

The University of Lisbon at CLEF 2006 ad hoc task

Nuno Cardoso, Mário J. Silva and Bruno Martins
 XLDB Group - Faculty of Sciences, University of Lisbon
 {ncardoso, mjs, bmartins} @xldb.di.fc.ul.pt
<http://xldb.di.fc.ul.pt>



Tumba at CLEF 2006 ad hoc

Objectives

- Develop an IR system for PT monolingual ad-hoc task, as a stable testbed to evaluate GIR approaches for the GeoCLEF task.
- Implement well-known algorithms in our IR system based on **tumba!**.

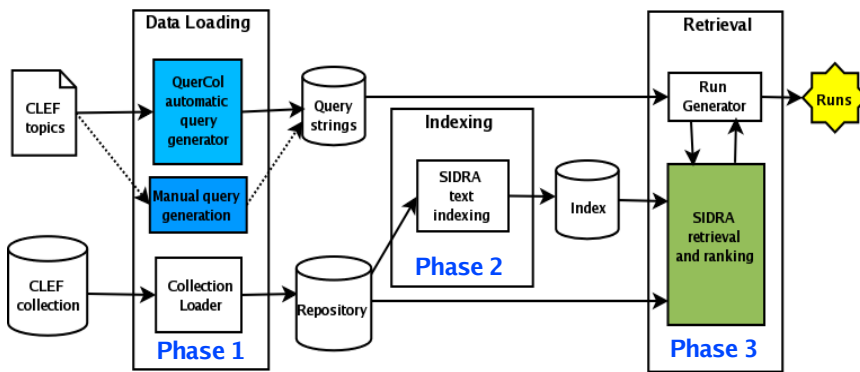
Improvements

- QuerCol** – Query expansion with blind feedback + Query generator.
- SIDRA** – Retrieval & ranking based on BM25 weighting scheme.

<http://www.tumba.pt>



Architecture



Phase 1: Data Loading

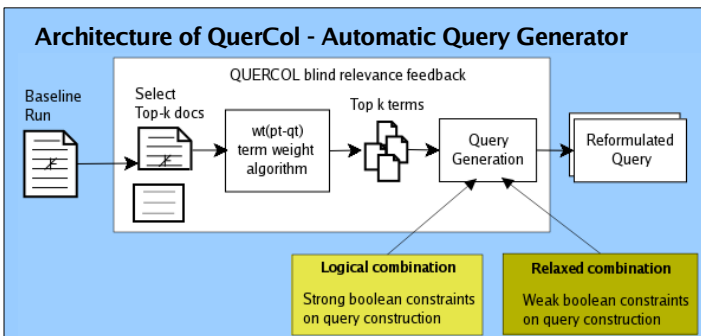
CLEF collection is loaded into repository.
 QuerCol automatically generates queries from CLEF topics.
 Additional queries created manually.

Phase 2: Indexing

Documents are indexed.
 SIDRA generates term indexes.

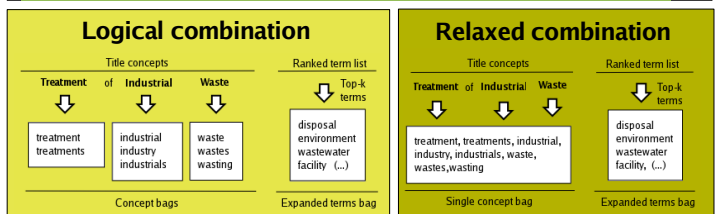
Phase 3: Retrieval

Queries submitted to SIDRA by a run generator.
 Runs created and submitted to CLEF.



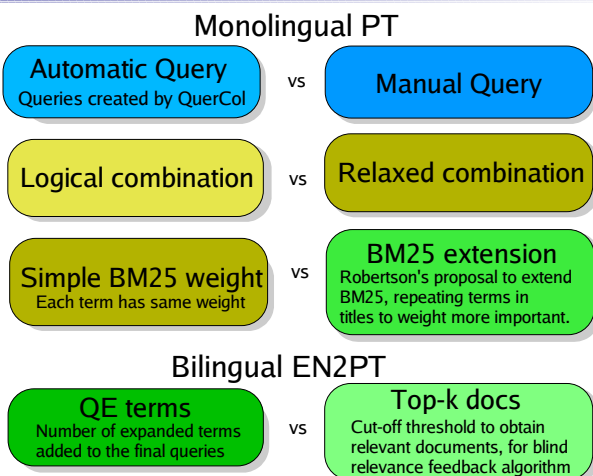
BM25 Weighting Scheme

$$BM25(t_i) = \frac{(k_1 + 1) \times term_freq(t_i)}{k_1 \times ((1 - k_2) + k_2 \times \frac{doc_len}{avg_doc_len}) + d} \log\left(\frac{N - doc_freq(t_i) + 0.5}{doc_freq(t_i) + 0.5}\right)$$

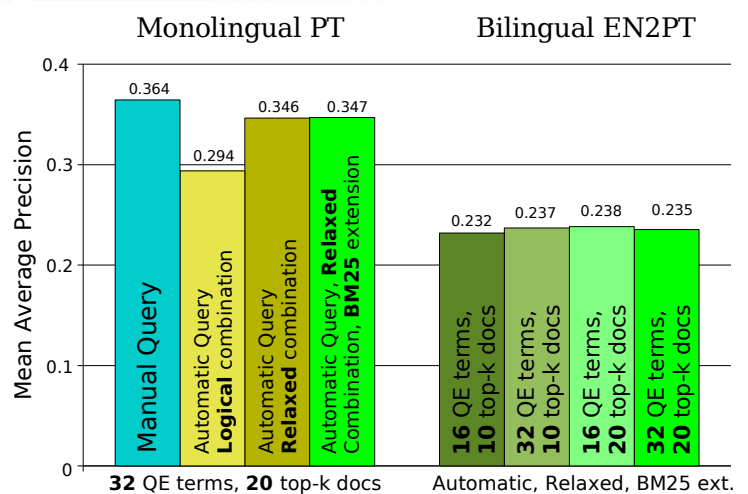


CLEF 2006 ad hoc evaluation

Experiments



Results



Conclusions

- 2006 results show **improvements** compared to 2005.
- Relaxed combination** generated queries for our best run.
- BM25 extension** to re-weight terms did not improved significantly the results.
- Performance of automatic runs is **close** to performance of manual run, and we only used the topic title terms.
- This IR system was used on GeoCLEF task. Now, for **GIR**, we can focus on the **G**, because the **IR** part is good.