Information Retrieval System Using Latent Contextual Relevance

Minoru Sasaki*, Hiroyuki Shinnou†

*Department of Computer and Information Sciences,
Faculty of Engineering, Ibaraki University, Ibaraki, Japan
E-mail: sasaki@cis.ibaraki.ac.jp
†Department of Systems Engineering,
Faculty of Engineering, Ibaraki University, Ibaraki, Japan
E-mail: shinnou@dse.ibaraki.ac.jp

Abstract

When the relevance feedback, which is one of the most popular information retrieval model, is used in an information retrieval system, a related word is extracted based on the first retrival result. Then these words are added into the original query, and retrieval is performed again using updated query. Generally, Using such query expansion technique, retrieval performance using the query expansion falls in comparison with the performance using the original query. As the cause, there is a few synonyms in the thesaurus and although some synonyms are added to the query, the same documents are retireved as a result. In this paper, to solve the problem over such related words, we propose latent context relevance in consideration of the relevance between query and each index words in the document set.

1. Introduction

A user usually applies natural language to express an own query. Using a search engine such as Lycos or Google on the Internet, the user represents queries which consists of a few words. If the user has some knowledge of the words typically, the user can describe a particular topic and represent a query exactly. However, if the user does not come up with the topic words, it is difficult to represent queries with the content to search. In an information retrieval (IR) system, without considering lexical and semantic ambiguity such as paraphrase representation, documents containing the input words are retrieved.

For the purpose of retrieving a document containing synonyms with a query, There are many research of automatic query expansion to help the user formulate what information is really needed. Query expansion is the process of IR system extracting and adding search terms to a user's weighted search based on the first retrieval result (Salton and Buckley, 1990) . Such retrieval processing method is called "relevance feedback", which is the most popular information retrieval model. This method has the effect of avoiding the ambiguity of the meaning in the original query (Kuriyama, 1998) .

To reduce response time for retrieval operation with IR system and output a document summarization(DS) result from a document with a high degree ofaccuracy, the use of thesaurus is a possible approach to expand and to focus the queries. A thesaurus is a compilation of words and phrases showing synonymous and hierarchical relationship and dependencies. It is often used for the support of retrieval approaches. However, to apply the thesaurus in the IR system, it is reported that the retrieval performance using expansion decreases in comparison with the performance using the original query. it is considered that the lack of the number of synonymous in the thesaurus is the cause of a performance decrease. Moreover, the system extracts some synonyms that contains the same word in the original query from the thesaurus so that the next retrieval finds the same document as the first.

To solve these problems about the thesaurus in such cases, it is necessary to construct the dictionary of the form which can be treated by the computer. Additionally, when a thesaurus is used for the support of IR and DS system, a complicated layered structure is not so required and it is possible to support these tasks with a thesaurus that has simpler layered structure. Therefore, it is necessary to construct the thesaurus of the exclusive use for the purpose of these objects. In our research, we improve Contextual Document Relevance(CDR) (Korpimies and Ukkonen, 1996) which is proposed as one of the related term extraction techniques and propose Latent Contextual Relevance(LCR) in consideration of the relevance between the index terms in the document set. Additionally, for the real-time extraction system, we construct IR system using a query expansion using LCR and evaluate the retrieval performance of this system.

When a query is given to the IR system, CDR method calculates a similarity between the query and each documents. In most methods of the related term extraction, these method apply a similarity calculation such as inner product. In this case, this similarity has a great influence of the quality of the extracted related words. Therefore, using the technique of calculating the similarity in higher accuracy, we consider that the system extracts the higher related terms which uses the distribution of the terms in the document. As the method of similarity calculation, we propose to use an approximated term-document matrix obtained by Singular Value Decomposition(SVD). By using SVD, this approximation associates a term with the other terms semantically for the terms which tend to cooccur in documents with the same contents (Deerwester et al., 1990).

2. Letent Context Document Relevance

For the processing that a user finds the information about the user's query effectively, the user extracts the index word that has a relation to the query and adds to the

original query. In this way, the user reduces the ambiguity of the meaning of the query. Generally, to find words that the concept resembles, thesaurus such as the synonym dictionary is used. However, thesaurus construction involves an immense amount of time and effort because there are quite a lot of index words. Therefore, as a method to construct a thesaurus automatically, we propose the CDR based on the latent semantic indexing named Latent Context Relevance(LCR).

In this section, we give the outline of the calculation of CDR and point out the problem of the CDR. To solve such a problem, we explain the detail of LCR to calculate a similarity between the query and each index terms.

2.1. Context Document Relevance

We show the alogrithm to calculate CDR which is one of the related term extraction techniques. A document is represented as a document vector $d_i = (w_{i1}, w_{i2}, \cdots, w_{it})$ and the element of d_i is equal to the weight w_{ij} of the index word t_j that appears in the document. A query is also represented as a query vector $Q = (q_1, q_2, \cdots, q_t)$ and the element of Q is equal to the frequency of the word t_j in the query.

Calculating the CDR, the similarity between the query Q and each document is calculated to extract relevant documents in order of the similarity. This similarity $rel(Q, d_i)$ is represented as follows:

$$rel(Q, d_i) = \frac{Q \cdot d_i}{|Q| \cdot |d_i|} = \frac{\sum_{j=1}^{t} w_{ij} q_j}{\sqrt{\sum_{j=1}^{t} {w_{ij}}^2 \sum_{j=1}^{t} {q_j}^2}}. \quad (1)$$

When the similarity of each document exceeds a constant threshold value, the term weight increases for all the index words that appear in the document. In case that the word t_j appears in a document, CDR between t_j and the query is equal to the product of the term weight w_{ij} in the document d_i and the similarity $rel(Q, d_i)$ as follows:

$$cdr(Q, t_j) = \sum_{i=1}^{n} w_{ij} rel(Q, d_i). \quad (2)$$

In this calculation, however, when the index word appears in any documents equally, the high weight is given to the word irrespective of whether the word is related to the query or not. In consideration of such a case, CDR needs to calculate the total of weight and normalize the vector as follows:

$$df_j = \sum_{i=1}^{n} w_{ij}. \quad (3)$$

So the normalized CDR is represented as follows:

$$ncdr(Q, t_j) = \frac{\sum_{i=1}^{n} w_{ij} rel(Q, d_i)}{df_j}. \quad (4)$$

The t_j is related to the query if the CDR value is high. Several related words are extracted in order of this value and expanded into the original query.

2.2. Letent Context Relevance

In the CDR mentioned above, the similarity between a query and a document is calculated in advance for a given query. Such a similarity calculation is used in almost the query expansion method based on the related term extraction proposed before. In this method, the similarity has a great influence to extract related terms. Therefore, we use the other similarity calculation technique with high retrieval performance to extract related terms. Taking the distribution of semantics in the document into consideration, it becomes possible to perform the effective extraction.

As one of the methods to extract related terms, we propose to apply the latent semantic indexing(LSI) in the similarity calculation (Berry et al., 1995) . LSI is an information retrieval model to get the latent semantic relation between terms and documents by using singular value decomposition(SVD) to reduce the dimension of term-document matrix. The dimension of the transformed space is reduced by selection of the highest singular values to approximate the original term-document matrix. Consequently, the major associative patterns are extracted from the document and the small patterns are ignored.

In this section that follow, we describe the algorithm of LCR which calculates a similarity between a query and each document. As well as CDR, a document is represented as a document vector $d_i = (w_{i1}, w_{i2}, \cdots, w_{it})$ and the element of d_i is equal to the weight w_{ij} of the index word t_j that appears in the document. A set of document vectors is represented as a term-document matrix A, where we suppose that the rank of the matrix A is r. The SVD of the matrix A is defined as the product of three matrices as follows:

$$A = U \Sigma V^T. \quad (5)$$

In this expression, the columns of U and V are the left and right singular vectors and the diagonal elements of Σ are called the singular values of the matrix A. The largest $k (k < r)$ singular values of A and the first k columns of the U and V matrices are extracted from these matrices to construct the closest rank k approximated matrix of A as follows:

$$A_k = U_k \Sigma_k V_k^T. \quad (6)$$

Now we describe the LCR calculation of a query Q for the another word t_l. The query is represented as a query vector $Q = (q_1, q_2, \cdots, q_t)$ and the element of Q is equal to the frequency of the word q_j in the query. the similarity between the query $Q^{(k)}$ and each document $d_i^{(k)}$ in the approximated k dimensional space is calculated to extract relevant documents in order of the similarity. This similarity $crel(Q, d_i)$ is represented as follows:

$$crel(Q, d_i) = \frac{Q^{(k)} \cdot d_i^{(k)}}{|Q^{(k)}| \cdot |d_i^{(k)}|}. \quad (7)$$

As well as the CDR mentioned above, LCR between the term t_l and the query Q is equal to the product of the term weight w_{il} in the document d_i and the similarity $rel(Q, d_i)$ as follows:

$$lcdr(Q, t_l) = \sum_{i=1}^{n} w_{il} \cdot crel(Q, d_i). \quad (8)$$

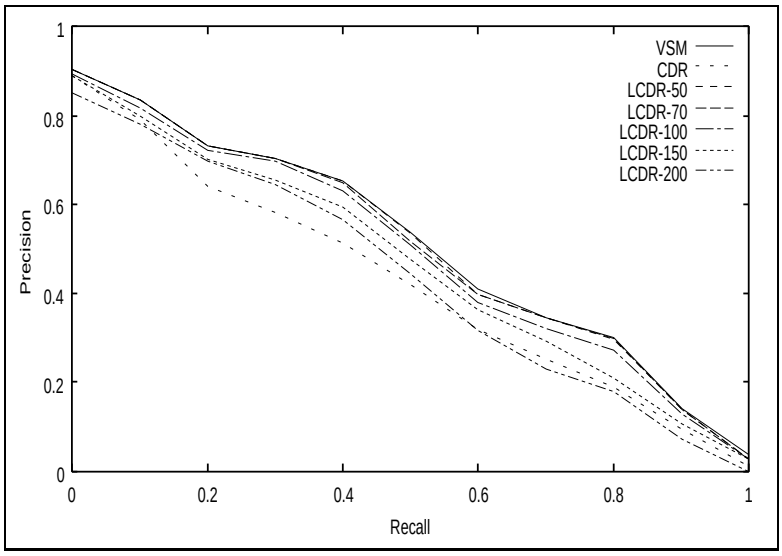

Figure 1: recall-precision curve using LCR

The normalized CDR, as well as the normalized CDR is represented as follows:

$$nlcdr(Q, t_l) = \frac{\sum_{i=1}^{n} w_{il} \cdot crel(Q, d_i)}{\sum_{i=1}^{n} w_{il}}. \quad (9)$$

3. Experiments

In this section, we describe our vector space information retrieval model using LCR and experimentally evaluate the efficiency of the model using the MEDLINE collection. The MEDLINE collection consists of 1033 documents from medical journals and 30 queries and relevancy judgments of the queries. We first preprocessed all the documents in the MEDLINE collection to remove all the stop words using a stop list of 439 common English words such as "a" or "about". We also remove words occurring in only one document after the stop word elimination. The remaining words were stemmed using the Porter's algorithm and 4329 index terms are obtained as a result of the preprocessing.

When each document is represented as a vector, the elements of a document vector d are assigned two-part values $d_{ij} = L_{ij} \times G_i$ (Chicholm and Kolda, 1998). In the experiments, the factor L_{ij} is a local weight that reflects the weight of term i within document j and the factor G_i is a global weight that reflects the overall value of term i as an indexing term for the entire document collection as follows:

$$L_{ij} = \begin{cases} 1 + \log f_{ij} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases}, \quad (10)$$

$$G_i = 1 + \sum_{j=1}^{n} \frac{\frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}}{\log n}, \quad (11)$$

where n is the number of documents in the collection, f_{ij} is the frequency of the i-th term in the j-th document, and F_i is the frequency of the i-th term throughout the entire document collection.

For the term-document matrix consisted of the term weight mentioned above, an approximated term-document matrix is obtained by the dimensionality reduction using SVD. Next, for the documents of which a similarity $crel(Q, d_i)$ to the query Q is 0.1 or more, our system calculates LCR value of each term. The system expands the top 15 index words of LCR into the query and performs retrieval using the updated query. As a retrieval result, the system outputs the top 50 documents in order of the similarity using the general vector space model.

To evaluate the performance of our system, we decided to measure performance in terms of recall and precision (Lewis, 1991) (Witten et al., 1994).

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}, \quad (12)$$

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}. \quad (13)$$

Evaluation of Information retrieval system is possible even if we use recall or precision individually. In this experiments, evaluation of the ranked output system results in a 11-pointed recall-precision curve generally, with points plotted that represent precision at various recall percentages. Typically, as average performance over a large set of queries, Average precision at each standard recall level across all queries is computed.

3.1. Experimental Results

Figure 1 shows results of experiments using LCR. In this figure, 'VSM' is a retrieval result for a convensional information retrieval model and 'CDR' is a result for the information retrieval model using CDR. 'LCDR-50', 'LCDR-70', 'LCDR-100', 'LCDR-150' and 'LCDR-200' are the result for the information retrieval model by reducing the specified number of dimension using LCR respectively.

In comparison with the precision of CDR, LCR improves the average precision in case that the dimension of vector space is 100 and below. In case that the dimension is 150, however, the precision of LCR is approximately

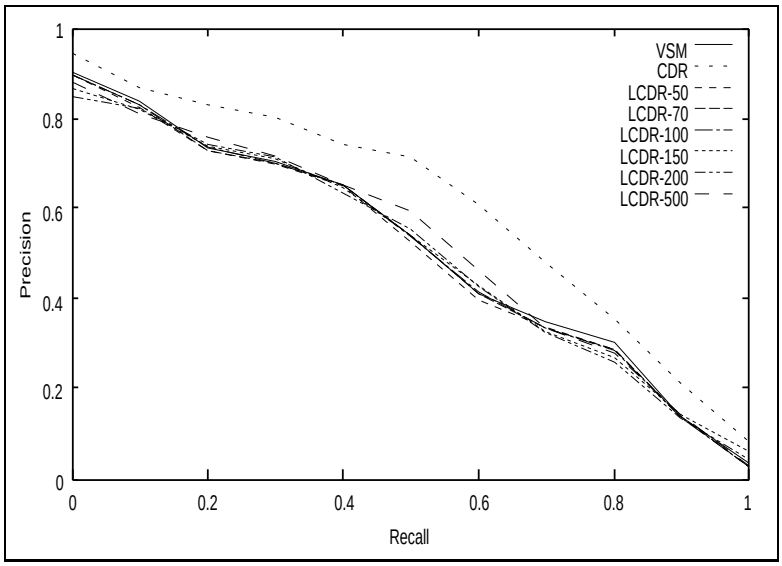

Figure 2: recall-precision curve using normalized LCR

the same as the precision of CDR. Moreover, in case that the dimension is 200, the precision of LCR is lower than the precision of CDR. By reducing the dimensionality of the term-document matrix, much of the noise is eliminated so that our system is able to extract related words more efficiently. However, the retrieval effectiveness of LCR is lower than using the convensional vector space model. Therefore, query expansion using LCR is unable to improve efficiency of searching results.

Figure 2 shows results of experiments using normalized LCR. In this figure, the precision of LCR is approximately the same as the precision in case of each dimension. In comparison with the precision of CDR, the precision of LCR is lower than that of CDR. As the reason of these results, It is possible that the weight of index words and the similarity veries by using the SVD.

4. Conclusion

In this paper, we propose LCR in consideration of the relevance between query and each index words in the document set. We construct IR system using the LCR and evaluate the information retrieval performance of this system by the MEDLINE data set.

As a result of this experiment, IR model using LCR improves the precision in comparison with the model using CDR on 100 dimensions and below. However, the precision using LCR is approximately the same as the precision using CDR on 150 dimensions and gets lower than the precision using CDR on 200 dimensions. By reducing the dimensionality of the term-document matrix, much of the noise is eliminated so that our system is able to extract related words more efficiently. However, the retrieval effectiveness of LCD is lower than using the convensional vector space model. Therefore, query expansion usin LCR is unable to improve efficiency of searching results.

Using the normalized LCR, the precision of LCR is approximately the same as the precision in case of each dimension. In comparison with the precision of CDR, the precision of LCR is lower than that of CDR. As the reason of these results, It is possible that the weight of index words and the similarity veries by using the SVD.

5. References

Berry, M. W., S. T. Dumais, and G. W. O'Brien, 1995. Using linear algebra for intelligent information retrieval. In *SIAM Review*, volume 37.

Chicholm, E. and T. G. Kolda, 1998. New term weighting formulas for the vector space method in information retrieval. Technical report, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Deerwester, S., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Korpimies, K. and E. Ukkonen, 1996. Searching for general documents. In *Proceedings of the 3rd International Conference on Frexible Query Answering Systems (FQAS98)*.

Kuriyama, Kazuko, 1998. Query expansion using thesauri. Technical report, IPSJ SIGNotes Fundamental Infology.

Lewis, D. D., 1991. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*.

Salton, G. and C. Buckley, 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297.

Witten, I. H., A. Moffat, and T. C. Bell, 1994. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York.