LDA-Based Retrieval Framework for Semantic News Video Retrieval

Juan Cao[1, 2],Jintao Li[1], Yongdong Zhang[1] , and Sheng Tang[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China

[2] Graduate University of the Chinese Academy of Sciences, Beijing 100039, China

{caojuan, jtli, zhyd ,ts }@ict.ac.cn

Abstract*

Topic-based language model has attracted much attention as the propounding of semantic retrieval in recent years. Especially for the ASR text with errors, the topic representation is more reasonable than the exact term representation. Among these models, Latent Dirichlet Allocation(LDA) has been noted for its ability to discover the latent topic structure, and is broadly applied in many text-related tasks. But up to now its application in information retrieval(IR) is still limited to be a supplement to the standard document models, and furthermore, it has been pointed out that directly employing the basic LDA model will hurt retrieval performance. In this paper, we propose a lexicon-guided two-level LDA retrieval framework. It uses the HowNet to guide the first-level LDA model's parameter estimation, and further construct the second-level LDA models based on the first-level's inference results. We use TRECVID 2005 ASR collection to evaluate it, and compare it with the vector space model(VSM) and latent semantic Indexing(LSI). Our experiments show the proposed method is very competitive.

Keywords: *ASR text, LDA, Topic-based model, Semantic video retrieval*

1. Introduction

In video retrieval, the users' need is not only the visually similar content, but the semantic similar content. So low-level features are now becoming insufficient to build efficient news video retrieval systems. Many works have done to bridge the gap between low-level visual features and semantic content, but this is still a challenging task in the future. In this paper, we try an alternative way to mine the video's semantic information from its automatic speech recognition (ASR) text. One reason is that ASR text is the direct semantic description about video content. The other is that the technologies of text retrieval are more mature than the ones of visual features process.

But the traditional vector space representation is far from the semantic retrieval's requirement，for it didn't consider the correlation between the terms, and can't generate an accurate representation of the inherent structure of the corpus. So researchers try to use the word clustering or higher order n-grams to capture the important relationships between terms. In these years, the topic model is definitely proposed to capture the inherent underling topical structure within the corpus. Especially for the ASR text with errors, the topic representation is more reasonable than the exact term representation. There are two important topic retrieval models: DeerWester's Latent Semantic Indexing(LSI)[1] and Blei's Latent Dirichlet Allocation(LDA)[2]. LSI maps the documents' high-dimensional term representation to its low-dimensional semantic representation (so-called latent semantic space) via the technique of singular-value decomposition(SVD). LSI has been applied with remarkable success in semantic retrieval field[4]. To strong the statistical foundation of the LSI, Haphman presents the Probabilistic Latent Semantic Indexing(PLSI)[3], and defines a proper generative model of the data based on the likelihood principle. To overcome the PLSI' drawbacks of lacking a probabilistic model for documents distribution, Blei presents LDA[2]. LDA is a generative probabilistic model for a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA has been effectively applied to general text retrieval as a supplement to the original document model [2][5], but also has been pointed out that directly employing the basic LDA model will hurt retrieval performance [5][6], for its topic representation of the documents is coarse as opposed to the exact term representation. Moreover, LDA fails

* **This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), Beijing Science and Technology Planning Program of China (D0106008040291), and National High Technology and Research Development Program of China (863 Program, 2007AA01Z416).**

0-7695-2997-6/07 $25.00 © 2007 IEEE
DOI 10.1109/ICSC.2007.26

to model the correlation between topics. This correlation information is very important to reduce the wrongs in the inference of the latent topical structure. Especially to the erroneous and abbreviated ASR texts, LDA's feasibility and effectiveness is still unknown. In TRECVID 2005 ASR texts, the word error rate(WER) is 33.8%, and the average length of shot document is about 14.5, while the counterpart in TREC's general document collection is 84.6(after pre-processing). These data are a challenge to LDA. To provide enough useful information for the LDA model, we utilize the HowNet's prior knowledge to guide the estimation of LDA model parameters, and design a two-level retrieval framework to inference the corpus' topic distribution by stages. The experiments show that the performance of this framework outperforms traditional term matching and LSI method.

The rest of this paper is organized as follows: In section 2, we give a brief review of LSI and LDA in IR; next we present the **Lexicon-guided Hierarchical LDA retrieval Framework** in the section 3. Then we set up the experimental circumstance to compare this framework to traditional methods in section 4 and discuss an open question in section 5. Finally, conclusion and future work are given in the last section.

2. Related Works

2.1 the LSI-based Retrieval Model

In natural language analysis field, a certain word can be interpreted in different ways within different contexts (polysemy); while the same concept can be described using different terms(synonymy). LSI was proposed to solve these problems. The key idea is clustering the **co-occurring keywords**, and mapping documents and queries into a lower dimensional space(latent semantic space)[1]. The advantage of retrieval in this space is that a query can be similar to a document even when they share no words.

The LSI technique makes use of the singular value decomposition(SVD) to mine the whole collection's semantic structure. Firstly, we represent the document collection as a term-document matrix $M_{t\times d}$, then use the SVD to decompose $M_{t\times d}$ into the product of three matrices:

$$M_{t\times d} = T_{t\times r} S_{r\times r} \left(D_{\mathrm{d}\times\mathrm{r}}\right)^T \qquad \textbf{(1)}$$

where ***t*** is the number of terms, ***d*** is the number of documents. ***r*** is the rank of M. T and D are the matrices with orthonormal columns. S is a diagonal matrix of M's singular values sorted in decreasing. The singular value is larger, the corresponding dimension is more important. By restricting the matrixes T,S,D to their first k rows , we can get a approximate matrix M' :

$$M'_{t\times d} = T'_{t\times k} S'_{k\times k} \left(D'_{\mathrm{d}\times k}\right)^T \qquad \textbf{(2)}$$

Where k< r **.**

Discarding the less important dimensions can remove much of the noise, and transform the term space to the reduced-dimension latent semantic space.

In retrieval period, through the formula(3) we transform the q to latent semantic space:

$$q' = q^T T'_{t\times k} {S^{-1}}_{k\times k} \qquad \textbf{(3)}$$

We use the standard cosine measure(4) to compute the similarity between the query and document:

$$LSI\left(q, doc_i\right) = \frac{q' \bullet \left(D'\right)_i}{\left|q'\right|\left|\left(D'\right)_i\right|} \qquad \textbf{(4)}$$

$\left(D'\right)_i$ is the i-th column of matrix D' .

Based on the LSI's theory, Hofmann further proposed probabilistic latent semantic indexing(PLSI).The details can be found in [3].

2.2 the Basic LDA Model for Semantic Retrieval

Hofmann's work is a useful step toward probabilistic modeling of text, but there is a serious overfitting problem. This problem essentially stems from the restriction that each future document exhibits the same topic proportions as were seen in one or more of the training documents. Blei's LDA overcomes this problem by treating the topic mixture weights as a ***k***-parameter hidden random variable rather than a large set of individual parameters which are explicitly linked to the training set(Actually, PLSI is in fact a maximum posteriori estimate of LDA[7]).

Given a corpus D containing V unique words and M documents, where each document containing a sequence of words that $\boldsymbol{d} = \{w_1, w_2, \ldots, w_n\}$. There is an assumption that the dimensionality k of the topic variable z is known and fixed. The LDA model defines two corpus-level parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Where $\boldsymbol{\alpha}$ is a k-vector of Dirichlet parameters. $\boldsymbol{\beta}$ is a $K\times V$ matrix of word probabilities, where $\beta_{ij} = p\left(w_j = 1 \mid z_i = 1\right)$, i=0,1,…,K; j=0,1,…, V.

The LDA model includes two parts, parameters estimation and posterior distribution inference. LDA uses an alternating *variational EM* algorithm to find the approximate empirical Bayes estimates for the LDA model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and in the iterations process maximizes a lower bound on the log likelihood of the data:

$$l(\alpha,\beta) = \sum_{i=1}^{M} \log p\left(d_i \middle| \alpha,\beta\right) \qquad \textbf{(5)}$$

The inference process mainly includes three steps:

a ***k***-dimensional Dirichlet random variable ***θ*** can take values from the following probability density function:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \quad \textbf{(6)}$$

Where $\Gamma\left(x\right)$ is the Gamma function, and we can get the distribution of topic variable z from θ, here ***p(z |θ)*** is defined simply by ***p(z^{i} |θ) =1*** where ***θ$_{i}$ =1***.

Then the word distribution can be get from the following function:

$$p(w \mid \theta, \beta) = \sum_{z} p\left(w \middle| z, \beta\right) p\left(z \middle| \theta\right) \quad \textbf{(7)}$$

Finally, the generative process for a document ***d*** is defined as a continuous mixture distribution:

$$p(d \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} p(w_n | \theta, \beta) \right) d\theta \quad \textbf{(8)}$$

In LDA model described above, the posterior distribution is intractable, So Blei presented a simple convexity-based variational approach for inference.

In the retrieval process, LDA model regard the given query ***Q***={q$_{1}$,q$_{2}$,……,q$_{m}$} as an unseen documents. Given a LDA model lda(*α,β*) and a document ***d***, the similarity between ***Q*** and ***d*** is defined as follows:

$$simi(Q, d) = p\left(Q \middle| d\right) = \prod_{i=1}^{m} p\left(q_i \middle| d, \alpha, \beta\right) \quad \textbf{(9)}$$

And the inference process of $p\left(q_i \middle| d, \alpha, \beta\right)$ is defined as follows:

$$p(q_i \mid d, \alpha, \beta) = \int p(\theta \mid d, \alpha) p\left(q_i \middle| \theta, \beta\right) d\theta \quad \textbf{(10)}$$

$$p(\theta \mid d, \alpha) = \frac{p(\theta, d \mid \alpha, \beta)}{p(d \mid \alpha, \beta)} = \frac{p(\theta \mid \alpha) \prod_{j=1}^{N} p(w_j \mid \theta, \beta)}{p\left(d \middle| \alpha, \beta\right)}$$

However, the LDA model is too coarse to be used as the only representation for IR, so many researchers try to linearly combine the LDA model with the original document model to perform retrieval:

$$simi\left(Q, d\right) = \prod_{q \in Q} \left(\lambda p_c\left(q \middle| d\right) + (1 - \lambda) p_{lda}\left(q \middle| d\right) \right) \quad \textbf{(11)}$$

Where $p_c\left(q \middle| d\right)$ is the original document model.

3. Lexicon-guided Hierarchical LDA retrieval Framework

It is a well known fact that the EM algorithm is sensitive to its initialization. In the parameter EM estimate step of the basic LDA model, the initial values of the word distribution over topics is get in two ways. One is get by a random function, the other is get by randomly sampling the documents from the corpus for every topic, and based on these samples to forecast the word distribution over the topics. Both methods ignore the correlation between topics, and regard their occurrence as independent. But in real data, many topics are highly correlated, and many repel one another. For example, NBA is likely co-occurrence with sports, game, player etc, but unlikely co-occurrence with politics, elect, and war. Blei further developed an correlated topic model(CTM) by using the logistic normal to model the correlation between topics in the basic LDA, and demonstrate that considering the correlation between topics gives a better fit than basic LDA[10].

Actually, the topics' correlations are exhibited through the words' correlations, and many correlated topics may share the same domain and the similar words distribution. Based on this knowledge, we get the correlations between words and several known domains from HowNet in advance, and utilize this information to guide the topic distribution estimation. This prior knowledge should provide a more informative prior as it relies on the term distribution over the domains but not on the entire collection. In addition, we design two layers of LDA models to perform retrieval, the first-level model is used to collect the documents in the same domain to subsets, and the second-level models are used to capture the topic structure underlying every domain. This structure is more in-depth, and can more accurately represent the document. We want to emphasize that the aim of using the clustering results is to collect the documents which sharing more commonness to a subset(semantic class), and perform the LDA on these optimized subsets. So we needn't the exact clustering and allow the subsets overlapping. The details of the retrieval framework are as following:

In **step1**, we predefine three domains of politics, sports and finance based on the characteristics of news video[7]. We get the word distribution over these domains via the HowNet's relevance calculator[11][12], and use the normalized results as the initial values of parameter β' to construct the first-level LDA model.

In **step2**, according to the first-level model's inference results, we get the document posterior over the three domains $\gamma_{ij} = p\left(Domain_j = 1 \mid d_i = 1\right)$, i=0,1,……M; j=1,2,3. Then we partition the corpus ***D*** to three subsets(C_1, C_2, C_3) based on γ, every subset is a collection of documents which are high relevant to one domain. One document is allowed to belong to more than one subset.

In **step3**, we construct the second-level LDA models with k topics separately on the three subsets, and get the documents' final topic representation.

In **step4**, in the retrieval process, we use the first-level LDA model to infer the distribution of query Q over three domains $<l_1, l_2, l_3>$. These values will be used as the confidence scores of each domain's model to Q. Then we further infer the likelihood of very document d generating query Q in the three second-level LDA models $p_{LDA\_i}\left(Q|\,d\right)$, i=1,2,3.Finally fuse these probabilities as follows:

$$p\left(Q|d\right)=\sum_{i=1,2,3} l_i p_{LDA\_i}\left(Q|\,d\right), i=1,2,3 \qquad \textbf{(12)}$$

Fig.1. displays the graphical representation of two-level LDA retrieval model. Where random variables are represented as nodes, and dependencies between them are represented as edges. Observed variables are represented as solid nodes and hidden, or unobserved, variables as open nodes.

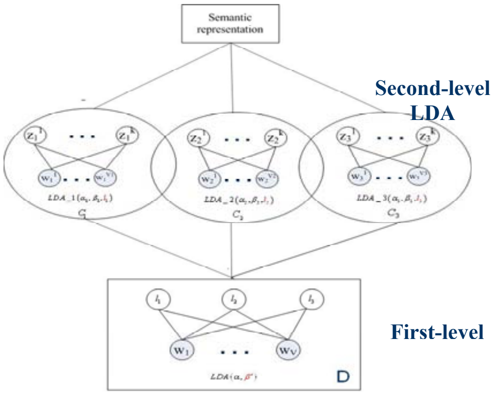

Fig.1. the Graphical representation of the two-level LDA model

4. Experiment

The basic LDA code that we use in this paper is available at http://www.cs.princeton.edu/~blei/lda-c, and we mend it in our method.

We choose the TRECVID 2005's English test collection as our experimental data. After pre-processing, the total number of the shot documents is 20932, the number of the unique terms is 8410, and 305529 non-zero entries in the term-document matrix.

We select 12 queries† from the 2005 search task to test our methods, and the choice criterion is the number of the relevant documents in the test collection bigger than 50. Table.1 is the queries' information:

Table 1. Statistics of topics

query	Rel-Num	Pre-Rank
149	72	9
160	90	17
161	255	12
162	162	22
163	95	14
164	64	5
165	130	13
166	159	18
168	548	7
169	286	11
170	135	19
172	487	16

Where the Rel-Num is the topics' relevant documents number in the test collection; and the Pre-Rank is the queries' mean average precision ranking among the 24 queries, the average ranking is 13.5.

The version of HowNet we used in our experiments includes 158849 words.

4.1 Evaluation Method

To determine the accuracy of the methods' retrieval results, we use *average precision*, following the standard in TRECVID evaluations. Let L^k = $\{l_1, l_2, \ldots\ldots, l_k\}$ be a ranked version of the retrieval results set S. At any given rank k let R_k be the number of relevant shots in the top k of L, where R is the total number of relevant shots in a set of size S. Then average precision is defined as:

$$\text{Average Precision} = \frac{1}{R}\sum_{k=1}^{A}\frac{R_k}{k} f\left(l_k\right) \qquad \textbf{(13)}$$

Where function $f\left(l_k\right)$ = 1 if $l_k \in$R and 0 otherwise. The average precision favors highly ranked relevant shots.

We use the truths provided by NIST to evaluate our methods. But we found that the truths are only subsets of the whole test collection, and many of the documents returned by our system were not judged for relevance. The same obstacle of evaluation is also encountered by[13]. So we only evaluate the performance of submitted results with relevance judgments.

†In TRECVID's search task, the query is named as topic. To distinguish this topic to the topic in LDA model, we instead it by query.

4.2 Pre-processing

Firstly, we matched the ASR texts with the shots based on the shot temporal boundaries, and expanded the shot boundaries to include up to 3 immediate neighbors on either side to compensate for speech and visual misalignment.

Then we removed the words in a standard stop word list with 524 terms. Additionally, we computed all the terms' document frequency (DF) occurred in the whole collection, and add the terms with highest DF to the user-defined stop word list.

To decrease the matrix total size and to enhance the efficiency, we extract only nouns and verbs. This step can decrease the term dimension from 8410 to 7159.

We also used the Porter's well-established stemming algorithm to unify terms, which allows several forms of a word to be considered as a unique term.

4.3 Primary Results

In our experiment, we realized five runs to compare our method's effect. Where:

Run1: the vector space model(VSM)
Run2: LSI(306)
Run3: LDA(100)
Run4: two-level LDA model(LDA(3)+LDA(200))
Run5:lexicon-guided two-level LDA model (LDA(3)+LDA(200))

The figures in the brackets are the number of topics for respective models, being selected after many tries. And the five methods have the same pre-processing.

Table2 displays the mean average precision(MAP) of above five runs to the 12 queries, and the differences in performance between Run5 and Run1, Run4.

Table 2. Evaluation Results in MAP

Runs	MAP
Run1	0.141
Run2	0.16
Run3	0.155
Run4	0.157
Run5	0.176
%Run5 over Run1	+24.895
%Run5 over Run4	+11.977

Run3's relatively low MAP once again validates that directly employing LDA is not advisable.

Run5's improvement on performance shows that **the** two-level LDA-based retrieval framework is effective.

Further more, after using lexicon to guide the initial word distribution over domains in Run5, we can get a posterior distribution which is more close to our expectation. Table.3 and table.4 separately display the top 10 terms over three domains in the Run5's first-level LDA model and in the Run4's. From table.3 we can easily find the underlying topical structure: politics, sports and finance; but the structure in table.4 is not so clear except politics, for the number of documents about politics in the test corpus is very large.

Notice that the terms in Table 3 and Table 4 are all the forms after stemming.

Table 3. Term occurrence with the highest Probability in Run5's first-level LDA model

1	**2**	**3**
presid	**time**	**dollar**
state	**call**	**american**
live	**peopl**	**percent**
famili	**start**	**number**
elect	**point**	**show**
peopl	**game**	**monei**
forc	**stori**	**make**
talk	**report**	**time**
hous	**season**	**rate**
countri	**nbc**	**bill**

Table 4. Term occurrence with the highest Probability in Run4's first-level LDA

1	**2**	**3**
presid	**time**	**forc**
state	**peopl**	**area**
elect	**dollar**	**citi**
issu	**famili**	**show**
hous	**report**	**point**
secur	**percent**	**make**
call	**live**	**american**
countri	**new**	**find**
servic	**start**	**game**
bill	**nbc**	**system**

On the other hand, LDA model is a relatively coarse representation of corpus. It can enhance the recall in retrieval process, but will lead to many irrelevant documents highly ranked in the result list. Table.5 displays the average hits of 12 queries at corresponding depths. The values of 1.75 to 1 at depth 10 can show that the guide of lexicon can alleviate the wrong inference of posterior topic distribution over the corpus, and increase the hits in the top ranking of retrieval list**.**

Table 5. The average hits at depths 10,50,100,300 and 500

	Top10	Top50	Top100	Top300	Top500
Run4	1	7.75	15.66	39.08	45.5
Run5	1.75	7.58	15.75	49.08	70.08

Finally, our experiments also show that adding the prior correlation information can speed up the convergence of EM estimation process. When constructing the first-level LDA model in Run4, we get the best estimative model parameters after 41 EM iterations, but only 18 is needed in run5. The improvement is more than one half.

5. Discussion

Selecting the appropriate number of topics for LDA model is an open research question. In generally we may think that a larger corpus needs more topics, but there is an interesting observation when we make comparisons between Run3 and Run5. We tried the different values of k with 50,100,200,300,400 and 500 for all the LDA models. The detail results as Fig2:

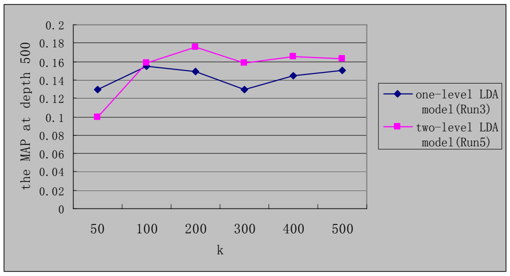

Fig. 2. The contrast between one-level LDA model and two-level LDA model on MAP with different number of topics(k)

In run3, the LDA model is constructed on the whole corpus with 20932 documents, and get the best performance when k=100. On the contrary, in the experiment of Run5, the three second-level models are constructed on the subsets with about 7000 documents, but get the best performance when k=200. After analyzing the posterior word distribution over topics in these LDA models, we found that the topics inferred from the whole corpus is abstract, and the document representation with these topics is relatively coarse. But the topics inferred from the subsets are more material and meaningful, it reflects a more embedded topical structure underlying the corpus, and can accurately represent the document. It shows that the problem of selecting k is not only correlative with the size of the corpus, but also correlative with the correlation between documents in the corpus.

6. Conclusion

In this paper, we present a lexicon-guided two-level LDA framework for semantic news video retrieval. The topic-based model is better fit the ASR texts with errors. Further more, through the guide of Hownet and the two level retrieval framework, we make good use of the LDA's abilities of mining knowledge, and well capture the embedded topical structure under the term representation. The experiments show that this retrieval framework is effective. In the future work, we'll extend LDA model to process the consecutive visual features data, and integrate texts features to better understand the semantic contents in the video.

References

[1] S. Deerwester, S. Dumais, G. Furnas, T. Lanouauer, and R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41 (1990), pp. 391-407.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.

[3] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, 50-57.

[4] B. Rosario, 2000. Latent Semantic Indexing: An overview [A]. INFOSYS 240 Spring 2000

[5] X. Wei, and W. Croft, LDA-based document models for ad-hoc retrieval, Proceedings of the 29th SIGIR Conference, 2006, pp. 178-185,

[6] L. Azzopardi, M. Girolami and K. van Rijsbergen. Topic Based Language Models for ad hoc Information Retrieval. In the Proceeding of IJCNN 2004 & FUZZ-IEEE 2004, July 25-29,2004, Budapest, Hungary.

[7] M. Girolami and A. Kaban. On an equivalence between plsi and lda. In 26th Annual ACM Conference on Research and Development in Information Retrieval, SIGIR, pages 433–434, Toronto, Canada, 2003.

[8] TS. Chua, SY. Neo, HK. Goh, M. Zhao, Y. Xiao, G. Wang "TRECVID 2005 by NUS PRIS" In TRECVID 2005, NIST, Gaithersburg, Maryland, USA, 14-15 Nov 2005.

[9] M. Girolami, and A. Kaban, On an equivalence between PLSI and LDA. In Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, 433-434.

[10] D. Blei and J. Lafferty (2006). Correlated topic models. InWeiss, Y., Schölkopf, B., and Platt, J., editors, Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA.

[11] Z. Dong and Q. Dong, HowNet, http://www.keenage.com/

[12] Q. Liu, SJ. Li, Computing word similarities based on HowNet． Computational Linguistics and Chinese Language Processing，2002，7(2)：59～76

[13] ST. Dumains Latent Semantic Indexing(LSI)and TREC-2[C]. In D. Harman, ed. The Second Text Retrieval Conference(TREC2),1994. 105～116