ments increase, which decrease the precision of the system. This problem can be solved by considering the strength of co-occurrency in irrelevant documents.

The weighting method for expanded query terms is based on the assumption that a term in relevant paragraphs that co-occurrate strongly with a initial query term is the closely related term of the query term in the topic. This method should be examined by comparing with the relevance weighting method to show its effectiveness.

As our future development, we will examine the above problems of our query term expansion method by comparing with the other query term expansion methods recently applied widely by other TREC participants.

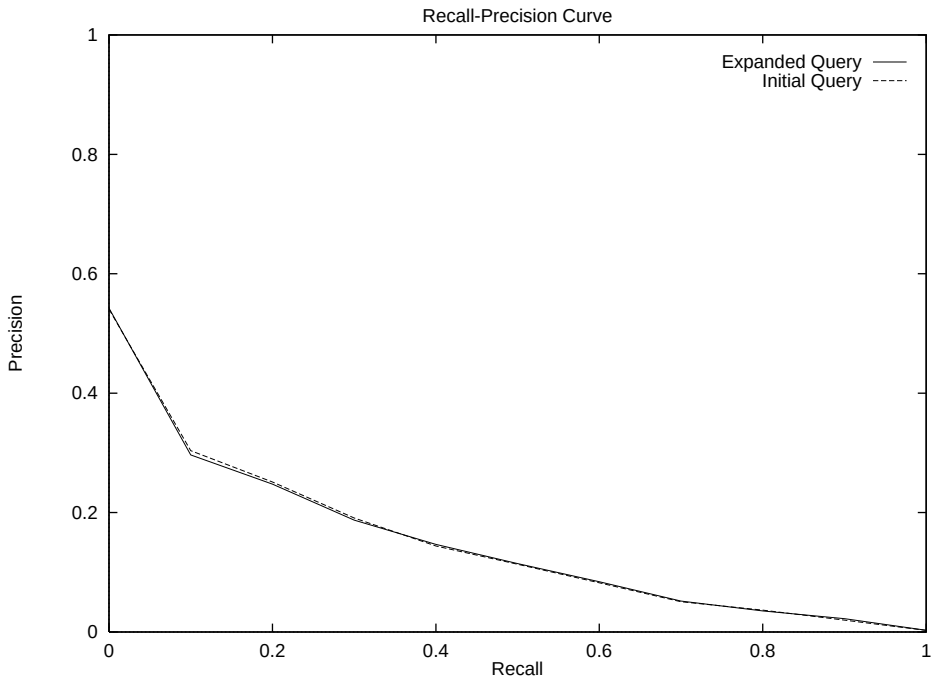
References

[1] Robertson, S.E. and Sparck Jones, K. Relevance weighting of search terms. Journal of the American Society for Information Science,27,pages 129-146,1976.

[2] http://potomac.ncsl.nist.gov/TREC/

[3] Kenji Satoh, Akitoshi Okumura, Kiyoshi YAMABANA. Information Retrieval System for TREC3. The Third Text REtrieval Conference(TREC-3),pages 311-318.

[4] Kenji Satoh, Susumu Akamine and Akitoshi Okumura. Improvements on Query Term Expansion and Ranking Formula. The Fourth Text REtrieval Conference(TREC-4),pages 475-481.



These values and curves show that there are small difference between the results obtained by using the initial queries and the expanded queries. For the TREC-6 routing test data, effectiveness of the query term expansion cannot be admitted from the above results.

6 Conclusion

We obtained results of our system by using two types of queries, the initial queries and the expanded queries for TREC-6 routing test data. However, effectiveness of the query term expansion for the TREC-6 routing test data cannot be admitted from these results. This results is the opposite of the another results, which we obtained by using training data.

The problem of the query term expansion for the TREC-6 routing test data is caused by the inexistence of expanded terms in test data extracted from training data.

A solution of this problem is to determine appropriate value of paragraphs(N_B) ,because this inexistence is caused by small value of $N_B = 15$used in query term expansion.

By increasing the value N_B, number of terms appeared frequently in irrelevant docu-

2. Expanded query generation

   (a) Extraction of paragraphs from relevant documents (to topics)

   (b) Ranking paragraphs by their score calculated from initial query

   (c) Extraction of noun phrases from the upper ranks of paragraphs

   (d) Selection of expanded query terms from noun phrases who co-occurred strongly with initial query terms in the upper ranks of paragraphs

   (e) Weighting of the expanded query terms according to the weights of initial query terms who co-occurred strongly with the expanded query terms

3. Ranking generation

   (a) Generation of the Routing Query by adding Initial Query and Expanded Query

   (b) Ranking of the documents according to their score calculated by routing query

5 Results

In this section, our results for TREC-6 routing test data, obtained by using two types of queries, the initial queries and the expanded queries, are compared.

Table 1. shows the values of Average Precision and R-precision calculated from the results obtained by using the initial queries and the expanded queries.

Table 1. Results of our systems

	Average Precision	R-precision
Initial Query	0.1337	0.1815
Expanded Query	0.1328	0.1823

Fig. 3. shows the Recall-Precision curve calculated from the results obtained by using the initial queries and the expanded queries.

Fig. 3. Results of our systems

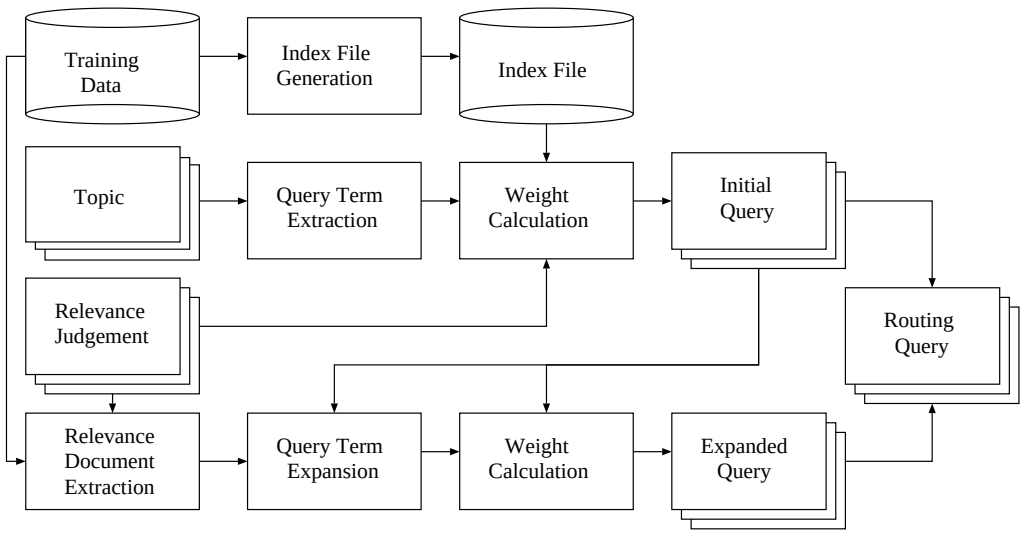


Fig.1 Qery Generation System

Fig.2 shows modules of the ranking generation system.

Here, test data are the initial data of the system. Index File Generation module generates index files for query term retrieval from test data.

Finally, Score Calculation module generates rankings of documents as output from routing queries and index files.

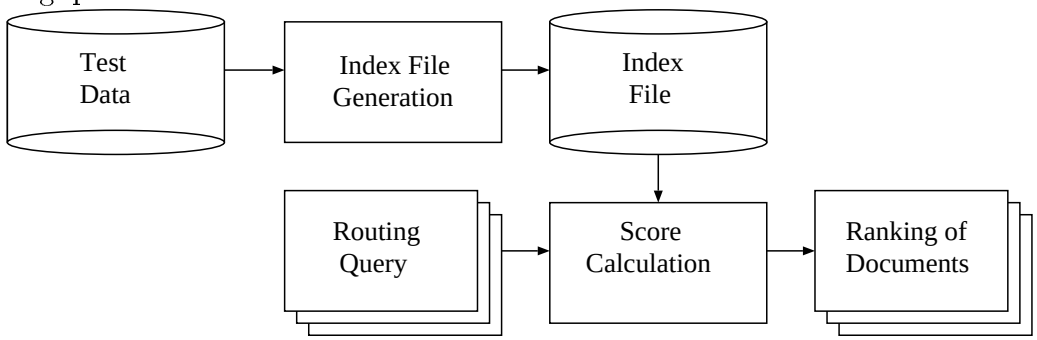


Fig.2 ranking generation system

4.2 Process Flow

Here, the process flow of the routing task is shown.

1. Initial query generation

   (a) Extraction of noun phrases from each topics as initial query terms

   (b) Weighting of query terms by formula (2.1)

Finally, the expanded query terms are weighted according to the weight value of the initial query terms which has the strongest co-occurrence with them. This weighting is based on the idea that an expanded query term which has strong co-occurrence with an initial query term is supposed to be the relational word(phrase) of the initial query term.

For the calculation of the weight of the expanded query term t^E, the weight value w_{t^I} of the initial query term t^I, which has the strongest co-occurrence with them, is used, i.e.

$$w_{t^E} = \beta w_{t^I}; t^I = argmax_{\{t^I\}} I(t^I, t^E), \tag{3.3}$$

where the parameter $\beta(\beta = 0.15)$ is introduced to control the overall weighting of the expanded query terms.

The values of the parameters $\beta and \gamma$ are determined to optimize the precision of the system, based on the rough estimation using training data.

4 Routing System

4.1 System Composition

Our system for routing task is composed of query generation system and ranking system. Fig.1 shows modules of query generation system.

Here, Training Data, Topic, and Relevance Judgment are initial data of the system.

First, Index File Generation module generates Index File for query term retrieval from training data.

Then, Query Term Extraction module and Weight Calculation module generate Initial Query as their output from Topic and Index File.

Next, Relevant Document Extraction module, Query Term Expansion module and Weight Calculation module generate Expanded Query as their output from Initial Query and Relevant Documents in Training Data.

Finally, Routing Query is obtained by adding Initial Query and Expanded Query.

where

N : the number of documents

n_i : the number of documents containing term i

R : the number of relevant documents

r_i : the number of relevant documents containing term i

Summation of weights w_i for the terms appeared in the document is calculated as the score, which is used to determine the ranking of relevant documents.

3 Expanded Queries Generation

Expanded queries are generated by extraction of expanded query terms and calculation of weight for each term as in the initial query generation process.

First, query terms are extracted from the upper ranked paragraphs of all the documents in training data. Ranking of paragraphs is determined according to the scores that are calculated by using initial queries for each paragraph. Here, proper noun phrases are extracted as expanded terms from the paragraphs, in the same way as in the initial query generation process.

Next, the extracted terms are selected as the extracted query terms which have strong co-occurred with the initial query terms in the upper ranked paragraphs. In this selection, the following mutual information is used to enumerate the strength of the co-occurrence. The mutual information $I(t^I, t^E)$ of the extracted term t^E with the initial query term t^I in the N_B paragraphs is;

$$I(t^I, t^E) = \log \frac{\frac{1}{N_B}\sum_{i=1}^{N_B} T_i(t^I)T_i(t^E)}{(\frac{1}{N_B}\sum_{i=1}^{N_B} T_i(t^I))(\frac{1}{N_B}\sum_{i=1}^{N_B} T_i(t^E))}, \tag{3.2}$$

where $N_B(=15)$ is the number of paragraph, and the function $T_i(t)$ is restricted to take value 1 or 0, according to the term t appears in the paragraph i or not.

After the calculation of the mutual information $I(t^I, t^E)$ of the extracted term t^E with all the initial query terms $\{t^I\}$, the extracted term t^E is selected as the expanded query term if it satisfies the condition $max_{\{t^I\}} I(t^I, t^E) > \gamma, here(\gamma = 2.0)$ is used.

To overcome this problem, query term expansion methods are known to be effective, which are query term generation methods that acuire expansion terms from thesaurus or from co-occurrent terms with initial query terms in the relevant documents.

Recently, we studied the method of extracting terms that co-occurred with initial query terms in relevant paragraphs as query term expansion method.

In our methods, paragraphs in the relevant documents are lanked by using initial query to extract terms from the upper ranked paragraphs with using term co-occurrence.

Compared to the methods using term co-occurrence in documents, the methods using term co-occurrence in paragraphs have both advantage and disadvantage: highly accurate treatment of term co-occurrence and huge amount of calculation for each combination of terms in paragraphs.

Our method eases the difficulty by ranking paragraphs with the initial query. Without using term co-occurrence in paragraphs, we could achieve the highly accurate treatment of term co-occurrence by small calculation.

In this paper, we report on our information retrieval system for English documents with using the data of TREC-6 [2] routing task.

First, we explain the query generation methods applied in our system, basic(initial) query generation method by using relevance weighting model [1] and the expanded query generation method by using query term expansion method.

Next, we explain the composisition of modules and process flow of our system.

Finally, we present our results and conclusion of the system using TREC-6 data.

2 Initial Query Generation

Initial queries are generated by extracting query terms from each topic and calculating weight of each term. Here, proper noun phrases are used as query terms, which are extracted from each topic by using dictionary of parsing system.

Term weight is calculated from term frequencies in training data. In the formula, the weight w_i is used for each query term i as shown below [1];

$$w_i = \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}, \tag{2.1}$$

Query Term Expansion based on Paragraphs of the Relevant Documents

Kai Ishikawa Kenji Satoh Akitoshi Okumura

C&C Media Research Labs.

NEC Corp.

Abstract

Recently, we studied the method of extracting terms that co-occurred with initial query terms in relevant paragraphs as query term expansion method. In our methods, paragraphs in the relevant documents are lanked by using initial query to extract terms from the upper ranked paragraphs with using term co-occurrence.

Our method eases the difficulty by ranking paragraphs with the initial query. Without using term co-occurrence in paragraphs, we could achieve the highly accurate treatment of term co-occurrence by small calculation.

The results of our system for TREC-6 routing test data, obtained by using the expanded queries generated by our query term generation method are compared with the results obtained by initial queries.

1 Introduction

In the field of information retrieval, methods of generating query with extracting query terms from topics(queries written in natural language) and calculating weight from term frequencies [1] are established.

Whereas in case a topic has few clear keywords or it is written in short description, these methods have a problem in composing accurate query with using the terms extracted only from the topic.