CONCEPT BASED QUERY AND DOCUMENT EXPANSION USING HIDDEN MARKOV MODEL

Jiuling Zhang, Zuoda Liu
Department of Electronic Engineering, Tsinghua University, Beijing, China

Beixing Deng, Xing Li
Department of Electronic Engineering, Tsinghua University, Beijing, China

Keywords: Hidden Markov model, WordNet, Word sense disambiguation.

Abstract: Query and document expansion techniques have been widely studied for improving the effectiveness of information retrieval. In this paper, we propose a method for concept based query and document expansion employing the hidden Markov model(HMM). WordNet is adopted as the thesaurus set of concepts and terms. Expanded query and document candidates are yielded basing on the concepts which are recovered from the original query/document term sequence by employing the hidden Markov model. Using 50000 web pages crawled from universities as our test collection and Lemur Toolkit as our retrieval tool, preliminary experiment on query expansion show that the score of top 20 retrieved documents have a 2.7113 average score increment. Numbers of documents with score higher than a given value also increased significantly.

1 INTRODUCTION

Expansion techniques aims at increase the likelihood of retrieving relevant documents, generally expansion method include query expansion and document expansion. Query expansion is widely studied and generally known, but research on document expansion is relatively few and in our paper it will be addressed.

Query expansion deals with the query term incompleteness or syntax malapropism problems in information retrieval. Query term incompleteness comes out frequently due to that users cannot provide a detailed enough query while catachresis or solecism lead to syntax malapropism, thus adding additional query terms to the original query or reformulating the original query is of necessity. Query expansion method generally helps in improving the recall ratio of information retrieval systems. But traditional structural query expansion method may lead to a decrease in precision. Qiu et al. proposed concept based query expansion in which terms expressing a common concept are selected and weighted before adding to the expansion term lists (Qiu and Frei, 1993). In another work queries were expanded using the term relationships defined in the WordNet thesaurus, which is a lexical-semantic knowledge base (Ellen, 1994). But the author have not taken the term context information into consideration. Hoeber pointed out that previous query expansion rely on a general thesaurus as the basis or rely on a specialized thesaurus which is constructed from the text of corpus (Hoeber, 2005). The author manually constructed a concept network based on which terms are selected to perform conceptual query expansion. But the quality of this expansion method depends highly on the quality of the concept thesaurus, and an automatic construction of high quality thesaurus and a good generation model is necessary. Thus an automatic method employing the hidden Markov model is proposed in our work.

Previously query expansion dominate the expansion field, but document should also be attached similar importance. Query expansion may be not a good choice as too many useless words may be introduced as the result of aimless expansion (Min Zhang, 2004). Similar problems to query may also occur to the documents such as the sentences incompleteness, syntax malapropism. The most important benefit is, by document expansion, users of IR systems have a potentially broader view from

Zhang J., Liu Z., Deng B. and Li X.
CONCEPT BASED QUERY AND DOCUMENT EXPANSION USING HIDDEN MARKOV MODEL.
DOI: 10.5220/0001842506880691
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

which more documents may be relevant to their query needs. By excavating all possible expressions of each document the IR system will be bound to provide a variety of candidate documents to a certain query.

Document expansion method has also been extensively studied recently. It was previously introduced in speech retrieval by selecting alternative terms from similar documents (Singhal and Pereira, 1999). Liu et al. proposed language models for cluster-based retrieval by using cluster centered around the document itself to smooth it (Liu and Croft, 2004). In contrast to the Liu's macroscopical document expansion method, we'd prefer a microscopical method, that is we expand the document by using the information contained in the document itself.

The remainder of this paper is organized as follows: Section 2 lays out how to recover the concept sequence using hidden Markov model. Section 3 carries out the query and document expansion approaches using the concept sequence. Preliminary experimental results are described in Section 4. Lastly, a conclusion is made in section 5.

2 RECOVER CONCEPT SEQUENCE USING HMM

A discrete hidden Markov model(HMM) is denoted by a five element array $<S,K,\Pi,A,B>$ (Manning and Schutze, 2005), which represents the following respectively:

1) $S=\{s_1,s_2,\cdots s_n\}$ is a finite set of states.

2) $K=\{k_1,k_2,\cdots k_m\}$ is the finite set of output symbols.

3) $\Pi=\{\pi_i\}, i\in S$ is the set of initial state.

4) $A=\{a_{ij}\}$ is the set of state to state transition probability.

5) $B=\{b_{ijk}\}$ is the set of observation symbol probability distribution given states in S.

Given predefined values of S,K,A,B and Π, the HMM can be used to model an observation sequence $\mathrm{O}=\{o_1,o_2,\cdots o_T\}$, with the state sequence being $X=\{X_1,X_2\cdots X_{T+1}\}$.

There are three basic problems of HMM which must be solved to make it useful in real-world applications, that is how to efficiently compute the probability $P(\mathrm{O}\mid\mu)$ of a specific sequence O given the mode $\mu=\{A,B,\Pi\}$, how to select an optimal state sequence $X=\{X_1,X_2\cdots X_{T+1}\}$ so that it can give the best explanation of a certain sequence, and how to adjust the model parameters $\mu=\{A,B,\Pi\}$ to maximize $P(\mathrm{O}\mid\mu)$.

All the above three question have been solved theoretically and what we take into consideration is about problem 2, as our target is to perform query expansion based on the synsets before which the synsets should be determined previously (Manning and Schutze, 1999). The Viterbi algorithm is an appropriate method to retrieve the synsets by employing a lattice structure, which can efficiently implement the computations.

Hidden Markov model have been applied over the recent two decades successfully in the field of speech recognition and a variety of other language recognition problems (John and Richard, 1994). In HMM applications the observed data is modelled as the symbol generated by some hidden state. In our work we propose a totally different idea. In our retrieval case, we assume the query terms to be observed symbols and the synsets to be hidden states.

In our work, we propose a comprehensive method to expand both queries and documents using HMM. We know that a person express his meanings through a sequence of words, but what he is considering about are represented by concepts. But we don't have a perfect expression mechanism to produce the optimal word sequence so that one's intention can be clearly and uniquely identified. From the same point of view, the process of generating words from certain concepts can be modelled as a symbol generation procedure. Words are observable in contrast with the invisible states which are hidden behind the expressions.

Our purpose is to recover the optimal hidden state associated with the given sequence of query terms. There are several criterion and algorithms can be utilized (Manning 2005). Here we use the Viterbi algorithm (Viterbi 1967). In order to discover the optimal state sequence $X=\{X_1,X_2\cdots X_{T+1}\}$ for the given observation sequence $\mathrm{O}=\{o_1,o_2,\cdots o_T\}$, we need to compute:

$$\arg\max_X P(X\mid\mathrm{O},\mu) \qquad (1)$$

We denote:

$$\delta_j(t)=\max_{X_1\cdots X_{t-1}} P(X_1\cdots X_{t-1},o_1\cdots o_{t-1},X_t=j\mid\mu) \qquad (2)$$

that is to say $\delta_j(t)$ stores the probability of the most possible sequence of symbols. In order to efficiently recover the state sequence and reduce the amount of computation, let $\varphi_j(t)$ record the state sequence associated with the above symbol sequence.

The integrated procedure of recovering the most probable state sequence is stated as follows (Viterbi 1967) :

1) initialization:

$$\delta_j(1) = \pi_j, \quad 1\le j \le N \tag{3}$$

2) computation and recursion:

$$\delta_j(t+1) = \max_{1\le i\le N} \delta_i(t) a_{ij} b_{ijo_t}, \ 1\le j \le N \tag{4}$$

$$\varphi_j(t+1) = \arg\max_{1\le i\le N} \delta_i(t) a_{ij} b_{ijo_t}, \ 1\le j \le N \tag{5}$$

3) termination and state(concept) backward retrieving

$$\hat{X}_{T+1} = \arg\max_{1\le i\le N} \delta_i(T+1) \tag{6}$$

$$\hat{X}_t = \varphi_{\hat{X}_{t+1}}(t+1) \tag{7}$$

Hitherto, we have recovered the optimal concept sequence given an original word sequence using the HMM. Then we would perform expansions to sentence in queries or documents.

3 CONCEPT BASED EXPANSION

As demonstrated in section 2, the probability of any concept sequence, and word sequences can be computed given the model's parameters. But most of those concept or word sequences have zero or very small probability, which implies that they are small probability events and will be sparse in practice thence can be neglected. Thereafter we only select the concept sequence recovered with the highest probability using the Viterbi backtracking algorithm. Since we have obtained the selected concept sequence, word sequence candidates can be gathered by adopting those sequences with higher observation probabilities.

After a sequence of concepts have been recovered from the original word sequence using the hidden Markov model, we can regenerate the most possible expanded word sequence basing on it. The word sequences can be determined by the following formula:

$$S = \{o_t \mid o_t \in \mathrm{O}, P(o_t \mid \mu, X) > T)\} \tag{8}$$

in which the probability $P(o_t \mid \mu, X)$ can be computed rapidly (Manning and Schutze, 2005).

Although the word sequences are generated from the same concept sequence, they may have different appearing probabilities. Sequences with higher probabilities means that they occur frequently in practice, they are more possible to represent a given sequence of concepts.

After expanded word sequence are collected, they are treated as new candidate sentence and added to query or document.

All the retrieved results of the expanded queries and documents are combined as the final results, thus, the number of relevant documents of all the queries is definitely larger than that of a single original query. That is to say, the recall of the expansion method is theoretically higher than without expansions.

4 PRELIMINARY EXPERIMENTAL RESULT

Though theoretically the document expansion method could improve the information retrieval effectiveness, practically it is extremely time-consuming to expand all the documents. However, since the processing of the documents is once for all thus document expansion is also desirable if the computational resource is abundant. Due to our limited computing power the document expansion is not implemented and evaluated in our experiment. Rather we perform the concept based expansion only to queries.

We have crawled approximately 50000 web pages from university websites to estimate parameters employed in our hidden Markov model. Word sense disambiguation is performed to the preprocessed free text, then the sense of each word in the text are disambiguated. By counting the occurrence times of the triple $< synset, word, synset >$, parameters a_{ij}, b_{ijo_t} are estimated.

We use Lemur Toolkit as our retrieval system. The original query are used as the input query, the retrieved documents are scored basing on the TFIDF retrieval model in Lemur. 10 original queries are expanded. Given the collection, Lemur firstly build the index, then rank the documents using the TFIDF model, and give final documents according to their scores which represent the relevancy to queries. Though the number of documents that is relevant to original query is not available, the relevancy of queries and documents are closely related with the scores, thus our result is still statistically important.

In order to illustrate that the expansion method at least provide more candidate document for the retrieval, the number of retrieved documents with score higher than the mean of top 20 retrieved documents without expansion is recorded, as shown in figure 1.

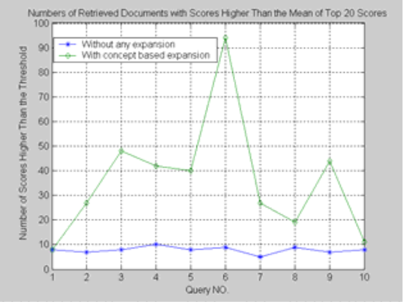

Figure 1: Numbers of retrieved documents with scores higher than the threshold.

Figure 1 shows that there are more high score candidate documents are retrieved. This implicitly means that we can obtain more possibly related documents using this expansion method.

To illustrate the average score improvement, the retrieval results of expanded queries are combined together. Then they are compared with the retrieval result of the original query as is shown in figure 2.

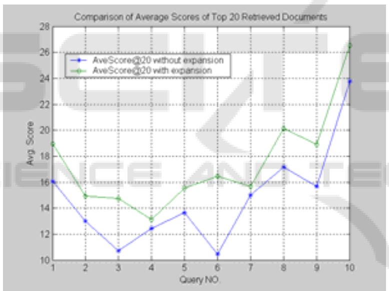

Figure 2: Comparison of average scores of top 20 retrieved documents.

Figure 2 shows the score of combined documents of expanded queries is obviously higher than without any expansion. The average score increment over the 10 queries is 2.7113.

The figures above show that concept based query expansion can not only provide more high score candidate documents, but also improve the average score of the top 20 retrieved documents.

5 CONCLUSIONS

In this paper, we proposed concept based query and document expansion method using hidden Markov model. Preliminary experimental result on query expansion show that our concept based method can not only retrieve more high score documents, but also improve the average score of top 10 retrieved documents.

Theoretically, document expansion can also improve the retrieval effectiveness by providing more candidate documents, and it will be part of our future work. Apart from that, using TREC collections to show how the expansion method reallyaffect the recall and precision is also imperative.

ACKNOWLEDGEMENTS

This work is supported by National Basic Research Program of China (973 Project, No. 2007CB310806).

REFERENCES

Yonggang Qiu, H.P. Frei, 1993. Concept based query expansion. In *SIGIR’93, 16th Int. ACM/SIGIR Conf. on R&D in Information Retrieval*, pages 160-167, Pittsburgh, PA, USA.

Ellen M. Voorhees, 1994. Query expansion using lexical-semantic relations. In *SIGIR’94, 17th Int. ACM/SIGIR Conf. on R&D in Information Retrieval*, pages 61-69, Dublin, Ireland.

Orland Hoeber, Xue-Dong Yang, Yiyu Yao, 2005. Conceptual query expansion. In *Proceedings of the Atlantic Web Intelligence Conference*. Lodz, Poland.

Manning, C. D., Schutze, H. 1999. *Foundations of statistical natural language processing*. Cambridge Massachusetts: MIT Press.

Min Zhang, Ruihua Song, Shaoping Ma, 2004. Document Refinement Based On Semantic Query Expansion. Journal of Computer, Vol.27, No.10, pp1395-1401.

Singhal and Pereira, 1999. Document Expansion for Speech Retrieval. In *SIGIR’99, 22th Int. ACM/SIGIR Conf. on R&D in Information Retrieval*. pages 34-41

Liu, X. and Croft, W. B, 2004. Cluster-based retrieval using language models. In *Proceedings of SIGIR '04*, pages 186-193.

John Makhoul, Richard Schwartz, 1994. *State of the art in continuous speech recognition*, National Academy Press, Washington, DC, USA.

David R. H. Miller, Tim Leek, Richard M. Schwartz, 1999. A hidden Markov model information retrieval system, In *proceedings of the 1999 ACM SIGIR Conf. on R&D in Information Retrieval*, page 214-221.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press. 2008

A. J. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithms. *IEEE Trans. Informat. Theory, vol. IT-13*, pp. 260-269.