Topic Discovery based on LDA_col Model and Topic Significance Re-ranking

Lidong Wang[1,2]
1. College of Computer Science and Technology, Zhejiang University, Hangzhou, China
2. Hangzhou Normal University, Hangzhou,China
Email: violet_wld@163.com

Baogang Wei
College of Computer Science and Technology, Zhejiang University, Hangzhou, China
Email: wbg@zju.edu.cn

Jie Yuan
College of Computer Science and Technology, Zhejiang University, Hangzhou, China
Email: jave_mc@163.com

***Abstract*—This paper presents a method to find the topics efficiently by the combination of topic discovery and topic re-ranking. Most topic models rely on the bag-of-words(BOW) assumption. Our approach allows an extension of LDA model—Latent Dirichlet Allocation_Collocation (LDA_col) to work in corpus such that the word order can be taken into consideration for phrase discovery, and slightly modify the modal for modal consistency and effectiveness. However, LDA_col results may not be ideal for user's understanding. In order to improve the topic modeling results, two topic significance re-ranking methods (Topic Coverage(TC) and Topic Similarity(TS)) are proposed. We conduct our method on both English and Chinese corpus, the experimental results show that the modified LDA_col discovers more meaningful phrases and more understandable topics than LDA and LDA_col. Meanwhile, topic re-ranking method based on TC performs better than TS, and has the ability of re-ranking the "significant" topics higher than "insignificant" ones.**

***Index Terms*—Topic model, LDA, Latent Dirichlet Allocation_Collocation, Topic significance re-ranking, Topic Coverage, Topic Similarity**

I. INTRODUCTION

Topic mining (topic analysis) is a critical part of representing the content of a text document, which is also important in many areas of natural language processing(e.g. machine translation, text mining, information retrieval). Recently, using topic models for document representation has also been an area of considerable interest in machine learning. However, most of the topic models, such as Latent Dirichlet Allocation(LDA)[1], consider the "topic" as individual words other than phrases(or collocations). According to people's perception, phrases contain more information than the sum of its individual word [21], in which word order is ignored, and it is a better way to represent a text document.

Assume that topic analysis is conducted on a large collection of research paper for text categorization, a topic model ranks very high word like "machine", "learning". However, these words do not have certain meanings. The document has the topic "machine" can be categorized in "industry", the document has the topic "learning" may be categorized as "education". Actually, the document has the topic word "machine learning" should mostly be categorized as "computer".

Besides, some topic models allow a document to belong to multiple topics. However, current approaches to topic discovery perform manual examination of their output to find meaningful and important topic, and many models haven't considered different users' demands. For example, the topics extracted by LDA model are randomly ordered, and users must navigate the topics to find the important topic they need. It is very inconvenient if there are a large number of topics. Therefore, the target of our paper is to find out the solution for phrases topic analysis and topic significance re-ranking.

In this paper, a topic discovery method is presented to find the topics efficiently according to the combination of topic discovery and topic re-ranking. Up until now, no intensive research has been developed in topic model for topic analysis, especially in Chinese corpus. Using NIPS proceeding[1] datasets and Chinese corpus[2], our approach allows an extension of LDA model—Latent Dirichlet Allocation_Collocation (LDA_col) to work in corpus such that the word order can be taken into consideration in phrase discovery. We also slightly modify the model to enforce consistency. The model parameters are inferred by Gibbs Sampling. Moreover, in order to enhancing the LDA_col topic modeling results, two methods are proposed to automatically select the most salient topics to meet user's interests.

This work is supported by National Science Foundation of China (No.60673088)

Corresponding author: Lidong. Wang, email: violet_wld@163.com

[1] http://www.cs.toronto.edu/~roweis/data.html
[2] http://www.nlp.org.cn/docs/doclist.php

doi:10.4304/jcp.6.8.1639-1647

The rest of our paper is organized as follows: Section 2 describes the related works. In Section 3, LDA_col model is introduced, and our method of topic extraction from LDA_col is discussed. We describe the algorithm that utilizes topic significance re-ranking to improve topic discovery in Section 4 before introducing our data set and evaluating our algorithm's performance in Section 5.

II. Related Works

Probabilistic topic modeling has been successfully applied to explore and predict the underlying structure of text document. Latent Dirichlet Allocation(LDA) has quickly become one of the most popular text modeling techniques and has inspired a lot of research papers[5-8], and also some papers have applied the LDA model in text categorization[2] and Information retrieval[3]. LDA overcomes the drawbacks of previous topic models such as PLSA[4]. Up until now, many variations of LDA have been proposed for different purpose. Nallapati[5] combines PLSA and LDA into a single framework that can be used to provide a user with highly influential blog postings on the topic of user's interest. In order to use both of function and content words, Griffiths[9] presents a composite model called HMMLDA, which can identify the role of words play in documents. But this method should not omit the stopwords, and the result is not effective enough. M.Blei[6] develops the correlated topic model to overcome the limitation of LDA, which has the inability of modeling topic correlation. Chang[8] develops the relational topic model (RTM), a hierarchical model of both network structure and node attributes, which can be used to summarize a network of documents, predict links between them, and predict words within them. Besides, there also have some topic models applied in different areas [14,15,16].

Collocation has long been studied by lexicographers and linguists in various ways. However, there is no intensive research at the area of phrase discovery. Wallach develops a bigram topic model[17] on the basis of the hierarchical Dirichlet language model[8]. BTM can automatically infer a separate topic for function words. But for the target of important topic discovery, it is not suitable for Chinese corpus because function words are meaningless and should be removed from the corpus. Later, Topical N-gram Model(TNG) is presented by Wang[3]for topic discovery with an application to information retrieval. The TNG model is built on the combination of BTM and LDA_col. The key contribution of the model is to decide whether to form a n-gram no matter what the context is. Although the model has better result in ad-hoc retrieval than BTM and LDA_col, the topic discovery procedure is more complicated than LDA_col, and the result in Chinese corpus is still unknown. In this paper, we use LDA_col model for topic discovery in both of Chinese corpus and English corpus.

III. Latent Dirichlet Allocation_Collocation Model

A. Model Description

First we describe some symbols used in this section:

a) T number of topics

b) D number of documents

c) W number of unique words

d) z_i the topic assigned to i^{th} word token

e) x_i the bigram status between $(i-1)^{th}$ word token and i^{th} word token

f) w_i the i^{th} word token

g) α, β, δ Dirichlet prior of θ, ϕ, ϑ

h) γ_0, γ_1 Beta prior of π

i) θ^d the multinomial distribution of topics in document d

j) ϕ^z the multinomial unigram distribution of words with respect to topic z

k) $\vartheta^{(w)}$ the multinomial bigram distribution of words w

l) $\pi^{(w)}$ the binomial distribution of status variable with respect to word w

Starting from the LDA topic model, Griffiths presents the extensive model LDA_col without any in-depth research[11]. The LDA collocation model incorporates collocations and introduces an additional variable x_i to indicate whether a word is part of a collocation. If $x_i = 1$, then w_i is generated from the distribution $p(w_i \mid w_{i-1})$. If $x_i = 0$, then w_i is generated from the distribution associated with topic $p(w_i \mid z_i)$. The graphical model corresponding to this generative model is shown in Fig.1.

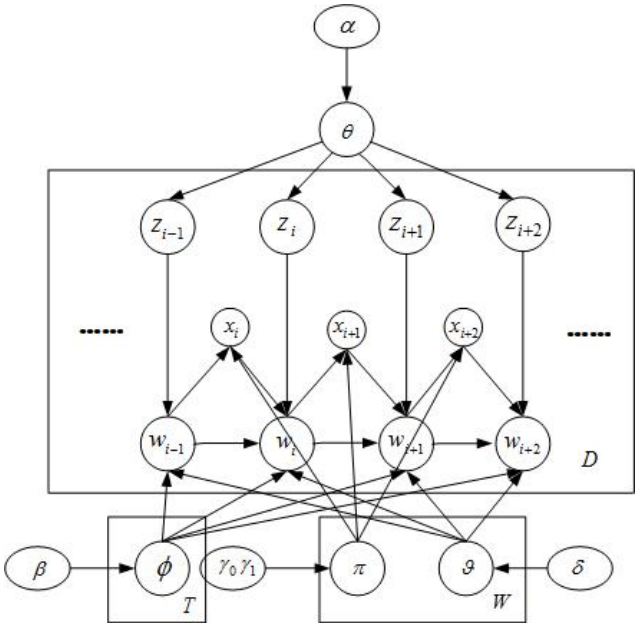


Figure 1. LDA_collocation Model. Circles represent variables, ellipses represent model parameters, and plates represent replications.

The generative process of the model can be described as:

1) For each topic z, choose $\phi^z \sim Dirichlet(\beta)$
2) For each word w,choose $\pi^{(w)} \sim Beta(\gamma_0, \gamma_1)$;
3) For each word w, choose $\vartheta^{(w)} \sim Dirichlet(\delta)$;
4) Choose $\theta^{(d)} \sim Dirichlet(\alpha)$; For each word w_i in document d:

a)Choose $x_i \sim binomial(\pi^{(w_i-1)})$;

b)Choose $z_i \sim multinomial(\theta^{(d)})$;

c)If $x_i = 1$, choose $w_i \sim multinomial(\vartheta^{(w_{i-1})})$; else choose $w_i \sim multinomial(\phi^{z_i})$

Thus, the model has the power to decide whether to generate a unigram or bigram, even trigram. It is more useful than BTM model which always generates a bigram. However, there is a problem if a bigram phrase or trigram phrase is formed by a word with its previous words, the word does not have the option to be assigned the same topic as previous words. Therefore, there is no close connection between bigrams or trigrams with its previous words. To solve this problem, we could slightly modify the model in plate D as follows:

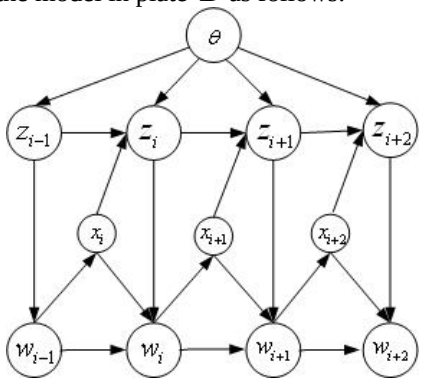


Figure 2. Modified LDA_collocation model

Accordingly, the generative process in step 4(b) can be modified as: If $x_i = 1$ (collocation), let $z_i = z_{i-1}$; else choose $z_i \sim discrete(\theta^{(d)})$. At this case, if a phrase is composed of w_{i-1} and w_i, the topic assignment for word w_i is directly assigned the same topic as w_{i-1}. In our experiment, the modified LDA_col modal is implemented to conduct topic discovery.

B. Gibbs sampling for modified LDA_collocation

Gibbs sampling is a special case of Markov-chain Monte Carlo(MCMC) simulation and often yields relatively simple algorithms for approximate inference in high-dimensional models[10]. Therefore we choose this approach although there exists many other inference algorithms.

The strategy of integrating out $\theta, \phi, \vartheta, \pi$ is often used in Gibbs Sampling. The target of inference is the distribution $p(z_i \mid w_i)$ and $p(x_i \mid w_{i-1})$. During Gibbs sampling, we draw the topic assignment z_i and the bigram status x_i iteratively for each word token according to Eq.(1)~(4).

If $x_i = 0$, z_i is drawn from the distribution

$$\begin{aligned} & p(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}, \mathbf{x}) \\ & \propto p(w_i \mid z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \mathbf{x}) \times p(z_i = j \mid \mathbf{z}_{-i}, \mathbf{x}) \\ & = \int p(w_i \mid z_i = j, \mathbf{x}, \phi^{(j)}) p(\phi^{(j)} \mid \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)} \\ & \quad \times \int p(z_i = j \mid \theta^{(d_i)}, \mathbf{x}) p(\theta^{(d_i)} \mid \mathbf{z}_{-i}) d\theta^{(d_i)} \quad (1) \\ & = \frac{n_{w_i}^{z_i=j} + \beta}{n_{\cdot}^{z_i=j} + W\beta} \frac{n_{z_i=j}^{d_i} + \alpha}{n_{\cdot}^{d_i} + T\alpha} \end{aligned}$$

If $x_i = 1$, z_i is drawn from the distribution according to the modification in Fig.2:

$$p(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}, \mathbf{x}) = p(z_{i-1} = j \mid \mathbf{z}_{-(i-1)}, \mathbf{w}, \mathbf{x}) \quad (2)$$

where $p(z_i = j)$ represents the probability of assigning word w_i to each topic j. $n_{w_i}^{z_i=j}$ is the number of words assigned to topic j that are the same as w_i, $n_{\cdot}^{z_i=j}$ is the total number of words assigned to topic j, $n_{z_i=j}^{d_i}$ is the number of words from document d_i assigned to topic j, and $n_{\cdot}^{d_i}$ is the total number of words in document d_i. $\mathbf{z}_{-i}$ denotes the topic assignments for all word tokens except word w_i. All counts are exclude the current case and only refer to the words for which $x_i = 0$.

For x_i, it is drawn from the distribution: if $x_i = 0$,

$$\begin{aligned} p(x_i \mid \mathbf{x}_{-i}, \mathbf{w}, \mathbf{z}) & \propto p(w_i \mid x_i, \mathbf{x}_{-i}, z_i) p(x_i \mid \mathbf{x}_{-i}) \\ & = \frac{n_{w_i}^{z_i=j} + \beta}{n_{\cdot}^{z_i=j} + W\beta} \frac{n_{x_i=0}^{w_{i-1}} + \gamma_0}{n_{\cdot}^{w_{i-1}} + \gamma_0 + \gamma_1} \quad (3) \end{aligned}$$

if $x_i = 1$,

$$\begin{aligned} p(x_i \mid \mathbf{x}_{-i}, \mathbf{w}, \mathbf{z}) & \propto p(w_i \mid x_i, \mathbf{x}_{-i}, w_{i-1}) p(x_i \mid \mathbf{x}_{-i}) \\ & = \frac{n_{w_i}^{w_{i-1}} + \delta}{n_{\cdot}^{w_{i-1}} + W\delta} \frac{n_{x_i=1}^{w_{i-1}} + \gamma_0}{n_{\cdot}^{w_{i-1}} + \gamma_0 + \gamma_1} \quad (4) \end{aligned}$$

all counts above exclude the current assignment, where $n_{x=0}^{w_{i-1}}$ is the number of times the word w_{i-1} has been drawn from a topic, $n_{x=1}^{w_{i-1}}$ is the number of times the

word has formed a collocation, $n_{w_i}^{w_{i-1}}$ is the number of times the word w_{i-1} follows w_i, $\mathbf{x}_{-i}$ represents the bigram status for all word tokens except word w_i.

Furthermore, the estimates of $\theta, \phi, \vartheta, \pi$ can be given as follows:

$$\phi_{w_i}^{z=j} = \frac{n_{w_i}^{z_i=j} + \beta}{n_{\cdot}^{z_i=j} + W\beta} \qquad \theta_{z=j}^{d_i} = \frac{n_{z_i=j}^{d_i} + \alpha}{n_{\cdot}^{d_i} + T\alpha}$$

$$\vartheta^{w_{i-1}} = \frac{n_{w_i}^{w_{i-1}} + \delta}{n_{\cdot}^{w_{i-1}} + W\delta} \qquad \pi^{w_{i-1}} = \frac{n_{x_i}^{w_{i-1}} + \gamma_0}{n_{\cdot}^{w_{i-1}} + \gamma_0 + \gamma_1} \tag{5}$$

IV. Topic Re-ranking

The result of the LDA_col model can be represented as two outputs: the term(unigram or phrase) distribution $p(w_i \mid z_i = j)$ for each topic j and topic distribution $p(z_i = j \mid d_m)$ for each document d_m. However, topics discovered by the model are randomly ordered if they are generated over a whole corpus. Therefore, it is necessary to help users to quickly find the topics that they are interested in. Based on the results of modified LDA_col, we can further improve the performance by re-ranking the topics, which assigns the most significant topic the highest ranking. In this paper, the topic discovery is completed through the combination of modified LDA_col model and topic re-ranking to enhance the modeling results.

Our methods on topic re-ranking are motivated by the observation that the more talked topics in a corpus tend to be labeled as more important topics. At this aspect, the rank of a topic would be higher if it covers more documents. Two re-ranking methods are presented in our paper. In Section 5, the comparison between these two methods is provided.

A. Topic Coverage

Based on the assumption that the topics covering significant portion of the corpus content are more important than those covering little texts, we can simply applying the topic coverage to determine the rank of each topic. If the topics have significant content coverage, then they are ranked higher.

Thus, we define:

$$\begin{aligned}\mu(p(z_i = j \mid d_m)) &= \sum_{m=1}^{M} N_m \cdot \theta_{z=j}^{(d_m)} / \sum_{m=1}^{M} N_m \\ &= \sum_{m=1}^{M} N_m \cdot \frac{n_{z_i=j}^{d_m} + \alpha}{n_{\cdot}^{d_m} + T\alpha} / \sum_{m=1}^{M} N_m\end{aligned} \tag{6}$$

N_m is the length of document m. Therefore, if the value computed is higher, such topic is ranked higher.

B. Topic Similarity

Our second re-ranking algorithm is proposed to select "significant" topics by topic similarity calculation. If the distribution of topic i is completely different with all other topics, it must not be considered as most talked topics. Topic coverage(TC) makes use of the document-topic distribution, here we employ topic-word distribution to compute topic similarity. Thus, our algorithm is designed as follows:

1) For each topic pair (i, j), probability distribution $p(w_i \mid z_j)$ is employed to represent topic distribution vector. Then the KL divergence between word-topic distributions i and j is calculated as follows:

$$KL(z_i \,||\, z_j) = KL(\phi^{z_i} \,||\, \phi^{z_j}) = \sum_{w=1}^{W} \phi_w^{z_i} \log_2 \frac{\phi_w^{z_i}}{\phi_w^{z_j}} \tag{7}$$

2) The KL divergence is not a proper distance measurement because it is not symmetric. Thus the symmetrised extension, the Jensen-Shannon distance is calculated:

$$Dist(z_i, z_j) = \frac{1}{2}[KL(z_i \,||\, z_j) + KL(z_j \,||\, z_i)] \tag{8}$$

3) For each topic i, calculate the average distance between i and all the other topics.

$$average\_dist(i) = \sum_{j \neq i,\, j=1}^{T} Dist(z_i, z_j) \Big/ T - 1 \tag{9}$$

Here, T is the topic number.

4)Descending sort the $average\_dist$ for each topic in a queue. The last element in the queue should be ranked the highest. In contrary, the first element in the queue is ranked the lowest.

V. Experiments

A. Data Sets

Most of topic models are conducted on English collection. In our paper, several experiments are conducted on both Chinese corpus and English corpus, which are listed as follows:

1) NIPS proceedings dataset, which consists of the full text of 13-years of proceedings from 1987 to 1999 NIPS conference. In addition to removing stopwords, we also remove the words appear less than five times. The dataset contains 1740 papers, 13649 unique words and 2301375 word tokens.
2) Chinese Corpus for categorization, which consists of 10 categories that are Environment, Computer, Transportation, Education, Economy, Military, Sports, Medicine, Art and Politics. Topic models are conducted on each category. The information for each category dataset is shown as follows:

TABLE I. DETAIL INFORMATION ABOUT CHINESE CORPUS

number / category	text	word token	unique word
环境（Environment）	200	24598	4006
计算机（Computer）	200	45535	3540
交通（Transportation）	214	20125	3259
教育（Education）	220	45628	4897
经济（Economy）	325	63882	5390
军事（Military）	249	28098	3929
体育（Sports）	450	32195	5034
医药（Medicine）	204	13233	3059
艺术（Art）	248	55893	7807
政治（Politics）	505	31959	3545

All of the data in table 1 are under the manipulation of word segmentation and part-of-speech tagging by ICTCLAS2009 tool[3] at first. Then we keep the nouns and remove other words, because the topic words are mostly determined by the nouns in Chinese expression. Moreover, the individual words are also removed, such as “县”, “市”, “区”, which are meaningless in expression.

B. Parameters

There are several parameters that need to be determined in our experiment. For the LDA_col and LDA estimation, the number of topics has to be tuned due to its influence on the model performance, and the number of iterations. For the number of iterations, it is one of the drawbacks with MCMC approaches. The target is to check if the Markov chain has converged. However, no realistic method can be applied on large corpus to determine the convergence of the chain. In consideration of a good trade-off between accuracy and running time, we set the iteration number to 100 in our final experiment. Other parameters are set as follows that are common settings in the literature:

$$\alpha = 50/T, \quad \beta = 0.01, \quad \delta = 0.1,$$

$$\gamma_0, \ \gamma_1 = 0.1, \max c = 4$$

where T is the topic number, maxc represents the maxium collocation length. In the following, we focus on the selection for topic number.

Different datasets should have different topic number. Paper[12] has proved that the topic model reaches optimum as the average similarity among topics reaches minimum. Based on this, we design a method to obtain the best topic number to make model reach optimum. Firstly, KL divergence is employed to calculate the topic similarity according to Eq.7 and Eq.8 under the condition of different topic number K varies from 10 to 150. We randomly select 200 papers from NIPS proceeding datasets and the “environment” collection from Chinese corpus for experiment, and the topic number selected can be used in other collections. Then, calculate the average similarity according to Eq.10.

$$avg\_dist = \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} dist(z_i, z_j)}{K \times (K-1)/2} \qquad (10)$$

If the value computed by Eq.10 reaches the highest, the average similarity among topics reaches minimum. At this case, we can choose the number. The experimental results are shown in Fig.3 and Fig.4.

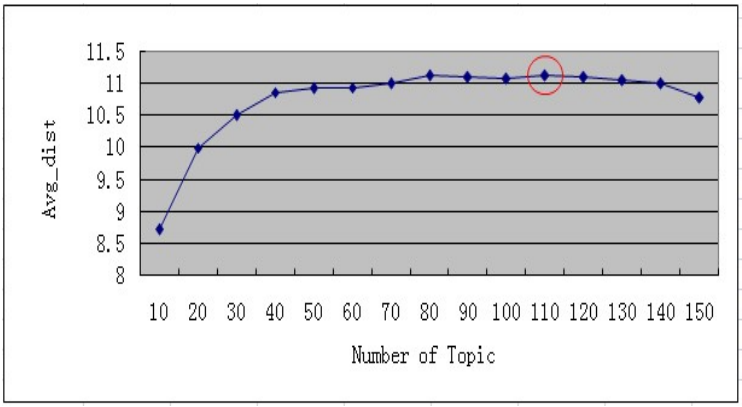


Figure 3. Experimental result of topic number selection on NIPS proceeding datasets

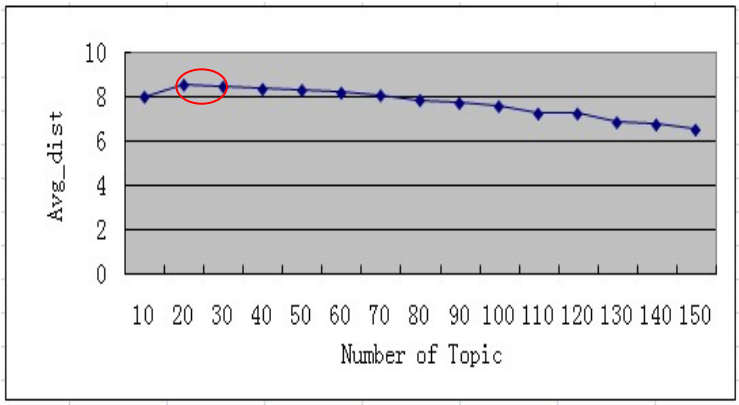


Figure 4. Experimental result of topic number selection on “Environment” collection

As shown in Fig.3 and Fig.4, we find that an appropriate number of topic is different for different datasets. In NIPS dataset, when the appropriate number of topic is selected in 110, the value computed by Avg_dist reaches maximum, which means the average similarity among topics reaches minimum and the model reaches optimum. In Chinese corpus, the model can get best performance when the number of topic is set to 20. Therefore, the best topic numbers are set to 110 and 20 for English corpus and Chinese corpus, respectively.

C. Experimental Results

With comparison to the corresponding topics found by LDA and LDA_col. Our experiments apply the modified LDA_col model(Fig.2) to the NIPS proceedings dataset and Chinese categorization corpus, respectively.

[3] http://ictclas.org/

Topic discovery

As shown in table 2, we run the model with the topic number 110 according to the analysis in Section 5.2. In "support vector machine" topic, it can be seen that the LDA_col model and modified LDA_col model provide extremely salient phrases("support_vector_machines","principal_component_analysis"), which can make users easily understand the topic. In "principal component analysis" topic, we can find the similar phenomena as well. Phrases ("principal component analysis", "covariance_matrix ") are more meaningful than individual words, such as "principal", "analysis". Moreover, some generic words(such as "data", "space") rank high in LDA topic list. On the contrary, they do not appear in modified LDA_col topic list.

TABLE II. THE TWO TOPICS DISCOVERED BY LDA , LDA_COL AND MODIFIED LDA_COL ON NIPS PROCEEDINGS DATASET. THE TOPIC NAME ABOVE THE WORD LISTS IS OUR OWN SUMMARY OF THE TOPIC. EACH TOPIC CONTAINS TOP 20 KEYWORDS.

Support Vector Machine			Principal Component Analysis		
LDA	LDA_col	Modified LDA_col	LDA	LDA_col	Modified LDA_col
kernel	machine	support	data	principal_components	principal
support	set	machine	space	projection	principal_component_analysis
vector	support	support_vector	clustering	pca	principal_components
margin	problem	svm	cluster	covariance_matrix	pca
kernels	svm	support_vector_machines	pca	subspace	principal_component
svm	support_vector_machines	machines	dimensional	eigenvalues	non-linear
training	support_vector	empirical	clusters	principal_component_analysis	covariance
set	algorithm	decision	projection	directions	covariance_matrix
machines	loss	algorithm	principal	reconstruction	projection
data	problems	vector	unsupervised	dimension	directions
adaboost	support_vectors	support_vectors	analysis	eigenvectors	eigenvalues
vapnik	machines	vapnik	dimensionality	orthogonal	dimension
sv	errors	risk	components	constraints	eigenvectors
cost	number	class	structure	constraint	orthogonal
convex	class	errors	algorithm	non-linear	components
test	utility	number	dimensions	projection_pursuit	component
working	algorithms	utility	component	constrained	eigenvalue
machine	choice	algorithms	dimension	find	decomposition
algorithm	comparison	choice	measure	found	analysis
smola	vapnik	comparison	reduction	principal_component	found

We can also find the effectiveness of modified LDA_col model. Words in ellipse represent related previous words that do not appear in the result of LDA_col. Take the "support vector machine" topic for example, the topic words "support", "vector","machines"and"support_vector_machines" are appeared in the same topic, which is the result of our modifying. The strong co-occurrence of these related previous words make the topics more understandable for users. The similar result can be found in "principal component analysis". However, as the result of LDA_col in topic "SVM", related previous word "vector" does not appear. That is because the word in LDA_col model does not have the option to be assigned the same topic as previous words. Meanwhile, some extremely related words(suchas"support_vector_machines","principal_component_analysis") rank higher in modified LDA_col model than in LDA_col.

There should be some standard measures to evaluate the quality of topic mining results. Some research [3,13,19]have applied the information retrieval results to evaluate the quality of topic discovery. However, what we really concern is whether the topic words extracted are extremely what we need. The end goal of topic discovery is to use topics to improve some end-user task, such as text categorization, text retrieval in search interface or digital library. So we define the "precision", which is calculated as the number of relevant topic words discovered by topic model divided by the total number of topic words. We manually classify the topic words into "relevant" and "irrelevant". The standard for "relevant" means the word can represent the category information clearly , and is salient for the topic extracted. However, the generic words should be considered as "irrelevant". For example, a topic about " 病 毒 "(virus) from "Computer" collection discovered by modified LDA_col contains keywords ("病毒"(virus), "计算机"(computer), "电子_邮件"(email), "文件"(file), "版本"(edition), "人们"(people), "网络"(network), "软件"(software), "邮件"(mail), "梅利莎"(melissa)), the "relevant" words are "病毒" (virus), "计算机" (computer), "电子_邮件" (email), "文件" (file), "网络" (network), "软件" (software), "邮件" (mail), "梅利莎" (melissa). Other two words are classified as "irrelevant" because we wouldn't expect them to be useful when searching or browsing, and also they may appear in other categories and cann't represent the topic "病毒"(virus).

According to the topic number selection mentioned above, we employ LDA, LDA_col and modified LDA_col to conduct topic discovery. Each topic contains top 10 keywords. The experimental result is shown in Fig.5. In comparison with the result of LDA and LDA_col, the modified LDA_col has a performance improvement in most of collections. Especially for

"Sports" and "Military" collection, it achieves an increment of 0.1 and 0.125 on the "precision" compared with LDA. In general, LDA_col and modified LDA_col perform better than LDA because of phrase discovery. Modified LDA_col model performs best although precision values of some collections are close to LDA_col. This comparison indicates that the modified LDA_col model can efficiently improve the ability of topic discovery.

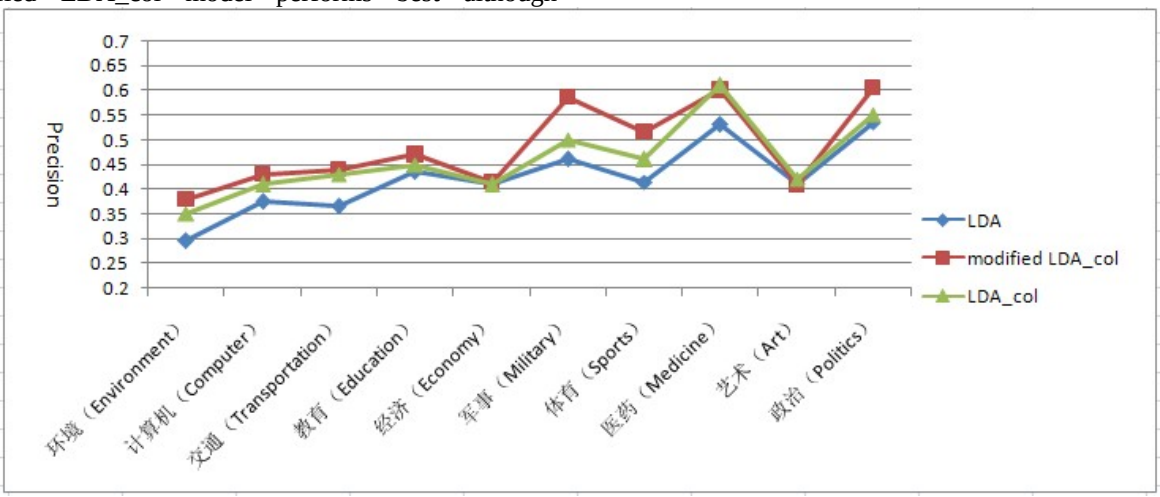


Figure 5. Comparison of LDA , LDA_col and modified LDA_col on Chinese Corpus

Topic Re-ranking

As the previous word for our project CADAL(China Academic Digital Associative Library, which has 100,0000 collections of Chinese books), we test our topic re-ranking techniques using 10 categories of Chinese corpus under the topic discovery result of modified LDA_col. Averge-r measure(Eq.11) is employed to evaluate the performance. From the experiment in Section 5.3.1, top 10 keywords are listed in a topic. For each topic, we define: if there are more than 6 keywords are considered as "relevant", then the topic is categorized as "significant"; otherwise, it is categorized as "insignificant". Here, Average-r is used for evaluation:

$$Average-r=\frac{1}{m}\sum_{r=1}^{m}\rho_r \qquad (11)$$

Where m is a given constant, we set m =5 in our experiment, ρ_r is the rank of No.r topic in the first m "significant" topics. Therefore the ideal value of Average-r is 3 .

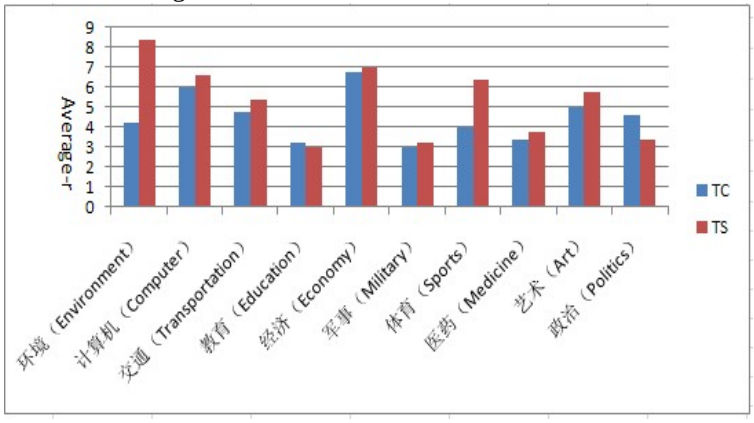


Figure 6. Topic significance re-ranking results on Chinese corpus

The topic re-ranking results for the Chinese corpus are shown in Fig. 6. The topic coverage(TC) method outperforms topic similarity(TS) method on most of collections except "Politics" and "Education" collection. Many collections have achieved an ideal value close to 3 under the method TC. Overall, TC method seems to capture important topics more easily, and the topics needed are ranked higher, which can enhance the topic modeling result. The detail of the TC based topic re-ranking results on "medicine" collection is shown in table 4.

Our experiment lists 5 "medicine" topics that gain 3 highest ranking and 2 lowest ranking in table 3 by TC method. This experimental result shows that topic re-ranking with TC method can further improve the topic discovery performance because, as is shown in table 3, all the top three 3 ranked topics are categorized as "significant". Some closely related topic words on "medicine" collection is ranked high, such as topic 20 , topic11 and topic 4. For example, each keywords in topic 20 have direct relation with the category "medicine". Furthermore, the number of "related" topic words reaches 10 in topic 20, which can represent the topic "伤口_感染" (wound infection) clearly. However, topic 17 and topic 13 receive the lowest significance ranking. These two topics is showed as "insignificant" because their topic words don't have close connection with the "medicine" collection, such as "children", "people", "worker", "measure" and so on.

In conclusion, the main goal of topic re-ranking method based on TS and TC is to select "significant" topics and rank them higher. Experimental results show that TC method could actually outperform TS method. Therefore, it is a better way to choose TC method for our furture research on text mining.

TABLE III. 5 TOPICS RE-RANKED BY TC(TOPIC SIMILARITY) METHOD

topic20(rank2)	Topic11(rank2)	topic4(rank3)	topic17(rank19)	topic13(rank 20)
伤口(wound) 绷带(splenium) 创伤(wound) 细菌(bacterium) 石膏(gypsum) 红药水 (mercurochrome) 肢体(limbs) 伤口_感染 (wound infection) 成人(adult) 破伤风_抗毒素 (antitoxinum tetanicum)	病人(patient) 手术(operation) 王成标_教授 (professor C.B.Wang) 腹部(stomach) 糖尿病(glycuresis) 尿道(urethra) 胰岛素(insulin) 弧形(arc) 肛门(anus) 病人_症状 (symptom of patient)	药物(medicine) 方法(method) 作用(affect) 血压 (blood pressure) 剂量(dosage) 高血压 (hypertension) 药液(soup) 副作用 (side-affect) 因素(factor) 抗生素(antibiotic)	生命(life) 刘金英(Jinying Liu) 孩子(children) 人们(people) 北京(Bei Jing) 北京_中医药 (chinese medicine in Bei Jing) 工人(worker) 领导(leadership) 时间(time) 地区(area)	计划_生育 (birth control) 局部(locality) 措施(measure) 党员_干部 (party cadre) 群众 (general public) 协会_会员 (academician) 政府(government) 问题(problem) 计划(plan) 全市(city wide)

VI. CONCLUSIONS

In this paper, we present a method to efficiently find the topics according to the combination of topic discovery and topic re-ranking. Up until now, no intensive research has been developed in LDA_col topic model for topic discovery analysis, especially for Chinese corpus. In order to enforce consistency, we slightly modify the graph structure of the model. We use Gibbs sampling to conduct approximate inference. Examples of topic found by modified LDA_col are more interpretable than LDA and LDA_col on NIPS dataset and Chinese corpus. Unlike some traditional topic discovery method, we present two topic re-ranking methods to enhance the topic modeling result by topic significance ranking. Among the two methods that we tested, the TC method performed better than TS.

Evaluating the performance of topic model is a big challenge. Another contribution of our paper is that we propose a formal evaluation for topic discovery model. Thus, we plan to use our method on our digital library for text (books) correlation mining and text(books) clustering, which is our next step for future work.

ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers for their careful and insightful reviews. The work described in this paper was supported by CADAL(China Academic Digital Associative Library).

REFERENCES

[1] D.M. Blei, A.Y. Ng, Jordan M.I.. “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol .3, pp.993-1022, 2003.

[2] D. Ramage, D. Hall, R. Nallapati, C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceeding of EMNLP, pp.248-256,2009.

[3] X. Wang, A. McCallum , X. Wei. Topical N-grams: Phrase and Topic discovery, with an Application to Information Retrieval. In:Proceedings of the 7th IEEE International Conference on Data Mining, pp. 697-702, 2007.

[4] T. Hoffman. Probabilistic latent semantic analysis. In Uncertainty in Artificial Intelligence, 1999.

[5] R. Nallapati , W. Cohen, Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In: Proceeding of AAAI, 2008.

[6] D. M. Blei, J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, vol. 1, pp. 17-35, 2007.

[7] D. M. Blei, J. D. McAuliffe. Supervised topic models. Advances in Neural Information Processsing Systems(NIPS), 2007.

[8] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistic*, vol.4, pp.124-150, 2010.

[9] T. Griffiths, M.Steyvers, D. M. Blei. Integrating topics and syntax. Advances in Neural Information Processing System, 2005.

[10] G. Heinrich. Parameter estimation for text analysis. In:Web: http://www. arbylon. net/publications/text-est. pdf ,2005.

[11] T. L. Griffiths, M. Steyvers, J. B. Tenenbaum. Topics in semantic representation. *Psychological Review*, vol.114, pp.211-244, 2004.

[12] J. CAO, Y. D. ZHANG , J. T. LI , S. TANG. A method of adaptively selecting best LDA model based on density. Chinese Journal of Computer. vol.31, pp. 1780-1787, 2008.

[13] X. Wei, W. Bruce. LDA-based Document Models for Ad-hoc Retrieval. In: Proceedings of SIGIR, pp.178-185, 2006.

[14] L. AlSumait, D. Barbara et al. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of IEEE International Conference on Data Mining, pp.3-12, 2008.

[15] X. Chen, C. Lu, Y. An, Achananuparp P.. Probabilistic models for topic learning from images and caption in online biomedical literature. In:Proceeding of CIKM, pp.495-504, 2009.

[16] H. Wallach. Topic modeling: beyond bag-of-words. In:Proceeding of ICML, pp.977-984, 2006.

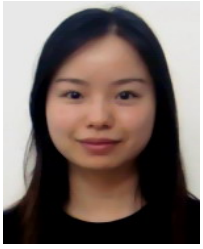
[17] J. Shi, M. Hu, X. Shi, G. Z. Dai. Text Segmentation based on Model LDA. *Chinese Journal of Computers*, vol.31, pp. 1866-1873, 2008.

[18] Y. Q. Song, S. Pan, S. X. Liu, M. X. Zhou, Qian W.H.. Topic and keyword re-ranking for LDA-based topic modeling. In: Proceeding of CIKM, pp.1757-1760, 2009

[19] D. Newman, Y. Noh, E. Tally. Evaluating topic models for digital libraries. In Proceeding of JCDL, pp. 215-224, 2010.

[20] L. AlSumait, D. Barbara, J. Gentle, Domeniconi C.. Topic significance ranking of LDA generative models. In: Proceeding of ECML/PKDD, pp.67-82, 2009.

[21]H. Chim, X. T. Deng. Efficient phrase-based document similartiy for clustering. *IEEE Transactions on Knowledge and Data Engineering*. Vol.20, pp.1217-1229, 2008.



Lidong Wang born in December,4, 1982. She received the M.S degree in computer Science from Ningbo University. She is currently pursuing the Ph.D. degree in College of Computer Science and Technology, Zhejiang University.

Her main research interests include image processing, machine learning and text mining.

Baogang Wei was born in Shenyang, China. He received the M.S. degree in computer software and the Ph.D. degree in computer application from Northwestern Polytechnical University, China, in 1993 and 1997 respectively. He worked as a post-doctor at Zhejiang University, China from October 1997 to September 1999.

Since 1999, he has been being members of Chinese Association for Artificial Intelligence. He is currently a professor at college of computer science and technology of Zhejiang University. So far, he has published more than 40 papers in international conference proceedings and journals. His main research interests include artificial intelligence, pattern recognition, image processing, machine learning, digital library, and information and knowledge management.

Jie Yuan received the M.S. degree in Zhejiang Normal University, Zhejiang province, China. He is currently pursuing the Ph.D. degree in College of Computer Science at Zhejiang University (ZJU), Hangzhou, Zhejiang Province, China.

His current research interests are image processing, pattern recognition and information retrieval.