



FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

## Adding Geographic Scopes to Web Resources

---

Mário J. Silva, Bruno Martins, Marcírio  
Chaves, Nuno Cardoso, Ana Paula Afonso

<http://xldb.fc.ul.pt>

---

### You heard it in the halls

- Many Web information sources are primarily relevant to geographically delimited communities.
- Users frequently use geographic terms as “filters” in queries

# Tumba!

- **National Web search engine**
  - Public service since 2002
- See it in action at **<http://tumba.pt>**



## Geographic Information on Web Pages/Logs

- Portuguese Web test collection (3.5 Mdocs)
- Counted # of Municipality references
  - 308 municipalities
  - 8,000 to 500000 voters each
- 2.17 times/document (average)
- occur in 4% of the queries



## Scope/Context definitions

- Orkut Buyukkokten et al., WebDB'99
  - **globality of entities**, conjectures that web linkage could be used.
- Junyan Ding, Gravano et al., VLDB'2000
  - **scope** is geographical area that the creator intends to reach
  - scopes defined both by linkage and content
- McKurley WWW10'2001
  - **geocoding HTML** or **geographic context** for web pages – no definition
- Jones et al., SIGIR'02
  - **geographical context** = “geographic information about a page”, later renamed **web page footprint**
- Amitay et al., SIGIR'04:
  - **focus** is the locality that a page discusses as a whole

## Geographical Scope of Web Resources

- Region where more people than average would find the resource relevant.

## Goal

- **Build and evaluate** a research search system for GeoIR.
  - Avoid getting spatial
  - Assign a scope (if it exists) to each page.

## Outline

- Motivation
- **Architectural Overview**
- Assigning Scopes to Web Pages
- Searching with Web scopes
- Evaluation plan
- Conclusion

# IR with Geographic Scopes

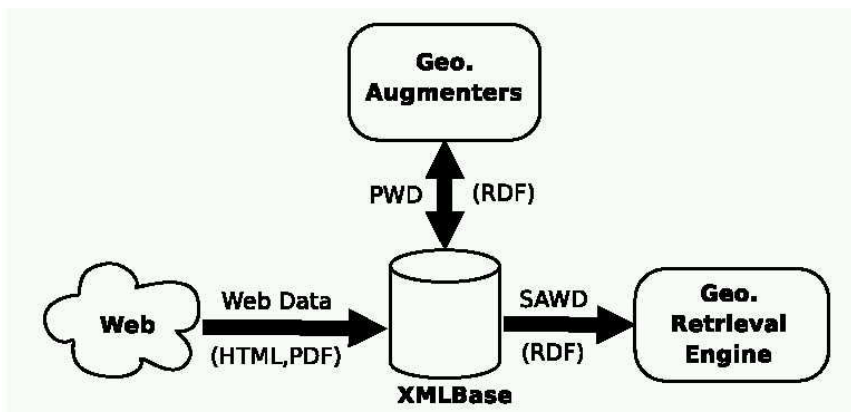
## Data Sources:

- Contents (text and meta-data)
- Linkage between pages
- **External Information Sources**
  - Mikheev et al., 1999: precision of geo-entities recognition w/o gazetteers is small.

## Goals:

- Rank / Cluster results by geo scope (**IF that matters**)
- Find resources nearby.
- Limit search to a particular region
- Display spatial distribution of the results

# Architectural Overview



Web Data



Purified Web Documents

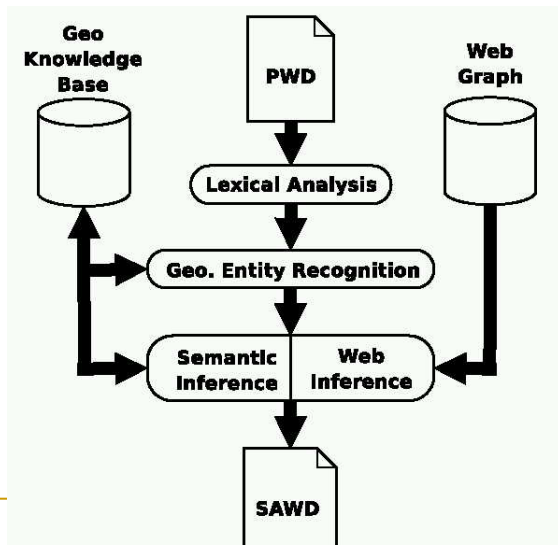


Scope Augmented Web Documents

## Outline

- Motivation
- Architectural Overview
- **Assigning Scopes to Web Pages**
- Searching with Web scopes
- Evaluation plan
- Conclusion

## Assigning Scopes



## Lexical Analysis

- Web data file ->PWD
  - Tokenization
  - HTML markup is important
  - HTML pages are not regular text documents

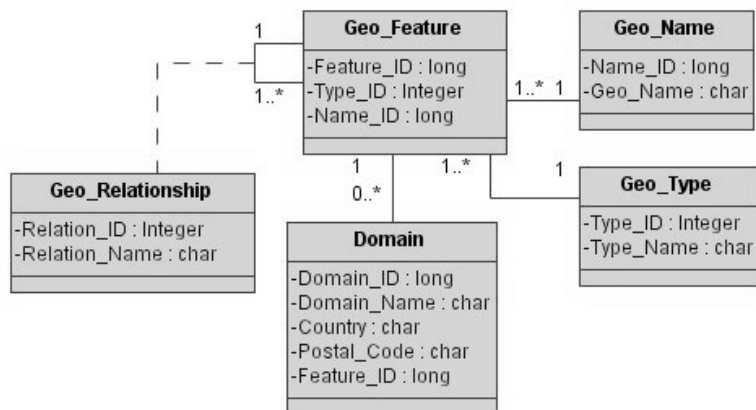
## Geo-Named Entity Recognition

- Machine Learning
  - Can tag names not in gazetteer
  - Needs large amounts of training data
- Simple pattern match
  - Detects
    - location expressions
    - GKB's place names
    - "Port of" + "New York"
  - Does not detect names not in gazetteer
    - We don't need them anyway!

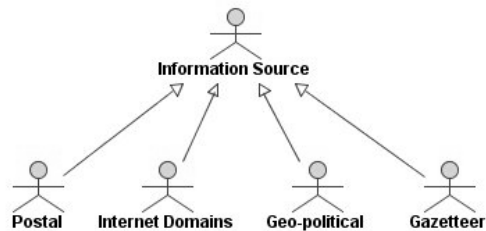
# GKB – Shared Knowledge Base

- DNS sub-domains information
  - Yahoo.com, foo.blogspot.com
- Names with known location
  - Cities, landmarks, postal codes, ...
- Relationships
  - Equivalent names
  - Hierarchical broader-narrower names
  - Adjacent names
  - Related names
- **Data quality** is an issue

# Prototype Class Diagram



## External Information Sources



- Postal codes database  
CTT, Portuguese Post Office
- .PT TLD registry information (postal codes only)  
FCCN, Registrar
- Geopolitical organization and demographics data  
INE, Statistics Bureau
- Phone plan numbering information  
ANACOM, National Telecom Authority

## Web Inference

- **Input:**
  - Web graph with nodes partially tagged
    - Many web pages don't have associated geo tags
- **Algorithm:**
  - Propagate tagged entities to adjacent pages
  - Pages from same site more closely associated
    - Example: small company site with "contacts" page
  - Set of tags shows spread of "scope"

## Semantic Inference

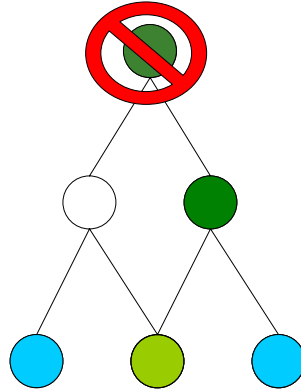
- **Inputs:**
  - Tagged pages
  - Tagged domains (whois registrants highly correlated to scope!)
  - Relationships among tagged entities (GKB)
  - Some GKB entities labelled as scopes
- **Output:**
  - Pages labelled with scopes
- **Algorithm:**
  - Weighted sum of tagged entities
  - Weights fine-tuned during evaluation

## Semantic Inference Algorithm (1)

- **Weights assigned to tagged entities**
  - Multipliers  $> 1$ : Anchors, title, meta, typeface, whois registrant
  - Multipliers  $< 1$ : propagated from same site, from different site
- **Weights “propagated” to GKB-related entities**
  - Equivalence: same score
  - Broader/narrower, Adjacent, Related: fraction of score

## Semantic Inference Algorithm (2)

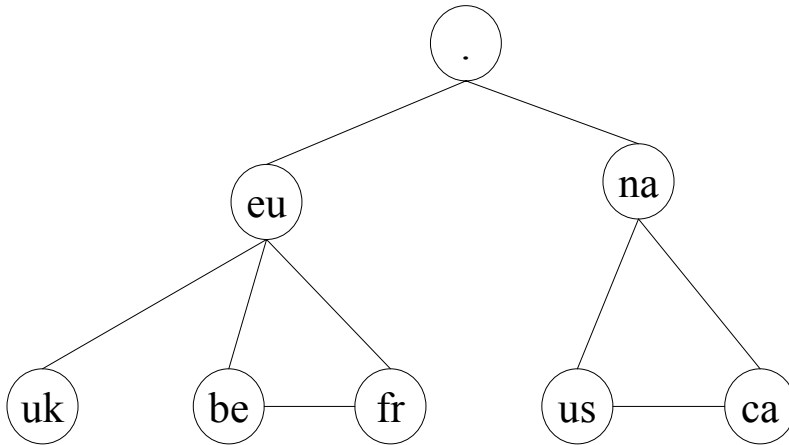
- Assign to page:
  - No scope if all < th
  - scope with highest score, if >> other scopes
  - parent scope, if it exists
  - scope with largest population count



## Outline

- Motivation
- Architectural Overview
- Assigning Scopes to Web Pages
- **Searching with Web scopes**
- **Evaluation plan**
- **Conclusion**

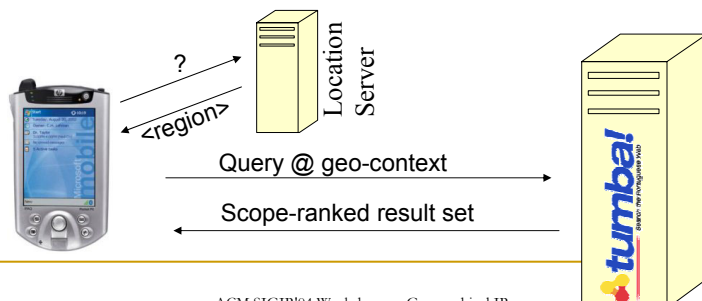
## Proximity measure



Topologic Distances Graph

## Search

- Translate user geo context (e.g. coordinates) into scope of query
- Detect global/local queries
- Incorporate proximity in global ranking function



## Evaluation

- Tagging – Extraction
  - Participation on joint evaluation on Portuguese NER
  - Use the consensus rules to derive heuristics
- Scopes Classification
  - Test collection
    - DMOZ data
      - Top:Regional:Europe:Portugal is very limited
    - Proprietary directories (not open)
- Retrieval Evaluation
  - Query becomes subject @ geo context
  - Relevance Judgments @ geo context
  - Joint evaluation task needed

## Conclusion

- Geographic search can be decomposed in several (still hard) problems
  - Ambiguity
  - Incorrect/incomplete data
  - Evaluation
- Much of the IR work on modelling and evaluation can be used as foundation