The University of Lisbon at GeoCLEF 2006

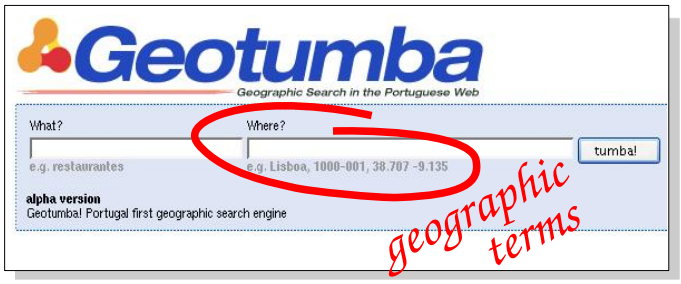
Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade and Mário J. Silva
XLDB Group – Faculty of Sciences, University of Lisbon
{bmartins, ncardoso, mchaves, leonardo, mjs} @xldb.di.fc.ul.pt
http://xldb.di.fc.ul.pt

Objectives at GeoCLEF 2006

Compare 2 strategies for Geographic IR against conventional IR approaches:

#1: Geographic Text Mining

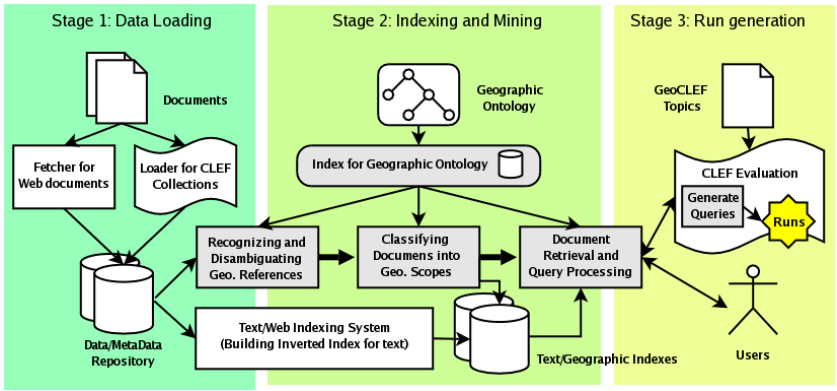
- **Mining** geographic references from text.
- Assigning geographic scopes to each **document**.
- Converting topics into <***what,relation,where***> triples.
- Assigning scopes to <***where***> terms.
- Using a ranking function that combines **scope similarities** with **BM25**.



#2: Augmenting Geographic Terms

- Converting GeoCLEF topics into <***what, relation. where***> triples.
- Augmenting <***where***> terms with terms from our geographic ontology
- Using a **BM25 term ranking**.

The Geographic IR System Architecture



Ranking Formula used on Geographic text mining strategy

$$DistSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is part of or parent of } scope_2 \\ 1 - \left(\frac{1+sign(D-D_{MBR}) \times (1-\exp(-\left|\frac{D-D_{MBR}}{\sigma_{MBR}}\right|^2))}{2}\right) & \text{otherwise} \end{cases}$$

Spatial Distance Similarity — 20%

$$OntSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is the same or equivalent to } scope_2 \\ \frac{2 \times NumCommonAncestors(scope_1, scope_2)}{NumAncestors(scope_1) + NumAncestors(scope_2)} & \text{otherwise} \end{cases}$$

Ontology Similarity — 50%

$$AdjSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is adjacent to } scope_2 \\ 0 & \text{otherwise} \end{cases}$$

Spatial Adjacency Similarity — 10%

$$PopSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is the same or equivalent to } scope_2 \\ \frac{PopulationCount(scope_1)}{PopulationCount(scope_2)} & \text{if } scope_1 \text{ is part of } scope_2 \\ \frac{PopulationCount(scope_2)}{PopulationCount(scope_1)} & \text{if } scope_2 \text{ is part of } scope_1 \\ 0 & \text{otherwise} \end{cases}$$

Population Similarity — 20%

$$NormBM25(doc, query) = \frac{\sum_{t_i \in doc} BM25(t_i) \times weight(query, t_i)}{\sum_{t_i \in doc} \log\left(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5}\right)(k_1 + 1)}$$

Text similarity

$$GeoSim(s_1, s_2) = (0.5 \times OntSim(s_1, s_2)) + (0.2 \times DistSim(s_1, s_2)) + (0.2 \times PopSim(s_1, s_2)) + (0.1 \times AdjSim(s_1, s_2))$$

Geographic similarity

$$Ranking(doc, query) = 0{,}5 \times Norm\,BM25(doc, query) + 0{,}5 \times Max(GeoSim(scope_{doc}, s))$$

Ranking Formula

- Combination parameters based on the intuition that **topology matters** and **metric refines**.
- Linear combination due to its simplicity
- All measures are normalized into [0,1] values.

GeoCLEF 2006 Evaluation

Runs

PT1/EN1: Baseline using manually generated queries from the topics and BM25 text retrieval.

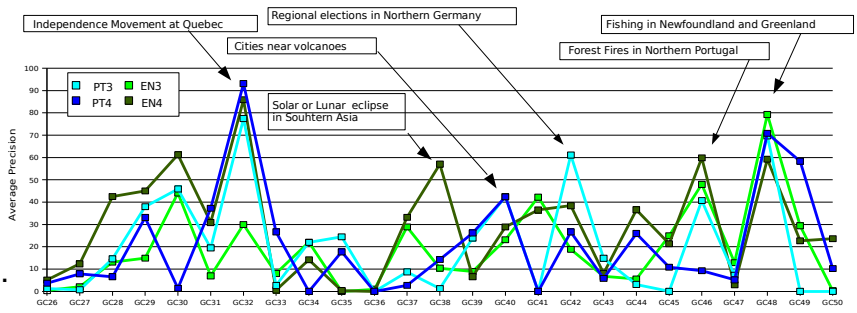
PT2/EN2: BM25 text retrieval. Queries were generated by blind-feedback expansion of <*what*> terms from topic title, together with the original <*where*> terms.

PT3/EN3 (Strategy #1): Geographic relevance ranking using geographic scopes. Queries were generated by blind-feedback expansion of <*what*> terms from topic title, and <*where*> terms matched into scopes.

PT4/EN4 (Strategy #2): Augmentation of <*where*> terms from the topic title, using top-10 related concepts from the geographic ontology. BM25 text retrieval. Queries were generated by blind-feedback expansion of <*what*> terms from topic title, together with augmented <*where*> terms.

Results

	Monolingual PT				Monolingual EN			
	PT1	PT2	PT3	PT4	EN1	EN2	EN3	EN4
num_ret	5232	23350	22617	10483	3324	22483	21228	10652
num_rel	1060	1060	1060	1060	378	378	378	378
num_ret_rel	607	**828**	519	624	192	**300**	240	260
MAP	**0,301**	0,257	0,193	**0,293**	**0,303**	0,158	0,208	**0,215**
R-Prec	**0,359**	0,281	0,239	**0,346**	**0,336**	0,153	0,215	**0,220**
bpref	**0,321**	0,254	0,208	**0,306**	**0,314**	0,140	0,191	**0,199**
gm-ap	**0,203**	0,110	0,074	**0,121**	**0,065**	0,027	0,024	**0,047**
P5	0,488	0,416	0,432	**0,536**	**0,384**	0,208	0,240	0,288
P10	**0,496**	0,392	0,372	0,480	**0,296**	0,180	0,228	0,240
P100	**0,218**	0,193	0,162	**0,218**	0,072	0,073	0,068	**0,084**



Conclusions

- **PT4/EN4** runs performed **better** than **PT3/EN3** runs.
- Text mining approach had some problems.
- Additional experiments are underway.