

Index Combinations and Query Reformulations for Mixed Monolingual Web Retrieval

Krisztian Balog and Maarten de Rijke

ISLA, University of Amsterdam
kbalog,mdr@science.uva.nl

Abstract. We examine the effectiveness on the multilingual WebCLEF 2006 test set of light-weight methods that have proved successful in other web retrieval settings: combinations of document representations on the one hand and query reformulation techniques on the other.

We investigate a range of approaches to crosslingual web retrieval using the test suite of the mixed monolingual CLEF 2006 WebCLEF track, featuring a stream of known-item topics in various languages. The topics are a mixture of manual (human generated) and automatically generated topics. We examine the robustness of well-known web retrieval techniques on this test set: compact document representations (titles or incoming anchor-texts), and query reformulation techniques. In Section 1 we describe our retrieval system as well as the approaches we applied. In Section 2 we describe our experiments, while the results are detailed in Section 3. We conclude in Section 4. For details on the WebCLEF collection and on the topics used we refer to [1].

1 System Description

Our retrieval system is based on the Lucene engine [5]. For ranking, we used the default similarity measure of Lucene, i.e., for a collection D , document d and query q containing terms t_i :

$$sim(q, d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where

$$tf_{t,X} = \sqrt{\text{freq}(t, X)} \quad idf_t = 1 + \log \frac{|D|}{\text{freq}(t, D)},$$
$$norm_d = \sqrt{|d|} \quad coord_{q,d} = \frac{|q \cap d|}{|q|}, \quad \text{and} \quad norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t^2}.$$

We did not apply any stemming nor did we use a stopword list. We applied case-folding and normalized marked characters to their unmarked counterparts, i.e., mapping á to a, ö to o, æ to ae, î to i, etc. The only language specific processing we did was a transformation of the multiple Russian and Greek encodings into an ASCII transliteration.

We extracted the full text from the documents, together with the title and anchor fields, and created three separate indexes: *content* (index of the full document text), *title* (index of all <title> fields), and *anchors* (index of all incoming anchor-texts). We created three base runs using the separate indexes and evaluated the base runs using the WebCLEF 2005 topics. Based on the outcomes we decided to use only the *content* and *title* indexes.

1.1 Run combinations

We experimented with combinations of *content* and *title* runs, using standard run combination methods Fox and Shaw [2]: **combMAX** (take the maximal similarity score of individual runs); **combMIN** (take the minimal similarity score of individual runs); **combSUM** (take the sum of the similarity scores of individual runs); **combANZ** (take the sum of the similarity scores of individual runs, and divide by the number of non-zero entries); **combMNZ** (take the sum of the similarity scores of individual runs, and multiply by the number of non-zero entries); and **combMED** (take the median similarity score of individual runs).

Fox and Shaw [2] found **combSUM** to be the best performing combination method. Kamps and de Rijke [3] conducted extensive experiments with the Fox and Shaw combination rules for nine European languages, and demonstrated that combination can lead to significant improvements. Moreover, they showed that the effectiveness of combining retrieval strategies differs between English and other European languages. In Kamps and de Rijke [3], **combSUM** emerges as the best combination rule, confirming Fox and Shaw’s findings.

Similarity score distributions may differ radically across runs. Therefore, we apply combination methods to normalized similarity scores only. We follow Lee [4] and normalize retrieval scores into a $[0, 1]$ range, using the minimal and maximal similarity scores (min-max normalization): $s' = (s - min)/(max - min)$, where s denotes the original retrieval score, and min (max) is the minimal (maximal) score over all topics in the run.

For the WebCLEF 2005 topics the best performance was achieved using the **combSUM** combination rule, which is in conjunction with the findings in [2, 3], therefore we used that method for the experiments on the WebCLEF 2006 topics.

1.2 Query reformulation

In addition to our run combination experiments, we conducted experiments to measure the effect of phrase and query operations. We tested query-rewrite heuristics using phrases and word n -grams.

As to phrases, we simply added the topic terms as a phrase to the topic. For example, for the topic WC0601907, with title “diana memorial fountain”, the query becomes: *diana memorial fountain* “*diana memorial fountain*”. Our intuition is that rewarding documents that contain the whole topic as a phrase, not only the individual terms, would be beneficial to retrieval performance.

In our approach to n -grams, every word n -gram from the query is added to the query as a phrase with weight n . This means that longer phrases get a

Table 1. Submission results (Mean Reciprocal Rank)

runID	all	auto	auto-u	auto-b	manual	man-n	man-o
Baseline	0.1694	0.1253	0.1397	0.1110	0.3934	0.4787	0.3391
Comb	0.1685 ¹	0.1208 ³	0.1394 ³	0.1021	0.4112	0.4952	0.3578
CombMeta	0.1947 ³	0.1505 ³	0.1670 ³	0.1341 ³	0.4188 ³	0.5108 ¹	0.3603 ¹
CombPhrase	0.2001 ³	0.1570 ³	0.1639 ³	0.1500 ³	0.4190	0.5138	0.3587
CombNboost	0.1954 ³	0.1586 ³	0.1595 ³	0.1576 ³	0.3826	0.4891	0.3148

bigger boost. Using the previous topic as an example, the query becomes: *diana memorial fountain* “*diana memorial*”² “*memorial fountain*”² “*diana memorial fountain*”³, where the number in the upper index denotes the weight attached to the phrase (the weight of the individual terms is 1).

2 Runs

To assess the effectiveness of index combinations and query reformulations, we created five runs using the WebCLEF 2006 mixed monolingual topics: *Baseline* (base run using the *content* only index); *Comb* (combination of the *content* and *title* runs, using the *combSUM* method); *CombMeta* (like the *Comb* run, but using the *target domain* meta field; we restrict our search to the corresponding domain); *CombPhrase* (the *CombMeta* run, but using the *Phrase* query reformulation technique); *CombNboost* (the *CombMeta* run, using the *N-grams* query reformulation technique).

3 Results

Table 1 lists our results in terms of mean reciprocal rank. Runs are listed along the left-hand side, while the labels indicate either all topics (*all*) or various subsets: automatically generated (*auto*)—further subdivided into automatically generated using the unigram generator (*auto-u*) and automatically generated using the bigram generator (*auto-b*)—and manual (*manual*), which is further subdivided into *new* manual topics and *old* manual topics; see [1] for details.

Significance testing was done using the two-tailed Wilcoxon Matched-pair Signed-Ranks Test, where we look for improvements at a significance level of 0.05 (¹), 0.001 (²), and 0.0001 (³). Significant differences noted in Table 1 are with respect to the **Baseline** run.

Combining the *content* and *title* runs (**Comb**) increased performance only for the manual topics. The use of the title tag does not help when the topics are automatically generated. Instead of employing a language detection method, we simply used the target domain meta field. The **CombMeta** run improved the retrieval performance significantly for all subsets of topics. Our query reformulation techniques show mixed, but promising results. The best overall score was achieved when the topic, as a phrase, was added to the query (**CombPhrase**).

The comparison of **CombPhrase** vs **CombMeta** reveals that they achieved similar scores for all subsets of topics, except for the automatic topics using the bigram generator, where the query reformulation technique was clearly beneficial. The n -gram query reformulation technique (**CombNboost**) further improved the results of the *auto-b* topics, but hurt accuracy on other subsets, especially on the manual topics. The **CombPhrase** run shows that even a very simple query reformulation technique can be used to improve retrieval scores.

Comparing the various subsets of topics, we see that the automatic topics are more difficult than the manual ones. Also, the new manual topics seem to be more appropriate for known-item search than the old manual topics. There is a clear ranking among the various subsets of topics, and this ranking is independent from the applied methods: $man - n \gg man - o \gg auto - u \gg auto - b$.

4 Conclusions

We investigated the effectiveness of known web retrieval techniques in the mixed monolingual setting. The only language-specific processing we applied was the transformation of various encodings into an ASCII transliteration. We found that the standard combination **combSUM** combination using Min-Max normalization is effective only for human generated topics; using the title field did not improve performance when the topics are automatically generated. Significant improvements (+15% MRR) were achieved by using the target domain meta field. Furthermore, even simple query reformulation methods can improve retrieval performance, but not consistently across all subsets of topics. Although it may be too early to talk about a solved problem, effective and robust web retrieval techniques seem to carry over to the mixed monolingual setting.

5 Acknowledgments

Krisztian Balog was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.-065.-120 and 612.000.106. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.000.106, 612.-066.302, 612.069.006, 640.001.501, and 640.002.501.

Bibliography

- [1] K. Balog, L. Azzopardi, J. Kamps, and M. de Rijke. Overview of WebCLEF 2006. In *This Volume*, 2007.
- [2] E. Fox and J. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. NIST Special Publ. 500-215, 1994.
- [3] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for European languages. In *Proc. ACM-SAC'04*, pages 1073–1077, 2004.
- [4] J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In *Proc. SIGIR '95*, pages 180–188. ACM Press, 1995.
- [5] Lucene. The Lucene search engine, 2005. <http://lucene.apache.org/>.