Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 9–18

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Exploring the use of latent topical information for statistical Chinese spoken document retrieval

Berlin Chen *

Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, No. 88, Section 4, Ting-Chow Road, Taipei 116, Taiwan, ROC

Received 1 October 2004; received in revised form 14 February 2005
Available online 29 August 2005

Communicated by E. Backer

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 probabilistic latent semantic analysis model, 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 Chinese collections (TDT-2 and TDT-3). Noticeable improvements in retrieval performance were obtained.

Keywords: Information retrieval; Topical mixture model; Probabilistic latent semantic analysis model; Vector space model; Latent semantic indexing model; HMM/N-gram retrieval model

1. Introduction

Due to the advent of computer technology and the proliferation of Internet activity, tremendous volumes of multimedia information, such as text files, Web pages, 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 (Voorhees and Harman, 2000). It is also obvious that speech is the primary and most convenient means of communication between people, as well as the most rich source of information for the great volumes of multimedia. Therefore, with the rapid evolution of speech recognition technology, substantial efforts and very encouraging results on recognition and retrieval of spoken documents have been reported in the last few years (Woodland, 2002; Gauvain et al., 2002; Beyerlein et al., 2002; Chen et al., 2002; Chang et al., 2002; Meng et al., 2004; Byrne et al., 2004).

The conventional information retrieval (IR) 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) 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

* Tel.: +886 2 29322411x203; fax: +886 2 29322378.
E-mail address: berlin@csie.ntnu.edu.tw
URL: http://berlin.csie.ntnu.edu.tw

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.06.010

of its simplicity and satisfactory performance (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999). The probability-based approach instead attempts to handle the retrieval problem within a probabilistic framework. The language model (Ponte and Croft, 1998; Song and Croft, 1999; Zhai and Lafferty, 2001) and the hidden Markov model (HMM) (Miller et al., 1999; Lo et al., 2003; Chen et al., 2004a) are good examples of it, and research at a number of sites has confirmed that such a modeling approach does provide a potentially effective and theoretically attractive probabilistic framework for studying information retrieval problems. Excellent survey articles on the use of the probability-based approach for information retrieval can also be found (Croft and Lafferty, 2003; Liu and Croft, 2005). However, most of these approaches often suffer from the problem of word usage diversity (or so-called vocabulary mismatch), which will make the retrieval performance degrade 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, and latent semantic indexing (LSI) model is one example. 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 document may have a high proximity value even if they do not share any words or terms in common (Furnas et al., 1988; Deerwester et al., 1990). On the other hand, from the perspective of learning capability, it is well known that VSM and LSI are based on linear algebra operations and can incorporate a wide range of term weighting schemes as well as query or document expansion formulae (Salton and Buckley, 1988; Sparck Jones et al., 1998; Singhal and Pereira, 1999; Mandala et al., 2000) to modify the representations of query or documents, or to improve the information retrieval performance. While the probability-based approach, such as HMM, follows solid statistical foundations for automatic model refinement or optimization (Makhoul et al., 2000; Liu and Croft, 2005; Chen et al., 2004a), and thus can be steadily improved by using a variety of machine learning algorithms in either supervised or unsupervised modes.

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 virtue of being able to perform concept matching as well. Various kinds of model complexities for the topical mixture model were extensively investigated. In addition, their retrieval capabilities were verified by comparison with the other retrieval models. Structures similar to the presented approach also have been investigated in the machine learning literature recently (Hofmann, 2001; Blei et al., 2003; Wang et al., 2005). There are several differences between the presented approach and the previous ones. First, we explicitly interpret the document as a mixture model used to predict the query, which can be easily related to the conventional HMM modeling approaches that have been widely studied in speech and language processing community, and quite a few of theoretically attractive model training algorithms or optimization criteria can be therefore applied (Chou and Juang, 2003, Chapter 1). Moreover, we measure the relevance between the query and documents directly under the likelihood criterion (or in the likelihood space), unlike the previous approach (Hofmann, 2001), which evaluates the relevance between the query and documents in the low-dimensional factor (topic) space and only reports results by linearly combining with the cosine measure score obtained by using the VSM retrieval model. Finally, in this paper, both the supervised and unsupervised model learning approaches are extensively studied, while in the previous work (Hofmann, 2001; Blei et al., 2003; Wang et al., 2005), only unsupervised learning was investigated. We find that the results obtained based on the supervised learning approach are much better than those based on the unsupervised one.

In this paper, all the experiments were performed on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). The TDT corpora have been used for cross-language spoken document retrieval (CL-SDR) in the Mandarin English Information (MEI) Project (Meng et al., 2004), which is an NSF sponsored project conducted at the Johns Hopkins University Summer Workshop 2000. Project MEI investigated the use of an entire English newswire story (text) as a query to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) in the document collection. In this paper, we study the monolingual spoken document retrieval task instead. 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) from the document collection. Such a retrieval context is termed query-by-example. This technique can help users to find the corresponding video or audio news reports, which could be more attractive and informative when they see a newswire text report. Most of the prior work on Chinese spoken document retrieval is focused on retrieving spoken documents by short queries (Wang, 2000; Bai et al., 2001; Chang et al., 2002).

The rest of this paper is organized as follows. The experimental corpus is introduced in Section 2. In Section 3, we explain the structural characteristics of the topical mixture model and briefly review the other retrieval models. Then, the experimental settings and a series of information retrieval experiments are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Experimental corpus

We used two Topic Detection and Tracking (TDT) collections (LDC, 2000) for this work. TDT is a DARPA-sponsored program where participating sites tackle tasks such as identifying the first time a news story is reported on a given topic; or grouping news stories with similar topics from audio and textual streams of newswire data. Both the English and Mandarin Chinese corpora have been studied in the recent past. The TDT corpora have also been used for cross-language spoken document retrieval (CL-SDR) in the Mandarin English Information (MEI) Project (Meng et al., 2004). In this paper, we use the Mandarin Chinese collections of the TDT corpora for the retrospective retrieval task, such that the statistics for the entire document collection is obtainable. The Chinese news stories (text) from Xinhua News Agency are used as our queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts are used as the spoken documents. All news stories are exhaustively tagged with event-based topic labels, which are merely served as the relevance judgments for performance evaluation and will not be utilized in the training of information retrieval models (see Section 3). Table 1 describes the details for the corpora used in this paper. The TDT-2 collection is taken as the development set, which forms the basis for tuning parameters or settings. The TDT-3 collection is taken as the evaluation set; i.e., all the experiments performed on it were conducted following the same training (or parameter) settings and model complexities optimized based on the TDT-2 development set. Therefore, the experimental results can validate the effectiveness of the proposed approaches on comparable real-world data.

The Dragon large-vocabulary continuous speech recognizer (Zhan et al., 1999) provided Chinese word transcriptions for our Mandarin audio collections (TDT-2 and TDT-3), such that the results reported in this paper may be compared with work done by other groups. To assess the performance level of the recognizer, we spot-checked a fraction of the TDT-2 development set (about 39.90 h) by comparing the Dragon recognition hypotheses with the manual transcriptions, and obtained error rates of 35.38% (word), 17.69% (character) and 13.00% (syllable). Spot-checking approximately 76 h of the TDT-3 test set gave error rates of 36.97% (word), 19.78% (character) and 15.06% (syllable). Notice that Dragon's recognition output contains word boundaries (tokenizations) resulting from its own 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 24k words extracted from Dragon's word recognition output, and used the augmented LDC lexicon (about 51k words) to tokenize the manual transcriptions in an automatic way for computing error rates. We also used this augmented LDC lexicon to automatically tokenize the text query exemplars in the following retrieval experiments.

3. Retrieval models

We will concentrate on elucidating 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 transformed to the following equation by applying the Bayes theorem:

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

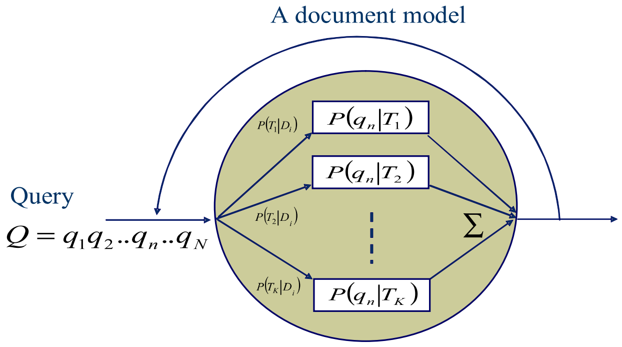
where $P(Q|D_i)$ is the probability of the query Q being generated by the document D_i, $P(D_i)$ is the prior probability of document D_i being relevant, and $P(Q)$ is the prior probability of query Q being posed. $P(Q)$ in Eq. (1) can be eliminated because it is identical for all documents and will not affect the ranking of the documents. Furthermore, because the way to estimate the probability $P(D_i)$ is still unknown, we may simply assume that $P(D_i)$ is uniformly distributed, or identical for all documents (Miller et al., 1999; Liu and Croft, 2005). In this way we can approximate the probability $P(D_i|Q)$ by means of the probability $P(Q|D_i)$ for the problem studied here. The query Q is treated as a sequence of input observations (terms or words), $Q = q_1q_2 \cdots q_n \cdots q_N$, 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 individual query terms generated by the document

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

In this research, each document D_i is interpreted as a mixture model as shown in Fig. 1, which is just a special case of

Table 1
Statistics of TDT-2 and TDT-3 collections used in this paper

	TDT-2 (development set) 1998, 02–06			TDT-3 (evaluation set) 1998, 10–12		
# Spoken documents	2265 stories, 46.03 h of audio			3371 stories, 98.43 h of audio		
# Distinct text queries	16 Xinhua text stories (topics 20,001–20,096)			47 Xinhua text stories (topics 30,001–30,060)		
	Min.	Max.	Mean	Min.	Max.	Mean
Doc. length (characters)	23	4841	287.1	19	3667	415.1
Query length (characters)	183	2623	532.9	98	1477	443.6
# Relevant documents per query	2	95	29.3	3	89	20.1



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

Fig. 1. The topical mixture model for a specific document D_i.

HMM. In the model, a set of K latent topical distributions characterized by unigram language modeling are used to predict the input query terms, and each of the latent topics is associated with a document-specific weight. That is, each document can belong to many topics. The relevance measure therefore can be further written as

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

where $P(q_n|T_k)$ denotes the probability with respect to 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 specifically, the topical unigram distributions, e.g., $P(q_n|T_k)$, 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. Thus, a query and a document may have a high relevance (likelihood) score even if they do not share any words or terms in common, which in fact exhibits some sort of concept matching.

During training, the K-means algorithm (Ball and Hall, 1967; Duda and Hart, 1973, p. 250) 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 C_k of each respective cluster k as well, which can be computed based on the cosine measures, $R(\vec{D}_i, \vec{C}_k)$, of the vector representations (see Section 3.2) of the document and cluster centroids and then can be transformed into a probability measure by the following equation:

$$p(T_k|D_i) = \frac{R(\vec{D}_i, \vec{C}_k)}{\sum_{r=1}^{K} R(\vec{D}_i, \vec{C}_r)}. \tag{4}$$

Moreover, the topical unigrams as well as the probabilities for each document generating the topics are further optimized by employing the expectation-maximization (EM) algorithm (Dempster et al., 1977). 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 (Manning and Schutze, 1999, p. 523):

$$\widehat{P}(q_n|T_k) = \frac{\sum_{Q \in [TrainSet]_Q} \sum_{D_i \in [Doc]_{R\,\mathrm{to}\,Q}} n(q_n, Q) P(T_k|q_n, D_i)}{\sum_{Q \in [TrainSet]_Q} \sum_{D_i \in [Doc]_{R\,\mathrm{to}\,Q}} \sum_{q_s \in Q} n(q_s, Q) P(T_k|q_n, D_i)}, \tag{5}$$

$$\widehat{P}(T_k|D_i) = \frac{\sum_{\substack{Q \in [TrainSet]_Q \\ \mathrm{st.}\, D_i \in [DOC]_{R\,\mathrm{to}\,Q}}} \sum_{q_s \in Q} n(q_s, Q) P(T_k|q_s, D_i)}{\sum_{\substack{Q \in [TrainSet]_Q \\ \mathrm{st.}\, D_i \in [DOC]_{R\,\mathrm{to}\,Q}}} |Q|}, \tag{6}$$

$$P(T_k|q_n, D_i) = \frac{P(T_k|D_i) P(q_n|T_k)}{\sum_{l=1}^{K} P(T_l|D_i) P(q_n|T_l)}, \tag{7}$$

where $[TrainSet]_Q$ is the set of training query exemplars, $[Doc]_{R \text{ to } Q}$ 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, $P(T_k|q_n, D_i)$ is the probability that the latent topic T_k occurs given the query term q_n and the document D_i, and $Q \in [TrainSet]_Q$ st. $D_i \in [\mathrm{DOC}]_{R \text{ to } Q}$ in Eq. (6) means that the query exemplar Q in the training query set can satisfy the condition that the document D_i in the collection is relevant to it. Notice here that the empirical frequency (or distribution) of words occurring in a specific document is not involved in the computation of the above equations (Eqs. (5)–(7)).

Structures similar to the presented approach also have been investigated in the machine learning literature recently (Hofmann, 2001; Blei et al., 2003; Wang et al., 2005). There are several differences between the presented approach and the previous ones. First, we explicitly interpret the document as a mixture model used to predict the query, which can be easily related to the conventional HMM modeling approaches that have been widely studied in speech and language processing community, and thus quite a few of theoretically attractive model training algorithms or optimization criteria can be applied (Chou and Juang, 2003, Chapter 1). For example, we can further employ the Minimum Classification Error (MCE) training algorithm to correctly discriminate the query observations for the best retrieval results (Chen et al., 2004a) rather than just to fit the distributions of the query observations, as done in EM training. Moreover, in this paper, we measure the relevance between the query and documents directly under the likelihood criterion (in the likelihood space), i.e., the likelihood a query is generated from a document that can satisfy the user's information need (Ponte and Croft, 1998), unlike the previous work, for example, the probabilistic latent semantic analysis (PLSA) retrieval model (Hofmann, 2001), which evaluates the relevance between the document and query in the low-dimensional factor (topic) space, and only reports results by linearly combining with the cosine measure score obtained by using the VSM retrieval model (see Section 4.4). Finally, in this paper, both the supervised and unsupervised model learning approaches are extensively studied, while only unsupervised learning was investigated in the previous work (Hofmann, 2001; Blei et al., 2003; Wang et al., 2005).

3.2. Vector space model (VSM)

For the vector space model, every document D_i is represented as a feature vector $\vec{D}_i$. Each component in the vector, $g(t)$, is associated with the statistics of a specific indexing term t, both within the document D_i and across all the documents in the collection,

$$g(t) = (1 + \ln(c(t)))\ln(N/N_t), \tag{8}$$

where $c(t)$ denotes the occurrence count of the term t within document D_i, and the logarithmic operation is to compress its distribution. The term weighting scheme, $1 + \ln(c(t))$, which is a variation to the conventional schemes, is used to measure the intra-document weight for the term t and its performance has been extensively studied previously (Singhal et al., 1998; Chen et al., 2002). The value of $\ln(N/N_t)$ is the inverse document frequency (IDF), where N_t is the total number of documents in the collection in which the specific indexing term t appears, and N is the total number of documents in the collection. IDF is to measure the inter-document discriminativity for the term t, reflecting that indexing terms appearing in many documents are less useful in identifying the relevant documents. A query Q is also represented by a vector $\vec{Q}$ constructed in the same way. The cosine measure is then used to estimate the query-document relevance

$$R_{\text{VSM}}(\vec{Q}, \vec{D}_i) = \left(\vec{Q} \cdot \vec{D}_i\right) \Big/ \left(\|\vec{Q}\| \cdot \|\vec{D}_i\|\right), \tag{9}$$

which apparently matches Q and D_i based on the terms literally. This model has been widely used because of its simplicity and satisfactory performance (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999).

3.3. HMM/N-gram model (HMM)

In our previous implementation of the HMM/N-gram retrieval model, each document is composed of a mixture of N-gram distributions (Miller et al., 1999; Liu and Croft, 2005), which also can be regarded as a special case of HMM and in essence has the hidden variables, i.e., the corresponding mixture sequence of N-gram components that generates the given observation sequence (the query) cannot be explicitly observed (or is non-deterministic). 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 (Chen and Goodman, 1999; Zhai and Lafferty, 2001; Bellegarda, 2004). This can also be viewed as a combination of information from a local source; i.e., the document, and a global source; i.e., the large text corpus (Liu and Croft, 2005). For example, the relevance measure for an HMM/Bigram retrieval model can be expressed as

$$\begin{aligned} P(Q|D_i) = {} & [m_1 P(q_1|D_i) + m_2 P(q_1|Corpus)] \\ & \times \prod_{n=2}^{N} [m_1 P(q_n|D_i) + m_2 P(q_n|Corpus) \\ & + m_3 P(q_n|q_{n-1}, D_i) + m_4 P(q_n|q_{n-1}, Corpus)], \end{aligned} \tag{10}$$

which again matches Q and D_i based on the terms literally. For training purpose, a general text corpus consisting of 40 million Chinese characters, which were mainly newswire texts collected from the Internet during January to June 2000, is used to estimate the background unigram and bigram language models (Chen et al., 2004a). The weighting parameters, m_i, are summed to 1 and tied among the entire document collection. They are optimized using the EM algorithm as well, given a training set of query exemplars with the corresponding query-document relevance information.

3.4. Latent semantic indexing (LSI)

The latent semantic indexing model is based on the concept matching strategy, which also tries to discover the latent topical information inherent in the query and documents. LSI starts with a term-document matrix W, describing the intra- and inter-document statistical relationships between all the terms and all the documents in the collection, and singular value decomposition (SVD) is performed on the matrix W 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 (Furnas et al., 1988; Deerwester et al., 1990; Wang et al., 2002).

4. Experimental results

4.1. Experimental Setup

The underlying probability distributions of the document mixture models were estimated by the EM updating formulae depicted in Eqs. (5)–(7) using an outside training query set consisting of 819 query exemplars with the corresponding query-document relevance information to the TDT-2 development set. The test results assuming manual transcriptions for the spoken documents to be retrieved are known (denoted as TD, text documents, in the result tables, Tables 2–6) are also shown for reference, compared to the results when only the erroneous transcriptions by speech recognition are available (denoted as SD, spoken documents, below). The retrieval results are expressed in terms of non-interpolated *mean* average precision (*m*AP) following the TREC evaluation (Harman, 1995), which is computed by the following equation:

$$m\mathrm{AP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{j}{r_{i,j}}, \tag{11}$$

where L is the number of testing queries, M_i is the total number of documents that are relevant to query Q_i, and $r_{i,j}$ is the position (rank) of the jth document that is relevant to query Q_i, counting down from the top of the ranked list.

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 outside text corpus, as mentioned earlier in Section 3.4, into Eq. (3) for probability smoothing

$$\widehat{P}(Q|D_i) = \prod_{n=1}^{N}\left[\alpha\left(\sum_{k=1}^{K}P(q_n|T_k)P(T_k|D_i)\right) + (1-\alpha)P(q_n|Corpus)\right], \tag{12}$$

where $P(q_n|Corpus)$ is the unigram probability of query term q_n occurring in the outside text corpus, 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 (denoted as TMM-S) trained with supervision and varying model complexities on the TDT-2 development set. The model parameters were trained using the 819 training query exemplars with their corresponding query-document relevance information to the TDT-2 development set. The retrieval results are shown in the upper part (TMM-S) of Table 2, where each column illustrates the retrieval results in both the TD and SD cases by using different numbers of latent topics for document modeling. As can be seen, the retrieval performance is steadily improved as the topic number increases. The best retrieval result of 0.7794 is obtained for the TD case when the document topic number is set to 256, while the best result is 0.6652 for the SD case with 128 topic mixtures. Notice that although the word error rate (WER) for the spoken document collection is higher than 35%, however, the average degradation in retrieval performance is much smaller. Such an observation indicates that the WER does not cause much adverse effect on retrieval performance, which is quite in parallel with those reported by other groups (Renals et al., 2000; Srinivasan and Petkovic, 2000; Federico, 2000). One possible reason is that, a query word or a keyword might occur repeatedly (more than once) within a broadcast news story and it is not always the case that all the occurrences of the word would be misrecognized totally as other words. For example, a word spoken by the studio anchor might have higher recognition accuracy than the same word spoken by the field reporter or the interviewee, which is mainly because for the anchor speech, the corresponding bandwidth variability, recording environment and speaking style, as well as the amount of acoustic training data, can be well controlled. Therefore, the true meaning of the word occurring within the spoken document could be still preserved for the following retrieval process.

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 query-document relevance information to be used for model training. Thus, in this paper, we study unsupervised model training for TMMs, and two kinds of unsupervised training approaches are preliminarily investigated. In the first approach (denoted as TMM-U1), we assume that there is no any query exemplar beyond the document collection. We instead use each individual document in the collection as a query exemplar to train its own mixture model, by using Eqs. (5)–(7) with some small modifications. For example, Eqs. (5) and (6) can be respectively modified as follows:

$$\widehat{P}(w_n|T_k) = \frac{\sum_{D_i\in[D]} n(w_n,D_i)P(T_k|w_n,D_i)}{\sum_{D_i\in[D]}\sum_{w_s\in D_i} n(w_s,D_i)P(T_k|w_n,D_i)}, \tag{13}$$

$$\widehat{P}(T_k|D_i) = \frac{\sum_{w_s\in D_i} n(w_s,D_i)P(T_k|w_s,D_i)}{|D_i|}, \tag{14}$$

Table 2
Retrieval results on the TDT-2 development set, achieved, respectively, with the TMM retrieval models (TMM-S) and the PLSA retrieval models (PLSA-S) trained in a supervised mode

Topic no.		2	4	8	16	32	64	128	256
TMM-S	TD	0.6362	0.6721	0.6750	0.6769	0.6823	0.6930	0.7243	0.7794
	SD	0.5759	0.5894	0.5918	0.5988	0.6255	0.6528	0.6652	0.6591
PLSA-S	TD	0.6012	0.6535	0.5679	0.5862	0.5898	0.5946	0.6132	0.6253
	SD	0.5493	0.5754	0.5510	0.5531	0.5399	0.5505	0.5626	0.5525

Table 3
Retrieval results on the TDT-2 development set, achieved, respectively, with the TMM retrieval models (TMM-U1 and TMM-U2) and the PLSA retrieval models (PLSA-U) trained in an unsupervised mode

Topic no.		2	4	8	16	32	64	128	256
TMM-U1	TD	0.6277	0.6332	0.6266	0.5973	0.5949	0.6267	0.6041	0.5878
	SD	0.5545	0.5659	0.5681	0.5503	0.5534	0.5664	0.5484	0.5831
TMM-U2	TD	0.6368	0.6458	0.6456	0.6498	0.6433	0.6555	0.5844	0.5872
	SD	0.5677	0.5696	0.5697	0.5799	0.5586	0.5699	0.5463	0.5459
PLSA-U	TD	0.5035	0.5190	0.5212	0.5331	0.5415	0.5565	0.5538	0.5538
	SD	0.4575	0.4930	0.4758	0.4674	0.4403	0.4939	0.5068	0.5104

where $[D]$ is the entire document collection, $n(w_n, D_i)$ is the number of times a term (or word) w_n occurring in the document D_i, $|D_i|$ is the length of document D_i. The retrieval results for the experiments carried out on the TDT-2 collection are shown in the upper part (TMM-U1) of Table 3. As it can be seen, the results are not always improved as the topic number increases. The best result of 0.6332 for the TD case is obtained when the document topic number is set to 4, while the best result of 0.5831 for the SD case when document topic number is 256. When comparing with the best results achieved in supervised training, there are at most about 0.15 (0.7794 vs. 0.6332) and 0.08 (0.6652 vs. 0.5831) decreases in precision, respectively, for the TD and SD cases. In the second approach (denoted as TMM-U2), with the TMMs trained according to the first approach (TMM-U1), we again exploit the 819 training query exemplars, but neglecting their respective query-document relevance information, to perform an initial retrieval. After the initial retrieval, for each query, a ranked list of documents according to the relevance between the query and the documents can be obtained. Then, the top M ranked documents of each query are assumed to be relevant to it and thus are selected for EM training, by using Eqs. (5)–(7). The middle part (TMM-U2) of Table 3 shows the results on the TDT-2 collection after such unsupervised training (with M being set to 3). As we compare the results with that of the TMM-U1, it can be found that the retrieval results are substantially improved as the topic mixture is smaller than or equal to 64, but are degraded as the topic mixture is larger than 64. It implies that the additional use of query exemplars for unsupervised training is helpful, but only for when the model complexity is lower. In summary, for the TMM retrieval model, given a training set of query exemplars with the corresponding query-document relevance information, the retrieval results obtained based on the supervised training approach are much better than those based on the unsupervised approach.

4.4. Comparisons with the PLSA retrieval model

In an attempt to access the performance level of the TMM retrieval model as well as the associated approaches for model training and relevance measure, we study here the use of the probabilistic latent semantic analysis (PLSA) retrieval model (Hofmann, 2001) for Chinese spoken document retrieval. The PLSA retrieval model adopts a graphical model representation and introduces a set of latent class variables to characterize the word-document co-occurrences. PLSA is thought to have a similar modeling structure as the TMM retrieval model proposed in this paper; however, it is instead trained in an purely unsupervised manner and utilizes the low-dimensional representations of the query and document in the factor (topic) space, e.g., $P(T_k|Q)$ and $P(T_k|D_i)$, to evaluate the query-document similarity

$$R_{\mathrm{PLSA}}(Q,D_i)=\frac{\sum_{k=1}^{K}P(T_k|Q)P(T_k|D_i)}{\sqrt{\sum_{k=1}^{K}P(T_k|Q)^2}\sqrt{\sum_{k=1}^{K}P(T_k|D_i)^2}},\qquad (15)$$

where $P(T_k|D_i)$ is the probability distribution of document D_i over the latent topics, which, as described previously in Section 4.3, can be trained beforehand in an unsupervised manner by using document D_i itself as the query exemplar and the EM training formula shown in Eq. (14); while $P(T_k|Q)$ is the probability distribution of query Q over the latent topics, which is not known at query time and should be trained (or folded-in) online by iteratively using Eq. (14) with some small modifications

$$P(T_k|Q)=\frac{\sum_{q_n\in Q}n(q_n,Q)P(T_k|q_n,Q)}{|Q|}.\qquad (16)$$

The final retrieval formula for PLSA is then expressed by linearly combining with the cosine measure score obtained from the VSM retrieval model (Hofmann, 2001)

$$\widetilde{R}_{\mathrm{PLSA}}(Q,D_i)=\lambda\cdot R_{\mathrm{PLSA}}(Q,D_i)+(1-\lambda)\cdot R_{\mathrm{VSM}}(\vec{Q},\vec{D}_i),\qquad (17)$$

where λ is a weighting parameter and is empirically set at an optimum value for each experimental condition in this study. The retrieval results of such a modeling approach (denoted as PLSA-U) on the TDT-2 collection are shown in the lower part (PLSA-U) of Table 3, where each column illustrates the retrieval results in both the TD and SD cases by using a different number of latent topics for PLSA modeling. The best retrieval result of 0.5565 is obtained for the TD case when the latent topic number is set to 64, while the best result is 0.5104 for the SD case with 256 topic mixtures. They are obviously worse than those achieved by the TMM trained either in supervised or unsupervised modes, as the results previously illustrated in the upper part (TMM-S) of Table 2 and in the upper (TMM-U1) and middle (TMM-U2) parts of Table 3. Besides, we also attempt to compute the probability distribution $P(T_k|D_i)$ by using the supervised training approach described in Section 4.2. The same set of 819 query exemplars and their corresponding query-document relevance information to the TDT-2 development set are employed here again to iteratively estimate $P(T_k|D_i)$ by using Eq. (6). The retrieval results of such an approach (denoted as PLSA-S) on the TDT-2 collection are shown in the lower part of Table 2, in which the best result of 0.6535 is obtained for the TD case when the document topic number is set to 4 and the best result of 0.5754 is obtained for the SD case with the same topic number. Such results are competitive to those obtained by using the TMM trained in an unsupervised manner (TMM-U1 and TMM-U2), but are still far worse than those obtained by using the TMM trained in a supervised manner (TMM-S). We can thus conclude that for the Chinese spoken document retrieval task studied here, the relevance measure performed under the likelihood criterion (in the likelihood space) seems to be more preferable than those done in the factor (topic) space.

4.5. Comparisons with other retrieval models

Moreover, we also compare TMM with the other three popular retrieval models mentioned previously. The retrieval results on the TDT-2 collection are listed in Table 4 for comparison, as the VSM, LSI and HMM models are respectively applied. VSM and LSI are implemented with the best parameter settings; while for HMM, both the unigram and bigram modeling strategies are used, and the corresponding models are trained with the same set of 819 query exemplars in a supervised manner. As compared with the results in Tables 2 and 3, it can be observed that TMM significantly outperforms all the other retrieval models when supervised learning is adopted (TMM-S). Even though TMM is trained in an unsupervised manner (TMM-U1 and TMM-U2), 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 degrades as the model structure becomes more sophisticated (e.g., from unigram to bigram modeling), whereas the retrieval performance of TMM almost becomes better as the topic number increased, when both models were trained in a supervised manner. Since the number of distinct words (51k) is large, the estimation of bigram probabilities for the HMM inherently suffers from the sparse data problem. The smoothing terms in Eq. (10) obtained from the 40M general text corpus seems not to work well, probably because the 40M general text corpus is still not large enough for word bigram training, and the word bigrams thus obtained even slightly disturbed the uni/bigrams obtained for each documents (Chen et al., 2004a).

4.6. Further evaluations on the TDT-3 collection

Finally, in order to validate the effectiveness of the proposed TMM retrieval model on comparable real-world data, we further conducted a series of corresponding information retrieval experiments on the TDT-3 evaluation set. The retrieval results achieved by using TMM and PLSA are shown in Table 5, while the results achieved by using the other retrieval models, such as HMM, VSM and LSI, are shown in Table 6. For TMM, PLSA and HMM, the training settings and/or model complexities for different experimental conditions (TD and SD cases) are set with the same configurations as those optimized using the TDT-2 collection; while for VSM and LSI, the model parameters are also set at the same optimum values tuned based on the TDT-2 collection as well. We first examine the TMM trained in a supervised manner (TMM-S). The retrieval results are shown in column 2 of Table 5, in which the document models were respectively trained by another outside training query set consisting of 731 query exem-

Table 4
Retrieval results on the TDT-2 development set, achieved with the HMM/N-gram-based model (HMM), vector space model (VSM) and latent semantic indexing model (LSI), respectively

Retrieval model	HMM/unigram	HMM/bigram	VSM	LSI
TD	0.6327	0.5427	0.5548	0.5510
SD	0.5658	0.4803	0.5122	0.5310

Table 6
Retrieval results on the TDT-3 evaluation set, achieved with the HMM/N-gram-based model (HMM), vector space model (VSM), and latent semantic indexing model (LSI), respectively

Retrieval model	HMM/unigram	HMM/bigram	VSM	LSI
TD	0.6569	0.6143	0.6505	0.6440
SD	0.6308	0.5808	0.6216	0.6390

Table 5
Retrieval results on the TDT-3 evaluation set, achieved with the topical mixture models (TMM) and the PLSA retrieval models (PLSA), which are trained in both supervised and unsupervised mode and with the best model complexities set by the TDT-2 development set, respectively

Retrieval model	Supervised training		Unsupervised training		
	TMM-S	PLSA-S	TMM-U1	TMM-U2	PLSA-U
TD	0.7870 (256 topics)	0.6882 (4 topics)	0.6585 (4 topics)	0.6728 (64 topics)	0.6471 (64 topics)
SD	0.7852 (128 topics)	0.6688 (4 topics)	0.6582 (256 topics)	0.6403 (16 topics)	0.5982 (256 topics)

plars together with the corresponding query-document relevance information to the TDT-3 evaluation set. A result of 0.7870 for the TD case is obtained with the document topic number set to 256, while a result of 0.7852 for the SD case with the document topic number set to 128. Comparatively speaking, these results are significantly better than the results achieved by the PLSA (as respectively shown in the third (PLSA-S) and rightmost (PLSA-U) columns of Table 5), as well as those obtained by the other retrieval models (as shown in Table 6). We then examine the retrieval performance of the TMM trained by the two unsupervised training approaches described in Section 4.3. As the results shown in columns 4 (TMM-U1) and 5 (TMM-U2) of Table 5, the TMMs respectively trained by these two unsupervised training approaches are slightly worse than the PLSA trained with supervision (PLSA-S), however, they still achieve better retrieval performance than the other models (including the PLSA trained without supervision (PLSA-U)) in most retrieval conditions, though the differences are not significant.

Based on the experimental results achieved from this and previous sections, it has been clearly demonstrated that the TMM trained in a supervised manner does achieve superior performance over the other retrieval models for the Chinese spoken document retrieval task studied here. All the experiments throughout this paper have been carefully designed to avoid "testing on training"; i.e., all the training (or parameter) settings and model complexities are tuned or optimized by using the TDT-2 development set and tested on both the TDT-2 development set and the TDT-3 evaluation set. Generally speaking, the training settings and model complexities tuned from the TDT-2 development set perform rather well in the TDT-3 evaluation set.

5. Concluding remarks

In this paper, we presented a framework for using the topical mixture model (TMM) 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. We found that given a set of training query exemplars, the retrieval performance could be significantly improved, which means that TMM can be steadily improved through use. Besides, the retrieval capabilities of TMM have been verified by comparison with the other retrieval models, and superior retrieval performance was evidenced. More in-deep investigation and analysis of the TMM retrieval approach as well as comparison to other sophisticated approaches are currently undertaken. Meanwhile, we also have applied such a modeling approach to the language model adaptation problem for Mandarin broadcast news transcription, with very good potential indicated (Chen et al., 2004b). A best result of 4.93% character error rate reduction and 39.17% perplexity reduction was initially obtained.

References

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley.

Bai, B.R., Chen, B., Wang, H.M., 2001. Syllable-based Chinese text/spoken document retrieval using text/speech queries. Internat. J. Pattern Recognition Artif. Intell. 14 (5), 603–616.

Ball, G.H., Hall, D.J., 1967. A clustering technique for summarizing multivariate data. Behav. Sci. 12, 153–155.

Bellegarda, J.R., 2004. Statistical language model adaptation: review and perspectives. Speech Comm. 42, 93–108.

Beyerlein, P., Aubert, X., Haeb-Umbach, R., Harris, M., Klakow, D., Wendemuth, A., Molau, S., Ney, H., Pitz, M., Sixtus, A., 2002. Large vocabulary continuous speech recognition of broadcast news—The Philips/RWTH approach. Speech Comm. 37, 109–131.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Machine Learning Res. 3, 993–1022.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.J., 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans. Speech Audio Process. 12 (4), 420–435.

Chang, E., Seide, F., Meng, H., Chen, Z., Shi, Y., Li, Y.C., 2002. A system for spoken query information retrieval on mobile devices. IEEE Trans. Speech Audio Process. 10 (5), 531–541.

Chen, B., Wang, H.M., Lee, L.S., 2002. Discriminating capabilities of syllable-based features and approaches of utilizing Them for voice retrieval of speech information in Mandarin Chinese. IEEE Trans. Speech Audio Process. 10 (5), 303–314.

Chen, B., Wang, H.M., Lee, L.S., 2004a. A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents. ACM Trans. Asian Language Inform. Process. 3 (2), 128–145.

Chen, B., Tsai, W.H., Kuo, J.W., 2004b. Statistical language model adaptation for Mandarin broadcast news transcription. In: Proc. Internat. Symp. on Chinese Spoken Language Processing, pp. 313–316.

Chen, S.F., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Comput. Speech Language 13, 359–394.

Chou, W., Juang, B.H. (Eds.), 2003. Pattern Recognition in Speech and Language Processing. CRC Press (Chapter 1).

Croft, W.B., Lafferty, J. (Eds.), 2003. Language Modeling for Information Retrieval. Kluwer-Academic Publishers.

Deerwester, S., Dumais, S.T., Harshman, R., 1990. Indexing by latent semantic analysis. J. Amer. Soc. Inform. Sci. 41 (6), 391–407.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39 (1), 1–38.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Wiley & Sons.

Federico, M., 2000. A system for the retrieval of Italian broadcast news. Speech Comm. 32, 37–47.

Furnas, G.W., Deerwester, S., Dumais S.T., Landauer T.K., Harshman, R., Streeter, L.A., Lochbaum, K.E., 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 465–480.

Gauvain, J.-L., Lamel, L., Adda, G., 2002. The LIMSI broadcast news transcription system. Speech Comm. 37, 89–108.

Harman, D. 1995. Overview of the fourth text retrieval conference (TREC-4). In: Proc. Fourth Text Retrieval Conf., pp. 1–23.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196.

LDC 2000. Project topic detection and tracking. Linguistic Data Consortium. Available from: <http://www.ldc.upenn.edu/Projects/TDT/>.

Liu, X., Croft, W.B., 2005. Statistical language modeling for information retrieval. In: Cronin, B. (Ed.), Annual Review of Information Science and Technology 39, Chapter 1. Information Today.

Lo, W.K., Meng, H., Ching, P.C., 2003. Cross-language spoken document retrieval using HMM-based retrieval model with multi-scale fusion. ACM Trans. Asian Language Inform. Process. 2 (1), 1–26.

Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., Srivastava, A., 2000. Speech and language technologies for audio indexing and retrieval. Proc. IEEE 88 (8), 1338–1353.

Mandala, R., Tokunaga, T., Tanaka, H., 2000. Query expansion using heterogeneous thesauri. Inform. Process. Manage. 36 (3), 361–378.

Manning, C., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

Meng, H., Chen, B., Khudanpur, S., Levow, G.A., Lo, W.K., Oard, D., Schone, P., Tang, K., Wang, H.M., Wang, J., 2004. Mandarin English information (MEI): Investigating translingual speech retrieval. Comput. Speech Language 18 (2), 163–179.

Miller, D.R.H., Leek, T., Schwartz, R., 1999. A hidden Markov model information retrieval system. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 214–221.

Ponte, J.M., Croft, W.B., 1998. A language modeling approach to information retrieval. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 275–281.

Renals, S., Abberley, D., Kirby, D., Robinson, T., 2000. Indexing and retrieval of broadcast news. Speech Comm. 32, 5–20.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic retrieval. Inform. Process. Manage. 24 (5), 512–523.

Salton, G., McGill, M.J., 1983. Introduction to Modern Information Retrieval. McGraw–Hill, New York.

Singhal, A., Pereira, F., 1999. Document expansion for speech retrieval. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 34–41.

Singhal, A., Choi, J., Hindle, D., Lewis, D., Pereira, F., 1998. AT&T at TREC-7. In: Proc. Seventh Text Retrieval Conf. (TREC-7), pp. 186–198.

Song, F., Croft, W.B., 1999. A general language model for information retrieval. In: Proc. ACM Conf. on Information and Knowledge Management, pp. 279–280.

Sparck Jones, K., Walker, S., Robertson, S.E., 1998. A probabilistic model of information and retrieval: Development and Status, University of Cambridge Computer Laboratory Computer Laboratory, Technical Report UCAM-CL-TR-446.

Srinivasan, S., Petkovic, D., 2000. Phonetic confusion matrix based spoken document retrieval. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 81–87.

Voorhees, E.M., Harman, D., 2000. Overview of the ninth text retrieval conference (TREC-9). In: Proc. Ninth Text Retrieval Conf., pp. 1–13.

Wang, C.J., Chen, B., Lee, L.S., 2002. Improved Chinese spoken document retrieval with hybrid modeling and data-driven indexing features. In: Proc. Internat. Conf. on Spoken Language Processing, pp. 1985–1988.

Wang, H.M., 2000. Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese. Speech Comm. 32, 49–60.

Wang, S., Schuurmans, D., Peng, F., Zhao, Y., 2005. Combining statistical language models via the latent maximum entropy principle. Machine Learning J. 59, 1–22.

Woodland, P.C., 2002. The Development of the HTK broadcast news transcription system: An overview. Speech Comm. 37, 47–67.

Zhai, C., Lafferty, J., 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proc. ACM SIGIR Conf. on R&D in Information Retrieval, pp. 334–342.

Zhan, P., Wegmann, S., Gillick, L., 1999. Dragon systems' 1998 broadcast news transcription system for Mandarin. Available from: <http://www.nist.gov/speech/publications/darpa99/pdf/sp350.pdf>.