The XLDB Group at GeoCLEF 2005

Nuno Cardoso, Bruno Martins, Marcirio Silveira Chaves, Leonardo Andrade and Mário J. Silva
XLDB Group – Computer Science Department, Faculty of Sciences, University of Lisbon
{ncardoso, bmartins, mchaves, leonardo, mjs} @xldb.di.fc.ul.pt
http://xldb.di.fc.ul.pt

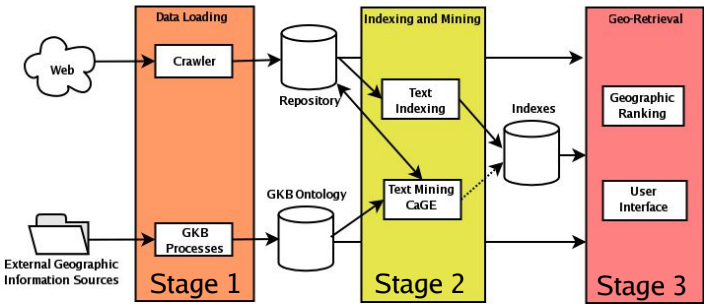
GeoTumba

tumba! + geographic reasoning = **Geotumba**!

- GeoTumba is a prototype location-aware search engine, handling queries with geographic terms.
- We participated in GeoCLEF with a Geo-IR system that uses components from GeoTumba.

The Geographic IR System

Architecture

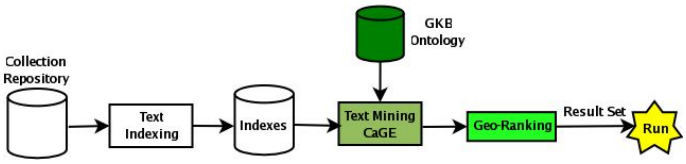


Stage 1: Data Loading
Web pages are harvested into a repository. Geographic knowledge is integrated into GKB ontology.

Stage 2: Indexing and Mining
CaGE detects geographic references and assigns scopes to documents. Term index and geo-scopes index are created.

Stage 3: Geo-Retrieval
Geographic query evaluated.
User interface presents results, ranked according to geographic criteria.

Software Configuration for GeoCLEF



1 - GKB
- Integrates geographic knowledge from two information sources (World Gazetteer, Wikipedia).
- Models geographic information as typed features and relationships.

2 – CaGE
- Mining module. Infers geographic context from documents.
- Identifies geographic references in document's text.
- Assigns a geo-scope to each document based on geographic references' frequency and relationships among them, using a graph-based algorithm.

3 - Geo-ranking
- Sorts result sets by geographic relevance.
- Weights documents using a scopes hierarchy, built from the Ontology.

GeoCLEF 2005 participation

Experiments

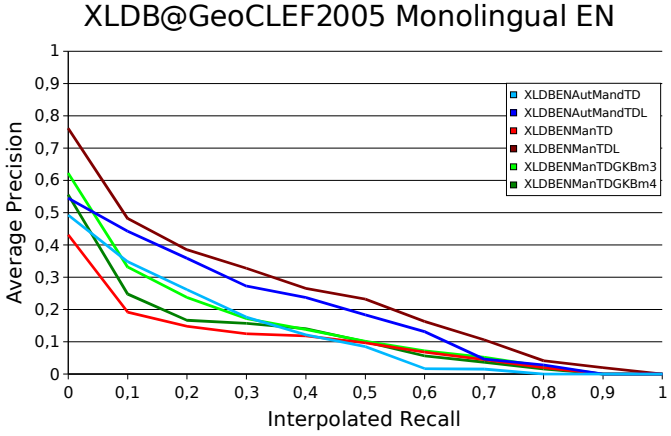
1) Scope Ranking – compare assigned geo-scopes vs geographic terms in the query strings.

2) Scope Assigning – compare the graph-based algorithm against a simpler baseline approach.

3) Location Terms Expansion – when not using geo-scopes, measure the contribution of expanding geographic terms.

4) Topic Translation – obtain initial results with EBMT translation system.

Results



Conclusions

1) Scope Ranking: geo-scopes not successful (We blame the assembled ontology).
2) Scope Assigning: graph-based algorithm always better (Scopes as expected, given the ontology).
3) Location Terms Expansion: manually generated queries performed better.
4) Topic Translation: monolingual runs better than bilingual.