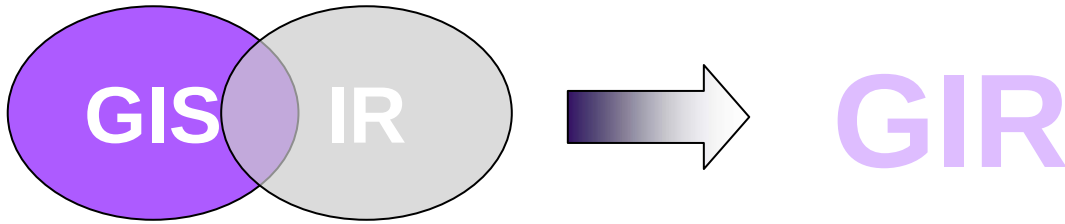
An Ontology-based Index to Retrieve Documents with Geographic Information

Miguel R. Luaces, José R. Paramá, Oscar Pedreira, Diego Seco

Database Laboratory
University of A Coruña
A Coruña, Spain

Motivation

- The Database Laboratory at the University of A Coruña works in two very active research fields:
  - Geographic Information Systems (GIS)
    - EIEL Project (http://www.dicoruna.es/webeiel)
  - Information Retrieval (IR)
    - Galician Virtual Library (http://bvg.udc.es)



Retrieval of geographically and thematically relevant documents in response to a query of the form <theme, location>

Motivation

- Many of the documents stored in digital libraries and document databases include geographic references
  - Example: *"…the hurricane touched land at Veracruz…"*
- Few index structures or retrieval algorithms take into account these geographic references
- Some proposals have appeared recently. But…
- There are some specific particularities of geographic space that are not taken into account by these proposals:
  - Hierarchical nature of geographic space
  - Topological relationships between the geographic objects

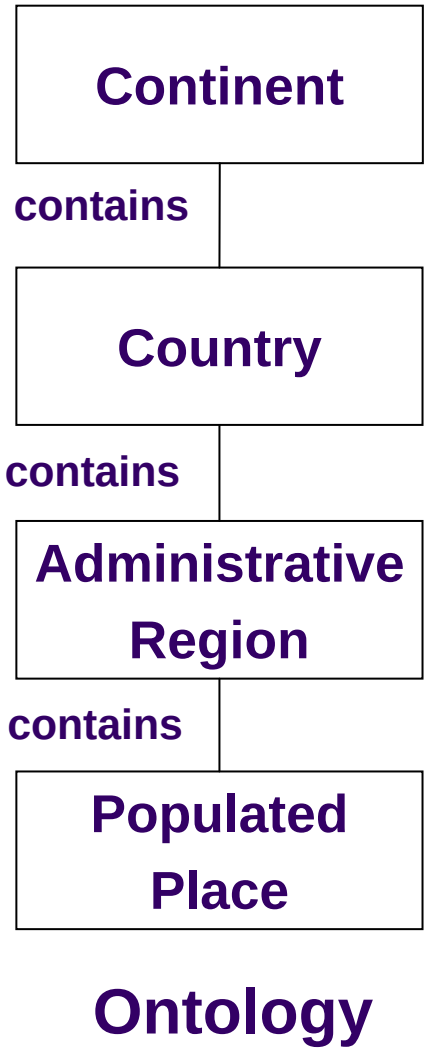
Contents

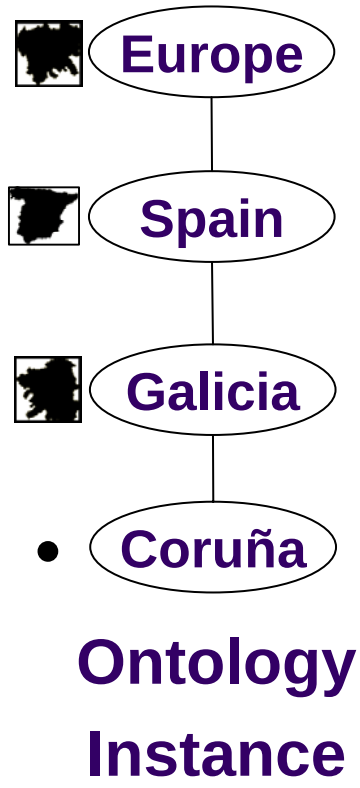
Related Work

- Many index structures and techniques have been proposed in each area:
  - IR: Inverted Index
    - Geographic references are completely ignored
  - GIS: R-Tree
    - These structures do not take into consideration the hierarchy of space
- The combination of both types of indexes (SPIRIT project):
  - *Text-First*
  - *Geo-First*
    - They do not take into account the relationships between the geographic objects that they are indexing
- An Ontology can describe the specific characteristics of geographic space:
  - Ontologies are used in query expansion, relevance rankings, and web resource annotation
    - Nobody has ever tried to combine it with other types of indexes to have a hybrid structure

Contents

Architecture





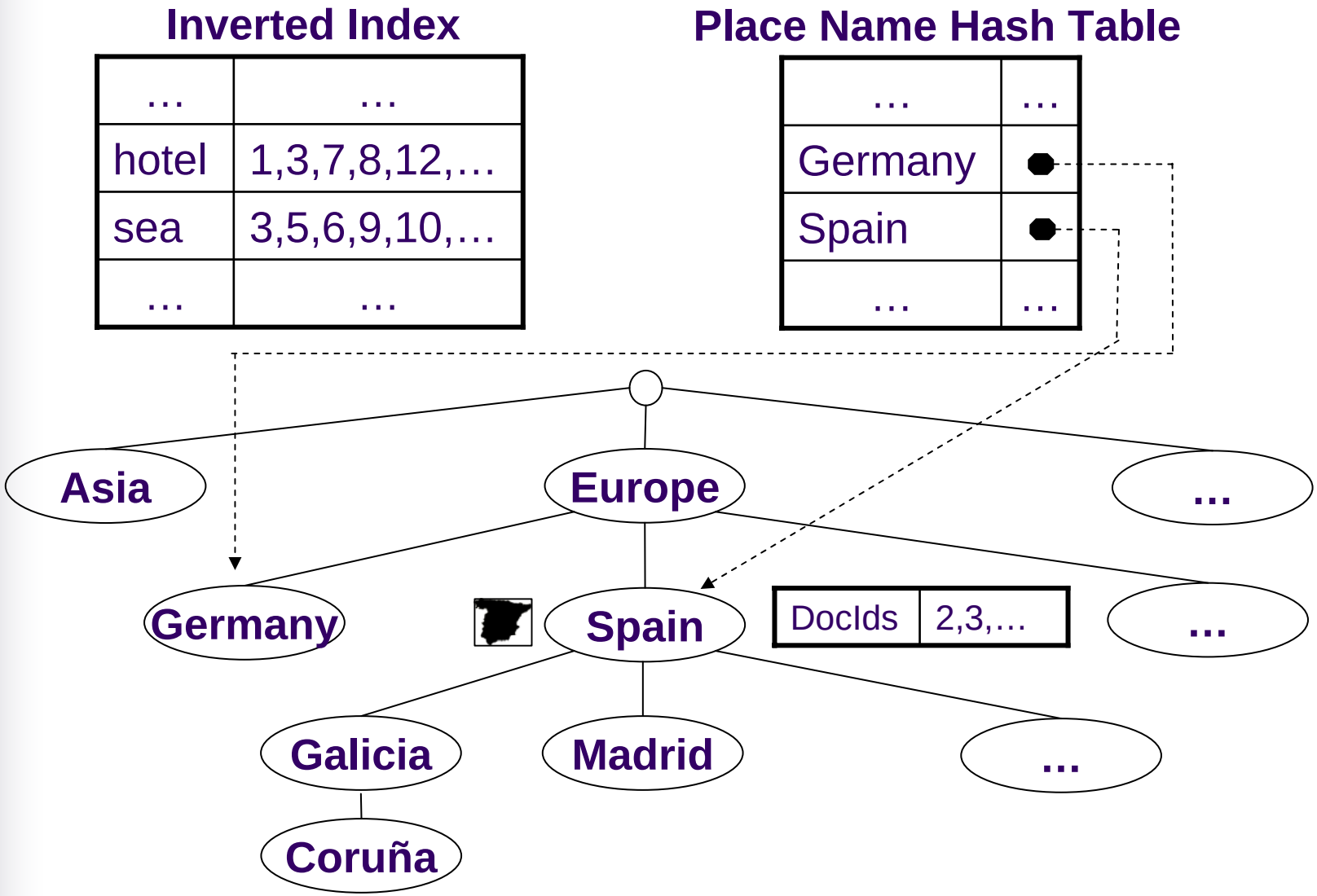
Architecture

- Document storage workflow:
  - Keyword extraction
    - Classic IR techniques (removing stopwords, stemmers)
  - Build the index structure
    - Natural Language Processing Techniques
    - Gazetteer Service
    - Geographic Space Ontology Service
- Processing services:
  - Query solving service
  - Web Map Service (WMS)
  - Geographic information retrieval module
- User Interface:
  - Administration (manage the document collection)
  - Query

Contents

- Motivation
- Related Work
- Architecture
- **Index Structure**
- Supported Query Types
- Experiments
- Conclusions and Future Work

Index Structure



Index Structure

- Based on a spatial ontology
  - It models both the vocabulary and the spatial structure of places for purposes of information retrieval
- Tree composed by nodes that represent place names
  - Each node:
    - Keyword (a place name)
    - Bounding box
    - Document identifiers
    - Children nodes
  - These nodes are connected by means of inclusion relationships
  - If the list of children nodes exceeds a threshold, an R-Tree is used
- Auxiliary structures:
  - Place name hash table
  - Traditional Inverted Index

Index Structure

- Advantages:
  - All textual queries and spatial queries can be efficiently processed
  - Queries combining textual and spatial aspects are supported
  - Updates and optimizations in each index are handled independently
- Drawbacks:
  - The tree that supports the structure is possibly unbalanced
  - The structure is static

Contents

- Motivation
- Related Work
- Architecture
- Index Structure
- **Supported Query Types**
- Experiments
- Conclusions and Future Work

Supported Query Types

- Pure textual queries
  - *"Retrieve all documents where the words* hotel *and* sea *appear"*
  - How do we solve it?
    - Inverted Index
- Pure spatial queries
  - *"Retrieve all documents that refer to the following geographic area"*
  - How do we solve it?
    - Descend the structure + refine the result
    - The same algorithm that is used with spatial indexes

Supported Query Types

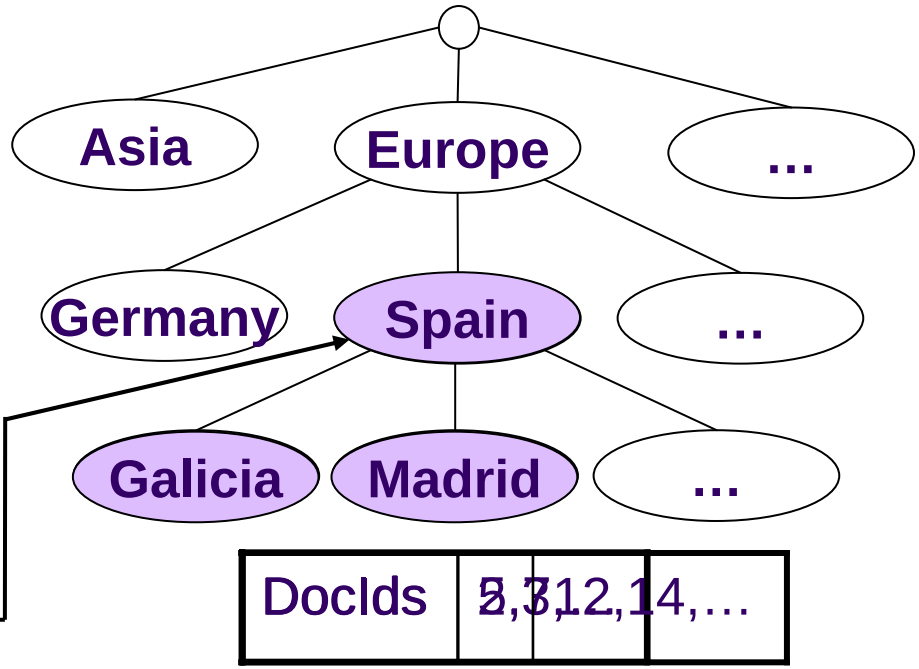
- Textual queries with place names
  - *"Retrieve all documents with the word* hotel *that refer to* Spain*"*
  - How do we solve it?
    - Example
  - We save some time by avoiding a tree traversal

Supported Query Types

Inverted Index

…	…
hotel	1,3,7,8,12,…
sea	3,5,6,9,10,…
…	…

Index Structure



Place Name Hash Table

…	…
Spain	
Germany	
…	…

Text Result	1,3,7,8,12,…
Spatial Result	2,3,5,7,12,14,…
Query Result	3,7,12,…

Supported Query Types

- Textual queries over a geographic area
  - *"Retrieve all documents with the word* hotel *that refer to the following area"*
  - How do we solve it?
    - Example
  - Geographic references can be given using place names

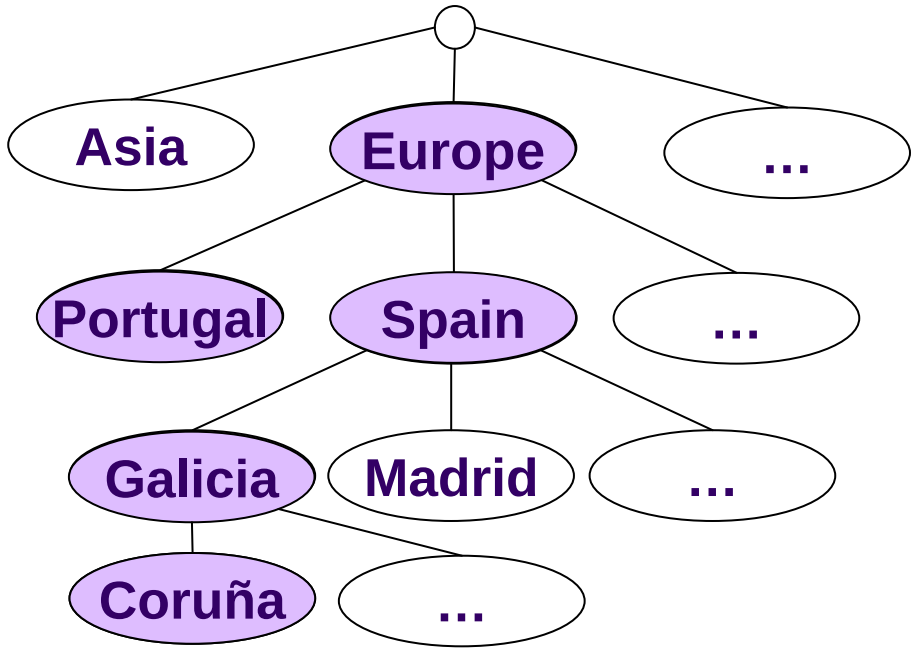
Supported Query Types

Inverted Index

…	…
hotel	1,3,7,8,12,…
sea	3,5,6,9,10,…
…	…

Query Window

Index Structure



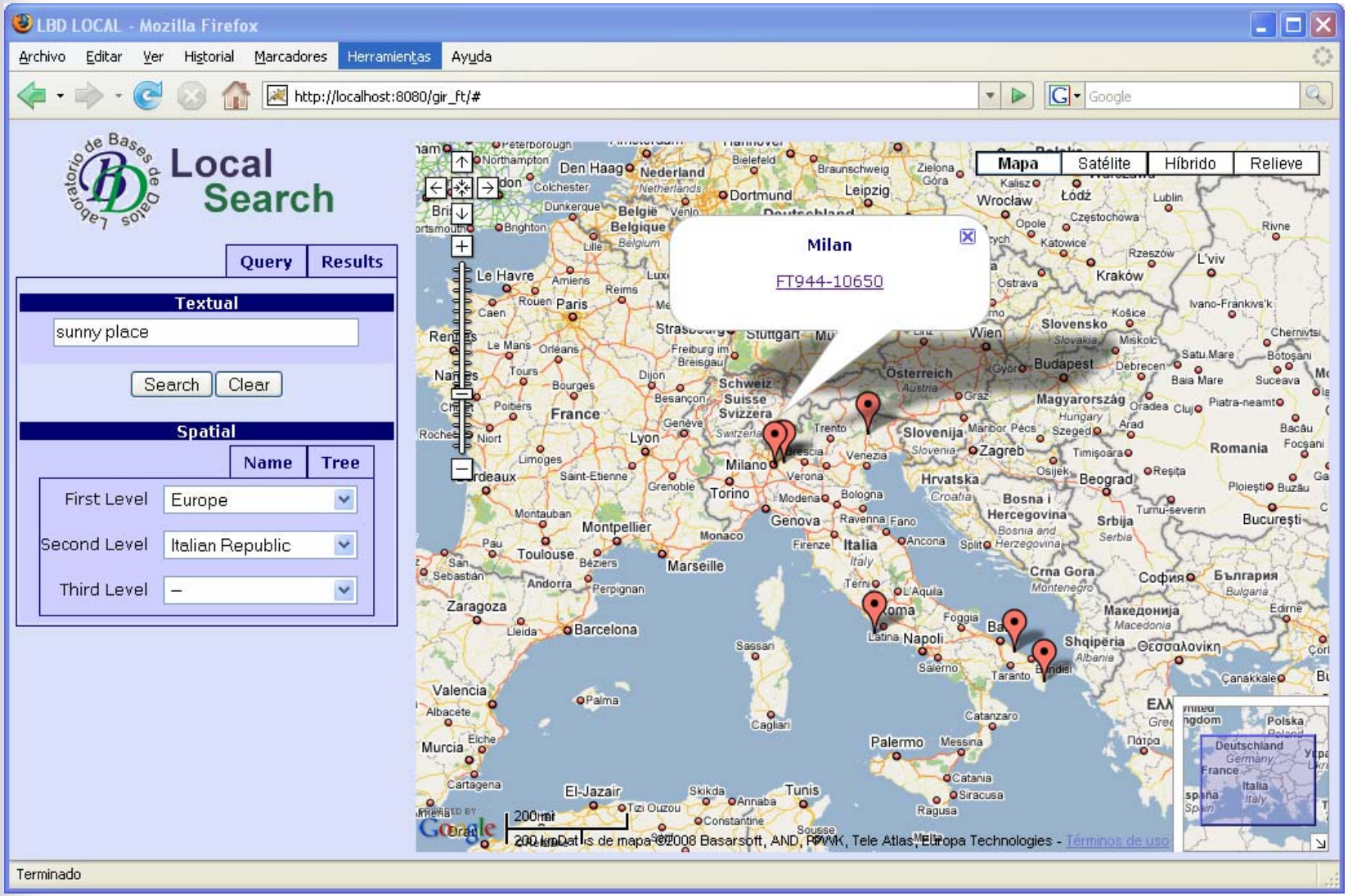
DocIds	12,14,…

Text Result	1,3,7,8,12,…
Spatial Result	12,14,…
Query Result	12,…

Supported Query Types

- Another improvement: *QUERY EXPANSION*
  - *"retrieve all documents that refer to Spain"*
  - How do we solve it?
    - The *Query Evaluation Service* discovers that *Spain* is a geographic reference
    - The *Place Name Hash Table* locates quickly the internal node that represents the geographic object *Spain*
    - All the documents associated to this node are part of the result
    - All the documents associated to the subtree are part of the result
  - The result contains not only those documents that include the term Spain, but also all documents that contain the name of a geographic object included in *Spain*

Demo

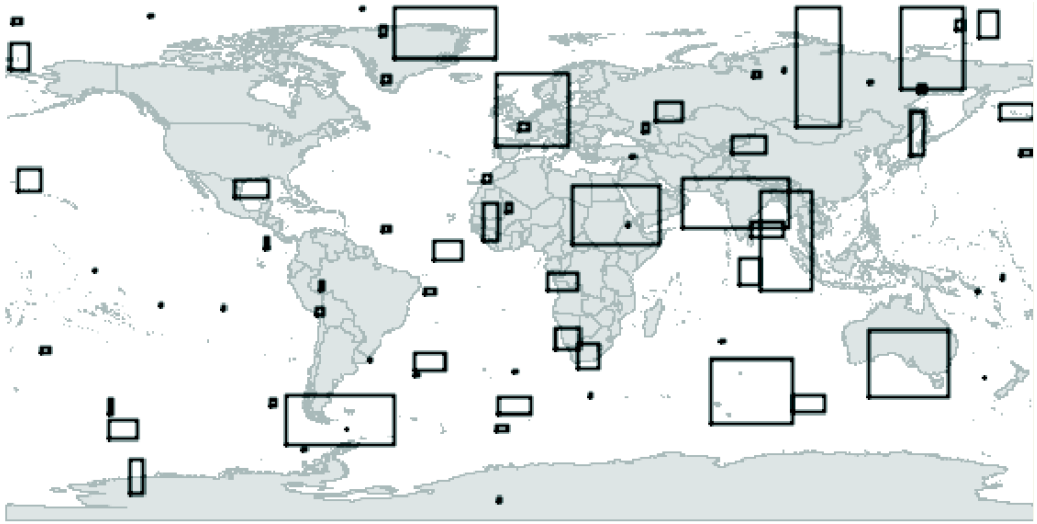

LBD LOCAL - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://localhost:8080/gir_ft/#
Local Search
Query
Results
Textual
sunny place
Search
Clear
Spatial
Name
Tree
First Level
Europe
Second Level
Italian Republic
Third Level
–
Milan
FT944-10650
Mapa
Satélite
Híbrido
Relieve
Terminado

Contents

- Motivation
- Related Work
- Architecture
- Index Structure
- Supported Query Types
- **Experiments**
- Conclusions and Future Work

Experiments

- Randomly generated pure spatial queries



- Ontology-based index versus R-Tree

Query area	0.001%	0.01%	0.1%	1%
Ontology-based index	0.013	0.017	0.052	0.360
R-Tree	0.010	0.016	0.057	0.370

Experiments

- Ontology-based index versus R-Tree (zones of high document density)

Query area	0.001%	0.01%	0.1%	1%
Ontology-based index	0.03	0.11	1.05	9.84
R-Tree	0.07	0.22	1.64	12.85

- Ontology-based index versus R-Tree (zones of low document density)

Query area	0.001%	0.01%	0.1%	1%
Ontology-based index	0.02	0.03	0.09	0.4
R-Tree	0.02	0.03	0.07	0.2

Contents

Conclusions

- We have defined an architecture for Geographic Information Retrieval systems
- It takes into account both textual and geographic references in the documents
- This is achieved by a new index structure that combines an inverted index, a spatial index and an ontology
- Traditional queries and new types of queries can be solved

Future Work

- Evaluate the performance of the index structure
- Explore the use of different ontologies
- Include other types of spatial relationships (e.g., adjacency)
- Improve the disambiguation process of place names
- Define algorithms to rank the retrieved documents

An Ontology-based Index to Retrieve Documents with Geographic Information

Miguel R. Luaces, José R. Paramá, Oscar Pedreira, Diego Seco

Contact: luaces@udc.es

Database Laboratory
University of A Coruña
A Coruña, Spain