First experiments with CLEF

Peter van der Weerd
pweerd@medialab.nl

Wilfred Blom
wblom@medialab.nl

Medialab BV, Schellinkhout, The Netherlands
http://www.medialab.nl

1. Goal

Our goal for participation in CLEF was to compare our information-retrieval engine with other engines, and to get more information about other approaches to the information retrieval problems. The Olympic spirit dominated MediaLab's first participation. To play is more important than to win.

2. Preparation

Due to the very limited time frame we had, we decided to use a straightforward approach. We decided to use our standard retrieval engine without any modifications, and to only parse the supplied queries by breaking them into words and feed the individual words into the retrieval engine.

Before we started the experiment it was recognized that our score would not be very large: the queries specified rather specific what to return and what not. Because we used **all** words from the query to find our results we knew on forehand that we would find the undesired results as well. However, we felt that it was nice to check out if we could deliver better results than the worst participants with this approach.

Two indexes were built:

- A title field consisting of the concatenation of the shorter fields: the TI, LE and OS fields from the articles.
- A body field consisting of the TE-field

3. Query processing

Besides normal search behavior we used the following add-ons to process the queries:

- Stop words
  We used the stop words supplied by the CLEF project
- Stemming.
  The built-in Dutch stemmer is based on the porter stemmer
- Compound term searching.
  MediaLab's search engine has built-in facilities to expand for example "roos" to "klaproos".

All fields from the query (<NL-title>, <NL-desc>, <NL-narr>) were stripped into words, sending each word into the engine, giving a ranked set of documents for each word. These individual results were combined using a bounded-add of their weights:

$$w = w1 + w2 - w1*w2$$

This final result was submitted to CLEF.

4. Results and future work

As one might expect the resulting precision was not very high: 22% of the queries scored above the median, the rest below. However, our results never were the worst! Realizing that the results of machine-driven approaches as we did and (a few) manual-driven queries are intermixed within the results, we are not at all discontented with the results.

The first thing we will do is to incorporate NLP into the query-process, and to use weighted queries. MediaLab is enthusiastic about CLEF and intends to allocate more resources for next year's participation.

5. Acknowledgements

The participation of MediaLab to CLEF would not be realized without the unlimited enthusiasm and push of Dennis Reidsma who visited us this year to get his degree at the University of Twente.