FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

The University of Lisbon at GeoCLEF 2007

Nuno Cardoso, David Cruz, Marcirio Chaves and Mário J. Silva

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

In 2006...

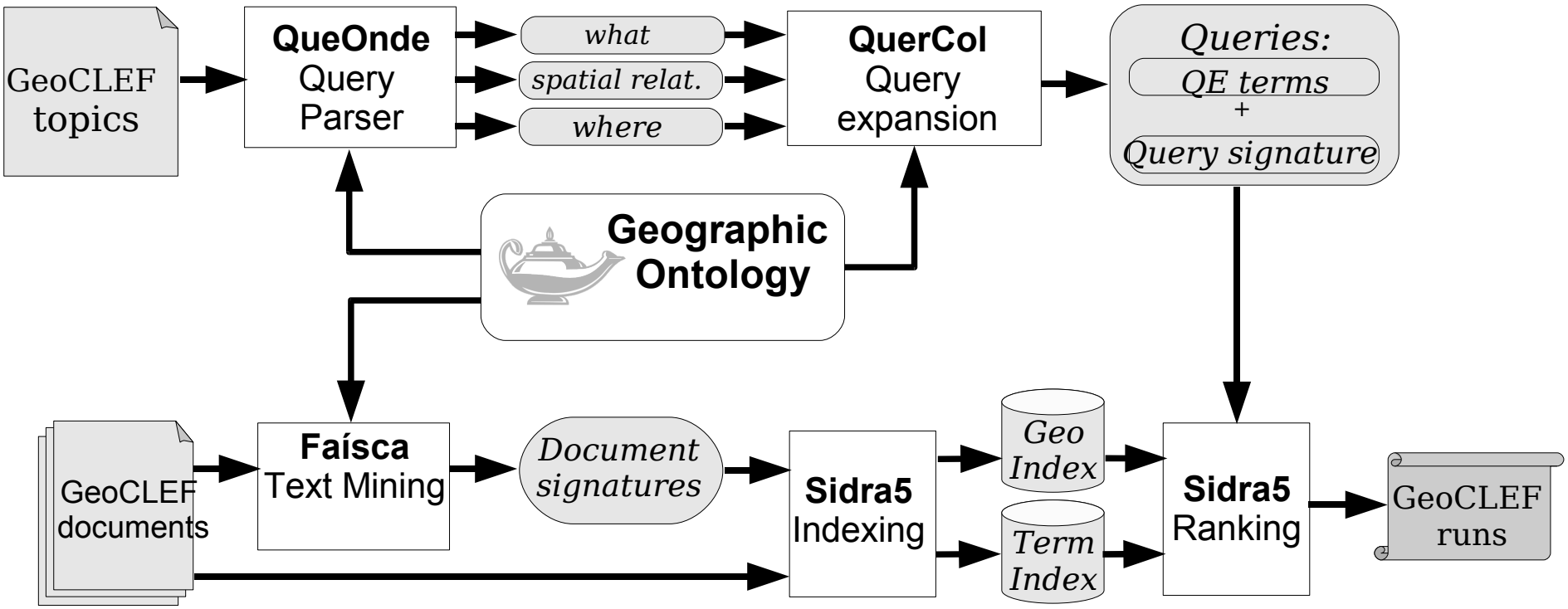
Results revealed some limitations:

- assigning one single geographic concept as a **scope** to each document limited the geo-ranking.
- some topics were not handled properly (e.g., "*diamond trade in Angola and South Africa*").
- classic IR approach still prevails!

For 2007:

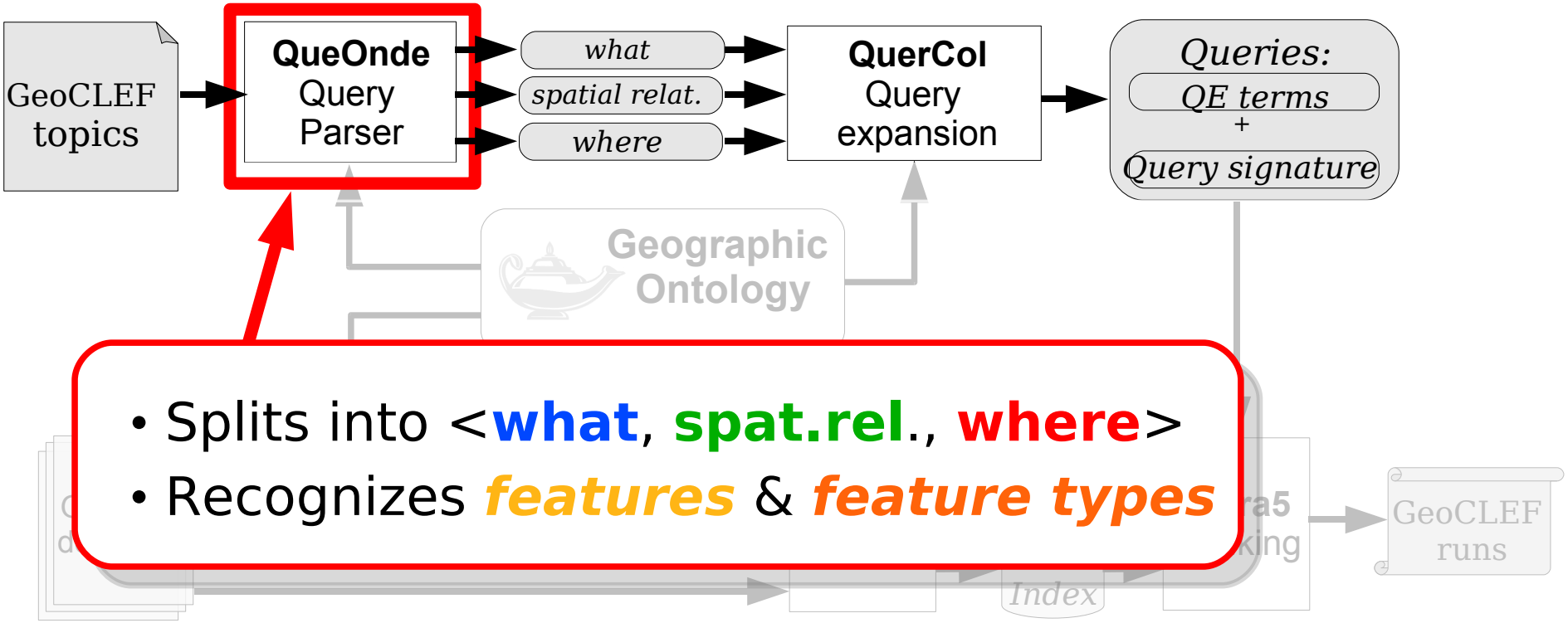
- Challenge: outperform classic IR.
- Generation of *geographic signatures* for both queries and documents.
- Geographic query expansion focused on *features*, *feature types* and *spatial relationships*.
- Geographic ranking improvements.

XLDB's 2007 GIR system



XLDB's 2007 GIR system

1. Query Processing



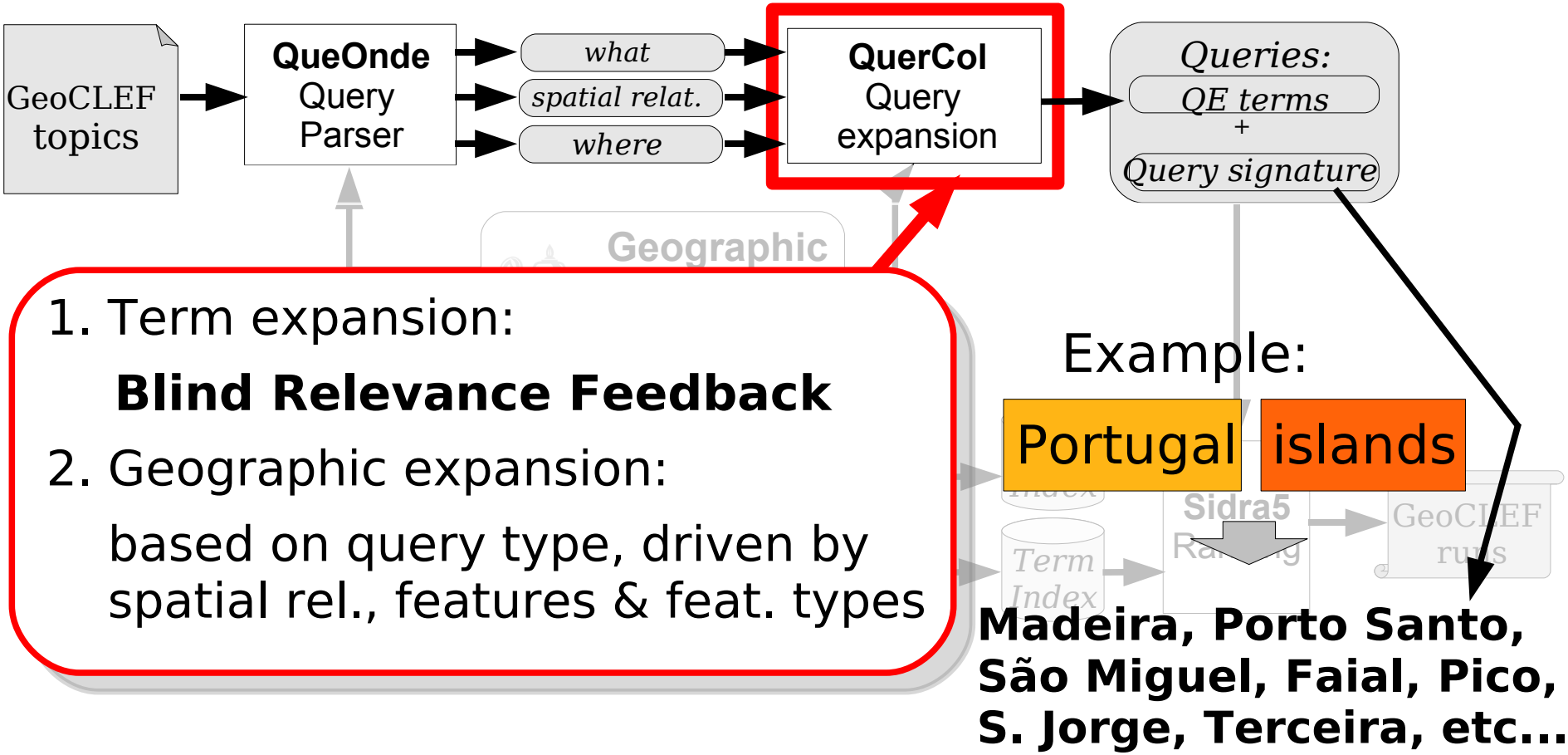
- Splits into <**what**, **spat.rel.**, **where**>
- Recognizes ***features*** & ***feature types***

Example: Sea traffic in Portuguese islands =

Sea traffic | in | Portugal | islands

XLDB's 2007 GIR system

1. Query Processing



1. Term expansion:

   Blind Relevance Feedback

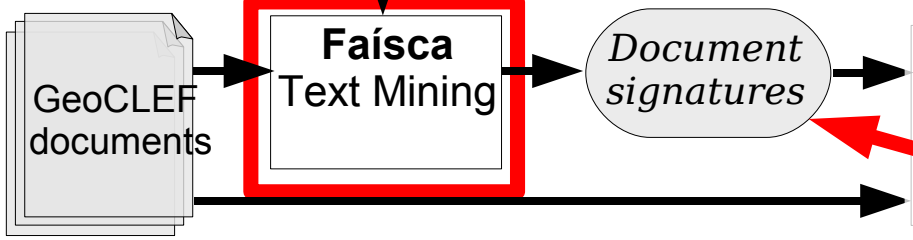
2. Geographic expansion:

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

XLDB's 2007 GIR system

2. Text Mining

Searching documents for geographic evidence in *<feat + feat types>* and *<feat types $ feat>* patterns.
Ex: **Lisbon Airport**
Airport of Lisbon



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]

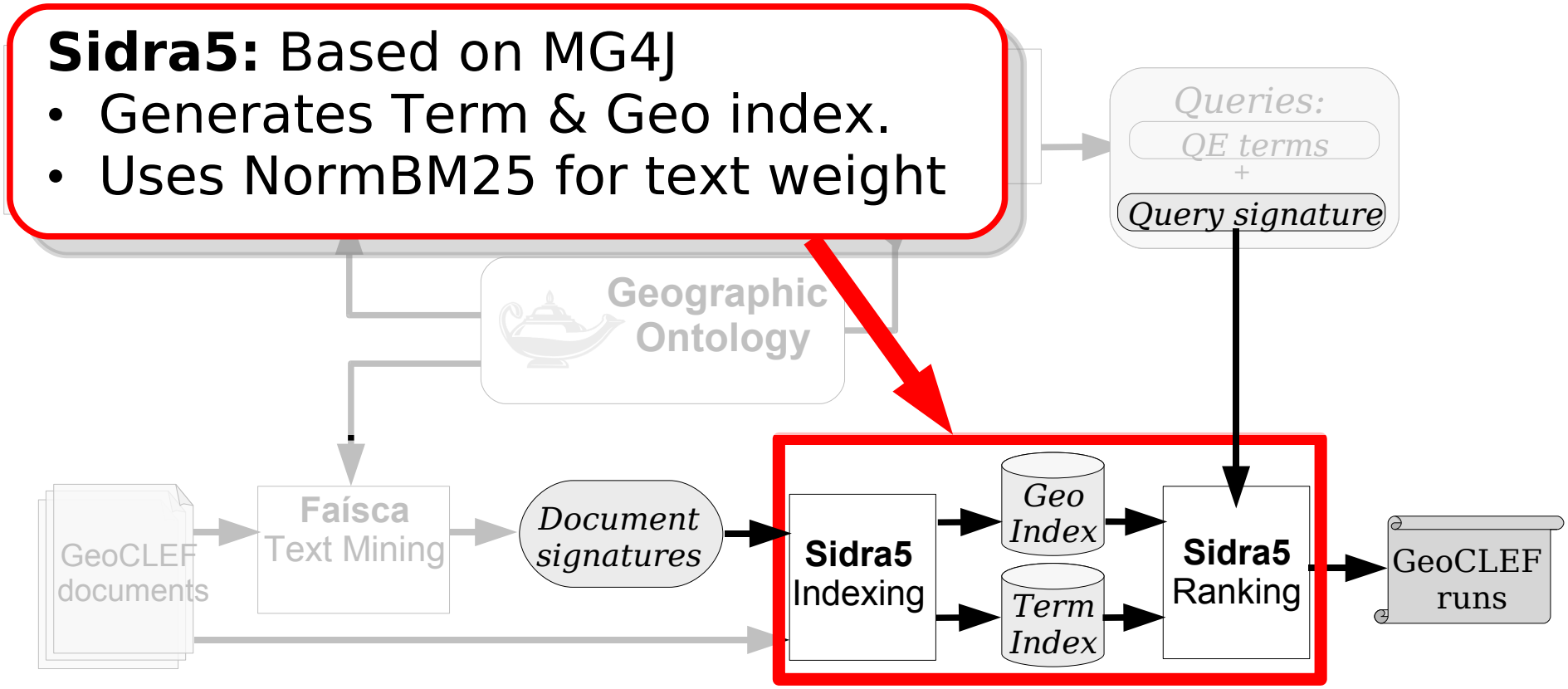
ID *ConfMeas*

Faísca generates *document signatures:*

XLDB's 2007 GIR system

3. Geographic Ranking

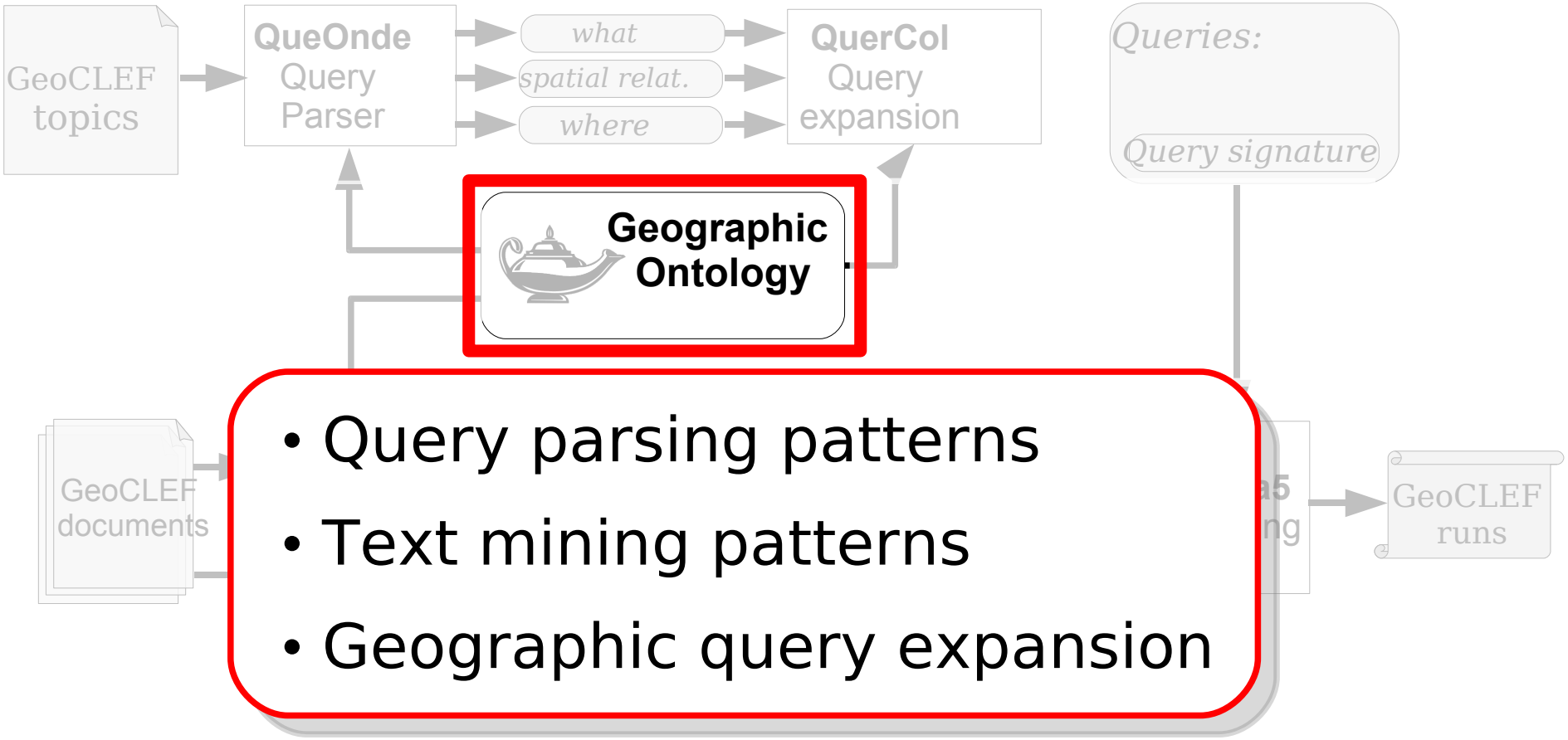
Sidra5: Based on MG4J
- Generates Term & Geo index.
- Uses NormBM25 for text weight



- For geographic weight... how to measure geographic relevance between query signatures and doc signatures?

XLDB's 2007 GIR system

4. Geographic Reasoning



- Query parsing patterns
- Text mining patterns
- Geographic query expansion

Blind Relevance Feedback

"sea traffic in Portuguese islands"

"(sea | ocean | overseas) & (traffic | routes | cruising | ...) & (boats | fishing | ...) in SCOPE

Initial query

Initial retrieval

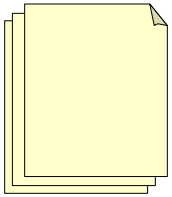
Initial run

blind rel. feedback

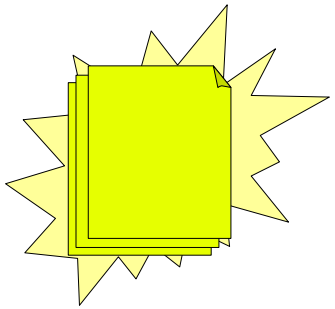
Final query

Final retrieval

Final run



+ ***Madeira, Porto Santo, Pico, Faial, S. Jorge, Graciosa, Terceira, ...***



Blind Relevance Feedback

Initial query

Initial retrieval

Initial run

blind rel. feedback

Final query

Final retrieval

Final run

? When is it better to insert the geographic restrictions?

Blind Relevance Feedback

When is it better to insert the geographic restrictions?

BEFORE Relevance Feedback?

"sea traffic in Portuguese islands"

Initial query

Initial retrieval

Initial run

blind rel. feedback

Final query

Final retrieval

Final run

"sea traffic in Madeira, Porto Santo, Pico, Faial, S. Jorge, Graciosa, Terceira, ...

"(sea | ocean | overseas) & (traffic | routes | cruising | ...) & (boats | fishing | ...) in Madeira, Porto Santo, Pico, Faial, S. Jorge, Graciosa, Terceira, ...

Blind Relevance Feedback

When is it better to insert the geographic restrictions?

or AFTER Relevance Feedback?

"sea traffic in Portuguese islands"

Initial query

Initial retrieval

Initial run

blind rel. feedback

Final query

Final retrieval

Final run

"sea traffic in Portugal

"(sea | ocean | overseas) & (traffic | routes | cruising | ...) & (boats | fishing | ...) in Madeira, Porto Santo, Pico, Faial, S. Jorge, Graciosa, Terceira, ...

Geographic relevance

- In 2006: **one** *GeoSim* for each pair (s_{query}, s_{doc}).
- In 2007: **multiple** *GeoSim* for each pair ($sign_{query}$, $sign_{doc}$).
  - Combination of multiple *GeoSim* values into a single *GeoScore*.

Geographic relevance (cont.)

GeoSim 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: **Hungary**

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
Document 2	0.60	0.60	0.00

Experiments

#	Description
1	Baseline using **classic** IR approach. Geographic QE before RF, but **just terms**: no GeoScore.
2	Geographic IR approach. Geographic QE **before** or **after** RF
3	Geographic IR approach. Test the **GeoSim** combinations.

Experiments

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

a) Results for the Portuguese monolingual subtask.

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

b) Results for the English monolingual subtask.

Questions raised:

- Geographic QE before blind RF seems to help. Why?
  - Shouldn't term query expansion be geographic-independent?
  - ...or are geographic terms also good thematic terms?
- Classic IR was outperformed by IR/GIR run: does it mean that we are finally using GIR the right way?
- Boolean and Maximum *GeoSim* combinations still inconclusive... and also dependent on the quality of the ontology.

Future Work

- Interesting results...
  - Why blind RF performs better with geographic criteria? Is it statistically significant?
  - Outperforming classic IR: coincidence... or not?
- Feature type-oriented query expansion has its merits.
- Next step: mature the GIR system for further experiments

The end.

- Thank you for you attention.
- Questions?

The University of Lisbon at GeoCLEF 2007