FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

Linguateca

To separate or not to separate? Reflections about current GIR practice

Nuno Cardoso and Diana Santos

ncardoso@xldb.di.fc.ul.pt, Diana.Santos@sintef.no

ECIR Workshop on Novel Methodologies for
Evaluation in Information Retrieval, NMEIR 2008

Glasgow. 30th March, 2008

Introduction of Geographic IR

(on a 'GeoCLEF-ish' perspective – Gey et al, 2006)

- Main objective: Retrieve relevant documents for queries with a geographic scope of interest.
  - “*Unemployment in the EU*”
  - “*Lakes with monsters outside Scotland*”
  - “*Rock concerts on the Mediterranean shores*”
  - “*UFO sights in deserts*”

Main challenges for GIR

- Capture the **thematic** and the **geographic** criteria given by the user.
- Retrieve documents that are relevant to the **subject** and within the desired **geographic scope** of interest.
- Reason over the geographic domain (e.g., isles of Scotland ⇢ *Islay*, *Jura*, *Skye*, *Unst*, ...)
- Understand spacial relationships ('*in the*', '*around*', '*between*', '*on the shores of*', ...)

Linguateca

Current GIR status

(based on past GeoCLEF editions):

- Classic IR still obtains better results :(
- Most research groups have not shown significant improvements compared to classic IR systems

Did we miss completely the goal of researching best practices for GIR in an organized way, taking for granted many (precipitated) assumptions, and rushing to conclusions?

Motivation for this work

- Understand why we are not achieving the expected results:
  - What have we learnt from our failures?
  - Are we insisting in dead-end GIR approaches?
  - Can we blame the approach or the system?
- Perform a thoroughly analysis of the topics, to understand where the GIR system drifts from the expected behavior.
- Our goal: get back in track with GIR, an **intrinsic natural language task**.

Overview of the Presentation

1) Dissecating GIR approaches

- GIR search subspaces
- A separation case study: XLDB's GIR system
- XLDB's GeoCLEF result analysis

2) To separate or not separate?

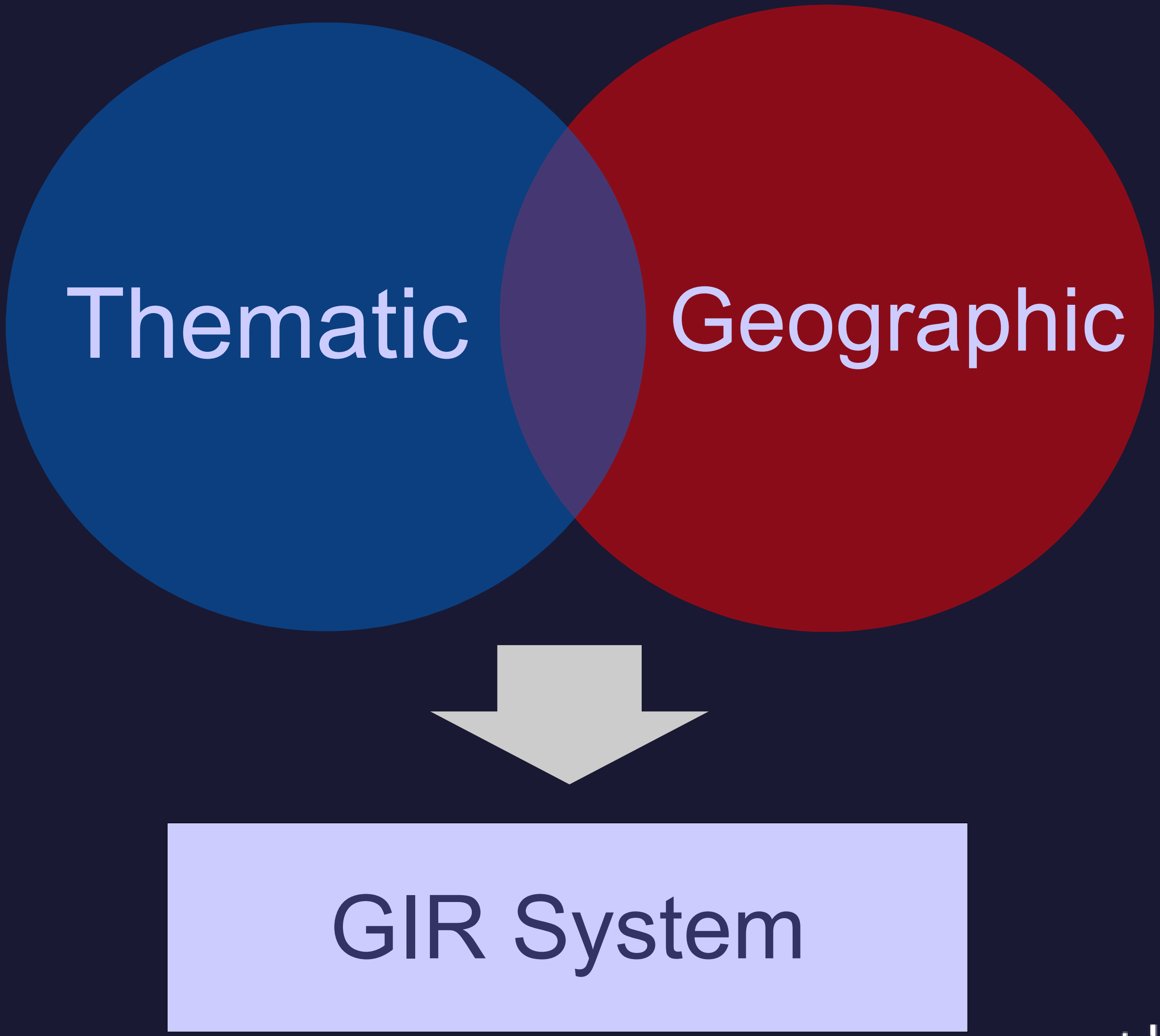
- Arguments for not separating
- Final remarks

GIR sub-spaces (Cai, 2002)

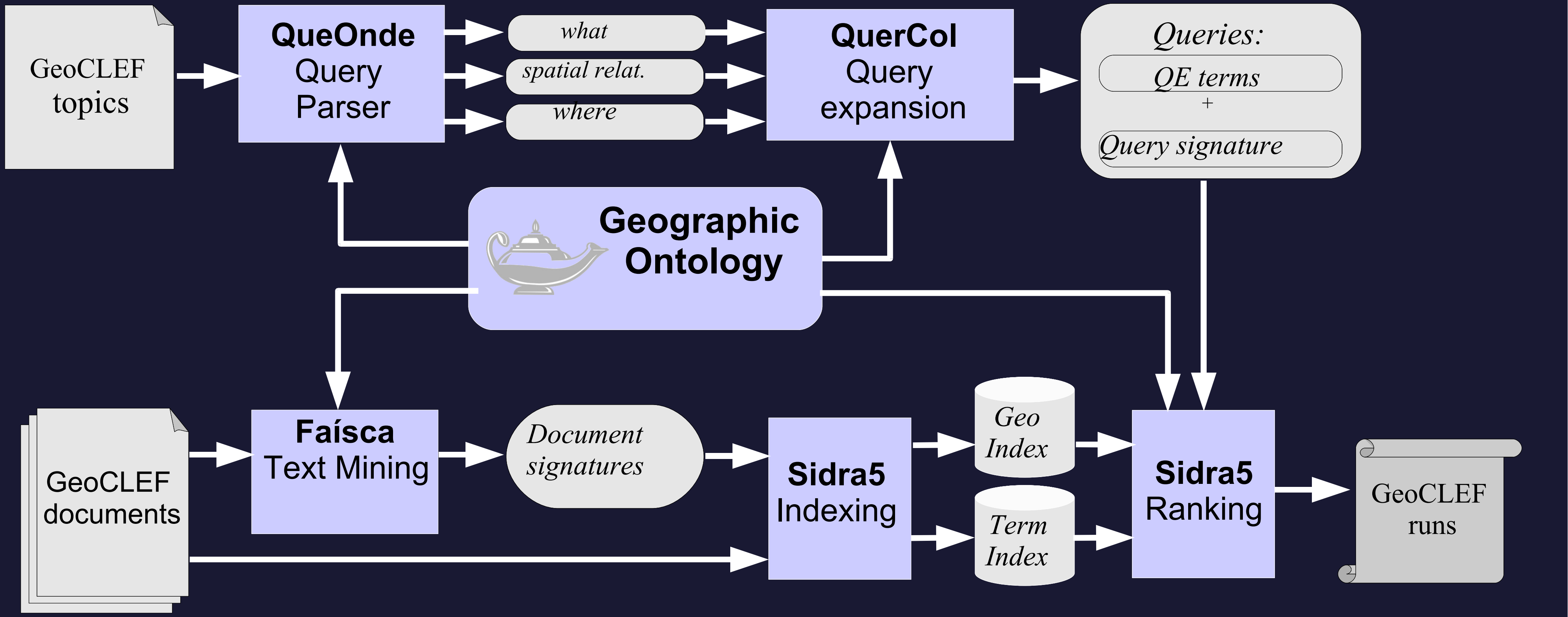
Typical GIR systems model the relevance in **two** search sub-spaces:

- **Term subspace**, the natural “habitat” for classic IR systems, well handled by statistic and probabilistic models;
- **Geographic subspace**, where computing relevance requires reasoning capabilities over a web of knowledge on the geographic domain.

GIR sub-spaces



Case Study: XLDB's GIR system

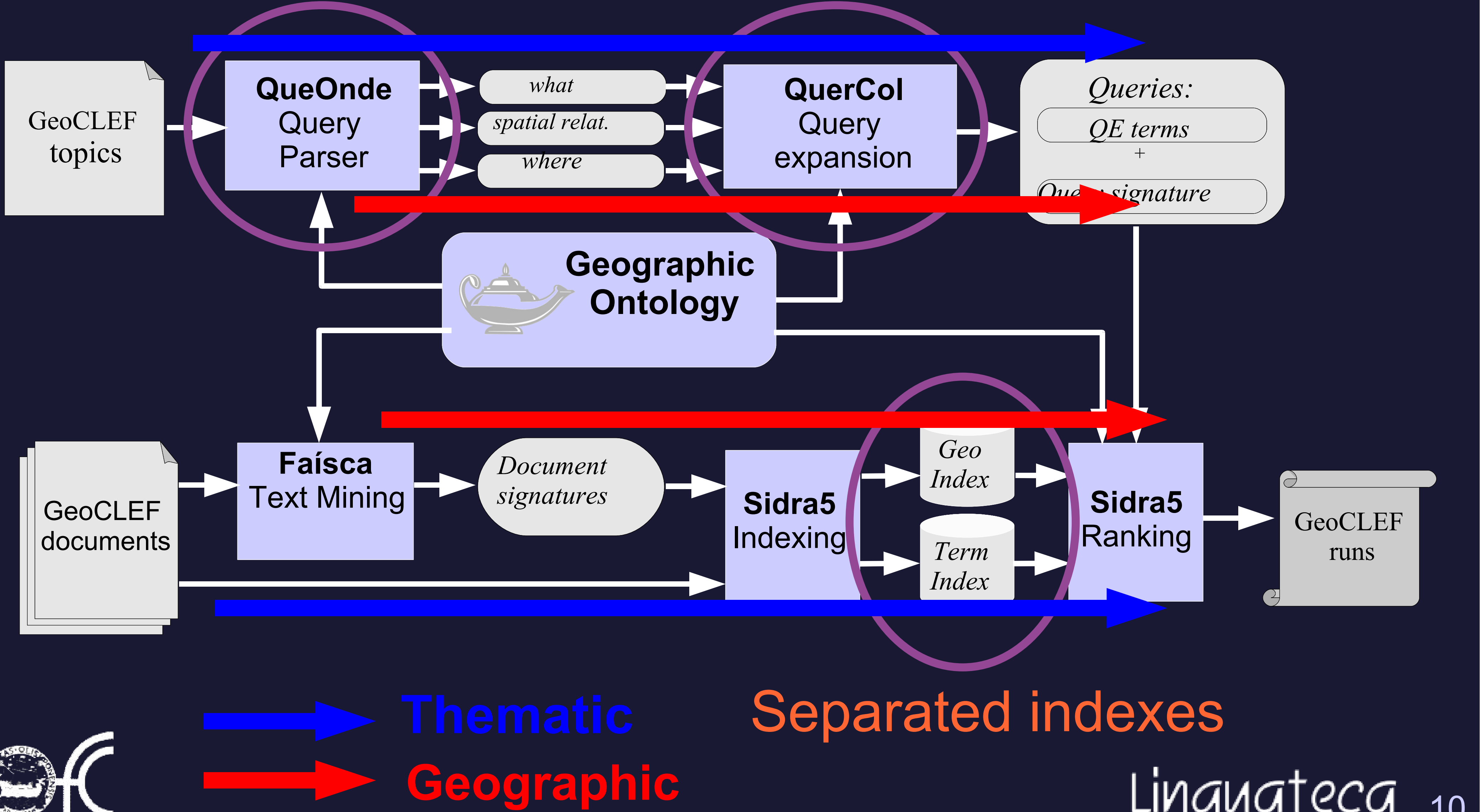


Linguateca

Case Study: XLDB's GIR system

Query Splitting

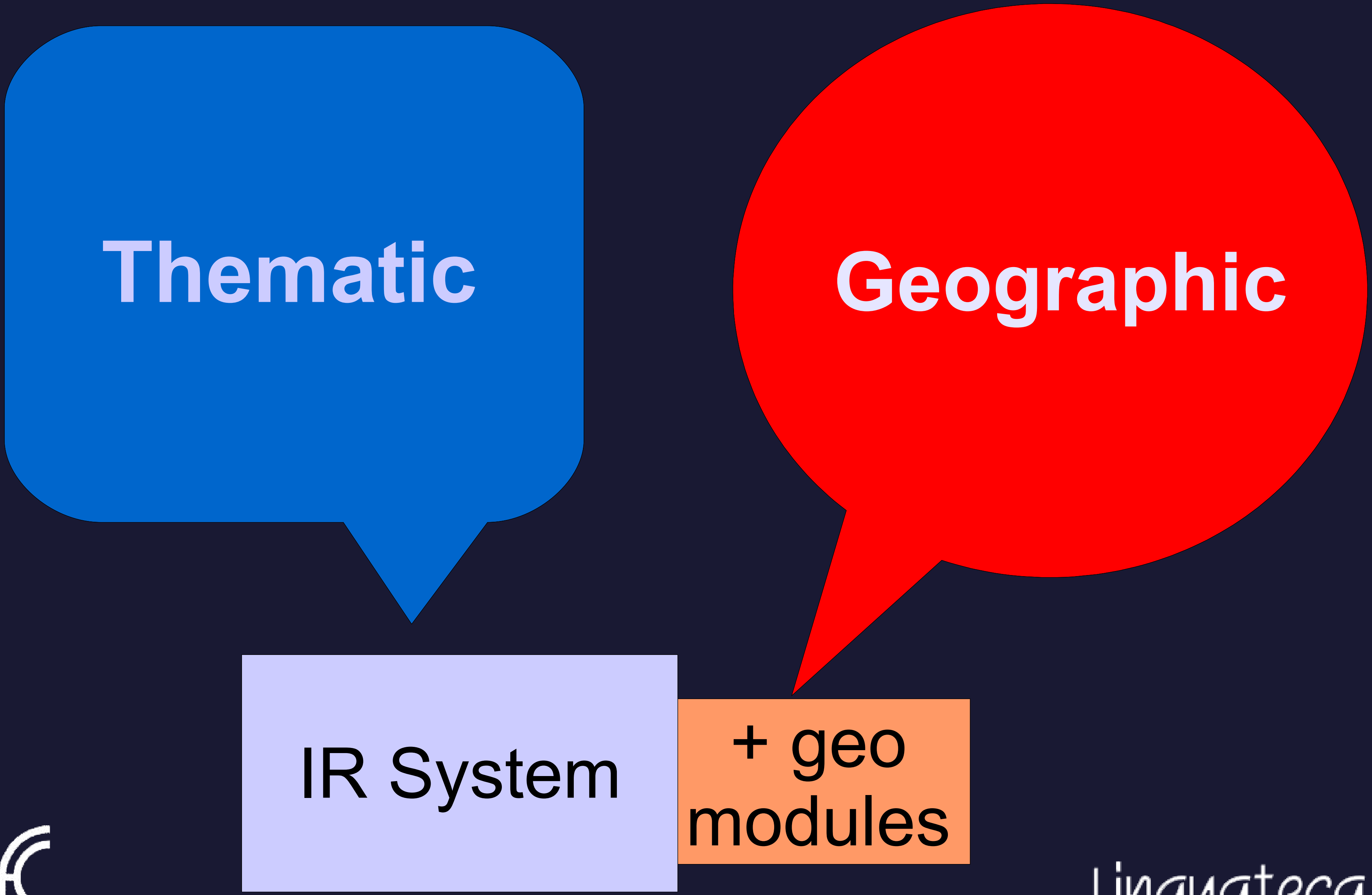
Term expansion and geographic expansion follows different paths.



Separated indexes

Linguateca

XLDB's GIR approach



Thematic

Geographic

IR System

+ geo modules

XLDB's GIR approach (cont.)

“Thematic and geographic relevance subspaces are best handled separatedly”

- Augment an existing IR system with custom geographic modules;
- Split queries into a thematic and geographic parts; each part “feeds” a different module;
- Query expansion: different strategies for each part;

XLDB's GIR approach (cont.)

- Mine document collection for ***place names***, disambiguate and ground into unique geographic features;
- Use additional geographic indexes to store and provide geographic information;
- Combine term weighting and geographic weighting into a single relevance score.

XLDB's GeoCLEF results

- Classic IR experiments had **better** MAP values than GIR experiments (!)
- A more detailed analysis revealed many interesting facts:

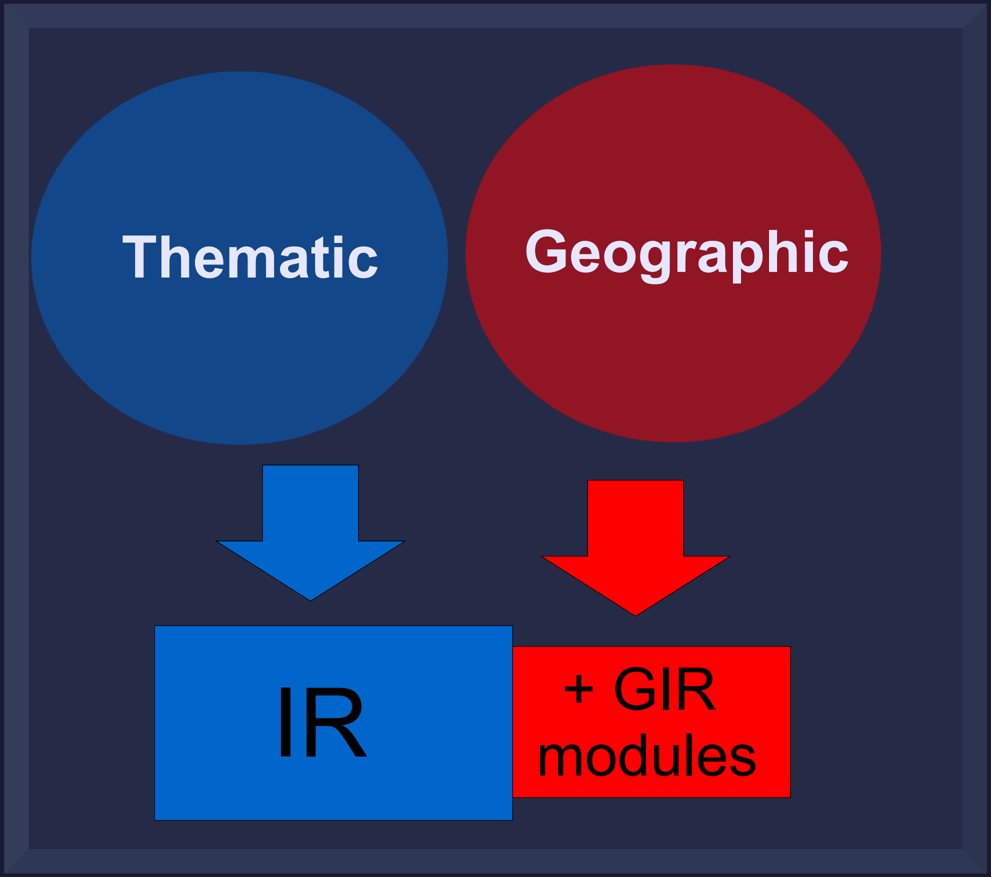
  1) Blind-relevance feedback **re-injected** geographic terms for the final query;

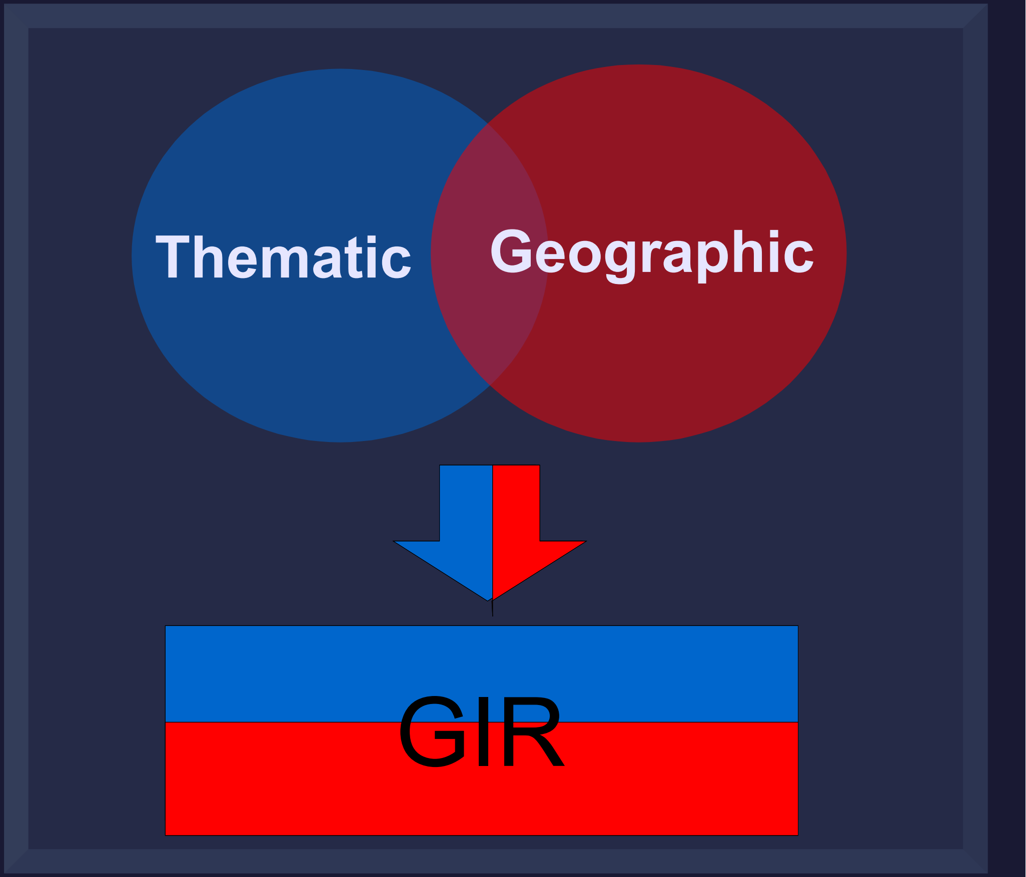
  2) Some document scopes were not sucessfully detected, because they were **hinted by other landmarks**.

Linguateca

Reflections in GIR practice

To separate or not to separate?





Arguments for not separating

1) Geographical themes

– *Where is Honolulu?* vs *What is Judo?*

2) Geographic part is contextual

- "*Find hotels near where I am standing*"...

3) Separate: is it at all possible?

- "*Cities near volcanos*"

- "*Lakes with monsters*" (≠ *monsters in lakes*)

Arguments for not separating (cont.)

3 cont.) Separate: is it at all possible?

- "*Events in St Paul Cathedral*" ⇢ How to expand "**events**"? Different subjects for cathedrals, islands and cities...
- "*Tourism in Northern Italy*" ⇢ "tourism", in the context of this query, is more likely to be expanded to "monuments, sightseeing", etc.
- "*Economy at the Bosphorous*" ⇢ Are we mainly interested in the economy, of in the Bosphorous area?

Arguments for not separating (cont.)

4) Search argument

Only place names convey geographic meaning?

- “*Hotels near Eiffel tower*”
- “*Accidents in Paris-Dakar*”

5) Is it **useful**?

- Geographic terms are also good retrieval terms;
- Geographic evidence can also be captured from the subject.

E.g., “*shipwrecks*” ⇢ in oceans or deserts?

Arguments for not separating (cont.)

6) Is it feasible? Is NLP mature enough?

– “*the Lisbon treaty*”, “*Brussels decided that*”...

– ”*the Portuguese architecture of the XV century*”

Real case for GeoCLEF query: “*Airplane crashes near russian cities*”

- retrieved document referring that a russian airplane crashed in China.
- simple place name capture and codification in a geographic index loses important context information.

Linguateca

Arguments for not separating (cont.)

6 cont.) Is it feasible? Is NLP mature enough?

HAREM – a Portuguese NER evaluation contest - showed that there is a significant proportion of "place" names that represent their governing heads **(Santos et al, 2006, Santos and Chaves, 2006)**

"*Washington declared...*" ⇢ Government

"*Brazil won the World cup*" ⇢ Soccer team

"*North Korea voted...*" ⇢ Government

"*Lisbon raised against...*" ⇢ People

See also

Lessons learnt

- Why insist in separate thematic and geographic subspaces? It seems a questionable practice...
  - place names are good terms for IR retrieval
  - geographic evidence on queries and documents is present in other named entities. "*Eiffel Tower*", "*Space Needle*", "*the Guggenheim Museum*", "*The Glasgow Stadiums*", "*Apple headquarters*", etc
- GIR task is more dependent on **message understanding** than IR. And doing it wrong can generate poor results.

Linguateca

Lessons learnt (cont.)

- GeoCLEF's GIR evaluation is vital to assess the fitness of our GIR systems.
  - But MAP values are ***not enough*** to assess the results of each topic.
- A thorough analysis of the results is a **key step** to understand:
  - why the GIR system fails;
  - how we should reposition our GIR research efforts.

Lessons learnt (cont.)

GeoCLEF's 2008 **GikiP** pilot task: “GeoCLEF task on Wikipedia documents”

- Overcome limitations that newspaper documents offer for some geographic queries
- Example: “Cities near volcanos”
  - Newspaper documents probably refer such cities when there are eruptions. “Quiet volcanos” are no good newspaper articles...
  - Wikipedia documents probably has the city page, the volcano page, the eruption page, etc.

Plea to GIR practitioners

- Try to measure specifically the impact of separation in **your** system
- Perform a detailed analysis by topic (a lot of work, but very rewarding!)
- Compare on further evaluations the separation / merging approaches
- Investigate the synergy between theme and geography

Acknowledgment

This work was jointly funded by the European Union (FEDER and FSE) and the Portuguese Government, under contracts ISFL/13/408 (FIRMS-FCT), 339/1.3/C/NAC (Linguateca) and PTDC/EIA/73614/2006 (GREASE II).
The first author acknowledges FCT grant SFRH/BD/29817/20

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER/FSE

FCCN
Fundação para a Computação Científica Nacional

fc

Linguateca

FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

Linguateca

The end. Questions?

To separate or not to separate? Reflections about current GIR practice

Nuno Cardoso and Diana Santos

ncardoso@xldb.di.fc.ul.pt, Diana.Santos@sintef.no

ECIR Workshop on Novel Methodologies for Evaluation in Information Retrieval, NMEIR 2008
Glasgow. 30th March, 2008

References

- **(Cai, 2002)** G. Cai. GeoVSM: An Integrated Retrieval Model for Geographic Information. In Proceedings of the Second International Conference on Geographic Information Science, GIScience’02, pages 65–79, London, UK, 2002. Springer-Verlag.
- **(Gey et al, 2006)** F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough., GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In C. Peters et al, editors, Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF’2005. Revised Selected papers, LNCS 4022, pp 908–919. Springer, 2006.
- **(Santos and Chaves, 2006)** D. Santos and M. S. Chaves. The place of place in geographical IR. In Proceedings of the 3rd Workshop on Geographic Information Retrieval, GIR’2006 (held at SIGIR’2006), pages 5–8, Seattle, WA, USA, 10 August 2006.
- **(Santos et al, 2006)** D. Santos, N. Seco, N. Cardoso, and R. Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, and D. Tapias, editors, Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC’2006, pages 1986–1991, Genoa, Italy, 22-28 May 2006.