

FACULDADE · DE · CIÊNCIAS | UNIVERSIDADE · DE · LISBOA

# Query Expansion through Geographical Feature Types

Nuno Cardoso and Mário J. Silva

{ncardoso, mjs}@xldb.di.fc.ul.pt

# Summary

- Introduction and motivation
- GIR prototype
- Evaluation results at GeoCLEF 2007
- New guidelines
- Conclusion and future work

# Introduction

- Work developed under the GREASE project, XLDB Group
- **Goal:** find the best approach for modelling geographic retrieval
  - provide geographic reasoning for tumba!, a Portuguese web search engine.

# GIR prototype features (2006)

- **Query Parsing:** convert queries into  $\langle \textit{what}, \textit{spat.rel}, \textit{where} \rangle$  triplets
- **Indexing & Ranking:** term index + geographic index.
  - Term weight computed by BM25.
  - Geographic weight calculated with a set of heuristics.
- **Geographic knowledge:** obtained from a geographic ontology

# GIR prototype features (2006)

- **Query parsing:** geoname matching and naïve disambiguation.
- **Geographic QE** based on geographic ontology; spatial relationships hardly used.
- **Feature types** used only for disambiguation purposes.
- **Grounding of geonames:** simple match into *ontological references*.

# GIR model (2006)

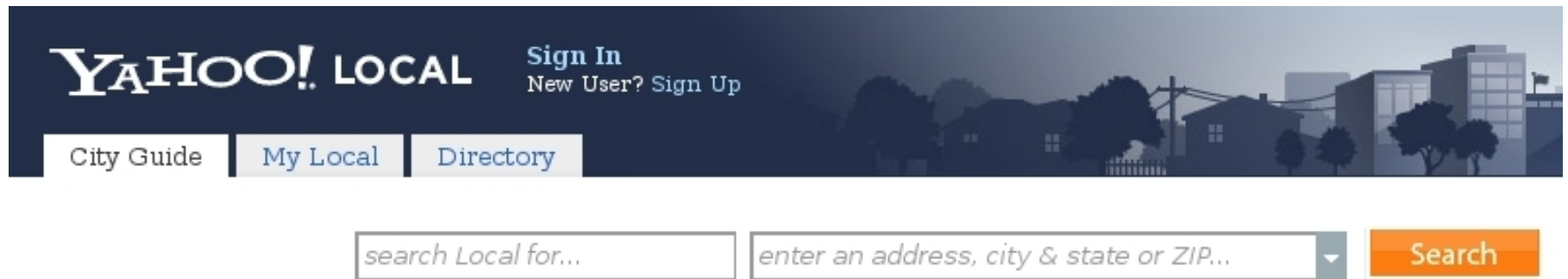
- Only **one** geographic concept per document scope.
- GeoCLEF 2006 evaluation revealed some limitations:
  - naïve and limited Geographic QE.
  - “*one sense per discourse*” assumption too restrictive for some docs.
  - Geographic topic titles not always follow the “restaurants in Lisbon” triplet pattern...

# GIR queries are harder...

- The user may resort to **indirect description** of geographic criteria  
*ex: (5 star hotel near the castle of S. Jorge)*
- The user may use **alias** for a group of geographic concepts  
*ex: (hotels in the Portuguese islands)*
- **Spatial relationships** may contain relevant information for the scope  
*ex: (Hotels on the coastlines of Portugal)*

# GIR queries are harder...

- The user **shouldn't** be forced to write queries in *<what, where>* format.



YAHOO! LOCAL Sign In  
New User? Sign Up

City Guide My Local Directory

search Local for... enter an address, city & state or ZIP... Search

- Geographic criteria are also described by **feature types** and **spatial relationships**.
- GIR systems should **understand user's needs** and **use all geonames**.

e.g., "10 market st, san francisco" or "hotels near lax"

Castles in Portugal

Search Maps

Search the map

Find businesses

Get directions

Google Maps

e.g., "10 market st, san francisco" or "hotels near lax"

Castles in Portugal

Search Maps

Search the map

Find businesses

Get directions

Search Results

My Maps New!

Text View Map View

Print Send Link to this page

Map Satellite Hybrid

Results 1-10 of about 70 for **Castles** near

**Portugal** - [Modify search](#)

Categories: [Museos](#), [Hotel](#)

**A** [Hacienda Benazuza El Bulli Hotel](#) -

[more info](#)

C/ Virgen De Las Nieves S/N, 41800 Sanlucar La Mayor, Spain  
+34 955 703 344 - ★★★★★

[Hacienda Benazuza elBulliho...](#)

... Hacienda Benazuza elBullihotel - Hotels, Country Estates and **Castles**, Sanlúcar la Mayor (Sevilla) ...  
[eventoplus.com](#)

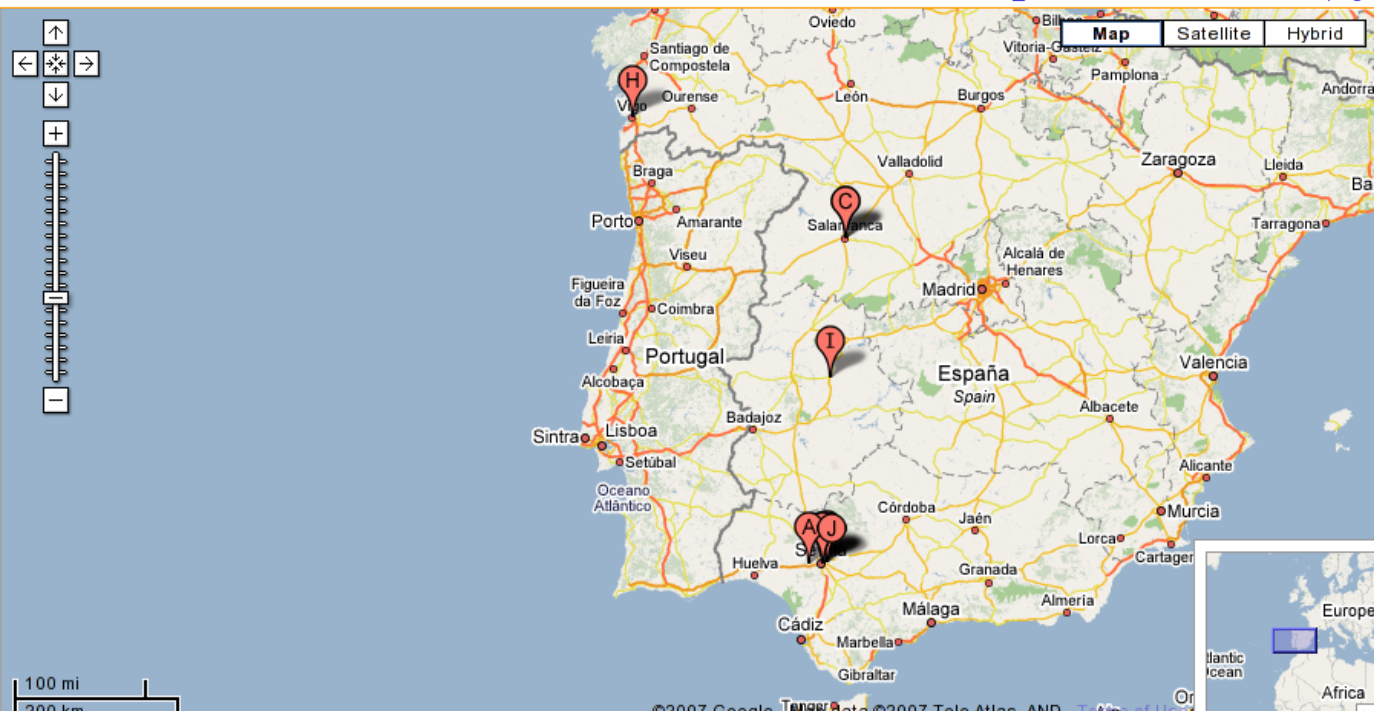
**B** [Hotel Torre Del Clavero](#) - [more info](#)

C/ Consuelo 21, 37001 Salamanca, Spain  
+34 923 280 410 - ★★★★★

[Salamanca, Spain Your Compl...](#)

... You Are Here:, Salamanca, Visiting the City Guide, Tourist Attractions & Sightseeing, **Castles**, Palaces & Historic Buildings ...  
[myareaguide.com](#)

**C** [Universidad De Salamanca](#) -



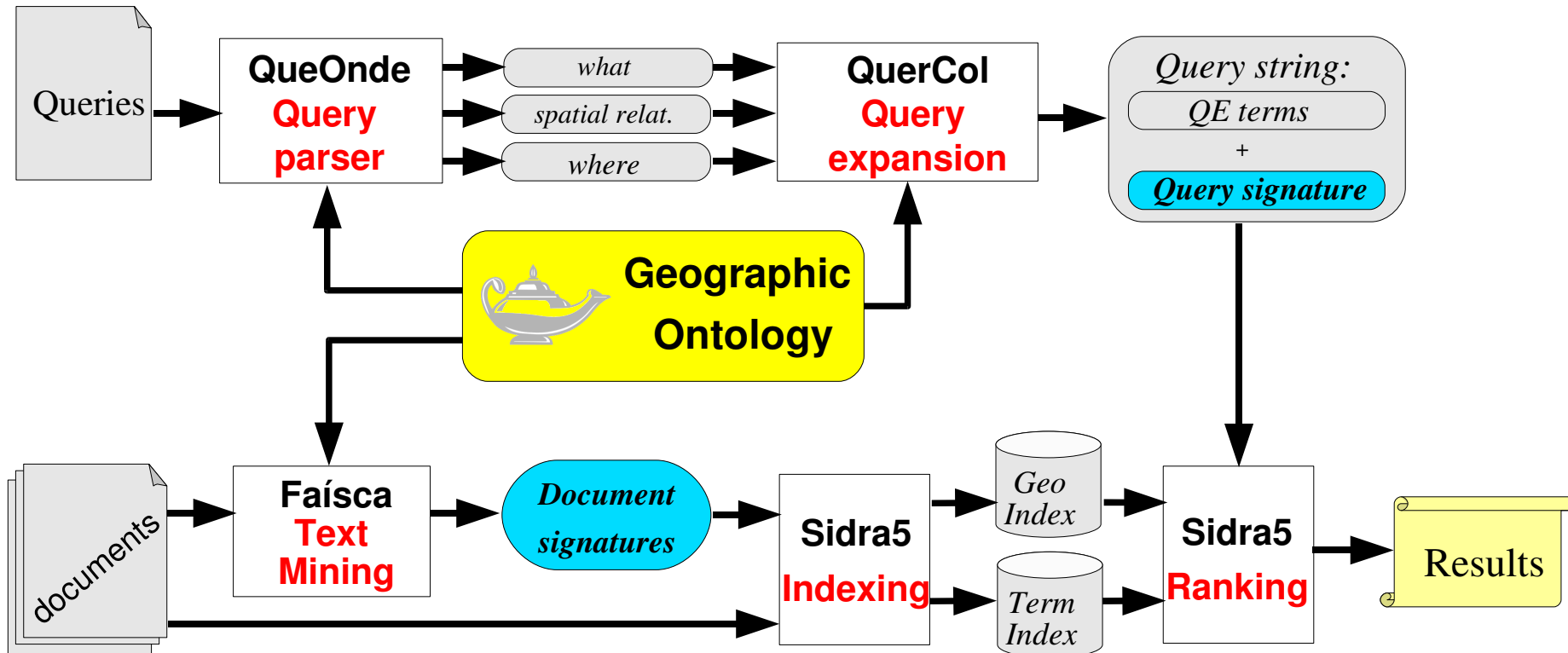
# Motivation for our work

- Document scopes best described by ***document signatures*** (list of geographic concepts).
- Query scopes also have ***query signatures***.
  - may be enhanced by geographic QE.
- **Geographic QE** may be driven by:
  - query type
  - feature type
  - spatial relationship

# Motivation for our work

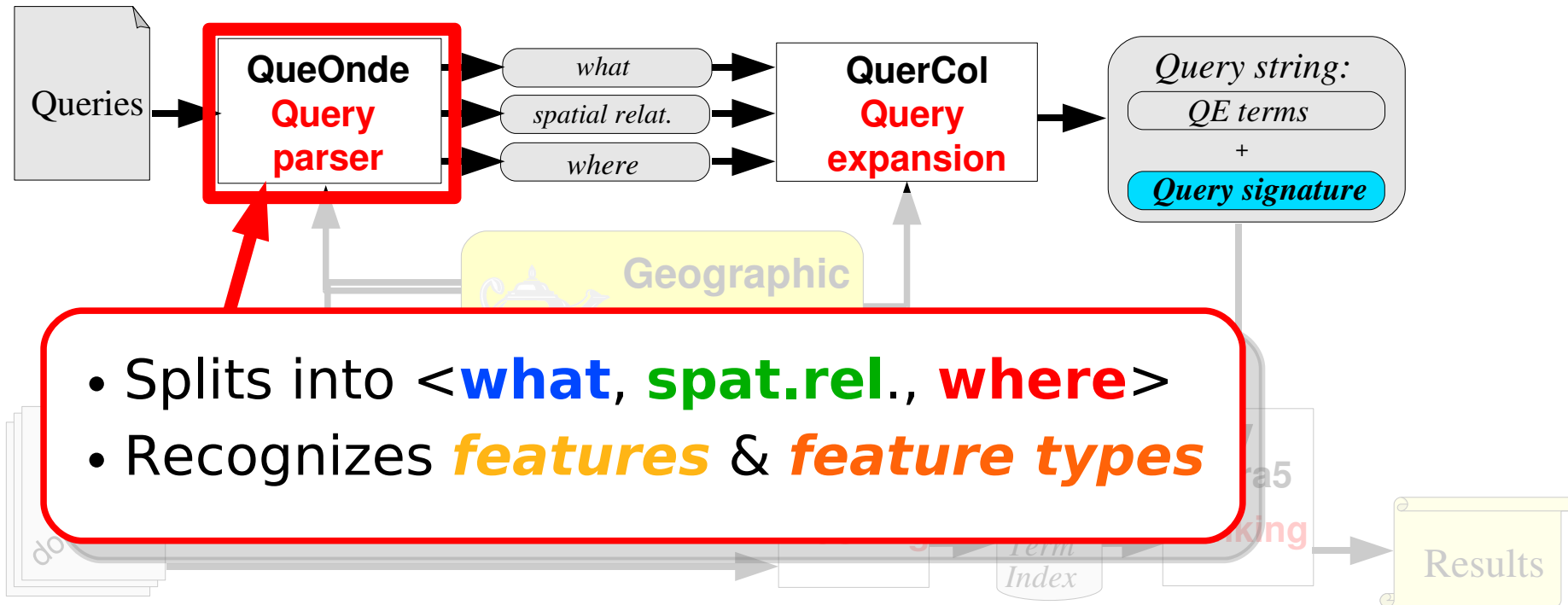
- Let's handle GIR queries in a smarter way.
- Use the **feature types** given in the queries for geographic QE.
- Test new approaches to calculate **geographic affinity** between **documents** and **queries**.
- Obtain first results on GeoCLEF evaluation.

# GIR prototype (2007)



# GIR prototype (2007)

## 1. Query parsing

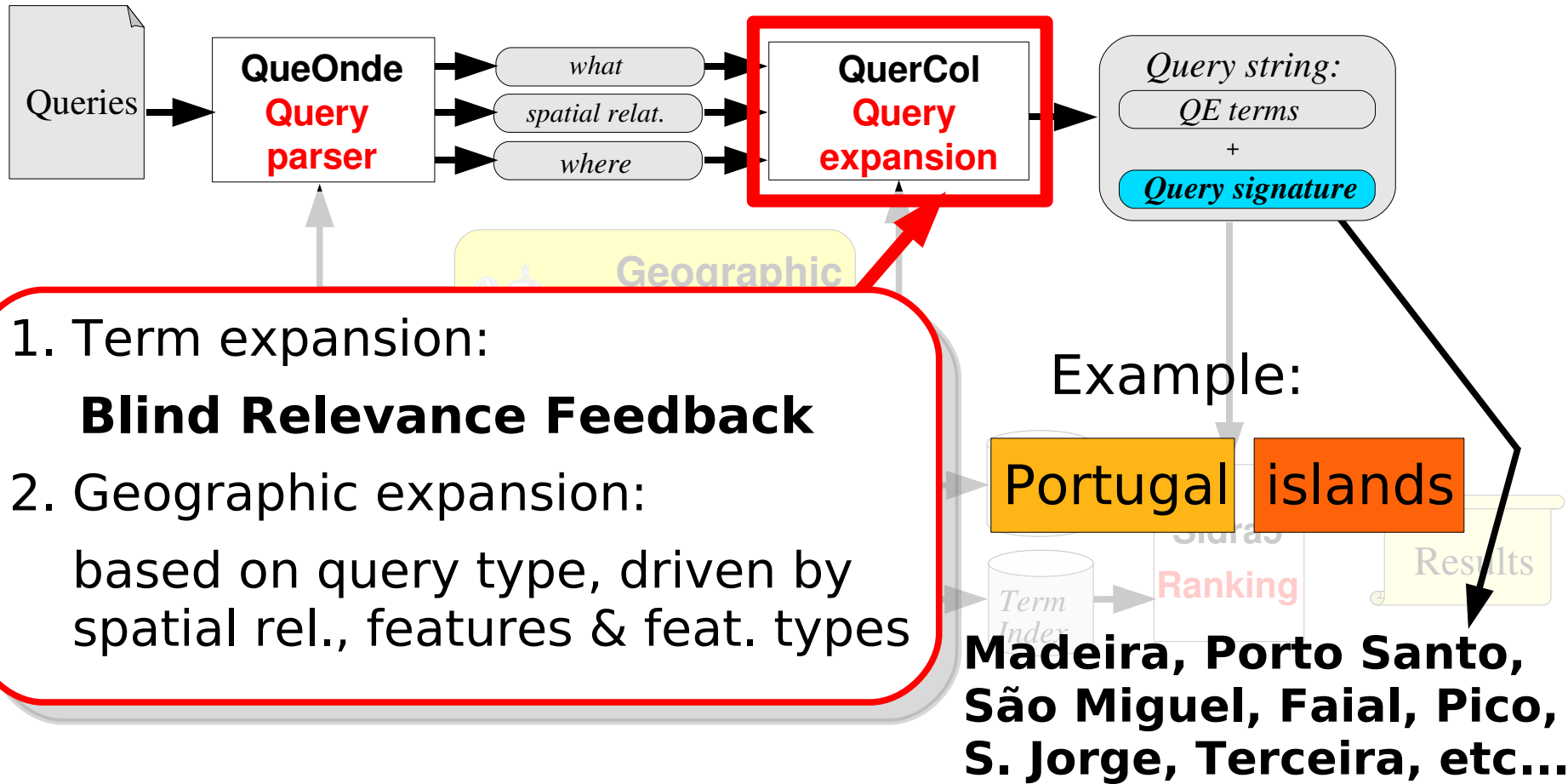


Example: Sea traffic in Portuguese islands =

Sea traffic in Portugal islands

# GIR prototype (2007)

## 2. Query Expansion



1. Term expansion:

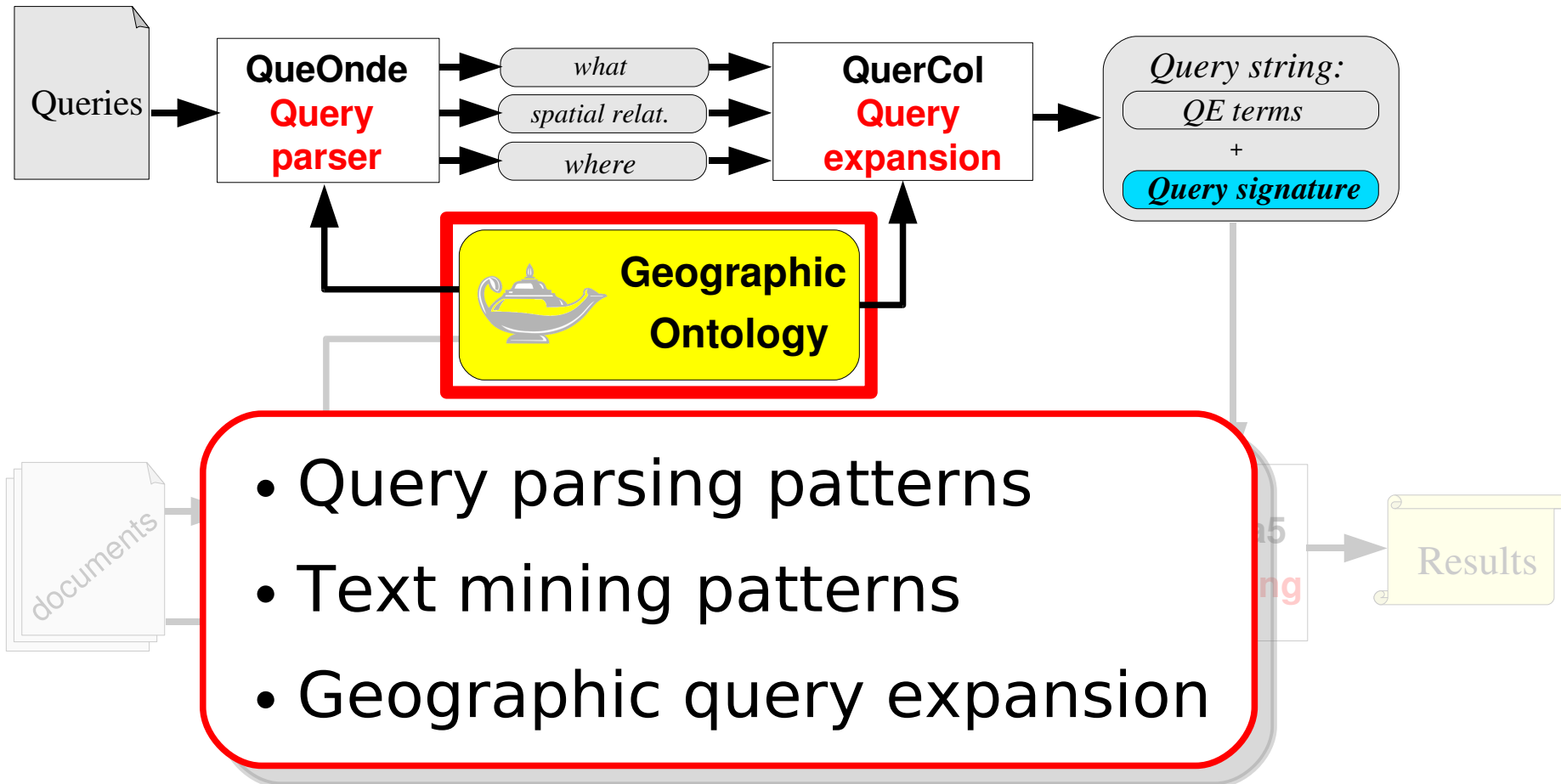
### Blind Relevance Feedback

2. Geographic expansion:

based on query type, driven by spatial rel., features & feat. types

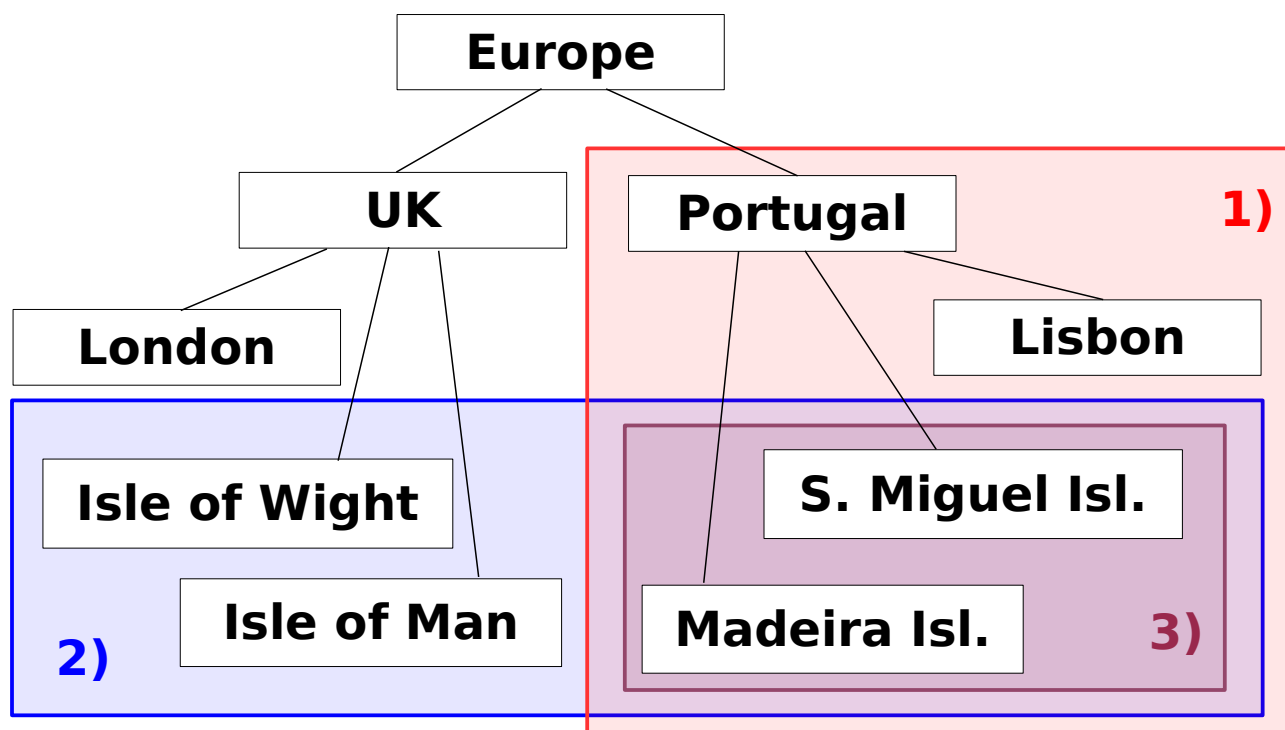
# GIR prototype (2007)

## 3. Geographic knowledge



# Geographic QE strategies

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



# Geographic QE strategies (cont.)

*sea traffic* around/along **Portugal**

*sea traffic* between **Portugal** and **USA**

*sea traffic* in the **Atlantic Ocean**,  
**except Azores**

{insert your weird geo-query here...}

- Map the **spatial relationships** into **ontology relationships**
- Use **features** and **feature types** to obtain **ontological concepts** within the scope

# Geographic relevance

- In 2006: **one** *geographic similarity* for each pair ( $scope_{query}$ ,  $scope_{doc}$ ).
- In 2007: **multiple** *geographic similarities* for each pair ( $sign_{query}$ ,  $sign_{doc}$ ).
- How to combine multiple *geographic similarity* values into a single geographic score?

# Geographic relevance (cont.)

*geographic similarity* combinations:  
**Mean, Maximum, Boolean.**

**Query:**

Tourist attractions in **Hungary**.

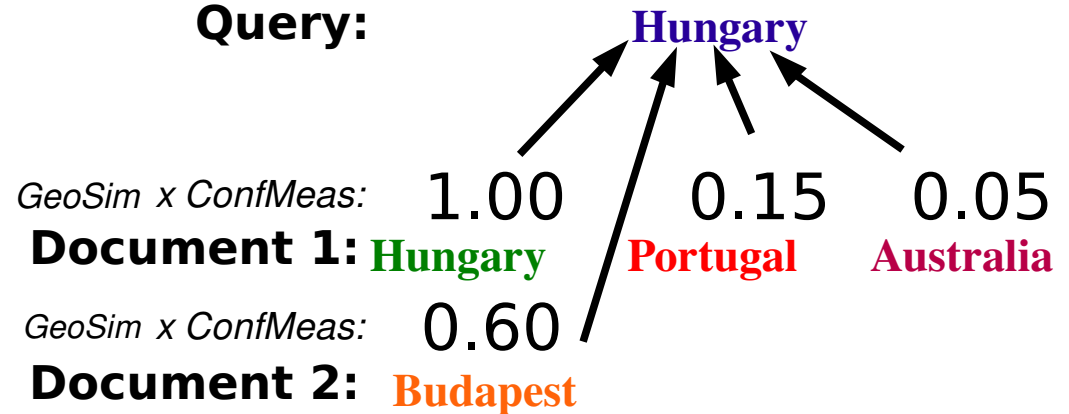
**Document 1:**

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

**Document 2:**

(...) there are many tourist attractions (...) in **Budapest**.

**Query:**



GeoScore	Mean	Max.	Bool.
<b>Document 1</b>	0.40	1.00	1.00
<b>Document 2</b>	0.60	0.60	0.00

# Experiments in GeoCLEF 2007

**Challenge:** outperform classic IR.

#	Description
<b>IR</b> 1	Baseline using <b>classic</b> IR approach. Geographic QE before RF, but <b>just terms</b> : no GeoScore.
<b>GIR</b> 2	Geographic IR approach. Geographic QE <b>before</b> or <b>after</b> RF
<b>IR/GIR</b> 3	Initial run: classic IR. Geographic IR approach after RF

# Results in GeoCLEF 2007

		IR	GIR		IR/GIR
	GeoScore	Terms only	Geo. QE before RF	Geo. QE after RF	Terms/GIR
Initial run		<b>0.210</b>	0.126	0.084	<b>0.210</b>
Final Run	Maximum		0.125	0.104	0.205
	Mean	<b>0,233</b>	0.022	0.021	0.048
	Boolean		<b>0.135</b>	<b>0.125</b>	<b>0.268</b>
	Null		0.115	0.093	0.021

a) Results for the Portuguese monolingual subtask.

Initial run		<b>0,175</b>	0.086	0.089	<b>0.175</b>
Final Run	Maximum		0.093	0.104	<b>0.218</b>
	Mean	<b>0.166</b>	0.043	0.044	0.044
	Boolean		<b>0.131</b>	<b>0.135</b>	0.204
	Null		0.081	0.087	0.208

b) Results for the English monolingual subtask.

# On IR vs GIR...

It seemed straightforward that:

- splitting the *what* and the *where* part at the beginning...
- ...expand separately each half...
- ...use two indexes (term and geographic indexes)...

...would produce better results than classic IR, but we could not confirm that.

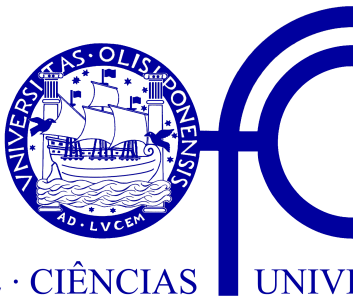
# An example:

## ***“shipwrecks of Portuguese boats”***

- *Shipwrecks* suggest coastal areas, near islands etc; not mountains or airports.  
*thematic term has geographic semantics*
- geographic terms are also good expansion terms (ex: the names of Portuguese islands)  
*geographic terms are good QE terms*

# Future Work

- Evaluate the “weight” of each query term in the thematic and geographic parts.
- Try mixed strategies (ex: sea traffic in Portugal @ Portugal)
- Feature type-oriented query expansion has its merits.
- Next step: mature the GIR system for further experiments



FACULDADE · DE · CIÊNCIAS | UNIVERSIDADE · DE · LISBOA

- Thank you for your attention.
- Questions?

# Query Expansion through Geographical Feature Types

Nuno Cardoso and Mário J. Silva

`{ncardoso, mjs}@xldb.di.fc.ul.pt`

