Merging Results by Using Predicted Retrieval Effectiveness

Wen-Cheng Lin and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN
denislin@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract. In this paper we proposed several merging strategies to merge the result lists of each intermediate runs in distributed MLIR. The prediction of retrieval effectiveness was used to adjust the similarity scores of documents in the result lists. We introduced three factors affecting the retrieval effectiveness, i.e., the degree of translation ambiguity, the number of unknown words and the number of relevant documents in a collection for a given query. The results show that the normalized-by-top-*k* merging with translation penalty and collection weight outperforms the other merging strategies except the raw-score merging.

1   Introduction

Multilingual Information Retrieval abbreviated as MLIR facilitates the uses of queries in one language to access documents in various languages. Most of the previous approaches (Oard and Diekema, 1998) focused on how to unify the language usages in queries and documents. The adaptation of traditional information retrieval systems has been considered. Query translation and document translation methods have been introduced. The resources used in the translation have been explored.

In real world, multilingual document collections are distributed from various resources, and managed by information retrieval system of various architectures. How to integrate the results from heterogeneous resources is one of the major issues in MLIR. Merging result lists of individual languages is a commonly adopted approach. Document collections of each language are indexed and retrieved separately, and the result lists of each document collection are merged into a multilingual result list. The goal of result lists merging is to include as more relevant documents as possible in the final result list and to make relevant documents have higher ranks. Several attempts were done on this problem (Peters, 2002). The simplest merging method is *raw-score merging*, which sorts all the documents by their original similarity scores, and then selects the top ranked documents. The second approach, *round-robin merging*, interleaves the results of each run based on the rank of each document. The third approach is *normalized-score merging*. For each topic, the similarity score of each document is divided by the maximum score in each result list. After adjusting scores, all results are put into a pool and sorted by the normalized score.

Lin and Chen (2002a, 2002b) proposed *normalized-by-top-k merging* to avoid the drawback of normalized-score merging. Translation penalty is also considered during merging result lists. The performance of normalized-by-top-*k* with translation penalty is similar to that of raw-score merging. Moulinier and Molina-salgado (2002) proposed collection-weighted normalized score to merge result lists. The normalized collection score is used to adjust the similarity score between a document and a query. Collection score only reflects the similarity of a (translated) query and a document collection. This method could fail if a query is not translated well. Savoy (2002) used logistic regression to predict the relevance probability of documents according to the document score and the logarithm of the rank. This method does not consider the quality of query translation either. Furthermore, the relationship between the rank and the relevance of a document is not strong. Braschler, Göhring and Schäuble (2002) proposed feedback merging that interleaves the results according to the propositions of the predicted amount of relevant documents in each document collection. The amount of relevant information was estimated by the portion of overlap between the original query and the ideal query constructed from the top ranked documents. The experimental results showed that feedback merging had little impact.

In this paper, we will explore several merging strategies. The basic idea of our merging strategies is adjusting the similarity scores of documents in each result list to make them more comparable and to reflect the confidence in retrieval effectiveness. We assume that the importance of each intermediate run depends on their retrieval performance. We introduced three factors affecting the retrieval effectiveness, i.e., the degree of translation ambiguity, the number of unknown words and the number of relevant documents in a collection for a given query. The rest of this paper is organized as follows. Section 2 describes our merging strategies. Section

3 shows the IR model and query translation technique. Section 4 discusses the experimental results. Section 5 concludes the remarks.

2   Merging Strategies

We aim to include as more relevant documents as possible in the final result list and to make relevant documents have higher ranks during merging. If a result list contains many relevant documents in the top ranks, i.e., it has good performance, the top ranked documents should be included in the final result list. On the other hand, if a result list has few or even no relevant documents, the final result list should not contain many documents from this list. Thus, the higher performance an individual run has, the more important it is. However, without the priori knowledge of a query in advance, it is challenging to predict the performance of an individual run for each document collection. The similarity score between a document and a query is one of a few clues that are common used. A document with higher similarity score seems to be more relevant to a specific query. Because document collections are various and the underlying IR systems may be different, the similarity scores of a query with different collections cannot be compared directly. The basic idea of our merging strategies is adjusting the similarity scores of documents in each result list to make them more comparable and to reflect the confidence in retrieval effectiveness. The characteristics of the underlying IR model, the effects of the query translation and the statistics of individual document collection will be addressed in the following subsections.

2.1   Normalized by Top K

Similarity scores reported by different information retrieval systems may differ a lot from each other. In vector-based IR model, the similarity score defined by cosine formula ranges from 0 to 1, but the score may be much larger than 1 when Okapi system (Robertson, *et al.*, 1998) is used. It is obvious that the scores cannot be compared directly. Thus, similarity scores have to be normalized to the same range to make them comparable at the first step. The approach of normalized-score merging maps the similarity scores of different result lists to the values within the same range. The major drawback is: if the maximum score is much higher than the second one in the same result list, the normalized-score of the document at rank 2 would be made lower even if its original score is high. Thus, the final rank of this document might be lower than that of the top ranked documents with similar original scores in another result list. A revised score normalization method is proposed as follows. The original score of each document is divided by the average score of top *k* documents instead of the maximum score. We call this *normalized-by-top-k* approach.

2.2   Translation Penalty

Similarity score reflects the degree of similarity between a document and a query. A document with higher similarity score seems to be more relevant to the desired query. However, if the query is not formulated well, e.g., inappropriate translation of a query, a document with high score may still not meet users' information needs. When the result lists are merged, those documents that have high, but incorrect scores should not be included in the final result list. Thus, the effectiveness of each individual run has to be considered in the merging stage.

When query translation method is used to deal with the unification of language usages in queries and documents, queries are translated into target language and then the target language documents are retrieved. We can predict the multilingual retrieval performance based on the translation quality. Intuitively, using English to access English collection is expected to have better performance than using it to access other collections. Similarly, using a bilingual dictionary of more coverage is expected to be better than using dictionary of less coverage. Less ambiguous queries have also higher tendency to achieve better translation than more ambiguous queries. Normalization in Section 2.1 just reflects the same comparison basis, but does not consider the above issues. Two factors, i.e., the degree of translation ambiguity and the number of unknown words, are used to model the translation performance. For each query, we compute the average number of translation equivalents of query terms and the number of unknown words in each language pair, and use them to compute the weights of each cross-lingual run. The following formula is proposed to determine the weights.

$$W_i = c_1 + \left[c_2 \times \left(\frac{51 - T_i}{50} \right)^2 \right] + \left[c_3 \times \left(1 - \frac{U_i}{n_i} \right) \right] . \tag{1}$$

where W_i is the merging weight of query *i* in a cross-lingual run,
T_i is the average number of translation equivalents of query terms in query *i*,
U_i is the number of unknown words in query *i*,
n_i is the number of query terms in query *i*, and
c_1, c_2 and c_3 are tunable parameters, and $c_1+c_2+c_3=1$.

2.3 Collection Weight of Individual Document Collection

How many relevant documents there are in a collection for a given query is also an important factor for measuring retrieval effectiveness. If a document collection contains more relevant documents, it could have more contribution to the final result list. Callan, *et al.* (1995) proposed CORI net to rank distributed collections of the same language for a query. Moulinier and Molina-salgado (2002) used collection score to adjust the similarity score between a document and a query.

In our approach, a collection weight of a document collection with a query is defined as follows. A document collection is viewed as a huge document and represented as a collection vector. The *i*th element in a collection vector is the document frequency *df* of the *i*th index term in the collection. Similarly, the *i*th element in a query vector is the frequency of the *i*th index term in the query. Since document collections are in different languages, we do not use inverse collection frequency *icf,* which is analogous to *idf.* Cosine similarity formula shown as follows is used to compute the collection weight, which is included to the merging weight.

$$W_i^{'} = W_i + c_4 \times CW_i \tag{2}$$

$$CW_i = \frac{\sum_{j=1}^{m} qtf_{ij} \times df_j}{\sqrt{\sum_{j=1}^{m} qtf_{ij}^2} \times \sqrt{\sum_{j=1}^{m} df_j^2}} \tag{3}$$

where W_i' is the new merging weight of query i in an intermediate run,
W_i is the merging weight described in Section 2.2,
CW_i is the collection weight of a target collection for query i,
c_4 is tunable parameter,
qtf_{ij} is the term frequency of index term j in query i,
df_j is the document frequency of index term j in a target collection, and
m is the number of index terms.

2.4 Predicting Retrieval Effectiveness by Linear Regression

Now three factors, i.e., the degree of translation ambiguity, the number of unknown words and the number of relevant documents for a given query in a document collection, are proposed to determine the retrieval effectiveness. We use linear regression to predict the retrieval effectiveness according to the three factors. The original similarity score of a document is normalized by normalized-by-top-*k* method, and the score of the predicted precision is added to the normalized score. Documents from all collections are sorted according to the adjusted similarity scores and the top ranked documents are reported.

3 Query Translation and Document Indexing

In the experiments, Okapi IR system was adopted to index and retrieve documents. The weighting function was BM25 (Robertson, *et al.*, 1998). The document set used in CLEF 2003 Small-multilingual task consists of English, French, German and Spanish. The numbers of documents in English, French, German and Spanish document sets are 169,477, 129,806, 294,809 and 454,045, respectively. The <HEADLINE> and <TEXT> sections in English documents were used for indexing. For Spanish documents, the <TITLE> and <TEXT> sections were used. While indexing French and German documents, the <TITLE>, <TEXT>, <TI>, <LD> and <TX> sections were used. The words in these sections were stemmed, and stopwords were removed. All letters were transformed to the lower cases. We adopted stopword lists and stemmers developed by University of Neuchatel[1] (Savoy, 2001).

English queries were used as the source language queries and translated into target languages, i.e., French, German and Spanish. A dictionary-based approach was adopted. For each English query term, we found its translation equivalents by looking up a dictionary. The first two translation equivalents with the highest occurrence frequency in the target language documents were considered as the target language query terms. If a query term does not have any translation equivalents, the original English query term was reserved in the translated query. The dictionaries we used are Ergane English-French, English-German and English-Spanish dictionaries. They are available at http://www.travlang.com/Ergane.

---

4 Experiments

We submitted four runs in CLEF 2003 Small-multilingual task. All runs used title and description fields. The details of each run are described in the follows.

1. NTUm4Topn

   The result lists were merged by normalized-by-top-k merging strategy. The average similarity score of the top 100 documents were used for normalization.

2. NTUm4TopnTp

   In this run, translation penalty was considered. The similarity scores of each document were first normalized by the average similarity score of the top 100 documents and then multiplied a weight determined by formula (1). The values of c_1, c_2 and c_3 were 0, 0.4 and 0.6, respectively. In query translation, an English query term which did not have any translation equivalents by dictionary lookup was also used to retrieve target language documents. If such an English term occurs in the target language documents, it can be viewed as a word similar to the other translated words when the merging weight is computed. Table 1 lists the number of English query terms that do not have translation, but occur in target language collection.

3. NTUm4TopnTpCw

   In this run, the collection weight was also considered. We used formula (2) to adjust the similarity score of each document. The values of parameters were the same as run NTUm4TopTp. The value of c_4 was 0.5.

4. NTUm4TopnLinear

   We used linear regression to determine the weights of the three variables, including the average number of translation equivalents of query terms, the portion of unknown words in a query and the collection weight of the target collection, to predict the performances of each intermediate run. CLEF 2001 and 2002 test sets were used as training data to estimate the parameters. The original similarity score of a document was normalized by normalized-by-top-k method, and the score of the predicted precision was added to the normalized score.

The results of official runs are shown in Table 2. To compare the effectiveness of our approaches with the past merging strategies, we also conducted several unofficial runs that used raw-score merging, normalized-score merging and round-robin merging strategies. The average precision of optimal merging proposed by Chen (2002) was regarded as an upper-bound, which was used to measure the performances of our merging strategies. The performances of the unofficial runs are also shown in Table 2. The performance relative to optimal merging is enclosed in parentheses. The results show that the performances of the merging strategies we proposed were better than normalized-score merging and round-robin merging, but worse than raw-score merging. The experimental results in CLEF 2002 and NTCIR3 (Lin and Chen, 2002a, 2002b) show that normalized-by-top-k merging overcomes the drawback of normalized-score merging. In CLEF 2003, the performance of normalized-by-top-k merging was still better than normalized-score merging. The performance dropped down slightly after considering translation penalty. From Table 3, the performance of English-Spanish run was worse than the other intermediate runs, but the merging weights of three cross-lingual runs were similar. This is because that the average number of translation equivalents and the number of unknown words of three cross-lingual runs did not differ too much. After considering the collection weights of each document collection, the performance was improved and was about 7.12% increase to normalized-by-top-k merging. The performance of using the merging weight predicted by linear regression was slightly better than normalized-by-top-k merging, but worse than normalized-by-top-k with translation penalty and collection weight.

Table 1. Number of English query terms without translation but in target language corpora

Language	French	German	Spanish
# query terms	891	891	891
# query terms without translation equivalents	326	322	251
# query terms without translation equivalents but in target language corpora	209 (64.11%)	230 (71.43%)	147 (58.57%)

Table 2. Performances of merging strategies

Run	Average precision
NTUm4Topn	0.1489 (60.97%)
NTUm4TopnTp	0.1478 (60.52%)
NTUm4TopnTpCw	0.1595 (65.32%)
NTUm4TopnLinear	0.1516 (62.08%)
Raw score merging	0.1691 (69.25%)
Normalized score merging	0.1366 (55.94%)
Round-robin merging	0.1412 (57.82%)
Optimum merging	0.2442

Table 3. Performances of intermediate runs

Run	# Topic	Average precision
English -> English	54	0.5063
English -> French	52	0.2568
English -> German	56	0.2574
English -> Spanish	57	0.0797

5   Conclusion

Merging problem is critical in distributed multilingual information retrieval. In this paper, we proposed several merging strategies to integrate the result lists of collections in different languages. We assume that the importance of each intermediate run depends on their retrieval performance. We introduced three factors affecting the retrieval effectiveness, i.e., the degree of translation ambiguity, the number of unknown words and the number of relevant documents in a collection for a given query. Normalized-by-top-k avoids the drawback of normalized-score merging. The experimental results show that considering translation penalty and collection weight improves performance. We also used linear regression to predict the retrieval effectiveness. The performance of using the merging weight predicted by linear regression is similar to normalized-by-top-k. The performances of our merging strategies were better than normalized-score merging and round-robin merging, but were worse than raw-score merging in single IR system environment. However, raw-scoring merging is not workable if different information retrieval systems are adopted.

References

Braschler, M., Göhring, A. & Schäuble, P. (2002). Eurospider at CLEF 2002. In Peters, C. (Ed.), *Working Notes for the CLEF 2002 Workshop*, (pp. 127-132).

Callan, J.P., Lu, Z. & Croft, W.B. (1995). Searching Distributed Collections With Inference Networks. In Fox, E.A., Ingwersen, P. & Fidel, R. (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 21-28).

Chen, A. (2002). Cross-language Retrieval Experiments at CLEF-2002. In Peters, C. (Ed.), *Working Notes for the CLEF 2002 Workshop*, (pp. 5-20).

Lin, W.C. & Chen, H.H. (2002a). NTU at NTCIR3 MLIR Task. In Kishida, K. & Ishida, E. (Eds.), *Working Notes of the Third NTCIR Workshop Meeting. Part II: Cross Lingual Information Retrieval Task*, (pp. 101-105). Tokyo, Japan: National Institute of Informatics.

Lin, W.C. & Chen, H.H. (2002b). Merging Mechanisms in Multilingual Information Retrieval. In Peters, C. (Ed.), *Working Notes for the CLEF 2002 Workshop*, (pp. 97-102).

Moulinier, I. & Molina-salgado H. (2002). Thomson Legal and Regulatory experiments for CLEF 2002. In Peters, C. (Ed.), *Working Notes for the CLEF 2002 Workshop*, (pp. 91-96).

Oard, D. and Diekema, A. (1998) Cross-Language Information Retrieval. *Annual Review of Information Science and Technology*, vol. 33, (pp. 223-256).

Peters, C. (Ed.) (2002). *Working Notes for the CLEF 2002 Workshop*.

Robertson, S.E., Walker, S. & Beaulieu, M. (1998). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In Voorhees, E.M. & Harman, D.K. (Eds.), *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, (pp. 253-264). Gaithersburg, MD: National Institute of Standards and Technology.

Savoy, J. (2001). Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In Peters, C., Braschler, M., Gonzalo, J. & Kluck, M. (Eds.), *Evaluation of Cross-Language Information Retrieval Systems*, (pp. 27-43). Lecture Notes in Computer Science, Vol. 2406. Springer.

Savoy, J. (2002). Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In Peters, C. (Ed.), *Working Notes for the CLEF 2002 Workshop*, (pp. 31-46).