

A Comparison of Different Approaches for Assigning Geographic Scopes to Documents

Ivo Anastácio, Bruno Martins, Pável Calado

September 11, 2009



Outline

- 1 Context
- 2 Related Work
- 3 Methods
- 4 Evaluation
- 5 Results
- 6 Conclusions
- 7 Future Work

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

- **Geographic Scope** – Region that best describes the geographic content of a document
- Several applications
 - Contextual Advertising
 - Geographical Information Retrieval
 - Text Mining
 - ...
- No previous cross-method comparison of existing algorithms

Related Work

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

- Place reference recognition
e.g., Turkey — Bird or Country?
- Place reference disambiguation [Leidner, 2007]
e.g., Lagos — Nigeria or Portugal?
- Web Services
e.g., Metacarta Geotagger, Yahoo! Placemaker

Methods

Overview

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

- Yahoo! Placemaker
- GIPSY
- Web-a-Where
- GraphRank
- Baselines

Methods

Yahoo! Placemaker

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

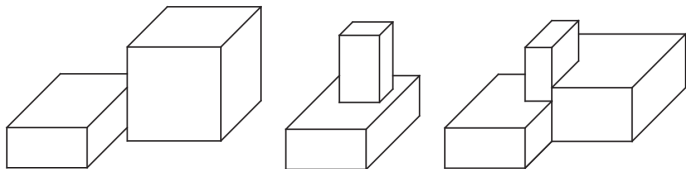
- No known implementation details



Methods

GIPSY [Woodruff and Plaunt, 1994]

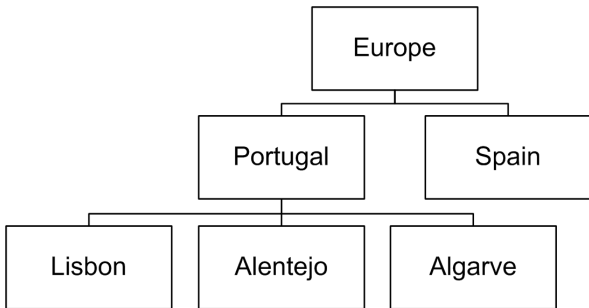
- Overlays the bounding boxes
- Highest bounding box is the scope



Methods

Web-a-Where [Amitay et al., 2004]

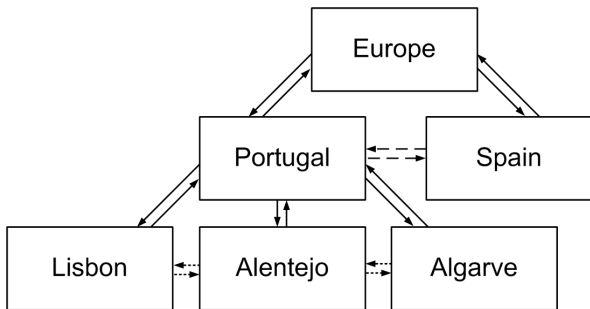
- Scores are propagated upwards in the hierarchy
- The scope has the highest accumulated score



Methods

GraphRank [Martins and Silva, 2005]

- Considers several relationship types
- Pagerank computes the scope



Methods

Baselines

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

- Most-Frequent location
- Cover area
- Cover area without outliers

Evaluation

- Yahoo! GeoPlanet has the gazetteer
- 6000 Web pages from the Open Directory Project (ODP)

	All Pages	Countries	States	Cities
Number of documents	6000	2000	2000	2000
Number of different topics	1440	692	665	303
Number of different regions	1127	1	51	1075
Average document length (bytes)	4143	4235	3841	4354
Average number of place references	9.2	9.1	9.1	9.4

- Metrics
 - Average distance
 - Average overlap
 - Accuracy (Exact match / Approximate match)

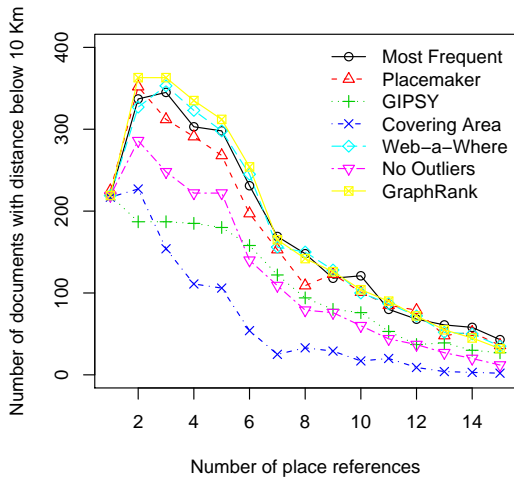
Results

	Avg. Distance (Km)	Avg. Overlap	Accuracy (D=0 Km)	Accuracy (D<100 Km)	Accuracy (O>0.75)
Placemaker Admin.	1030	0.42	0.37	0.45	0.39
GIPSY	1265	0.25	0.14	0.4	0.19
Web-a-Where	955	0.48	0.47	0.54	0.47
GraphRank	1083	0.48	0.47	0.53	0.48
Most Frequent	1093	0.49	0.37	0.55	0.43
Covering Area	2655	0.25	0	0.21	0.18
Non-outliers	1740	0.36	0.24	0.39	0.34

	Std.dev. Distance (Km)	Std.dev. Overlap
Placemaker Admin.	1460	0.49
GIPSY	2247	0.41
Web-a-Where	1890	0.49
GraphRank	1955	0.49
Most Frequent	2331	0.49
Covering Area	3009	0.38
Non-outliers	2826	0.46

- Web-a-Where and GraphRank have the best overall results
- The Most-Frequent baseline is very competitive

Results



Conclusions

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

- Cross-method comparison of geographic scope assignment algorithms
- Test pages from the ODP directory
- Web-a-Where, GraphRank and Most-Frequent had the best performance

Future Work

Context

Related Work

Methods

Evaluation

Results

Conclusions

Future Work

- Optimize parameters
- Experiment with machine learning approaches
- Measure how the quality of place name disambiguation affects the results

The End

Context
Related Work
Methods
Evaluation
Results
Conclusions
Future Work

Thank You!

Questions?