

The University of Lisbon at GeoCLEF 2007



Nuno Cardoso, David Cruz, Marcirio Chaves and Mário J. Silva
 XLDB Group – Faculty of Sciences, University of Lisbon
 {ncardoso, dcruz, mchaves, mjs} @xldb.di.fc.ul.pt



FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA

<http://xldb.di.fc.ul.pt> <http://local.tumba.pt>

Geographic Search in the Portuguese Web

Query Processing: all geographic information present on a query is captured and subject to proper **geographic query expansion**.

► **Feature types** and **spatial relationships** guide the **geographic query expansion**.

Text Mining: Discourse context narrowed to the **sentence level**. Generated **geographic signatures** for each document.

Geographic ranking: evaluate relevance considering queries and documents with **multiple geographic concepts as a scope**.

Geographic Signatures

= a list of **geographic concepts** that characterize a document scope - each document may have multiple geographic contexts.

Document signature:

LA072694-0011: 5668[1.00]; 2230[0.33]; 4555[0.33]; 4556[0.33]; 4557[0.33]
 LA072694-0012: 5388[1.00]; 5389[1.00]; 5390[1.00]; 12097[1.00]; 6653[0.67]
 LA072694-0013: 3691[1.00]; 225[0.33]; 452[0.33]; 7[0.33]; 367[0.33]; 137[0.33]
 LA072694-0014: 6653[1.00]; 6654[1.00]; 347[1.00]

DocID | Geo. concept #6653 ID | Confidence Measure for #6653

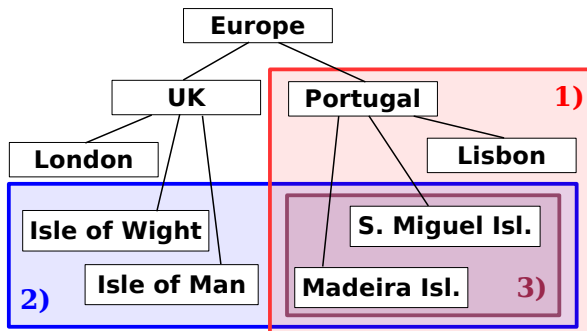
Query signature:

Sea traffic in Portuguese Islands → sea traffic @ 15 16 17 18 19 20 21 22 23 24
Geo. Concept Ids for all Islands part-of Portugal

Geographic Query Expansion

Geo. queries = <what, spatial relat., where>, where = feature + feature types.

- 1) sea traffic in Portugal
- 2) sea traffic in islands
- 3) sea traffic in Portuguese islands



Evaluation objectives

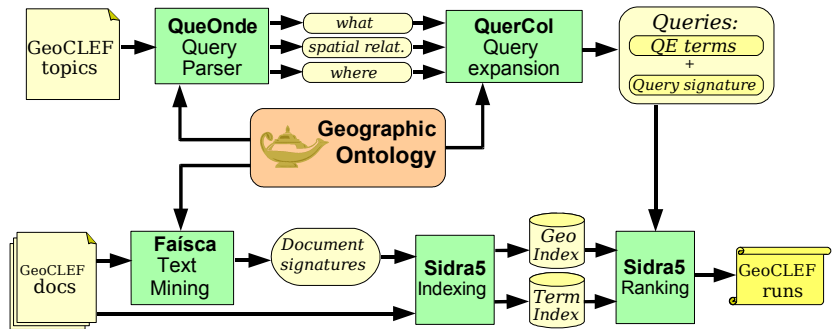
- 1) Evaluate if current GeoIR approach of **handling geonames in a separate geographic ranking** obtains better results than **handling geonames as terms in a standard IR approach**;
- 2) Determine which **GeoScore combination metrics is best**: Mean, Maximum or Boolean.
- 3) Compare **geographic query expansion before and after** the blind relevance feedback.

Questions Raised

- Classical IR was **outperformed** by Terms/GIR approach, but not full-GIR approaches... why?
- Blind relevance feedback **benefits** from early geographic QE? Needs stat. significance analysis.

Geographic QE with feature types has its merits. Yet, the GIR system should mature before more thorough experiments on these topics.

The Geographic IR System Architecture



Geographic Ranking

Query: Tourist attractions in **Hungary**.

Document 1: (...) there are many tourist attractions (...) in **Hungary**, (...) near **Portugal**, and (...) in **Australia**.

Document 2: (...) there are many tourist attractions (...) in **Budapest**.

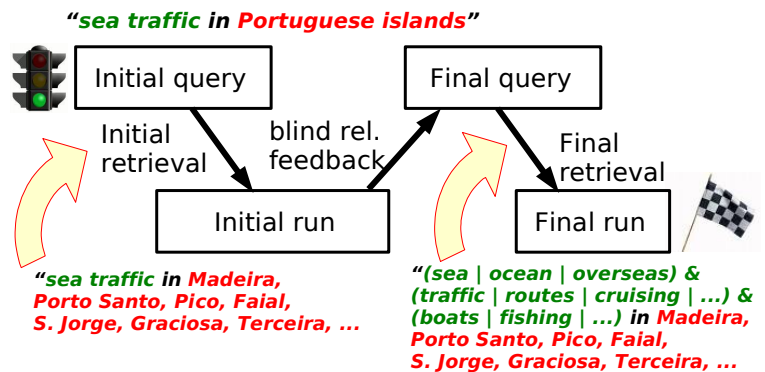
GeoSim x ConfMeas: 1.00 0.15 0.05
Document 1: Hungary Portugal Australia

GeoSim x ConfMeas: 0.60
Document 2: Budapest

	GeoScore	Mean	Max.	Bool.
Document 1	0.40	1.00	1.00	1.00
Document 2	0.60	0.60	0.00	0.00

GeoSim: geographic similarity value between a pair (s_q, s_d), norm. to [0,1]
ConfMeas: confidence measure on geo. concepts in D_{sig} , normalised to [0,1].
GeoScore: Combination strategies for the multiple $GeoSim$. s_q in Q_{sig}, s_d in D_{sig}

Blind Relevance Feedback



Experiment Results

	GeoScore	IR		GIR		IR/GIR
		Terms only	Geo. QE before RF	Geo. QE after RF	Terms/GIR	
PT	Initial run	0.210	0.126	0.084	0.210	
	Maximum		0.125	0.104	0.205	
	Final Run		0.022	0.021	0.048	
	Mean	0.233	0.135	0.125	0.268	
	Boolean		0.115	0.093	0.021	
a) Results for the Portuguese monolingual subtask.						
EN	Initial run	0.175	0.086	0.089	0.175	
	Maximum		0.093	0.104	0.218	
	Final Run		0.043	0.044	0.044	
	Mean	0.166	0.131	0.135	0.204	
	Boolean		0.081	0.087	0.208	
b) Results for the English monolingual subtask.						

- **Terms only:** Baseline using **classic IR approach**. Geographic query expansion before RF, but **just terms**: no **GeoScore**.
- **Geo. QE before/after RF:** Geographic IR. Initial retrieval for the initial run **uses/doesn't use** query signatures obtained from Geo. QE.
- **Terms/GIR:** **Initial run** are the same as Terms Only run (**classic IR**). Generation of **final run** the same as Geo. QE before RF (**GIR**).