Fusion of Probabilistic Algorithms for the CLEF Domain Specific Task(Extended Abstract)

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley, USA
ray@sims.berkeley.edu

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Performance, Measurement

Keywords

Cheshire II, Logistic Regression, Data Fusion

1 Extended Abstract

This extended abstract describes the Berkeley 1 participation in the Domain Specific task for CLEF 2005. This year we submitted the minumum number of entries for each subtask (3 monolingual runs, 6 bilingual runs, and 3 multilingual runs). In our runs we employed retrieval algorithms data fusion methods that have performed relatively well in some other retrieval contexts, but which will almost surely be abandoned in later attempts at CLEF. The main technique being tested is the fusion of multiple probabilistic searches against different XML components using both Logistic Regression (LR) algorithms and a version of the Okapi BM-25 algorithm. We also combine multiple translations of queries in cross-language searching. In the following paragraphs we will briefly describe the the indexing and term extraction methods used, followed by a description of the retrieval algorithms and data fusion methods. Since this is the first time that the Cheshire system has been used for CLEF, this approach can at best be considered a very preliminary base testing of some retrieval algorithms and approaches.

For both the monolingual and bilingual tasks we indexed the documents using the Cheshire II system. The document index entries and queries were stemmed using the Snowball stemmer. Text indexes were created for separate XML elements (such as document titles or dates) as well as for the entire document. The techniques and algorithms used for the DS task were essentially identical to those that we used for the GeoCLEF task, but without the special geographic indexes used for GeoCLEF (our GeoCLEF track paper describes the algorithms and approaches in detail).

For the bilingual and multilingual search tasks we used combinations of up to three different MT systems for query translation, using the L&H PC-based system, SYSTRAN (via Babelfish), and PROMT. Each of these translations was combined into a single probabilistic query. The hope was to overcome the translation errors of a single system by including alternatives. However, for translation to Russian from German and English, only the PROMT MT system was available.

We tried only a single primary approach for searching, using only the topic text from the title and desc elements. In all cases the different indexes mentioned above were used, and probabilistic searches were carried out on each index with the results combined using the CombMNZ data fusion algorithm algorithm developed by Shaw and Fox [1]. The CombMNZ algorithm merges result lists, normalizing the scores in each list and increasing scores for items based on the number of result lists that they appear in, while penalizing items that appear in only a single list. For all searches we used both the Berkeley TREC3 and the Okapi BM-25 algorithms, with the results from each algorithm were also combined using the CombMNZ algorithm.

The results did not have very good performance. Relative to our German and English result, the Russian results look fairly good (we suspect that this may be due to the smaller number of participants). We consulted with the Berkeley2 group (who were using a different system and algorithms) to find what the primary differences were. Among the beneficial techniques used in those runs are 1) query expansion from the thesaurus, 2) automatic de-compounding of German words and 3) application of blind relevance feedback. We are conducting further tests adding these techniques and hope to have results to report at the meeting. The official submitted runs can be considered preliminary baselines that, we hope, will be improved upon in the future.

(*NOTE: Extended Abstract Only, additional comments in our GeoCLEF paper)

References

[1] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215*, pages 243–252, 1994.