Statistical Chinese Spoken Document Retrieval Using Latent Topical Information

Berlin Chen[1], *Jen-Wei Kuo*[1], *Yao-Min Huang*[1], *Hsin-min Wang*[2]

[1] National Taiwan Normal University, Taiwan
[2] Institute of Information Science, Academia Sinica, Taiwan
{berlin, rogerkuo, angus}@csie.ntnu.edu.tw, whm@iis.sinica.edu.tw

Abstract

Information retrieval which aims to provide people with easy access to all kinds of information is now becoming more and more emphasized. However, most approaches to information retrieval are primarily based on literal term matching and operate in a deterministic manner. Thus their performance is often limited due to the problems of vocabulary mismatch and not able to be steadily improved through use. In order to overcome these drawbacks as well as to enhance the retrieval performance, in this paper we explore the use of topical mixture model for statistical Chinese spoken document retrieval. Various kinds of model structures and learning approaches were extensively investigated. In addition, the retrieval capabilities were verified by comparison with the conventional vector space model and latent semantic indexing model, as well as our previously presented HMM/N-gram retrieval model. The experiments were performed on the TDT-2 Chinese collection. Noticeable improvements in retrieval performance were obtained.

1. Introduction

Due to the advent of computer technology and the proliferation of Internet activity, tremendous volumes of multimedia information, such as text files, broadcast radio and television programs, digital archives and so on, are continuously growing and filling our computers and lives. Development of intelligent and efficient retrieval techniques to provide people with easy access to all kinds of information is now becoming more and more emphasized [1]. Meanwhile, with the rapid evolution of speech recognition technology, substantial efforts and very encouraging results on spoken document retrieval also have been reported in the last few years [2]-[4].

The conventional information retrieval approaches in principle can be characterized from two major perspectives: the matching strategy and the learning capability. There are two matching strategies frequently used to determine the degree of relevance for a document with respect to a query, namely, literal term matching and concept matching. The vector space model (VSM) [5] and probability-based model approaches are primarily based on literal term matching. VSM, which takes the vector representations of the query and documents, has been widely used because of its simplicity and satisfactory performance. The probability-based approach instead attempts to handle the retrieval problem within a statistical framework. The language modeling approach [6] and the hidden Markov model (HMM) approach [7] are good examples of it. Excellent survey articles of using the probability-based approach for information retrieval can also be found in [8]. However, these approaches often suffer from the problem of word usage diversity (or so-called vocabulary mismatch), which will degrade the retrieval performance severely as a given query and its relevant documents are using quite a different set of words. In contrast, concept matching is based on discovering the latent topical information embedded in the query and documents. The latent semantic indexing (LSI) model is one example [9]. LSI transforms the high-dimensional vector representations of the query and documents into a lower dimensional space (the so-called latent semantic space). Then the similarity measure can be estimated in the reduced space, where a query and a relevant document may have a high proximity value even if they do not share any words or terms in common. On the other hand, from the perspective of learning capability, it is well known that VSM and LSI are mainly based on linear algebra operations, and thus are much more deterministic and lack for a solid statistical foundation for automatic model refinement, while the probability-based approach, such as HMM, can be steadily improved by using a variety of machine learning algorithms in either supervised or unsupervised modes [10]. Based on these observations, in this paper we study the use of topical mixture model for statistical Chinese spoken document retrieval, which in essence belongs to the probability-based approach and has the virtue of being able to perform concept matching as well. Various kinds of model structures and learning approaches for the topical mixture model were extensively investigated. In addition, the retrieval capabilities were verified by comparison with the other retrieval models.

In the following, all the experiments were tested on the task involving the use of an entire Chinese newswire story (text) as a query, to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) in the document collection. Such a retrieval context is termed *query-by-example* [4].

2. Experimental corpus

We used the Topic Detection and Tracking collection (TDT-2) for this work. The Chinese news stories (text) from Xinhua News Agency were used as our queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. Table 1 describes the details for the corpus used in this paper. The Dragon large-vocabulary continuous speech recognizer [11] provided Chinese word transcriptions for our Mandarin audio collection, such that the results here may be compared with works done by other groups. We have spot-checked a fraction of the TDT-2 collection (of 39.90 hours) by comparing the Dragon recognition hypotheses with the manual transcriptions, and obtained error rates of 35.38% (word),

	TDT2 1998, 02~06		
No. of Spoken documents	2,265 stories, 46.03hrs of audio		
No. of Distinct text queries	16 Xinhua text stories (Topics 20001~20096)		
	Min.	Max.	Mean
Doc. Length (characters)	23	4841	287.1
Query length (characters)	183	2623	532.9
No. of relevant documents per query	2	95	29.3

Table 1: Statistics of the TDT-2 collection used in this paper.

17.69% (character) and 13.00% (syllable). Notice that Dragon's recognition output contains word boundaries (tokenizations) resulting from its language models and vocabulary definition while the manual transcriptions are running texts without word boundaries. Since Dragon's lexicon is not available, we augmented the LDC Mandarin Chinese Lexicon with the 24k words extracted from Dragon's word recognition output, and used the augmented LDC lexicon in tokenizing the manual transcriptions for computing error rates. We also used this augmented LDC lexicon in tokenizing the text query exemplars in the retrieval experiments.

3. Retrieval models

We will explain the structural characteristics of the topical mixture model studied in this paper and briefly review the other retrieval models.

3.1. Topical mixture model (TMM)

Given a query Q, a document D_i can be ranked according to the probability that D_i is relevant, conditioned on the fact that the query Q is observed; i.e., $P(D_i|Q)$, which can be rewritten as the following equation using Bayes theorem [12]:

$$P(D_i|Q) = \frac{P(Q|D_i)P(D_i)}{P(Q)}, \quad (1)$$

where $P(Q|D_i)$ is the probability of the query Q being generated by the document D_i, $P(D_i)$ is the prior probability of document D_i being relevant, and $P(Q)$ is the prior probability of query Q being posed. $P(Q)$ in Equation (1) can be eliminated because it has no influence on the final ranking. Furthermore, because there is no general way to estimate the probability $P(D_i)$, we can simply set it to unity for simplicity and approximate the probability $P(D_i|Q)$ by means of $P(Q|D_i)$. The query Q is treated as a sequence of input observations (terms or words), $Q = q_1 q_2 .. q_n .. q_N$, where the query terms are assumed to be conditionally independent given the document D_i. Therefore, the relevance measure $P(Q|D_i)$ can be decomposed as a product of the probabilities of the query terms generated by the document:

$$P(Q|D_i) = \prod_{n=1}^{N} P(q_n|D_i). \quad (2)$$

In this research, each individual document D_i is interpreted as a mixture model as shown in Figure 1, which is just a special case of HMM. In the model, a set of K latent topical

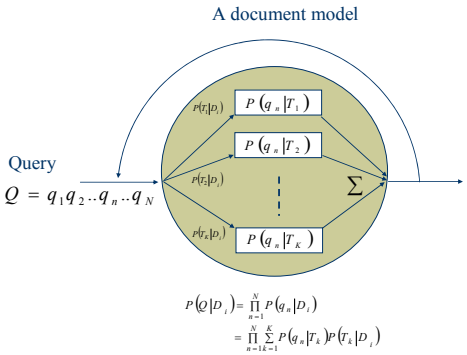

Figure 1: The topical mixture model for a specific document D_i.

distributions characterized with unigram language modeling are used to predict the query terms, and each of the latent topics is associated with a document-specific weight. The relevance measure therefore can be further written as:

$$P(Q|D_i) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(q_n|T_k) P(T_k|D_i), \quad (3)$$

where $P(q_n|T_i)$ denotes the probability of the query term q_n occurring in a specific latent topic T_k, and $P(T_k|D_i)$ is the posterior probability (or weight) of topic T_k conditioned on the document D_i, with the constraint $\sum_{k=1}^{K} P(T_k|D_i) = 1$ imposed. More precisely, the topical unigram distributions, e.g. $P(q_n|T_i)$, are tied among the entire document collection, while each document D_i has its own probability distribution over the latent topics, e.g. $P(T_k|D_i)$. The key idea we wish to illustrate here is that the relevance measure of a query term q_n and a document D_i is not computed directly based on the frequency of q_n occurring in D_i, but instead based on the frequency of q_n in the latent topic T_k as well as the likelihood that D_i generates the respective topic T_k, which in fact exhibits some sort of concept matching.

During training, the K-means algorithm [13] is first used to partition the entire document collection into K topical classes. Hence, the initial topical unigram distribution for a cluster topic can be estimated according to the underlying statistical characteristics of the documents being assigned to it, and the probabilities for each document generating the topics are measured according to its proximity to the centroid of each respective cluster as well. Then, the topical unigrams as well as the probabilities for each document generating the topics are optimized by employing the expectation-maximization (EM) algorithm [14]. Given a training set of query exemplars with the corresponding query-document relevance information, the document mixture models can be iteratively updated using the following three equations:

$$\hat{P}(q_n|T_k) = \frac{\sum_{Q \in [TrainSet]_Q} \sum_{D_i \in [Doc]_{R\,to\,Q}} n(q_n, Q) h(q_n, T_k, D_i)}{\sum_{Q \in [TrainSet]_Q} \sum_{D_i \in [Doc]_{R\,to\,Q}} \sum_{q_s \in Q} n(q_s, Q) h(q_s, T_k, D_i)}, \quad (4)$$

$$\hat{P}(T_k|D_i) = \frac{\sum_{\substack{Q \in [TrainSet]_Q \\ \text{st. } D_i \in [DOC]_{R\,to\,Q}}} \sum_{q_s \in Q} n(q_s, Q) h(q_s, T_k, D_i)}{\sum_{\substack{Q \in [TrainSet]_Q \\ \text{st. } D_i \in [DOC]_{R\,to\,Q}}} |Q|}, \quad (5)$$

$$h(q_n, T_k, D_i) = \frac{P(T_k|D_i)P(q_n|T_k)}{\sum_{l=1}^{K} P(T_l|D_i)P(q_n|T_l)}, \quad (6)$$

where $[TrainSet]_Q$ is the set of training query exemplars, $[Doc]_{R\ to\ Q}$ is the set of documents that are relevant to a specific training query exemplar Q, $n(q_n, Q)$ is the number of times a query term q_n occurring in the query exemplar Q, $|Q|$ is the length of query Q, and $h(q_n, T_k, D_i)$ is the expectation of the latent topic T_k generating query term q_n conditioned on the document D_i.

Structures similar to that shown in Figure 1 have also been investigated in machine learning literature [15]-[16]. The main differences between the proposed model and the previous ones are that we explicitly interpret the document as a mixture model used to predict the query, and both supervised and unsupervised learning approaches are extensively studied.

3.2. Vector space model (VSM)

For the vector space model, every document D_i and query Q is represented as a feature vector. Each component in the vector, $g(t)$, is associated with the statistics of a specific indexing term t,

$$g(t) = (1 + \ln(c(t)))\ln(N / N_t), \quad (7)$$

where $c(t)$ is the occurrence count of the term *t* within the document D_i or query Q, and $\ln(N/N_t)$ is the inverse document frequency (IDF). The popular cosine measure is used to estimate the query-document relevance [5].

3.3. Latent semantic indexing (LSI)

The latent semantic indexing model starts with a term-document matrix, and singular value decomposition (SVD) is applied to reduce the dimension and construct the latent semantic space, in which the original documents and indexing terms are properly represented, and queries or documents which are not part of the original matrix can be folded-in by matrix multiplication. The postulation is that indexing terms which occur in similar context will be near each other in the latent semantic space even if they never co-occur in the same document. The degree of relevance between a query and a document is then estimated by computing the cosine measure in the latent semantic space [9].

3.4. HMM/N-gram model (HMM)

In our previous implementation of the HMM/N-gram retrieval model, each document is represented as a single state discrete HMM composed of weighted N-gram distributions [10]. The N-gram distributions are estimated based on the frequency of words (for unigram modeling) or word pairs (for bigram modeling) occurring in the document and are smoothed using linear interpolation with background unigram or bigram language models estimated from a large outside text corpus. For example, the relevance measure for a HMM/Bigram retrieval model can be expressed as:

$$P(Q|D) = [m_1 P(q_1|D) + m_2 P(q_1|Corpus)] \times \prod_{n=2}^{N} [m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D) + m_4 P(q_n|q_{n-1}, Corpus)]. \quad (8)$$

The weighting parameters, m_i, are summed to 1 and tied among the entire document collection. They are optimized using the EM algorithm with the constraint $\sum_{i=1}^{4} m_i = 1$.

4. Experimental results

4.1. Experimental setup

The underlying probability distributions of the document mixture models were estimated by the EM updating formulas depicted in Equations (4)-(6) using an outside training query set consisting of 819 query exemplars with the corresponding query-document relevance information. The test results with manual transcriptions of the spoken documents (denoted as TD in the result tables below) are also shown for reference, compared to the results with the erroneous transcriptions obtained from speech recognition (denoted as SD below). The retrieval results are expressed in terms of *mean* average precision (*m*AP) [17].

In this study, when the TMM is employed in evaluating the relevance between a query and a document, we additionally incorporate the unigram probability of a query term occurring in the document into Equation (3) for probability smoothing:

$$\hat{P}(Q|D_i) = \prod_{n=1}^{N} \left[\alpha \left(\sum_{k=1}^{K} P(q_n|T_k)P(T_k|D_i) \right) + (1-\alpha) P_{ML}(q_n|D_i) \right], \quad (9)$$

where $P_{ML}(q_n|D_i)$ is the unigram probability of query term q_n occurring in the document D_i, and α is a weighting parameter whose value also can be optimized using the EM algorithm.

4.2. Experiments on supervised training

We first evaluate the retrieval performance of the topical mixture models by varying the model complexities. The model parameters were trained in a supervised manner using the 819 training query exemplars with their corresponding query-document relevance information. The retrieval results are shown in Table 2, where each column illustrates the retrieval results for both the TD and SD cases by using a different number of latent topics for document modeling. As can be seen, the retrieval performance is steadily improved as the mixture number increases. The best retrieval result of 0.7090 is obtained for the TD case when the document topic number is set to 32, while the best result is 0.6564 for the SD case with 64 topic mixtures. Notice that although the word error rate for the document collection is higher than 35%, the performance for the SD cases is only slightly lower than that for the TD cases. Such an observation is quite in parallel with those reported by other groups [18]-[19].

4.3. Experiments on unsupervised training

In most real-world applications, it is not always the case that the retrieval systems can have query exemplars correctly labeled with the relevance information to be used for model training. Thus, in this paper the unsupervised model training for TMM is also exploited. We use each individual document in the collection as a query exemplar to train its own mixture model, by using Equations (4)-(6) with some small modifications. The retrieval results are shown in Table 3. As can be seen, the results are not always improved as the topic

Topic No.	2	4	8	16	32	64
TD	0.6371	0.6852	0.6894	0.6934	0.7090	0.7081
SD	0.5773	0.5986	0.6068	0.6230	0.6481	0.6564

Table 2: Retrieval results achieved using the topical mixture models trained with a training set of query exemplars in a supervised mode.

Topic No.	2	4	8	16	32	64
TD	0.6271	0.6301	0.6320	0.6135	0.6130	0.6306
SD	0.5586	0.5703	0.5707	0.5536	0.5643	0.5714

Table 3: Retrieval results achieved using the topical mixture models trained by using the documents themselves as the query exemplars in an unsupervised mode.

Retrieval Model	HMM /Unigram	HMM /Bigram	VSM	LSI
TD	0.6327	0.5427	0.5548	0.5663
SD	0.5658	0.4803	0.5122	0.5362

Table 4: Retrieval results of the HMM/N-gram-based model (HMM), vector space model (VSM) and latent semantic indexing model (LSI).

number is increased. The best result of 0.6320 for the TD case is obtained when the document topic number is set to 8, while the best result is 0.5714 for the SD case when the document topic number is 64. As compared with the best results achieved in supervised training, there are at most about 0.08 and 0.09 absolute deceases in precision, respectively, for the TD and SD cases.

4.4. Comparison with other retrieval models

Finally, we compare TMM with the three popular retrieval models mentioned previously. The retrieval results, as the VSM, LSI and HMM models are respectively applied in the same retrieval task, are listed in Table 4 for comparison. VSM and LSI are implemented with the best parameter settings; while for HMM [10], both the unigram and bigram modeling strategies are used, and the corresponding model parameters are trained with the same 819 query exemplar set in a supervised manner. As compared with the results shown in Tables 2 and 3, it can be observed that TMM significantly outperforms all the other retrieval models when supervised training is adopted. Even though TMM is trained in an unsupervised manner, its retrieval performance is still far better than that of VSM and LSI, and achieves quite competitive results to that of the HMM trained in a supervised manner. It is interesting that the retrieval performance of HMM degraded as the model structure became more sophisticated (e.g., from unigram to bigram modeling), whereas the retrieval performance of TMM almost performed better as the topic number increased, when both models were trained in a supervised manner.

5. Concluding remarks

In this paper we have presented a framework for using the topical mixture model for statistical Chinese spoken document retrieval. We have extensively tested such a retrieval model by varying its model complexities and by using both the supervised and unsupervised training approaches. Besides, the retrieval capabilities of the topical mixture model have been verified by comparison with the other retrieval models. Very encouraging retrieval performance was obtained.

6. References

[1] Text REtrieval Conference (TREC), http://trec.nist.gov/.

[2] B. Chen, H. M. Wang, and L. S. Lee., "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", *IEEE Trans. on Speech and Audio Processing*, July 2002.

[3] E. Chang, F. Seide, H. Meng, Z. Chen, Y. Shi, and Y. C. Li, "A System for Spoken Query Information Retrieval on Mobile Devices," *IEEE Trans. on Speech and Audio Processing*, November 2002.

[4] H. Meng, B. Chen, S. Khudanpur, G. A. Levow, W. K. Lo, D. Oard, P. Schone, K. Tang, H. M. Wang, J. Wang, "Mandarin–English Information (MEI): Investigating Translingual Speech Retrieval," *Computer Speech and Language*, April 2004.

[5] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[6] J. M. Ponte, W. B. Croft, "A Language Modeling Approach to Information Retrieval", in *Proc. ACM SIGIR 1998*.

[7] D. Miller, T. Leek, and R. Schwartz, "A Hidden Markov Model Information Retrieval System", in *Proc. ACM SIGIR 1999*.

[8] W. B. Croft (editor) and J. Lafferty (editor), *Language Modeling for Information Retrieval*, Kluwer-Academic Publishers, 2003.

[9] S. Deerwester, S. T. Dumais, and R. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, Vol. 41, 1990.

[10] B. Chen, H. M. Wang, and L. S. Lee., "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," conditionally accepted by *ACM Transactions on Asian Language Information Processing*, February 2004.

[11] P. Zhan, S. Wegmann, and L. Gillick, "Dragon Systems' 1998 Broadcast News Transcription System for Mandarin," in *Proc. of the DARPA Broadcast News Workshop*, 1999.

[12] F. Jelinek, *Statistical Methods for Speech Recognition*, the MIT Press 1997.

[13] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, the MIT Press, 1999.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Society B, Vol. 39, 1977.

[15] T. Hoffmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, Vol. 42, 2001.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, 2003.

[17] D. Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)," 1995.

[18] S. Srinivasan and D. Petkovic, "Phonetic Confusion Matrix Based Spoken Document Retrieval," in *Proc. ACM SIGIR 2000*.

[19] M. Federico, "A System for the Retrieval of Italian Broadcast News," *Speech Communication*, Vol. 32, 2000.