Neurocomputing 72 (2009) 1775–1781

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A density-based method for adaptive LDA model selection

Juan Cao [a,b,*], Tian Xia [a,b], Jintao Li [a], Yongdong Zhang [a], Sheng Tang [a]

[a] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[b] Graduate University of the Chinese Academy of Sciences, Beijing 100039, China

ARTICLE INFO

Article history:
Received 1 August 2007
Received in revised form
20 December 2007
Accepted 18 June 2008
Communicated by T. Heskes
Available online 28 August 2008

Keywords:
Latent Dirichlet allocation
Topic model
Topic

ABSTRACT

Topic models have been successfully used in information classification and retrieval. These models can capture word correlations in a collection of textual documents with a low-dimensional set of multinomial distribution, called "topics". However, it is important but difficult to select the appropriate number of topics for a specific dataset. In this paper, we study the inherent connection between the best topic structure and the distances among topics in Latent Dirichlet allocation (LDA), and propose a method of adaptively selecting the best LDA model based on density. Experiments show that the proposed method can achieve performance matching the best of LDA without manually tuning the number of topics.

1. Introduction

Statistical topic models have been successfully applied in many tasks, including information classification [1,16,3], information retrieval [14,4], and data mining [15,8], etc. These models can capture the word correlations in the corpus with a low-dimensional set of multinomial distribution, called "topics", and find a relatively short description for the documents.

Latent Dirichlet allocation (LDA) is a widely used generative topic model [1,4,11,14]. In LDA, a document is viewed as a distribution over topics, while a topic is a distribution over words. To generate a document, LDA firstly samples a document-specific multinomial distribution over topics from a Dirichlet distribution; then repeatedly samples the words in the document from the corresponding multinomial distribution.

The topics discovered by LDA can capture the correlations between words, but LDA cannot capture the correlations between topics for the independency assumption underlying Dirichlet distribution. However, topic correlations are common in real-word data, and ignoring these correlations limits LDA's abilities to express the large-scale data and to predict the new data. In recent years many researchers have explored some more complicated and richer structures to model the topic correlations. One example is the correlated topic model (CTM) [2]. Like the LDA, CTM represents each document as a mixture of topics, but the mixture proportion is sampled from a logistic normal distribution. CTM captures the correlations between every pairs of topics by the covariance matrix. To capture the correlations with a more flexible structure, Li et al. [10] proposed Pachinko allocation model (PAM). PAM uses a directed acyclic graph (DAG) to model the semantic structure. Each leaf node in the DAG represents a word in the vocabulary, and each interior node corresponds to a topic. PAM expands the definition of topic to be not only a distribution over words (just like the other topic models), but also a distribution over other topics, called "Super Topic".

Although these models can describe the topic correlations flexibly, they all face the same difficulty to determine the number of topics (parameter K). This parameter will determine the topic structure extracted by the topic model. Y. Teh et al. [13] found an application of hierarchical Dirichlet process (HDP) to automatically learn the number of topics in LDA model. Moreover, Li et al. [9] proposed a nonparametric Bayesian prior for PAM based on a variant of the HDP.

This work is based on the nonparametric nature of the Bayesian analysis tool known as the Dirichlet process (DP) mixture model. But this method needs constructing a HDP model and a LDA model for the same dataset. In this paper, we propose a new method to adaptively select the best LDA model based on topic density, and integrate this task and the model parameter estimation into the same framework. By modeling the generation process of a new topic, we find that the words connecting several topics are likely to generate the new topics. Furthermore, the model's best K is not only determined by the size of dataset, but is also sensitive to inherent correlations in the document collection. After computing the density of each topic, we find the most

* Corresponding author at: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China. Tel.: +86 10 62600659; fax: +86 10 82611846.
E-mail address: caojuan@ict.ac.cn (J. Cao).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.06.011

unstable topics under the old structure, and iteratively update the parameter K until the model is stable.

The rest sections of this paper are organized as follows. In Section 2, we review the basic principles of LDA and the model selection method based on HDP. In Section 3, we study the meaning of the parameter K, and deeply analyze the inherent connection between the topic correlations and the LDA model performance. In Section 4, we propose our approach, and show the experimental results in Section 5. Finally we draw conclusions and give our future work in Section 6.

2. Related work

2.1. Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model, including a three-level structure with word, topic and document. In LDA, documents are viewed as a distribution over topics while each topic is a distribution over words. To generate a document, LDA firstly samples a document-specific multinomial distribution over topics from a Dirichlet distribution. Then it repeatedly samples the words from these topics. LDA and its variants have been successfully applied in many works [2,10,15,16].

Fig. 1 is the graphical model representation of LDA. Given a corpus $\boldsymbol{D}$ containing $\boldsymbol{V}$ unique words and $\boldsymbol{M}$ documents, where each document containing a sequence of words $\boldsymbol{d}$ $\{w_1, w_2, \ldots, w_{\mathbf{Nd}}\}$.

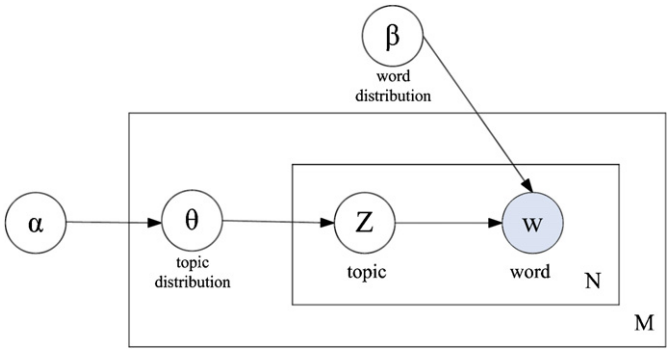

Fig. 1. Graphical model representation of LDA [1].

Given an appropriate topic number K, the generative process for a document $\boldsymbol{d}$ is as following:

(a) Sample a K-vector θ_d from the Dirichlet distribution $p(\theta|\alpha)$, where θ_d is the topic mixture proportion of document $\boldsymbol{d}$.
(b) For $i = 1 \ldots \boldsymbol{N_d}$, sample word $\boldsymbol{w_i}$ in the $\boldsymbol{d}$ from the document-special multinomial distribution $p(w_n|\theta_d, \beta)$.
where $\boldsymbol{\alpha}$ is a k-vector of Dirichlet parameters, and $p(\theta|\alpha)$ is given by

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\ldots\theta_k^{\alpha_k-1} \quad (1)$$

$\boldsymbol{\beta}$ is a $K \times V$ matrix of word probabilities, where $\beta_{ij} = p(w_j = 1|z_i = 1)$, $i = 0, 1, \ldots, K$; $j = 0, 1, \ldots, V$.

LDA assumes the topic proportions are randomly drawn from a Dirichlet distribution, which implies the independence between topics. But these correlations are very common in real-word data. For example, the topic "NBA" is often discussed together with "sports", but unlikely co-occurs with "disease". The inconsistency between assumption and reality makes the LDA be sensitive to the parameter K. CTM replaces the Dirichlet distribution with Logistic Normal one. After getting the correlation between every pair of topics through the covariance matrix, CTM can predict not only the words generated by the same topic, but also the words generated by the correlated topics. Compared with LDA, CTM is less sensitive to the K. But both cannot automatically select the number of topics.

2.2. The method of selecting best K for LDA based on HDP

Teh et al. [13] proposed to determine the best K in LDA by HDP. HDP is intended to model groups of data that have a pre-defined hierarchical structure. Each pre-defined group is associated with a DP whose base measure is sampled from a higher-level DP. Based on the similarity between HDP and LDA in structure, Teh et al. [13] used the nonparametric nature to resolve the problem of selecting appropriate number of topics for LDA. HDP replaces the finite topic mixture in LDA with a DP, and gives the different mixing proportions to each document-specific DP. In the experiments of [13], Teh et al. constructed both the LDA model and the HDP model on one corpus, and obtained the results shown in Fig. 2. The posterior sample of the number of topics used by HDP in the

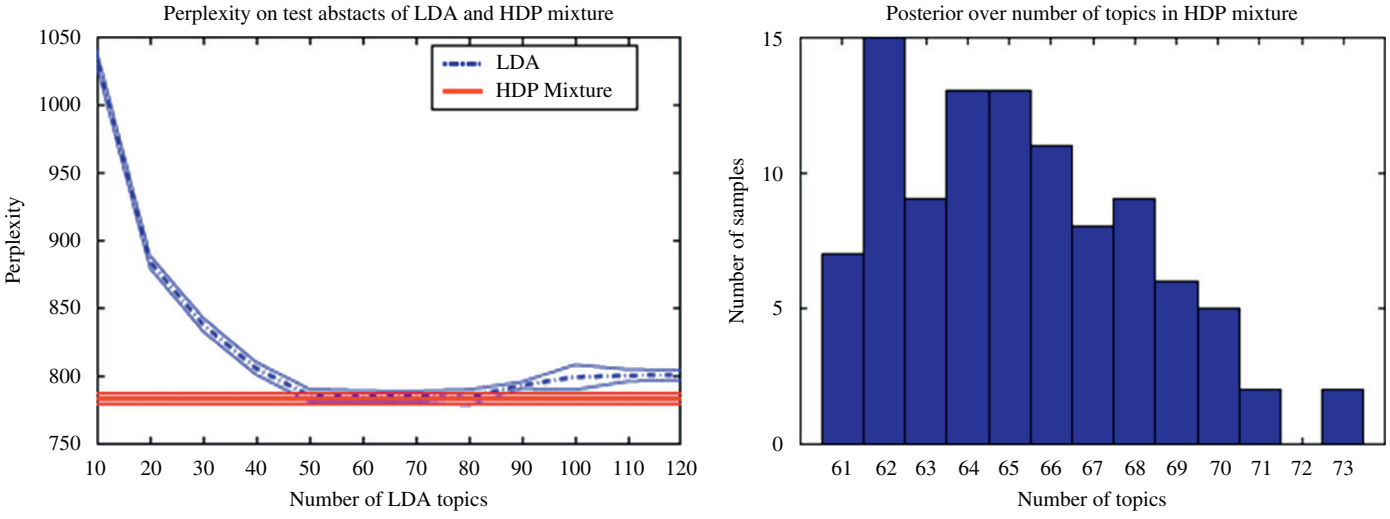

Fig. 2. (Left) Comparison between LDA and HDP. (Right) Histogram of the number of topics the HDP mixture used over 100 posterior samples [13].

right histogram is just consistent with the best parameter K of the LDA model in the left figure (the best number of topics is 50–80).

Being different from the HDP, our idea is to find the connection between the LDA model performance and the topic correlations, and adaptively guide the generation of the topics by the topic density statistics in the parameter estimation process.

3. The relationship between the best K and the topic correlations

Topic model can extract the latent topic structures by analyzing a large scale of statistical data. These structures are hierarchical and corpus-specific. In a good topic structure of LDA, every topic is an understandable, meaningful and compact semantic cluster, and is exclusive to each other. The higher layer needs fewer topics, but the topics are abstract and overlap with each other, which results in too many correlations to retain the discriminability; On the other hand, the lower layer needs more topics, and the topics are more concrete, then the information implicated in one topic is too little (every topic is a sparse vector in the large word space) to retain the discriminability. The number of topics determines the layer of the topic structure. So find the best K is important to the topic model.

3.1. The meaning of parameter K

We show the influence of K on the topic model with two graphs. The following topic structures are extracted from a corpus with five unique words. We denote the topics as open nodes, the words as solid nodes, and the dependencies between them as edges.

Fig. 3 describes the case when K is too small ($K = 2$). Z_1 and Z_2 overlaps over three words. Moreover, their dependence degrees on W_1 and W_2 are close. In this structure, the discrimination between Z_1 and Z_2 is small, and model the corpus with this topic structure will lose much important information in original data.

Fig. 4 describes the case when K is too large ($K = 4$). We find that the Z'_2 and Z'_3 have strong correlation in nature from Fig. 3. (In Fig. 3, the distribution proportion of W_3 on $\{Z_1, Z_2\}$ is $\{1, 0\}$, and W_4 is $\{0.9, 0.1\}$.) But LDA cannot capture it while inferring the document posterior distribution over topics. So this topic structure cannot represent the original data accurately. On the other hand, CTM can obtain the correlation between Z'_2 and Z'_3 from the covariance matrix, and predict W_3 from W_4. So CTM can support more topics.

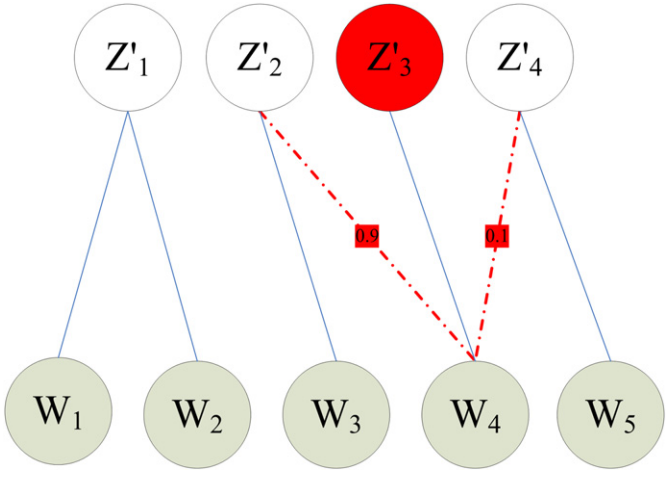

Fig. 3. Topic structure when K is too small.

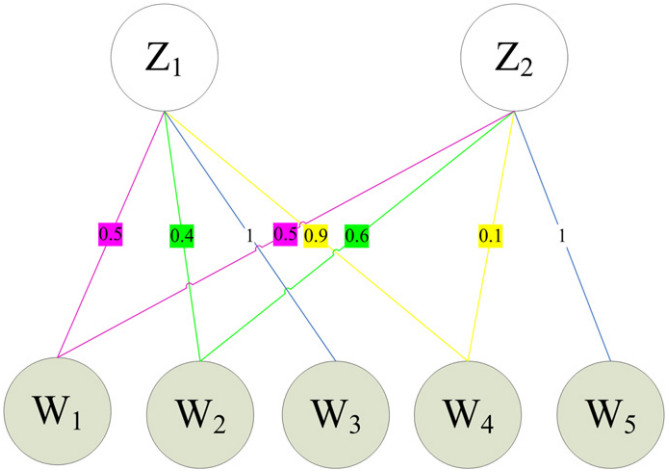

Fig. 4. Topic structure when K is too large.

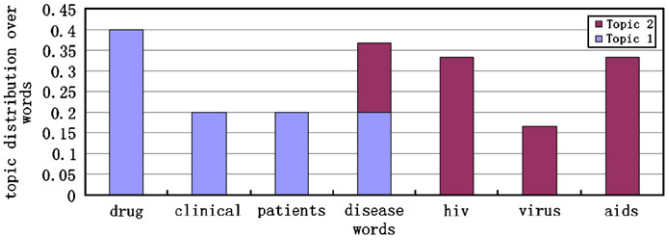

Fig. 5. Topic distribution over words when $K = 2$.

Table 1
Word assignment over the topics ($K = 2$)

Topics	Words
1	Drug clinical patients disease
2	Aids HIV virus

When we update the structure with the red dashed line, the correlations between topics are deduced, and every topic can imply more inherently correlative information.

3.2. Generation of a new topic

In this section we built three LDA models for a corpus with $K = 2,3,4$, and we will observe the correlation between a new topic's generation and the topic distribution over words.

This corpus includes four documents and seven unique words:

Doc1: drug clinical patients
Doc2: drug disease
Doc3: HIV virus aids
Doc4: aids HIV disease

Fig. 5 is the topic distribution over words when $K = 2$. The two topics overlay on the word "disease", and the distribution proportion is close. It results in a strong correlation between the two topics, and "disease" is an unstable factor in this topic structure.

Table 1 shows the word assignment in this topic structure (W_n belongs to topic $I = \text{argmax}_i(p(W_n|Z_i))$).

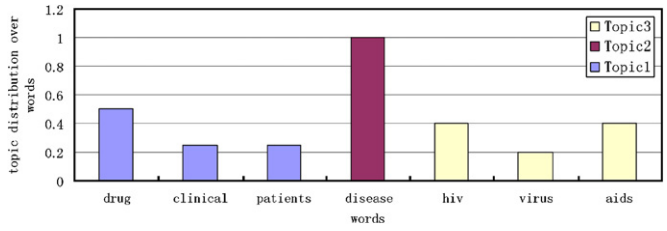

Fig. 6. Topic distribution over words when $K = 3$.

Table 2
Word assignment over the topics ($K = 3$)

Topics	Words
1	Drug clinical patients
2	Disease
3	Aids HIV virus

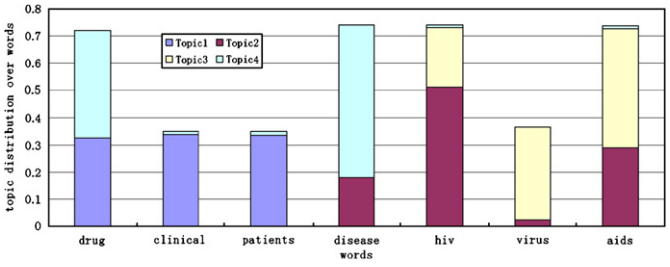

Fig. 7. Topic distribution over words when $K = 4$.

Table 3
Word assignment over the topics ($K = 4$)

Topics	Words
1	Clinical patients
2	HIV
3	Aids virus
4	Disease drug

Fig. 6 is the topic distribution over words when $K = 3$. The unstable factor in Fig. 5 has been separated from topic 1, and generates a new topic. In this new structure, the topic distribution over words has few overlap, and the structure is relatively stable (Table 2).

Fig. 7 shows that the overlaps in the topics distribution become serious when $K = 4$. Furthermore, the word assignment over the topics in Table 3 tells us that the topics in this structure are not so meaningful (for example, the information implying in topic 2 is few).

From the above analysis, we conclude that a new topic of LDA is generated from the words connecting several topics (just the overlap words in the old topic distribution).

We use standard cosine distance to measure the correlation between topics

$$\mathrm{corre}(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \sqrt{\sum_{v=0}^{V} (T_{jv})^2}} \tag{2}$$

$\mathrm{corre}(T_i, T_j)$ is smaller, the topics are more independent.

We use the average cosine distance between every pair of topics to measure the stability of topic structure:

$$\mathrm{ave\_dis(structure)} = \frac{\sum_{i=0}^{K} \sum_{j=i+1}^{K} \mathrm{corre}(T_i, T_j)}{K \times (K-1)/2} \tag{3}$$

A smaller ***ave_dis*** shows that the structure is more stable. The ***ave_dis*** of above three topics structure are 0.1195, 0.00014 and 0.279, respectively. Obviously the structure when $K = 3$ is most stable.

4. The density-based method for adaptive LDA model selection

We have validated that the best K of LDA is correlated with the distances between topics in Section 3. In this section, we integrate the idea of clustering based on density [5] into our method, and propose to adaptively select the appropriate number of topics in LDA based on topic density. The aim of clustering based on density is that the similarity will be as large as possible in the intra-cluster, but as small as possible between inter-clusters. This aim just fit the standard of selecting best topic structure in LDA. A topic is equivalent to a semantic cluster. A larger similarity in intra-cluster shows that this cluster can represent a more explicit meaning, and a smaller one between intra-cluster shows that the topic structure is more stable.

For the convenience of describing our method, we introduce three definitions first:

Definition 1. (*Topic density*). Given a topic ***Z*** and the distance ***r***, by computing the average cosine distance (Eq. (1)) between ***Z*** and the other topics, the number of topics within the radius of ***r*** from ***Z*** is the density of ***Z***, called ***Density*(*Z, r*)**.

Definition 2. (*Model cardinality*). Given a topic model ***M*** and a positive integer ***n***, the number of topics whose topic densities are less than ***n*** is the cardinality of ***M***, called ***Cardinality*(*M, n*)**.

Definition 3. (*Reference sample*). Given a topic ***Z***, radius ***r*** and threshold ***n***, if $\mathrm{Density}(Z, r) \leqslant \boldsymbol{n}$, then call the word distribution vector of ***Z*** as a reference sample of topic ***Z***.

The reference sample is not a document vector in the real dataset, but a virtual point over the word distribution.

Based on these definitions, we describe our method as follows:

(1) Given an arbitrary K_0, initialize the sufficient statistics by ***random*** [6] method, and use the *variational EM* algorithm [2] to estimate the model parameters, and get the initial model LDA (α, β);
(2) Regarding the topic distribution matrix â of the old model as a cluster result, we sequentially compute the model's average cosine distance $\boldsymbol{r1} = \mathrm{ave\_dis}(\beta)$, the densities of all the topics Density(Z, ***r1***), and the cardinality of the old model C = Cardinality(LDA,0);
(3) Re-estimating the model parameter K based on the C. The updating formula is as follows:

$$K_{n+1} = K_n + f(r) \times (K_n - C_n) \tag{4}$$

$f(\boldsymbol{r})$ is the changing direction of ***r***. If the direction is negative (being opposite to the former), then $f_{n+1}(\boldsymbol{r}) = -1 * f_n(\boldsymbol{r})$, else $f_{n+1}(\boldsymbol{r}) = f_n(\boldsymbol{r})$. $f_0(\boldsymbol{r}) = -1$.
When the convergence direction is negative, we ascending sort the topics by the densities, and extract the former K' topics as the reference samples to initialize the sufficient statistics. When the convergence direction is positive, we initialize the sufficient statistics by ***seeded*** [6] method.
(4) Repeat (2) and (3), until the average cosine distance and cardinality of the LDA model both converge.

5. Experiments

5.1. Experimental data

We build three datasets on the English ASR texts corpus of TRECVID2005 [7]. All the texts are pre-processed by the SMART's English stoplist and by Porter's stemming algorithm [12].

$\boldsymbol{D_0}$ is the whole English ASR texts corpus, including 20932 shot documents and 8410 unique words;

$\boldsymbol{D_1}$ is made up of three judged collections of 0168, 0160 and 0169 in the search task, including 3754 shot documents and 5535 unique words. We divide it into $\boldsymbol{D_1}$_train and $\boldsymbol{D_1}$_test by 10:1;

$\boldsymbol{D_2}$ is made up of three judged corpus of 0168, 0165 and 0172 in the search task, including 4129 shot documents and 5681 unique words. We divide it into $\boldsymbol{D_2}$_train and $\boldsymbol{D_2}$_test by 10:1.

Following is the detail of the above five search queries:

0160 = "Find shots of something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible";

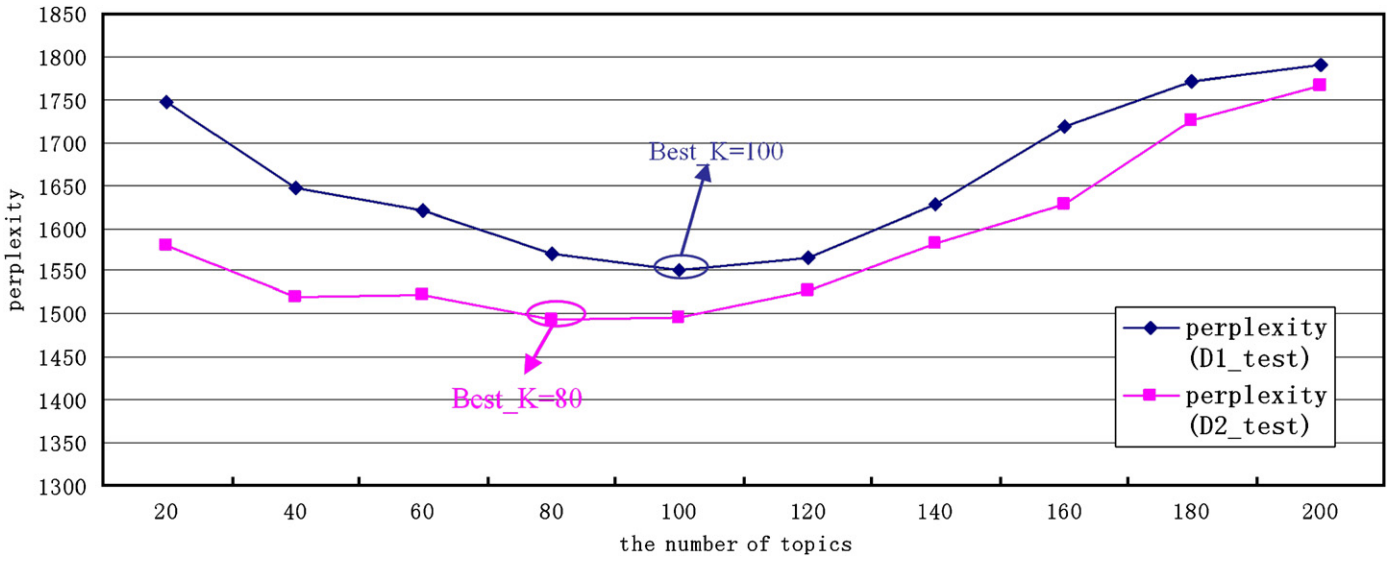

Fig. 8. Comparison of perplexity results of LDA models in D1 and D2.

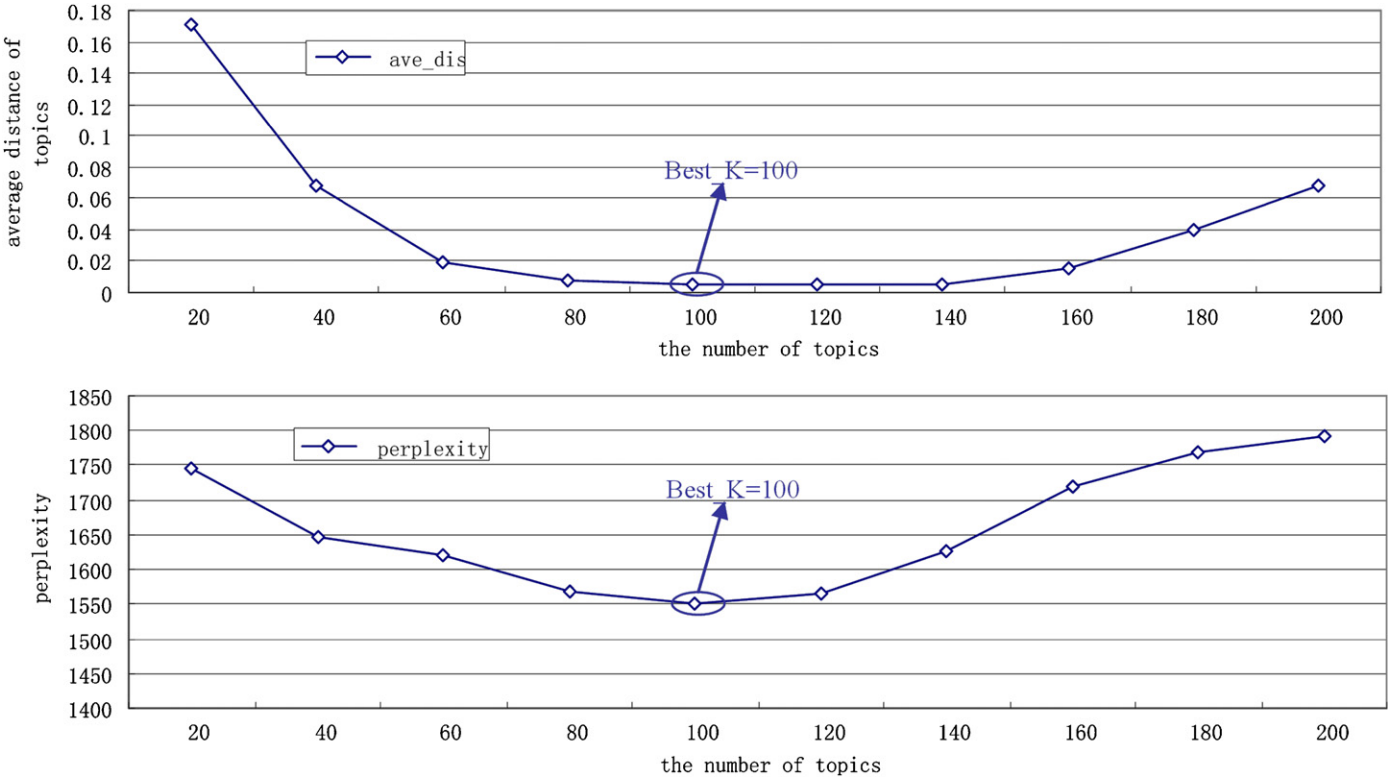

Fig. 9. Comparison between the cures of ave_dis and perplexity.

0165 = "Find shots of basketball players on the court";
0168 = "Find shots of a road with one or more cars";
0169 = "Find shots of one or more tanks or other military vehicles";
0172 = "Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people".

Among the three datasets, the sizes of $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ are equal, but the inherent correlations in $\boldsymbol{D_1}$ are strong. The size of $\boldsymbol{D_0}$ is greater than those of $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, but the documents in it are more noisy.

In particular, we computed the ***perplexity*** [2] of a held-out test set to evaluate the topic models. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraicly equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance. The perplexity of a test set $\boldsymbol{D_{test}}$ including $\boldsymbol{M}$ documents is given by

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} p(d_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (5)$$

where N_d is length of document d; $p(d_d)$ is the probability of the document $\boldsymbol{d}$ generated by the model.

Meanwhile, we measure the stability of the models by the average cosine distance ***ave_dis*** described in Section 3.

5.2. Experimental results

We designed three experiments to validate the points proposed above:

Experiment 1. We compared the best **K**s of different datasets in two groups of experiments.

Fig. 8 is the contrast between the best **K**s in two datasets with the same size but with different document correlation degrees. The best number of topics is 100 in $\boldsymbol{D_1}$, and is 80 in $\boldsymbol{D_2}$. It shows that the LDA model need more topics when the inherent correlations in the corpus are stronger (more topics can make the topics more material and independent). Moreover, the ***perplexity*** curve of $\boldsymbol{D_1}$ is wholly higher than that of $\boldsymbol{D_2}$, i.e. the LDA performs worse in the corpus with stronger correlation between documents. It stems from the limitation of LDA that it cannot model the topic correlation.

Meanwhile, we also test the best K in $\boldsymbol{D_0}$. We train several LDA models with $K = 10, 30, 50, 100, 200, 300, 400$ and 500 and the best K of $\boldsymbol{D_0}$ is 30. It again validates that the best K is not only relevant with the size of corpus, but also sensitive to the correlations in the corpus.

Experiment 2. We experiment the inherent connection between the best K and the average cosine distance of topics in $\boldsymbol{D_1}$.

In Fig. 9, we observe the changing trend of the average distance and perplexity of the model with K. The two curves change with the same rules, and reach the best values at the same K. When the average distance of the topics reaches the minimum, the corresponding model performs best.

Experiment 3. We realize our method of adaptively selecting best K based on density in $\boldsymbol{D_1}$ with six experiments. The initial K's are 10, 50, 100, 200, 300 and 500. All can stop at the best K ($K = 100$) after several iterations. If the initial values are closer to the best K, the iterations needed are less (Table 4).

Table 4
The results of the algorithm adaptively selecting best K based on density

Initial K	Best K	Iterations
10	97	19
50	99	26
100	102	2
200	109	3
300	106	5
500	102	34

6. Conclusions

In the topic model, the number of topics is crucial to the performance, but finding appropriate value for it is very difficult. In this paper, motivated by the limitation that LDA ignores the topic correlation, we further study the connection between the LDA performance and the topic correlation, and demonstrate that the LDA model performs best when the average cosine distance of topics reaches the minimum. We integrate the selecting best K into the estimation process of model parameters, and propose a new method of adaptively finding the best number of topics based on topic density. Experiments show that this method is effective.

However, our method is based on statistics of the whole corpus, with no straightforward extension for out-of-sample example. Therefore, an interesting future work is how to extend the topic structure when a new document appears. We could record the topic density information with a tree, and dynamically update the density tree as the corpus enlarges. This work is significant for the large-scale dataset.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), the National Nature Science Foundation of China (60873165), and the Beijing New Star Project on Science & Technology (2007B071).

References

[1] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[2] D. Blei, J. Lafferty, Correlated Topic Models. Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006.
[3] L. Cai, T. Hofmann, Text categorization by boosting automatically extracted concepts, Proc. SIGIR 9 (2003) 182–189.
[4] J. Cao, J.T. Li, Y.D. Zhang, S. Tang, LDA-based retrieval framework for semantic news video retrieval, IEEE Int. Conf. Semantic Comput. (ICSC) 9 (2007) 155–160.
[5] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, 1996, pp. 226–231.
[6] ⟨http://www.cs.princeton.edu/~blei/lda-c/⟩.
[7] ⟨http://www-nlpir.nist.gov/projects/t01v/⟩.
[8] V. Jain, E. Learned-Miller, A. McCallum, People-LDA: anchoring topics to people using face recognition, in: International Conference on Computer Vision (ICCV), 2007.
[9] W. Li, D. Blei, A. McCallum, Nonparametric Bayes Pachinko allocation, UAI (2007).
[10] W. Li, A. McCallum, Pachinko allocation: DAG-structured mixture models of topic correlations, in: International Conference on Machine Learning (ICML), 2006.
[11] F.-F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 2005, pp. 524–531.

[12] M.F. Porter, An algorithm for suffix stripping, Program 14 (1980) 130–137.
[13] Y. Teh, M. Jordan, M. Beal, D. Blei, Hierarchical Dirichlet processes, J. Am. Stat. Assoc. 101 (476) (2007) 1566–1581.
[14] X. Wei, W. B. Croft, LDA-based document models for ad-hoc retrieval, in: Proceedings of the 29th SIGIR Conference, 2006, pp. 178–185.
[15] E. Xing, R. Yan, A. Hauptmann, Mining associated text and images with dual-wing harmoniums, in: Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI-05), AUAI press, 2005.
[16] J. Yang, Y. Liu, E. P. Xing, A. Hauptmann, Harmonium-based models for semantic video representation and classification, in: Proceedings of the Seventh SIAM International Conference on Data Mining, 2007.

Juan Cao, born in 1980 (Ph.D. Candidate). Her research interests focus on multimedia retrieval and machine learning.

Tian Xia, born in 1980 (Ph.D. Candidate). His research interests focus on multimedia retrieval and machine learning.

Jintao Li, born in 1962 (Professor, Ph.D. Supervisor). His major field includes multimedia processing and VR technology.

Yongdong Zhang, born in 1973 (Ph.D. Associate Professor). His major field includes image processing and video processing.

Sheng Tang, born in 1972, (Ph.D). His major field includes multimedia retrieval and video processing.