [8], the proposed TWC-LDA achieves higher classification accuracy than the basic LDA.

It is worthwhile to compare TWC-LDA with some related researches. First, note that while the previous correlated topic model [2] aims at capturing the correlation between the occurrences of latent topics, TWC-LDA focuses on incorporating the weak correlation between the topics themselves (i.e. between topic-word distributions). These two approaches complement to each other. Second, it is recently found in [7] that LDA using asymmetric dirichlet prior over document-topic distributions can be robustness to stop-words. Its main motivation is that some topics are assumed *a priori* to occur more frequently in each document; these more frequently used topics are thus forced to absorb stop-words after model learning. This modeling motivation is different from TWC-LDA, which directly places a weak-correlated prior over topic-word distributions, thus makes that the topic-word distributions are less-overlapped and each topic has distinctive semantic meaning. Although the seeming consequence of the LDA in [7]

[1] This work is supported by NSFC (61075020) and 863 program (2006AA01Z149).

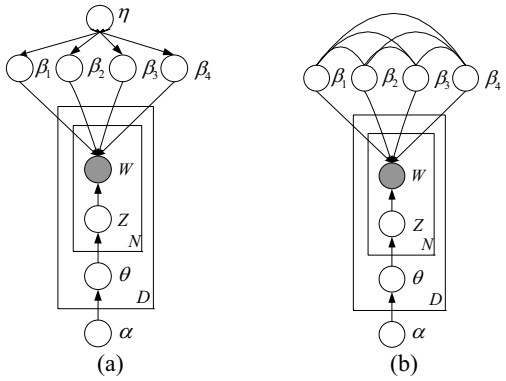

Figure 1. Graphical model representation for (a) basic LDA , and (b) TWC-LDA. Here we set the number of topics to be four for drawing convenience.

and TWC-LDA is similar - being robustness to stop-words, their modeling motivation are from different aspects. Moreover, the model in [7] employs computational-intensive Gibbs sampling, while TWC-LDA uses efficient variational inference.

The rest of paper is organized as follows. Section II describes the new TWC-LDA model and the variational inference. The experimental results on both synthetic and real-world corpus are shown in Section III. Finally, we present the conclusions in Section IV.

II. TOPIC-WEAK-CORRELATED LDA

A. The basic LDA

The basic LDA [1] shown in Fig. 1(a) assumes that in the corpus, each document $d = \{w_{dn} \mid n = 1, \cdots, N_d\}$ arises from a mixture distribution over latent topics. Each word w_{dn} is associated with a latent topic z_{dn} according to the document-specific topic proportion vector θ_d, whose prior is dirichlet with parameter α. The word w_{dn} is sampled from the topic-word distribution parameterized by a $K \times V$ matrix β, where each row, $\beta_n, 1 \le n \le K$, is independently drawn from an exchangeable dirichlet with parameter η. Here K and V denotes the number of topics and the vocabulary size respectively.

The generative process for the basic LDA is as follows.

1. for each document d, $\theta_d \sim Dir(\alpha)$;

2. for each of N_d word in document d

   Choose a topic $z_{dn} \sim Mult(\theta_d)$;

   Choose a word $w_{dn} \sim Mult(\beta_{z_{dn}})$.

B. TWC-LDA model formulation

Note that the topic term in the LDA is more a metaphor, with no epistemological claims. The learned topics are usually named after we inspect the top words from the learned topic-word distributions. So if we can reduce the overlapping between the topic-word distributions, it will make the learned topics more interpretable in the sense that each topic-word distribution can be clearly associated to a distinctive semantic meaning. The above two considerations motivate us to propose the topic-weak-correlated LDA (TWC-LDA) for topic modeling as shown in Fig.1 (b), which constrains different topics (i.e. topic-word distributions) to be weak- correlated.

This is technically achieved by placing a special prior over the topic-word distributions, which exponentially decreases as the correlation between different topics increases. This special prior is a non-conjugate prior over the parameters β, encoding our special prior knowledge.

$$p(\beta) = \frac{1}{Z} \exp\left\{ -\rho \sum_{m \ne n} \beta_m \beta_n^T \right\} \tag{1}$$

where $\rho > 0$ controls the strength of the prior, Z is the normalizing constant. The negative-log of the prior density is proportional to the sum of all the inner products for every pair of different rows in β matrix. The larger the correlation between different topics is, the smaller the prior is. In this way, the prior incorporates the interaction of different topics and forces them to have weak correlations. An approximate formula for ρ is given in section II-D.

C. Variational inference for TWC-LDA

Here for formula simplicity, we illustrate the posterior inference for a single document. The inference problem for the TWC-LDA is to compute the posterior $p(\theta, z, \beta \mid d)$, which is intractable in general. The basic idea of variational inference is to use a tractable distribution q to approximate the true posterior distribution p, and then to minimize the Kullback-Leibler divergence between the two distributions as measured by $KL(q \mid p) = \int q \log(q/p)$. Here we use the mean-field approximate distribution $q(\theta \mid \gamma_d) q(z_{1:N} \mid \phi_{d,1:N}) q(\beta)$, where $\gamma_d, \phi_{d,1:N}$ are the variational parameters for document d. The resulting variational update equations are as follows:

$$\gamma_{dk} \propto \alpha_k + \sum_{n=1}^{N_d} \phi_{dnk} \tag{2}$$

$$\phi_{dnk} \propto \exp\left\{ \psi(\gamma_{dk}) - \psi\left(\sum_{k=1}^{K} \gamma_{dk} \right) \right\} \cdot E_{\beta_k}\left[\log \beta_{k, w_{dn}} \right] \tag{3}$$

$$q(\beta_{k'}) \propto \prod_{j=1}^{V} \left(\beta_{k'j} \right)^{\lambda_{k'j}} \exp\left(-\rho B_{k'j} \beta_{k'j} \right) \tag{4a}$$

where $\lambda_{k'j} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{z_{dnk'}} w_{dn}^j$ (4b)

$$B_{k'} = \sum_{k=1, k \ne k'}^{K} E[\beta_k] \tag{4c}$$

and w_{dn}^{j} is the indicator function defined as $w_{dn}^{j} \triangleq 1(w_{dn} = j)$

From the equations above, we can see how the non-conjugate prior works. Considering topic k' and word j, if the occurring probabilities β_{kj} of word j in other topics (i.e. $k \neq k'$) are large, it will lead to a high value for $B_{k'j}$ in (4c), which subsequently encourages a low value of $\beta_{k'j}$ by (4a). Therefore, in TWC-LDA model, the word-probabilities for a given word in different topics suppress each other.

When $\rho = 0$, (4a) gives a dirichlet distribution, and we can easily compute (3) using the digamma function as in the basic LDA. Otherwise, if $\rho > 0$, computing the expectation in (3) using (4a) is intractable. For this reason, we further constrain

$$q(\beta_k) = \delta(\beta_k - \hat{\beta}_k) \quad (k = 1, 2...K) \tag{5}$$

and perform the maximum a-posterior (MAP) estimate for β_k's. As a result, the expectations in (3) and (4c) can be easily computed using the MAP point estimates. We use the line-search technique to calculate the mode of (4a) for MAP estimate.

For learning with the TWC-LDA over multiple documents, the variational updates of (2) and (3) are iterated until the convergence for each document, while (4) is iterated for the corpus scale. The empirical Bayes estimate for parameter α is the same as in the basic LDA model.

D. An approximate formula for ρ

As said in section II-B, the influence of weak-correlated prior is adjusted through the strength parameter ρ, when doing posterior inference for β. Considering the likelihood lower-bound with regard to β:

$$L(\beta) = \sum_{i=1}^{K}\sum_{j=1}^{V}\left(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j}\right)\log\beta_{ij} - \rho\sum_{i\neq j}\beta_i\beta_j^T - \sum_{i=1}^{K}\lambda_i\left(\sum_{j=1}^{V}\beta_{ij} - 1\right) \tag{6}$$

It can be seen that ρ weights the "weak correlated item" $-\sum_{i\neq j}\beta_i\beta_j^T$, which encourages weak-correlated topics. We define T as the total number of words in the corpus. We will see below that an approximate formula for ρ is related to the topic number K, the vocabulary size V and the word number T.

If we assume uniform distribution for every topic-word distribution in β, and the occurrences of every topic in the documents are uniform, then we have $\phi_{dni}w_{dn}^{j} = \frac{1}{KV}$. For the uniform β matrix, we have $\sum_{i=1}^{K}\sum_{j=1}^{V}\log\beta_{ij} = KV\log(1/V)$. Thus we obtain the first item in $L(\beta)$ as

TABLE 1: TOPICS LEARNED BY LDA (LEFT) AND TWC-LDA (RIGHT) RESPECTIVELY. EACH COLUMN IS THE TOP-TEN WORDS IN THE LEARNED TOPIC-WORD DISTRIBUTION.

LDA				TWC-LDA			
topic 1	**topic 2**	**topic 3**	**topic 4**	**topic 1**	**topic 2**	**topic 3**	**topic 4**
5	61	78	10	2	342	261	184
40	7	79	17	99	385	284	175
78	2	61	95	78	368	297	155
23	82	83	47	43	361	202	117
98	98	82	26	95	390	247	187
99	11	236	67	47	313	213	178
119	46	37	99	44	321	286	112
37	19	64	344	10	380	209	163
12	79	8	83	11	302	295	185
70	95	20	59	46	354	208	103

$$L_1 = \sum_{i=1}^{K}\sum_{j=1}^{V}\left(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j}\right)\log\beta_{ij} \approx T\log(1/V) \tag{7}$$

and the "weak correlated item" in $L(\beta)$ as

$$L_2 = -\sum_{i\neq j}\beta_i\beta_j^T = -K(K-1)*V*\frac{1}{V}*\frac{1}{V} \approx -K^2/V \tag{8}$$

Finally, we obtain an approximate formula for ρ, which says ρ is proportionally to a value defined by K, V and T as follows,

$$\rho_{approximate} \propto \frac{L_1}{L_2} = \frac{T\log(1/V)}{K^2/V} = \frac{TV}{K^2}\log(1/V) \tag{9}$$

In practice in order to obtain a reasonable value for ρ, we only need to tune the proportion factor first in a small-scale experiment, and then fixed in later large-scale experiments.

III. EXPERIMENTAL RESULTS

A. Synthetic dataset

Suppose that there is an imaginary vocabulary of 400 words, with the word-id being from 0 to 399. The 400 words are equally divided into 4 topics. Every word is hard assigned to one topic, and every topic has its own 100 words. Specifically, the topic-word assignments are that: word 0~99 for topic 1, word 100~199 for topic 2, word 200~299 for topic 3, word 300~399 for topic 4.

The word assignment probabilities over 100 words in each topic are randomly generated. The topic 1 that consists of word 0~99, is chosen to be the simulated topic of syntactic-words which occur more frequently. Therefore we set a relatively larger hyperparameter for topic 1 in the dirichlet prior ($\alpha_1 = 5$, $\alpha_2 = \alpha_3 = \alpha_4 = 0.5$). A total of 6000 documents (30 words per document) are generated.

Using the above synthetic dataset, we learn the four topics with LDA and TWC-LDA respectively. It can be seen from Table 1 that the TWC-LDA can successfully learn the four topics of the simulated model[2], but the basic LDA fails. The

[2] The order of the learned topics is not guaranteed. For example, the learned topic 2 by TWC-LDA actually corresponds to the real topic 4.

four topics learned by the basic LDA are almost occupied by the words from topic 1 (i.e. word 0~99), which represent simulated syntactic-words.

B. *Real-world text corpus*

For all the following experiments on real-world corpus, we set the number of topics K to be 30, hyperparameter α to be 0.5.

1) *Qualitative assessment of the learned topics*

We use the subset of the TREC AP corpus [5] containing 16333 articles with 23075 unique terms, which was the same as the corpus used in [1], and Year 1994 China daily newspaper (CDN) corpus.

We remove the stop-words in TREC AP corpus before running topic modeling. We use the raw CDN corpus, all the words are kept. The topics learned are shown in Table 2- 5.

Table 2 and 4 shows the topics learned by the basic LDA. Without deleting the stop-words, the topics learned by the basic LDA are mostly occupied by the syntactic words (see Table 2), and thus it is difficult to tell the semantic meaning of the topics. In experiments with deleting the stop-words, such problem is alleviated to some extent. The "topic 2" in Table 4 has the clear semantic meaning of "law", while "topic 3" and "topic 4" in Table 4 are still occupied by some semantic-vague words which are marked in red, such as "I, two, years, people, last".

Table 3 and 5 shows topics learned by TWC-LDA. Table 3 and 5 show that whether deleting the stop-words or not, the TWC-LDA can successfully learn the topics with clear semantic meanings. Incorporating weak-correlation among topics makes that each topic has its own distinctive semantic meaning. Moreover, the TWC-LDA can also discover the 'topics' with different syntactic functions. For example, "topic 3" in Table 3 includes preposition words and "topic 4" in Table 3 includes numeral words.

To make clear the semantic meanings of learned topics by deleting predefined stop-word list is subjective and non-adaptive. The stop-word list may be corpus-specific. Weak-correlated topics improve this problem by constraining the structure among topics. For example, preposition words have high probabilities in topic 3 of Table 3. The weak correlation between topic 3 and other topics prevents preposition words spreading into other topics, and thus helps other topics to have clearer semantic meanings.

2) *Quantitative Analysis of the learned topics*

We conduct quantitative analysis to see whether TWC-LDA learn more distinctive topics than LDA. We compare the correlation between topics extracted by LDA and TWC-LDA. We define the confusion matrix $C = \beta\beta^T$, whose off-diagonal elements represent the value of the cross-correlation between different topics, and $W = \sum_{m \neq n} \beta_m \beta_n^T$ which is the sum of all the off-diagonal elements in the confusion matrix C. The measurement of W gives an overall evaluation of the correlation between different topics. It is clear from Table 6 and Fig. 2 that the learned topics of TWC-LDA have significantly weaker correlation than that of LDA.

TABLE 2 TOPICS LEARNED BY BASIC LDA FROM CDN (RAW CORPUS). EACH CLOLUMN IS THE TOP-EIGHT WORDS IN THE LEARNED TOPIC-WORD DISTRIBUTION.

topic 1	topic 2	topic 3	topic 4
的 's	的's	的 's	的 's
是 is	人 people	体育 sports	是 is
了 -ed	到 to	了 -ed	在 in
产品 product	和 and	和 and	和 and
在 in	有 have	比赛 game	艺术 art
和 and	他 he	训练 train	观众 audience
企业 enterprise	来 come	有 have	音乐 music
市场 market	是 is	到 to	了 -ed

TABLE 3: TOPICS LEARNED BY TWC-LDA FROM CDN (RAW CORPUS) . EACH CLOLUMN IS THE TOP-EIGHT WORDS IN THE LEARNED TOPIC-WORD DISTRIBUTION.

topic 1	topic 2	topic 3	topic 4
犯罪 crime	文化 culture	的 's	十 ten
机关 office	出版 publish	在 in	二 two
案件 case	历史 culture	和 and	三 three
治安 safe	读者 reader	上 on	八 eight
公安 police	时代 epoch	中 mid	百 hundred
打击 attack	传统 tradition	有 have	九 nine
法院 court	读者 reader	对 to	七 seven
法律 law	书 book	为 for	千 thousand

TABLE 4: TOPICS LEARNED BY BASIC LDA FROM AP-CORPUS (STOP-WORDS REMOVED). EACH CLOLUMN IS THE TOP-EIGHT WORDS IN THE LEARNED TOPIC-WORD DISTRIBUTION.

topic 1	topic 2	topic 3	topic 4
i	court	soviet	government
years	case	gorbachev	president
new	attorney	new	people
first	trial	i	national
two	judge	air	new
like	charge	people	communist
just	prison	two	congress
people	sentence	africa	years
last	federal	flight	last

TABLE 5: TOPICS LEARNED BY TWC-LDA FROM AP-CORPUS (STOP-WORD REMOVED). EACH CLOLUMN IS THE TOP-EIGHT WORDS IN THE LEARNED TOPIC-WORD DISTRIBUTION.

topic 1	topic 2	topic 3	topic 4
i	court	soviet	bill
new	case	united	senate
years	drug	government	committee
people	judge	military	budget
two	attorney	states	congress
state	trial	president	tax
last	charges	war	rep
time	prison	foreign	sen
first	investigation	official	house

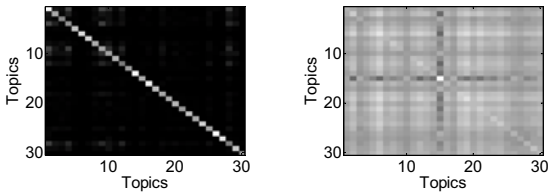

Figure 2. Confusion matrix for TWC-LDA (left), and LDA (right) (Lower value was showed in darker color)

TABLE 6: COMPARISON OF CORRELATIONS BETWEEN TOPICS

Dataset	W of LDA	W of TWC-LDA
TREC-AP	0.0416	0.0078
China Daily Newspaper	3.2922	0.0113

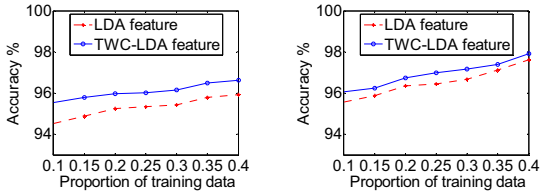

Figure 3. Classification Accuracy of Topic EARN vs. NOT EARN (left), GRAIN vs. NOT GRAIN (right)

3) Document Classification

The document-specific posterior topic proportions can be used as a reduced-dimensional representation of the document, which serves as the feature for document classification.

We conduct binary text classification using Reuters-21578 dataset. After removing the space character, non-English characters, and a list of 512 stop words, we obtain a dataset with 13476 documents and 32531 words.

In the experiment, we divide the whole dataset into training set and test set. Utilizing the SVMLight package [9], we train two support vector machines with the feature provided by LDA and TWC-LDA respectively. The classification experiments focused on two main categories in Reuters-21578 dataset - "EARN" and "GRAIN".

Experimental results in Fig.3 show that the document classification accuracies of the basic LDA we implemented are comparable to the results reported in the origin paper of LDA [1]. For different training data proportions, the proposed TWC-LDA model consistently achieves higher classification accuracies than the basic LDA model.

IV. CONCLUSIONS

In this paper, we propose the topic-weak-correlated LDA (TWC-LDA) for topic modeling, which constrains different topics (i.e. topic word-distributions) to be weak-correlated. Such weak correlation forces each topic to have clear and distinctive semantic meaning. Without manually deleting stop-words, the TWC-LDA can discover topics with clear semantic meanings. In the task of document classification, the proposed TWC-LDA model achieves higher classification accuracies than the basic LDA model.

Topic modeling has been used in computer vision to learn natural scene categories [10]. However, it becomes harder for researchers to define appropriate stop-patches list for images' topic modeling. Although this paper focuses on text analysis, the new TWC-LDA model can also be applied in other applications. It is worthwhile further studying the application of weak-correlated topic modeling for computer vision.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003

[2] D. M. Blei and J. D. Lafferty (2007). A correlated topic model of Science. Annals of Applied Statistics, 1(1):17–35.

[3] D. Mimno, H. Wallach, and A. McCallum. (2008) Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. NIPS Workshop on Analyzing Graphs, 2008

[4] Tam, Y.-C., & Schultz, T. (2007). Correlated latent semantic model for unsupervised LM adaptation. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing

[5] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. (2003). Hierarchical topic models and the nested Chinese restaurant process. Advances in Neural Information Processing Systems, 16, MIT Press

[6] C.Wang, B. Thiesson, C. Meek, and D. Blei. Markov topic models. In The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), pages 583–590, 2009.

[7] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems,2009

[8] http://www.daviddlewis.com/resources/testcollections

[9] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999

[10] Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. IEEE Computer Vision and Pattern Recognition, pages 524–531