A Combined Query Expansion Approach for Information Retrieval

Hisao Imai hisao@is.s.u-tokyo.ac.jp
Nigel Collier nigel@is.s.u-tokyo.ac.jp
Jun'ichi Tsujii tsujii@is.s.u-tokyo.ac.jp
Department of Information Science, Graduate School of Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

1 Introduction

In this paper we aim to contribute to meeting the information access needs of biology researchers who wish to retrieve abstracts from MEDLINE. The volume of research papers is growing daily due to factors such as increased research into the online database the genome and also a world-wide move to migrate information to online sources. We consider that one way to do this is by exploring the combination of query expansion methods.

2 Method

Currently there exist several methods of query expansion, such as local relevance feedback [2] and thesaurus expansion [1, 3]. However, so far few studies have explored the combination of these approaches which use different knowledge sources. In this paper we apply two query expansion methods in sequence to reformulate the query so that it will suit to the user's needs more appropriately. One method we applied is similarity thesaurus based expansion [3], and the other is local feedback method.

The similarity thesaurus we use, based on [3], calculates the relevance between terms and queries and is constructed by interchanging the role of documents and terms in retrieval model. The relevance of a term in the similarity thesaurus to the concept of the query is the sum of the weighted relevance of the term to each term in the query. The queries are expanded by adding top n relevant terms, which are most similar to the concept of the query, rather than selecting terms that are similar to the query terms.

The Local feedback method is similar to traditional relevance feedback method [2], which modifies queries by using the result of the initial retrieval, except that the latter uses the judgment set for calculating re-weighting while the former assumes that the terms in the top ranked n documents are relevant to the user's request. Queries are expanded by adding the weight of terms in relevant documents and reducing the weight of terms in last m documents of the initial retrieval.

We modify the traditional Rocchio expansion equation to include the query expanded by the thesaurus method and to include negative evidence from the lowest ranked documents rather than non-relevant documents. The new query Q_{new}, including thesaurus expansion, can be defined as the following:

$$Q_{new} = \alpha_1 Q_{org} + \alpha_2 Q_{te} + \beta \sum_{top} D_i - \gamma \sum_{last} D_j$$

Here, Q_{org} is a initial query, Q_{te} is a query expanded by the similarity thesaurus based method, $\sum_{top} D_i$ represents terms in top ranked documents retrieved in the initial run, and $\sum_{last} D_j$ is terms in low ranked documents. The parameters α_1, α_2, β and γ represent the importance of each item. Currently, these parameters are given by human experientially. For the initial retrieval, we used the queries expanded by thesaurus method.

Num. Terms	0				30				60			
Num. Top Doc	0	5	10	15	0	5	10	15	0	5	10	15
Avg. Precision	**0.551**	0.624	0.652	0.664	0.604	0.683	0.696	0.698	0.617	0.679	0.700	0.701
#% change	—	13.2	18.3	20.5	9.6	24.0	26.3	26.7	12.0	23.2	27.0	27.2

Figure 1: Improvement using expanded queries.

In this study, we set the parameters as following: $\alpha_1 = 1$, $\alpha_2 = 0.5$, $\beta = 0.6$, and $\gamma = 0.3$. We used last 40 documents as the non-relevant documents in the local feedback.

3 Results and Discussions

In our experiments, we used the MED test collection[1] for evaluation. This collection contains 1033 documents, which are MEDLINE abstracts, and 30 queries. This test collection is often used in researches for information retrieval. We evaluate the performance of the retrieval by average precision measure. Precision is the ratio of the number of relevant documents retrieved to the total number retrieved[2]. The average precision of a query is the average of precisions calculated when a relevant document is found in the rank list. All the query's average precisions are averaged to evaluate a experiment.

Fig. 1 shows the improvement of retrieval performance using expanded queries. The number in the row indexed "Num. Terms" is the number of terms added by the thesaurus method. The number in the second row, indexed "Num Top Doc", is the number of top ranked documents used in the local feedback procedure. The row "Avg. Precision" shows the average precision when the query is expanded using the above conditions. We can see that the baseline average precision, when using 0 terms and 0 documents in expansion, is 0.551. The bottom row shows the difference between the baseline result and the results using expansion.

Because the size of the test collection is small, we should be cautious of judging this result. Nevertheless, it suggests some trend. As Figure 1 shows, the performance of the combined expansion method is better than that of applying each method by itself. The thesaurus expansion improves the performance of initial retrieval, so the result of the retrieval in combination with the local feedback also improves. As for the local feedback retrieval with 15 top ranked documents, for instance, the average precision improved by 5.6% when we compare the retrieval results with 0 and 60 thesaurus terms. This implies that both methods can compensate for each other's weakness.

One of the weak points in this expansion model is that the result of the expansion depends on the parameters such as $\alpha_{1,2}$, β, etc., and it is difficult to predict the optimal value of parameters. How to determine the reliable values is our future work. In addition, we should apply this model to some larger test collection such as OHSUMED.

References

[1] Jing, Y. and Croft, W.B., An association thesaurus for information retrieval, *RIAO '94*, 146–160, 1994.

[2] Rocchio, J.J. Jr., *The SMART Retrieval System Experiments in Automatic Document Processing*, Chapter: Relevance Feedback in Information Retrieval, 313–323, Prentice Hall, 1971.

[3] Qiu, Y. and Frei, H.P., Concept based query expansion, *SIGIR '93*, 160–169, 1993.

[1] MED test collection is available via ftp://ftp.cs.cornell.edu/pub/smart/med.

[2] $Precision = \frac{\{numberofrelevantdocumentsretrieved\}}{\{numberofdocumentsretrieved\}}$. If a relevant document is not retrieved, the precision is 0.