Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim

René Hackl, Ralph Kölle, Thomas Mandl, Alexandra Ploedt, Jan-Hendrik Scheufen, Christa Womser-Hacker

University of Hildesheim, Information Science, Marienburger Platz 22
D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract.
Fusion and optimization based on relevance judgements have proven to be successful strategies in information retrieval. In this year's CLEF campaign we applied these strategies to multilingual retrieval with four languages. Our fusion experiments were carried out using freely available software. We used the snowball stemmers, internet translation services and Lucene's and the new MySQL's text retrieval tools.

1 Introduction

In the CLEF 2002 campaign, we tested an adaptive fusion system based on the MIMOR model within the GIRT track (Hackl et al. 2002). For CLEF 2003, we applied the same model to multilingual retrieval with four languages. We chose English as our source language because most of the web based translation services offer translations to and/or from English. Our experiments were carried out fully automatically.

2 Fusion in Information Retrieval

Fusion in information retrieval delegates a task to different retrieval engines and considers all the results returned. The single result lists are combined into one final result. Fusion is motivated by the observation that many retrieval systems reach comparable quality, however, the overlap between their ranked lists is often low (Womser-Hacker 1997). The retrieval status values (RSV) are combined by taking the sum, the minimum or the maximum of the results from the individual systems. Linear combinations assign a weight to each method which determines its influence on the final result. These weights may be improved for example by heuristic optimization or learning methods (Vogt & Cottrell 1998).
There has been a considerable interest in fusion algorithms in several areas of information retrieval. In web information retrieval, for example, link analysis assigns an overall quality value to all pages based mainly on the number of links which point to that page (Henzinger 2000). This quality measure needs to be fused with the retrieval ranking based on the document's content (e.g. Plachouras & Ounis 2002). Fusion is also investigated within image retrieval for the combination of evidences which stem from different representations like color, texture, and forms. In XML retrieval fusion is necessary to combine the ranks assigned to a document by the structural analysis and the content analysis (Fuhr & Großjohann 2001).

3 MIMOR as Fusion Framework

MIMOR (Multiple Indexing and Method-Object Relations) represents a learning approach to the fusion task which is based on results of information retrieval research which show that the overlap between different systems is often small (Womser-Hacker 1997, Mandl & Womser-Hacker 2001). Furthermore, relevance feedback is considered a very promising strategy for improving retrieval quality. As a consequence, the linear combination of different results is optimized through learning from relevance feedback. MIMOR represents an information retrieval system managing poly-representation of queries and documents by selecting appropriate methods for indexing and matching (Mandl & Womser-Hacker 2001). By considering user feedback about the relevance of documents, the model learns and adapts itself by assigning weights to the different basic retrieval engines. MIMOR can also be individualized, however, such personalization in information retrieval is difficult to evaluate within evaluation initiatives. MIMOR could train an individual or group based optimization of the fusion. However, in evaluation studies, a standardized notion of relevance exists.

4 CLEF Retrieval Experiments with MIMOR

The tools we employed this year include Lucene 1.3[1], MySQL 4.0.12[2] and Java™-based snowball[3] analyzers. Most of the data pre-processing was carried out by Perl-scripts. In a first step, customized snowball stemmers were used to stem the collections. Stopwords were also eliminated[4]. Then, the collections were indexed by Lucene and MySQL. Lucene needed less than half the time that MySQL needed for indexing the collections of 1321 MB. A second step involved the translation of the English topics into French, German and Spanish. The translation was carried out with the free internet services FreeTranslation, Reverso and Linguatec[5].
The decision to select these tools, was based on a heuristic evaluation of several services. The queries of CLEF 2001 were used to gather data for a comparison of the translations. Examining the different translations, it became apparent that the quality of the machine translations is certainly not quite satisfying. At the same time, the translation systems usually exhibited different weaknesses. Because of that, we decided to use more than one translation system and merge the results. The tools which performed best and showed significantly different results at our evaluation were chosen.
The topics were also stemmed with snowball and stopwords were removed. The translated and processed queries for each language were then merged by joining the three translations while eliminating dublettes. We did not try to identify any phrases.

Table 1. Results of the test runs

	Number of retrieved multilingual documents	Average precision	Average document precision
Data from 2001			
Lucene	5167 / 6892	0.2880	0.3248
MySQL	2873	0.1037	0.1359
1:1 merged	3975	0.1856	0.2206
4:1 merged	4984	0.2673	0.3094
9:1 merged	5101	0.2830	0.3248
17:3 merged	5056	0.2764	0.3189
Data from 2002			
Lucene	4454 / 6996	0.2876	0.2769
MySQL	2446	0.0913	0.0951
9:1 merged	4543	0.2851	0.2762
17:3 merged	4533	0.2787	0.2709
7:1 merged	**4553**	0.2822	0.2742
33:7 merged	4511	0.2740	0.2670

Before working on the official runs, both retrieval systems employed were tested. Using the data (collections and relevance assessments) from 2001 we carried out several runs. Despite their dissimilar stand-alone performances, the systems were granted equal weights for the fusion process at first (1:1). After four runs, the weights strongly favoured Lucene and we went on experimenting with the 2002 data. The peak performance of the fusion was reached at a ratio of 7:1 (= 0.875:0.125) favouring Lucene's results. This suggests that some of MySQL's best relevant results helped the overall precision. (cf. table 1). Despite the low retrieval quality of MySQL, it still contributed to the fusion. Note, however, that we did not include the Italian collections and that we used the "perfect", that is, monolingual, queries in our tests, so there may be some bias. Italian was part of the 2001 and 2002 campaign, but it is not part of the multilingual-4 track in CLEF 2003.

[1] Lucene: http://jakarta.apache.org/lucene/docs/index.html
[2] MySQL: http://www.mysql.com/
[3] Snowball: http://jakarta.apache.org/lucene/docs/lucene-sandbox/snowball/
[4] We employed the stopword lists at http://www.unine.ch/Info/clef/ and manually added some words.
[5] Linguatec Personal Translator: http://www.linguatec.net/online/ptwebtext/index.shtml
Reverso: http://www.reverso.net/, Free Translation: http://www.freetranslation.com/

According to our analysis, MySQL's weak performance can be partly explained by the following factors. When building a fulltext index, MySQL automatically filters the content of the table column(s) to be indexed, i.e. it also tries to remove stopwords and words with a length of three or less characters. If not explicitly changed, an English stopword list is used by default. This signifies that the indices of the English collections might have been altered slightly compared to the indices of other collections. The removal of very small words might have had a bigger impact on all collections. Because stemming was carried out separately before importing the data into the database, it is very likely that several stemmed words did have a length of three characters or less and therefore were not included in the index. As we later found out, the variable controlling the length of the words to be discarded can be changed, in which case the indices would have to be rebuild. We estimate that MySQL ignored approximately 5 to 10 % of the query and document terms due to this setting.

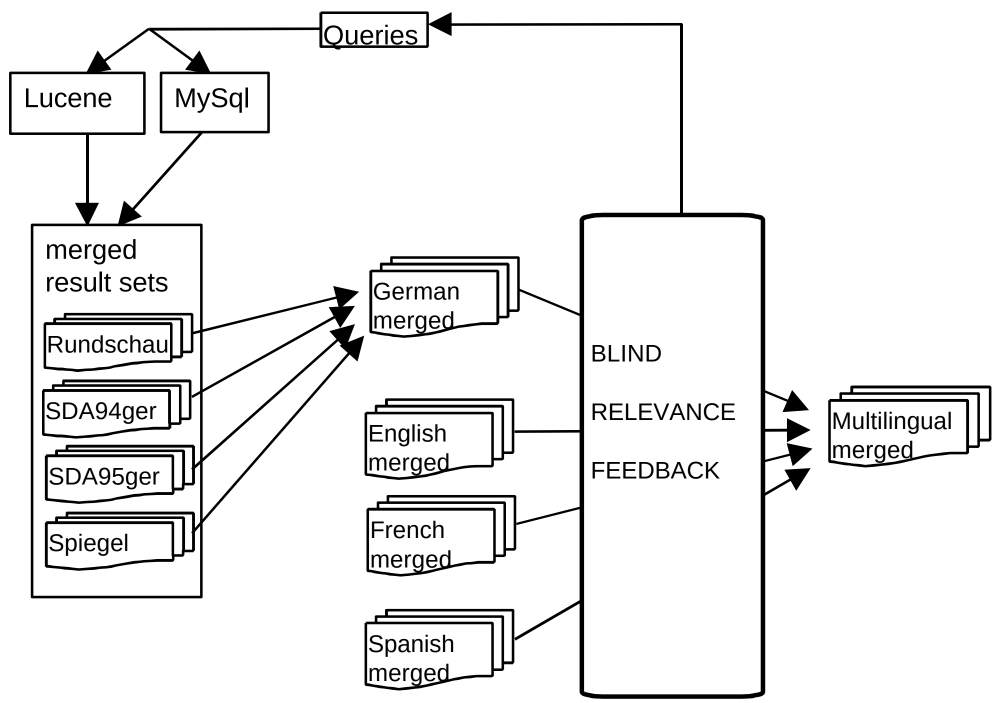

Figure 1. Experimental setup

To further improve retrieval quality, blind relevance feedback (BRF) was implemented. We selected expansion terms with either the Robertson selection value (RSV) or the Kullback-Leibler (KL) divergence measure (Carpineto et al. 2001). Results can be seen in Table 2. The precision could be improved and the number of retrieved documents was boosted (+14.8% compared to the best merged run for KL, +17.1% compared to Lucene).

Table 2. Query expansion

	Documents retrieved	Average precision	Average document precision
Lucene BRF RSV 5 10	5059	0.3017	0.3107
Lucene BRF KL 5 10	5216	0.3138	0.3277
7:1 BRF RSV 5 10	5157	0.3049	0.3174
7:1 BRF KL 5 10	5227	0.3127	0.3264

Due to time constraints, we could not determine the best parameters for BRF. A sample run without BRF took 4+ hours on our dual Pentium III 800Mhz, 1GB RAM, SCSI 160 HDD machine. A run with BRF taking the top five documents and adding ten terms commonly took more than twelve hours. Unfortunately, some instabilities in MySQL-DB caused further delay for our experiments.

All our submitted runs apply BRF KL 5 20, on behalf of the multilingual task, R1 uses the 7:1 merging scheme, yet R2 is a lucene-only run. Both monolingual runs are rather a by-product obtained in the course of our main (multilingual) task. The processing sequence chosen allowed for a efficient extraction of the monolingual data.
In our test runs, we were able to show that fusion helped raise at least the recall, although the results for 2003 could not confirm this finding. The Lucene-based runs generally outperform the fusion runs, except for a marginally better recall in the merged monolingual run (Table 3).

Table 3. Results 2003

	Documents retrieved	Average precision
UHImlt4R1	3944 / 6145	0.2849
UHImlt4R2	4137	0.3057
UHImnenR1	951 / 1006	0.3630
UHImnenR2	945	0.3797

5 Outlook

As in 2002, we included stand-alone retrieval systems in our MIMOR fusion approach. Again, one system performed much worse than the other. Next year, we would like to expand to the multilingual-8 track and continue to work on merging schemes, query expansion and term weighting.

Acknowledgements

We would like to thank the Jakarta and Apache projects' teams for sharing Lucene with a wide community as well as the providers of MySQL and snowball. Furthermore, we acknowledge the work of several students from the University of Hildesheim who implemented MIMOR as part of their course work.

References

Carpineto, C.; de Mori, R.; Romano, G.; Bigi, B. (2001): An Information-Theoretic Approach to Automatic Query Expansion. In: ACM Transactions on Information Systems. 19 (1). pp. 1-27.

Fuhr, N.; Großjohann, K. (2001): XIRQL: A Query Language for Information Retrieval in XML Documents. In: 24th Annual Intl Conf on Research and Development in Information Retrieval 2001. pp. 172-180.

Hackl, René; Kölle, Ralph; Mandl, Thomas; Womser-Hacker, Christa (2002): Domain Specific Retrieval Experiments at the University of Hildesheim with the MIMOR System. In: Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael (eds.): Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2002 Workshop. Springer [LNCS]

Henzinger, Monika (2000): Link Analysis in Web Information Retrieval. In: IEEE Data Engineering Bulletin, 23(3): pp. 3-8.

Mandl, T.; Womser-Hacker, C. (2001): Probability Based Clustering for Document and User Properties. In: Ojala, T. (ed.): Infotech Oulo International Workshop on Information Retrieval (IR 2001). Oulo, Finnland. Sept 19-21 2001. pp. 100-107.

McCabe, M. Catherine; Chowdhury, A.; Grossmann, D.; Frieder, O. (1999): A Unified Framework for Fusion of Information Retrieval Approaches. In: Eigth ACM Conference on Information and Knowledge Management (CIKM) pp. 330-334.

Plachouras, Vassilis; Ounis, Iadh (2002): Query-Based Combination of Evidence on the Web. In: Workshop on Mathematical/Formal Methods in Information Retrieval, ACM SIGIR Conference, Tampere, Finland.

Womser-Hacker, C. (1997): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Habilitationsschrift. Universität Regensburg, Informationswissenschaft.