IMPROVED SPOKEN DOCUMENT RETRIEVAL WITH DYNAMIC KEY TERM LEXICON AND PROBABILISTIC LATENT SEMANTIC ANALYSIS (PLSA)

Ya-chao Hsieh, Yu-tsun Huang, Chien-chih Wang and Lin-shan Lee

Graduate Institute of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
speech@speech.ee.ntu.edu.tw

ABSTRACT

Spoken document retrieval will be very important in the future network era. In this paper, we propose using a ”dynamic key term lexicon” automatically extracted from the ever-changing document archives as an extra feature set in the retrieval task. This lexicon is much more compact but semantically rich, thus it can retrieve relevant documents more efficiently. The key terms include named entities and others selected by a new metric referred to as the term entropy here derived from probabilistic latent semantic analysis (PLSA). Various configurations of retrieval models were tested with a broadcast news archive in Mandarin Chinese and significant performance improvements were obtained, especially with the new version of PLSA models based on a key term lexicon rather than the full lexicon.

1. INTRODUCTION

In the future network era, digital content over the network will include all information activities for human life. Apparently, the most attractive form of the network content will be in multi-media including speech information, and such speech information usually tells the subjects, topics and concepts of the multi-media content. As a result, the spoken documents associated with the network content will become the key for retrieval. In other words, the network content may be retrieved not only by its text, but possibly by the associated spoken document as well, and the user instructions may also be entered by spoken queries. [1]

The conventional Vector Space Model (VSM) for retrieving either text or spoken documents has been very successful, but this approach can only literally match the terms (or words) used by the user query and those by the documents directly; thus it very often suffers from the problem of word usage diversity (or vocabulary mismatch), i.e., very often the query and its relevant documents are using quite different sets of words. In order to be able to retrieve text/spoken documents semantically relevant to the query but not necessarily ”look like” or ”sound like” the user query, concept matching strategies were developed to discover latent topical information inherent in the query and documents, based on which the retrieval is performed; the Latent Semantic Indexing (LSI) model and the Probabilistic Latent Semantic Analysis (PLSA) are two typical examples [2] [3].

LSI starts with a ”term-document” matrix describing the relationships between all the terms and all the documents in the archives. Singular value decomposition (SVD) is then used to construct a latent semantic space in which the retrieval is actually performed. On the other hand, PLSA tries to establish the probabilistic framework for the above latent topical approach [3] by introducing a set of latent topic variables,$\{T_k, k = 1, 2, ...K\}$, to characterize the ”term-document” co-occurrence relationships as shown in Figure 1. A query Q is treated as a sequence of n observed terms, $Q = t_1t_2...t_j..t_n$, while the document d_i and a term t_j are both assumed to be independently conditioned on an associated latent topic T_k. The conditional probability of observing a term t_j in a document d_i thus can be parameterized by:

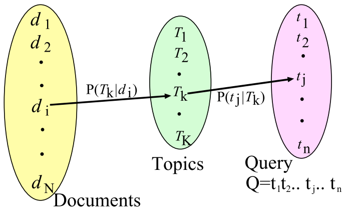

Fig. 1. Probabilistic Latent Semantic Analysis (PLSA) modeling

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j|T_k)P(T_k|d_i) \qquad (1)$$

When the terms in the query Q are further assumed to be independent given the document, the relevance score between the query and document can then be expressed as:

$$P(Q|d_i) = \prod_{j=1}^{n} \left[\sum_{k=1}^{K} P(t_j|T_k)P(T_k|d_i) \right] \qquad (2)$$

Notice that this relevance score is not obtained directly from the frequency of the respective term t_j occurring in d_i, but instead through the frequency of t_j in the latent topic T_k as well as the likelihood that d_i addresses the latent topic T_k. A query and a document thus may have a high relevance score even if they do not share any terms in common, which is therefore concept matching. The PLSA model can be optimized with the EM algorithm by maximizing a total likelihood function.

All the retrieval models mentioned above can be equally applied to text or spoken documents with text or spoken queries. The primary extra difficulties for the spoken documents and/or queries are the inevitable speech recognition errors including the problems with spontaneous speech such as pronunciation variation and disfluencies, and the out-of-vocabulary (OOV) problem for words outside the vocabulary of the speech recognizer. Extra approaches to handle these problems have been developed, good examples include developing more robust indexing terms for audio signals, using multiple recognition hypotheses obtained from N-best lists or word graphs [4].,

1-4244-0469-X/06/$20.00 ©2006 IEEE

using improved scoring methods based on different confidence measures [5][6], use of subword units or segments of them rather than words as the indexing terms [4][5][7], as well as various approaches of query and document expansion.

In this paper, we propose to use a ”dynamic key term lexicon” automatically extracted from the ever-changing dynamic network content as an additional feature set in the retrieval task. These key terms carry more semantic or topical information than other words, thus can retrieve the relevant documents more efficiently. Better approaches to utilize the PLSA modeling adequately in the retrieval task were also developed. Significantly improved retrieval performance was obtained with these proposed approaches in preliminary experiments. The rest of this paper is organized as follows. In section 2, we will briefly summarize the proposed approach, while the detailed approach for constructing the dynamic key term lexicon is presented in Sections 3, 4 and 5. Section 6 introduces the initial prototype system and the experimental conditions, and section 7 describes the experimental results. Section 8 is the conclusion.

2. PROPOSED APPROACH

The basic idea of using a dynamic lexicon of key terms is that the desired information in the user's mind can very often be described by one or a few key terms which carry significant semantic information. If these key terms can be identified in the query, they should be adequately exploited in the retrieval process. But the network content changes every day, so the key terms must be automatically extracted and the lexicon of key terms must be dynamic, so the proposed approach of using a dynamic key term lexicon and PLSA modeling is shown in Figure 2. This approach has been successfully implemented as a working system.

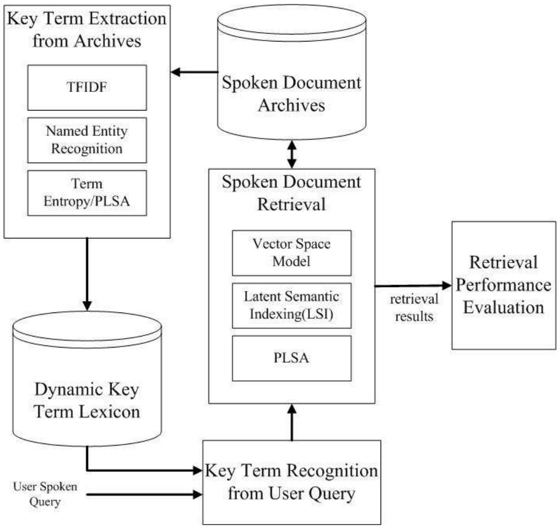
Key Term Extraction from Archives
TFIDF
Named Entity Recognition
Term Entropy/PLSA
Spoken Document Archives
Spoken Document Retrieval
Vector Space Model
Latent Semantic Indexing(LSI)
PLSA
retrieval results
Retrieval Performance Evaluation
Dynamic Key Term Lexicon
Key Term Recognition from User Query
User Spoken Query

Fig. 2. System Diagram for the Proposed Approach

As can be found in Figure 2, an important part of the proposed approach is the automatic key term extraction from the archives as shown on the top left of the figure. This can be achieved by at least three approaches: by TF/IDF scores, by named entities and by the term entropy based on PLSA modeling, as proposed here in this paper. The TF/IDF scores of the terms in the spoken document archives is not difficult to estimate, in which the term frequency (TF) of a term in the archives should be normalized by the number of documents which include the term, such that the situation of different key terms appearing in different number of documents can be equalized. The named entities used here include three types: person names, location names and organization names. Automatic recognition of such named entities from spoken documents will be summarized below. Term entropy based on PLSA modeling is a new metric for identifying key terms proposed in this paper, which will also be presented below. The key terms obtained by one or more approaches mentioned above are then used to construct the dynamic key term lexicon in the lower left part of the figure.

The second important part of the proposed approach is the key term recognition from the user spoken query as shown in the bottom of the figure. Here special approaches to recognize correctly the key terms in the user query were developed, including emphasizing the possible key term candidates during search through the phone lattice, and key term matching using a phone similarity matrix including two different distance measures, as will be presented below in detail.

The next important part of the approach is spoken document retrieval in the middle of the figure right below the spoken document archives. All the different retrieval models, including VSM, LSI and PLSA, as mentioned above, can be used here; in each case two different versions of models can be used: one based on the general lexicon including all terms except those stop terms deleted in advance, and the other based on the much smaller but semantically rich key term lexicon. As will be shown in the experiments below, the new version of PLSA models based on the key terms instead of all terms in the lexicon are very helpful in offering improved retrieval performance. The different retrieval models can also be integrated easily by summing the weighted relevance scores for each document in the archive obtained with different models. For example, in all the experiments presented below, we always integrate the LSI or PLSA model with the VSM model, such that both concept matching and literal term matching can be achieved simultaneously.

3. NAMED ENTITY RECOGNITION

Named entities (NEs) are certainly useful key terms for spoken document retrieval, not only because such names are usually the key for the content of the documents and the fact that many of them are out of vocabulary (OOV) which require special approaches to handle, but also because many heuristic rules and carefully designed algorithms are available to recognize named entities from spoken documents. As a result, the NEs recognized from spoken documents are usually more reliable feature elements than other terms.

In our NE recognition module in Figure 2, two special approaches were developed [8]. The first is to recognize the NEs from a text document (or the transcription of a spoken document) using global information extracted from the entire collection of documents in addition to the local (internal and external) evidences. The basic idea is that very often an NE is difficult to identify in a single sentence. But if the scope of observation can be extended to the entire document, it will be found that this entity appears several times in several different sentences, and has a higher likelihood of being an NE when all those occurrences in different sentences can be considered jointly. The PAT tree data structure was found to be very useful in recording such global information for the entire text document.

The second special approach used here is for spoken documents, to recover the OOV NEs using external knowledge. Each spoken document was first transcribed into word graphs, for which not only the NE extraction approaches for text documents including usage of PAT trees as mentioned above can be applied, but also words with higher confidences scores can be identified. Possibly relevant text documents (in the case of broadcast news in the prototype system and the preliminary experiments reported here, the text news published in the same time period) available over the Internet were then automatically retrieved using queries constructed from those words with higher confidence measures on the transcribed word graphs. Named entity recognition was then performed on these retrieved text documents including using the global information extracted from the PAT trees as mentioned above, and a set of possible NE candidates was obtained. The NE matching can then be performed between the word graphs of the spoken documents and the NE candidates obtained above. The basic idea is that those path segments in the word graphs with relatively lower confidence measure are likely to be recognition errors due to OOVs. So the phone lattices for such segments can be matched with the NE candidates obtained above. This matching process is exactly the same as the matching of the query phone lattice with the key terms as mentioned in section 2, so it is not further repeated here.

4. KEY TERM EXTRACTION BY TERM ENTROPY BASED ON PLSA

As mentioned in section 2, an important new approach proposed here in this paper is to extract the dynamic key terms using a new metric referred to as term entropy based in this paper on PLSA. This is briefly explained below. As mentioned above, the probability $P(t_j|T_k)$ of observing a term t_j for a latent topic T_k as in equation (1) can be obtained from PLSA modeling. We can first estimate $P(T_k|t_j)$ by:

$$P(T_k|t_j) = \frac{P(t_j|T_k) \times P(T_k)}{P(t_j)} \approx \frac{P(t_j|T_k)}{P(t_j)} \tag{3}$$

Where the probability $P(T_k)$ is left out because a good approach to estimate it is not yet available, while $P(t_j)$ can be obtained from a large corpus. The term entropy $H(t_j)$ for a term t_j can then be obtained as:

$$H(t_j) = -\sum_{k=1}^{K} P(T_k|t_j) log P(T_k|t_j) \tag{4}$$

Apparently higher entropy here implies the term is frequently observed in many different topics, or is less specific semantically. Lower entropy, on the other hand, indicates that the term is focused on very few topics, and thus possibly is a key term for these few topics. Key terms can therefore be extracted using this term entropy measure.

5. KEY TERM RECOGNITION FROM USER QUERY

In order to correctly recognize the key terms in the user query, special approaches were developed. The user spoken query is transcribed not only into a word graph as usual recognition process, but into a phone lattice as well. We then match the phone lattice with the phone sequences of the key terms in the dynamic lexicon using dynamic programming. If the similarity measure is higher than a threshold, we assume the query may include the key term and therefore emphasize the likelihood score of the key term by a factor such that the key term has a higher probability to be recognized. The price paid here is of course the overall word error rate may be increased. But a good tradeoff is possible because correctly recognized key terms are much more helpful in the retrieval task than other normal words. In order to perform the matching between two phone sequences, we defined a phone similarity matrix which included both the phonemic distance (distance between two acoustic models) and the pronunciation distance (the distance is smaller if a phone is more likely to be pronounced as another phone) [9]. The phone sequence matching is then based on the total distance normalized with the number of phones in the sequence. After an utterance verification process performed on the key term obtained above, the key term is finally recognized from the query.

6. INITIAL PROTOTYPE SYSTEM AND THE EXPERIMENTAL CONDITIONS

An initial prototype system has been successfully developed at National Taiwan University. The broadcast news are taken as the example spoken/multi-media documents. All the modules shown in Fig 2 and presented in Section 2, 3, 4 and 5 were successfully implemented, and this is the experimental platform for all the experimental results reported below.

The broadcast news archives to be retrieved in the prototype system includes roughly 130 hours of about 9836 news stories, all in Mandarin Chinese. They were all recorded from radio/TV stations in Taipei from Feb 2002 to May 2003. The word, character and syllable error rates of 27.96%, 14.29% and 8.91% respectively were achieved in the transcriptions. These transcriptions of the 9836 news stories including errors were used to train the latent semantic space of LSI and the PLSA model. 32 topics were used in PLSA modeling. A lexicon of 61521 word was used here. 1000 news stories among the above 9836 were used in the retrieval performance tests reported below as the archives to be retrieved, in order to evaluate the recall/precision rates, a total of 50 natural language queries as human-generated and then respectively produced by two male and two female speakers. The length of the queries is roughly 8-11 words. Computer-aided human labeling methods were used to identify the target relevant news stories for each query among the 1000 mentioned above as the reference to evaluate the recall/precision rates. The key terms in the dynamic lexicon were extracted from the whole archives of 9836 news stories. A total of 1708 NE were obtained. For key terms selected with term entropy, we first obviated single character words and preposition phrases, and then picked up the top 2000 terms ranked by term entropy to be included in the dynamic key term lexicon. Among the 2000 ,133 are NEs already recognized in NE recognition, while others are key terms which are not NEs.

7. EXPERIMENTAL RESULTS

First, consider the key term extraction from user query as discussed in Section 5. There is apparently a tradeoff with the proposed approach that although the key term recognition accuracy can be improved, the error rate for other words will be increased too. Preliminary tests were performed for queries produced by a single speaker to see this tradeoff, and the results are listed in Table 1, where the baseline is the results when the normal recognition process was performed on the queries, while the proposed approach included those mentioned in Section 5. As can be found from Table 1, the error rate for other words was indeed increased, but the F-measure was improved as well. So the proposed approach is actually helpful for retrieval. Next, we also compared the retrieval performance of LSI and PLSA, both based on full lexicon, both integrated with a baseline of VSM, in another preliminary experiment using another set of 20 short queries. The F-measure obtained for LSI and PLSA are 0.526 and 0.547 respectively. PLSA is apparently better. So only PLSA was tested in all the following experiments.

Table 1. The tradeoff of using the proposed key term extraction approach from the user queries.

	baseline	proposed key term extraction approach
Error rate for other words	0.457	0.479
Error rate for key terms	0.433	0.414
F-measure	0.399	0.405

Table 2. Retrieval results for the various retrieval model configurations.

model configurations	Precision/Recall	F-measure
(1)VSM_0=Baseline	0.422/0.394	0.407
(2)VSM_0+VSM_{ENT}	0.446/0.399	0.421
(3)VSM_0+VSM_{NE}	0.454/0.420	0.436
(4)VSM_0+VSM_{ENT+NE}	0.473/0.439	0.455
(5)VSM_0+$PLSA_0$	0.453/0.422	0.436
(6)VSM_0+$PLSA_{ENT}$	0.462/0.427	0.443
(7)VSM_0+$PLSA_{NE}$	0.471/0.433	0.451
(8)VSM_0+$PLSA_{ENT+NE}$	0.482/0.454	0.467

The final retrieval results in terms of precision/recall rates and F-measures for the various retrieval model configurations are listed in Table 2. The baseline in row (1) (VSM_0) used indexing terms based on words, characters and syllables and was shown to be very successful in the past [10]. The next three rows (2)(3)(4) are respectively the results when an extra VSM-based model using key terms selected by term entropy alone (VSM_{ENT}), by named entities (VSM_{NE}), and by both (VSM_{ENT+NE}) was integrated with the baseline. Significant improvements were obtained in all cases. It can be found that named entities are more efficient in improving the retrieval performance than the key terms selected by the term entropy (row (3) VS rows (2) and (1)), but the term entropy can identify many key terms which are not named entities. This is why the key term lexicon obtained by both approaches offered much better performance (row (4) VS rows (3)(2)(1)). $PLSA_0$ in row (5) is the approach of "plain PLSA", in which all words in the lexicon were used in PLSA modeling for retrieval. Integration of $PLSA_0$ with the baseline (VSM_0) gave very significant improvement (row (5) VS row (1)). The next three rows (6)(7)(8) are then respectively the results when the new versions of PLSA models proposed here, i.e., the PLSA model was based on the key terms obtained with term entropy, named entities or both ($PLSA_{ENT}$, $PLSA_{NE}$, $PLSA_{ENT+NE}$ in rows (6)(7)(8)) instead of all words in the lexicon as in $PLSA_0$, were integrated with the baseline. Exactly the same trend as observed above were obtained here (named entities better than term entropy, and using both is better, rows (6)(7)(8)). Furthermore the new versions of PLSA models were always much better than the conventional "plain PLSA" (rows (6)(7)(8) VS row (5)), and the new versions of PLSA models are apparently better than the new version of VSM models, both based on key terms (rows (6)(7)(8) VS rows (2)(3)(4)). The best results obtained here is in row (8), which represented a significant improvement in precision (0.482 VS 0.422), recall (0.454 VS 0.394) and F-measure (0.467 VS 0.407) as compared to the baseline of VSM_0. Note that the best result of row (8) in Table 2 (VSM_0 + $PLSA_{ENT+NE}$) was obtained by carefully weighting the relevant scores, and the different results for different weighting parameters are shown in Figure 3.

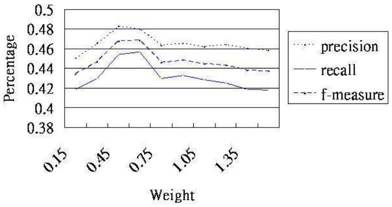

Fig. 3. Retrieval results of VSM_0+$PLSA_{ENT+NE}$ for row (8) of Table 2 for different weighting parameters

For the above experiment, we see that key terms extracted by term entropy and named entities did complement each other. For example, there do exist many important key terms which are not named entities. These key terms very often can be found by the term entropy. We thus calculated the following mutual information (MI) for two key terms t_i and t_j:

$$MI = \frac{P(t_i, t_j)}{P(t_i)P(t_j)} \quad (5)$$

We found that the mutual information when t_i is a named entity and t_j is not a named entity but extracted by term entropy is 1.227. But if t_i and t_j are simply randomly selected from the key term lexicon in the system in rows (4) or (8) of Table 2 (ENT+NE), regardless of which group they belong to, the mutual information is 1.320. This also indicated that the key terms extracted by term entropy and the named entities actually complemented each other.

8. CONCLUSION

This paper presents an improved approach for spoken document retrieval using dynamic key term lexicon and PLSA. The integration of the new version of PLSA based on the dynamic key term lexicon with the conventional VSM model gives very significant improvements. Experimental results show that the combination of entropy-based and NE-based key term lexicons moreover provides better results.

9. REFERENCES

[1] L. S. Lee and B. Chen, "Spoken document understanding and organization," *in Special Section, IEEE Signal Processing Magazine*, 2005.

[2] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 2001, pp. 19–25.

[3] T. HofMann, "Probabilistic latent semantic analysis," *Uncertainty in Artificial Intelligence*, 1999.

[4] B. Chen, H.W. Wang, and L.S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing then for voice retrival of speech information in mandarin chinese," *IEEE Trans.Speech Audio Processing*, vol. 10, no. 5, pp. 303–314, 2002.

[5] K. Ng and V.W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[6] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 2000, pp. 81–87.

[7] 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.Speech Audio Processing*, vol. 10, no. 8, pp. 531–541, 2002.

[8] Y.C. Pan, Y.Y. Liu, and L.S. Lee, "Named entity recognition from spoken documents using global evidences and external knowledge sources with applications on mandarin chinese," in *Proc. ASRU*, 2005.

[9] M.Y. Tsai and L.S. Lee, "Pronunciation variation analysis based on acoustic and phonemic distance measures with application examples on mandarin chinese," in *Proc. ASRU*, 2003, pp. 117–122.

[10] C.J. Wang, B. Chen, and L.S. Lee, "Improved chinese spoken document retrieval with hybrid modeling and data-driven indexing features," in *Proc. ICSLP*, 2002, pp. 16–21.