Dublin City University at CLEF 2005: Multilingual Merging Experiments

Adenike M. Lam-Adesina, Gareth J. F. Jones
School of Computing, Dublin City University, Dublin 9, Ireland
{adenike,gjones}@computing.dcu.ie

Abstract

This year the Dublin City University group participated in the CLEF 2005 Multilingual merging task. We tested different a range of standard merging techniques for merging the provided ranked result lists and show that the success of these techniques can sometimes be dependent on the retrieval system used.

Categories and Subject Descriptors

H.3 Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; H3.4 Systems and Software – Distributed systems; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Multilingual information retrieval, Retrieved list merging

1 Introduction

Multilingual information retrieval (MIR) refers to a process of retrieving relevant documents from collections in different languages in response to a user request in a single language. Standard approaches to MIR involve either translating the topics into the document languages or the document collections into the expected topic language. In CLEF 2003 we showed that translating the document collections into the query language using a standard machine translation system and then merging them to form a single collection, can result in better retrieval performance than translating the topics [1]. However, this method is not always practical, particularly if the collection is very large or the translation resources are limited. For the second method whereby the topics are translated, the topics are used to retrieve ranked lists of potentially documents from the separate collections. These result lists then need to be merged together to form a single ranked list for the system output. The different statistics of the individual collections and the varied topic translations mean that the scores of documents in the separate lists will generally be incompatible, and thus that merging is a non-trivial problem. The CLEF 2005 Multilingual merging task aims to encourage researchers to focus directly on the merging problem, since merging strategies explored previously for multilingual retrieval tasks at CLEF and elsewhere have generally produced disappointing results. Previously work on multilingual merging has been combined with the document retrieval stage, the idea of the CLEF merging task is to explore the merging of provided precomputed ranked lists to enable direct comparison of the behaviour of merging strategies between different retrieval systems.

Many different techniques for merging separate result lists to form a single list have been proffered and tested in recent years. All of the techniques suggest that making an assumption that the distribution of relevant documents in the results sets of retrieval from individual collections is similar is not true [2]. Hence, straight merging of relevant documents from the sources will result in poor combination. However, none of the proposed more complex merging techniques have really been demonstrated to be consistently effective.

For our participation in the merging track at CLEF 2005 we applied a range of standard merging strategies to the two provided sets of ranked lists. Our aim was to compare the behaviour of these methods for the two sets of ranked documents in order to learn something about concepts that might be consistently useful or poor when merging ranked lists.

This paper is organized as follows: Section 2 overviews the merging techniques explored in this paper, Section 3 gives our experimental results, and Section 4 draws conclusions and considers strategies for further experimentation.

2 Merging Strategies

The aim of a merging strategy is typically to include as many relevant documents at the highest ranks in the merged list. This section overviews the merging strategies used in our experiments. The basic idea is to modify the scored weight of each document to take account of some aspect of the maximum and minimum values of the matching scores or the distribution of scores in the lists to improve the compatibility of scores to form a more effective ranked list. The schemes used in our experiments were as follows:

$$p = doc\_wgt \tag{1}$$

$$t = doc\_wgt * rank \tag{2}$$

$$d = \frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt} \tag{3}$$

$$r = (\frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt}) * rank \tag{4}$$

$$q = (\frac{doc\_wgt - g\min\_wt}{g\max\_wt - g\min\_wt}) * rank \tag{5}$$

$$b = \frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt * rank} \tag{6}$$

$$m1 = (\frac{doc\_wgt - gmean\_wt}{gstd\_wt}) + (\frac{gmean\_wt - g\min\_wgt}{gstd\_wt}) \tag{7}$$

$$m2 = (m1) * rank \tag{8}$$

where *p, d, r, q, b, m1 and m2* are the new document weight for all document in all collections and corresponding results are labelled * where * can be *p, d, r, q, b, m1 and m2* depending on the merging scheme used

doc_wgt = the initial document weight
gmax_wt = the global maximum weight, i.e the highest document weight from all collections for a given query
gmin_wt = the global minimum weight, i.e the lowest document weight from all collections for a given query
gmean_wt = the global median weight, i.e the mean document weight from all collections for a given query

$$gmean\_wt = \frac{\sum_{i=0}^{n} doc\_wgt1_i}{totdocs}$$

max_wt = the individual collection maximum weight for a given query
min_wt = the individual collection minimum weight for a given query
rank = a parameter to control the effect of size of collection - a collection with more documents gets a higher rank (value ranges between 1.5 and 1).

Method *p* is used as a baseline using the raw document scores from the retrieved lists without modification. A useful merging scheme should be expected to improve on the performance of the *p* scheme. The *rank* was adjusted using the 20 training topics provided for the merging task.

3 Experimental results

Results for our experiments using these merging schemes are shown in Tables 1 and 2. Our official submissions to CLEF 2005 are marked *.

Run-id	P10	% chg.	P30	% chg.	MAP	% chg.	Rel. Ret.	chg.
dcu.hump*	0.5175	-	0.3958	-	0.2086	-	2982	-
dcu.humd	0.3725	-28.0	0.3467	-12.4	0.1775	-14.9	2965	-17
dcu.humr	0.4550	-12.1	0.3642	-8.0	0.1932	-7.4	2964	-18
dcu.humq	0.4575	-11.6	0.3633	-8.2	0.2005	-3.9	2752	-230
dcu.humb	0.3200	-32.2	0.2925	-26.1	0.1596	-23.5	2950	-32
dcu.humt*	0.4075	-21.3	0.3275	-17.3	0.1734	-16.9	2442	-540
dcu.humm1*	0.4800	-7.2	0.3817	-3.6	0.1988	-4.7	2873	-109
dcu.humm2*	0.4650	-10.1	0.3625	-8.4	0.1846	-11.5	2846	-136

Table 1: Merging results using the provided Hummingbird ranked lists.

Run-id	P10	% chg.	P30	% chg.	MAP	% chg.	Rel. Ret.	chg.
dcu.Prositqgp*	0.4500	-	0.4458	-	0.3103	-	4404	-
dcu.Prositqgd	0.4850	+7.7	0.4442	-0.4	0.2931	-5.5	4552	+148
dcu.Prositqgr	0.4950	+10.0	0.4458	0.0	0.3011	-3.0	4544	+140
dcu.Prositqgq	0.4650	+3.3	0.4462	+0.1	0.3192	+2.9	4469	+65
dcu.Prositqgb	0.4725	+5.0	0.4408	-1.1	0.2834	-8.7	4538	+134
dcu.Prositqgt*	0.4600	+2.2	0.4458	0.0	0.3201	+3.2	4477	+73
dcu.Prositqgm1*	0.4750	+5.6	0.4592	+3.0	0.3241	+4.5	4486	+82
dcu.Prositqgm2*	0.4700	+4.4	0.4608	+3.4	0.3286	+5.9	4512	+108

Table 2: Merging results using the provided Prosit ranked lists from the University of Neuchatel.

Tables 1 and 2 show merging results using runs provided by Hummingbird and the University of Neuchatel respectively. Results are shown for precision at cutoff of 10 and 30 documents, Mean Average Precision (MAP) and the total number of relevant documents retrieved. The raw score merging scheme *p* is taken as a baseline and changes for each scheme are shown for each data set with respect to the reported metrics.

The most obvious results are that the more complex merging schemes are shown in Table 2 to generally improve performance by a small amount for the Prosit data, but in Table 1 in all cases reduce performance for the Hummingbird data with respect to both the precision measures and the number of relevant retrieved. This appears to offer an answer to one of the questions associated with the CLEF merging task, namely whether the same merging techniques will always be found to be effective for different sets of ranked lists for a common merging task generated using alternative information retrieval systems. The reasons for this difference in behaviour need to be investigated. This analysis will hopefully provide insights into the selection of appropriate merging strategies or the development of merging strategies which will operate more consistently when merging different sets of ranked lists. There are some other observations of consistent behaviour which can be made be. It can be seen that there is no consistent relationship between the variation in precision measures and the number of relevant documents retrieved for the different merging schemes. Schemes with better precision can be accompanied by lower relevant retrieved and vice versa. This is most notable for the *b* results where good relevant retrieved (in relative terms) is accompanied by a large reduction in MAP for both data sets.

4 Conclusions

Results of our merging experiments for CLEF 2005 indicate that the behaviour of merging schemes varies for different sets of ranked lists. The reasons for this behaviour are not obvious and further analysis is planned to attempt to better understand this behaviour as a basis for the extension of these techniques for merging or the proposal of new ones.

References

[1] A.M.Lam-Adesina and G.J.F.Jones. Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval, Proceedings of the CLEF 2003 Workshop on Cross-Language Information Retrieval and Evaluation, Trondheim, Norway, 2003.

[2] Jacques Savoy, Report on CLEF-2003 Multilingual Tracks,Proceedings of the CLEF 2003 Workshop on Cross-Language Information Retrieval and Evaluation, Trondheim, Norway, 2003.