TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering

Daniel Ferrés and Horacio Rodríguez

TALP Research Center
Universitat Politècnica de Catalunya

CLEF 2007, 21 September, Budapest, Hungary

Outline

1. Introduction
2. System Overview
3. Document Retrieval
4. Experiments
5. Conclusions

TALPGeoIR

TALPGeoIR
Daniel Ferrés
Introduction
System Overview
Geographical Resources
Geographical Thesaurus
Collection Pre-processing
Shape Files Toolbox
Document Retrieval
Thematic IR
Geographical IR
Document Filtering
Experiments
Results
Conclusions
Future Work

- GIR system that combines thematic and geographical searches.
- An improved version of TALPGeoIR 2006 *[ferres-2006]*.
- Motivation at GeoCLEF 2007:
  - Using a state-of-the-art IR: *Terrier [Ounis-2006]*.
  - Using geographical knowledge to improve standard IR results.

System Overview

1 Introduction

2 System Overview
- Geographical Resources
- Geographical Thesaurus
- Collection Pre-processing
- Shape Files Toolbox

3 Document Retrieval

4 Experiments

5 Conclusions

Geographical Knowledge Base

- Geographical Gazetteers:
  - **GEOnet Names Server** (GNS). 5.3 million entries
  - **Geographic Names Information System** (GNIS). 39,906 entries (US. Concise subset)
  - **GeoWorldMap** (Geobytes Inc.). 40,594 entries
  - **World Gazetteer**: 29,924 cities

Geographical Thesaurus

- Information for each **geographical entry**: feature name, feature type base, geo-ontology parent, coordinates, (population).
- Alexandria Digital Library (ADL) **Feature Type Thesaurus**: 575 features *[hill-2000]*.
- **Disambiguation Hierarchy**: continent, sub-continent, capital, country, region (state),sea , summit, river, county (province), other.

Collection Pre-processing

TALPGeoIR
Daniel Ferrés
Introduction
System Overview
Geographical Resources
Geographical Thesaurus
Collection Pre-processing
Shape Files Toolbox
Document Retrieval
Thematic IR
Geographical IR
Document Filtering
Experiments
Results
Conclusions
Future Work

- Linguistic Pre-processing:
  - *Part-of-speech (POS) tags*. TnT *[brants-2000]*.
  - *Lemmas*. WordNet Lemmatizer *[fellbaum-1998]*.
  - *Named Entities*. Maximum Entrophy-based NERC (CoNLL 2003 English Dataset for training).
- Geographical Preprocessing with GeoKB.
- Indexing:
  - **Geographical Index**: feature type and geo-ontology path information and coordinates.
  - **Textual Index**: lemmatized content of the documents without added extra geographical information.

Shape Files Toobox

- *[pouliquen-2004]* propose the use of a publicly available database of 'shape files' for countries.
- 'shape files': encoding polygons that representing the 'border' of the area.
- Our main features with shape files:
  - 9-grid zone division. (North, East, North-East,...)
  - Close/Near points around a point P.

Document Retrieval

1 Introduction

2 System Overview
- Geographical Resources
- Geographical Thesaurus
- Collection Pre-processing
- Shape Files Toolbox

3 Document Retrieval
- Thematic IR
- Geographical IR
- Document Filtering

4 Experiments
- Results.

5 Conclusions
- Future Work

Terrier Configuration

- Thematic document retrieval over Terrier.
- All keywords are used for search (only stopwords removal).
- Lemma searching.

Selection of schemas based on experiments over the GeoCLEF 2006 data set:

- **TF-IDF** vs DFR vs BM25
- **Porter Stemmer** vs No stemmer
- **Blind Relevance Feedback** (docs=10;terms=40) vs No Relevance Feedback

Geographical IR using GKBs

- Obtains the set of documents that are geographically rellevant.
- Uses the geographical places and geographical feature types detected in the topics to perform the search.
- The feature types can be expanded with a list of synonyms extracted from GNS.
- Relaxed geographical search policy (e.g. a query that contains U.S. retrieves documents that contain New York).

Document Filtering

TALPGeoIR

Daniel Ferrés

Introduction

System Overview

Geographical Resources

Geographical Thesaurus

Collection Pre-processing

Shape Files Toolbox

Document Retrieval

Thematic IR

Geographical IR

Document Filtering

Experiments

Results

Conclusions

Future Work

- Documents retrieved by Terrier that have been also retrieved by the GKBs had priority over the other documents retrieved by Terrier.

GeoCLEF 2007 Experiments

TALPGeoIR
Daniel Ferrés
Introduction
System Overview
Geographical Resources
Geographical Thesaurus
Collection Pre-processing
Shape Files Toolbox
Document Retrieval
Thematic IR
Geographical IR
Document Filtering
Experiments
Results
Conclusions
Future Work

Table: 1. Description of the TALPGeoIR Experiments at GeoCLEF 2007.

Runs	IR System	Relevance Feedback	Border Filtering
TD1	Terrier	yes	-
TD2	Terrier & GeoKB	yes	-
TDN1	Terrier	yes	-
TDN2	Terrier & GeoKB	yes	-
TDN3	Terrier & GeoKB	-	yes

Global Results

TALPGeoIR
Daniel Ferrés
Introduction
System Overview
Geographical Resources
Geographical Thesaurus
Collection Pre-processing
Shape Files Toolbox
Document Retrieval
Thematic IR
Geographical IR
Document Filtering
Experiments
Results
Conclusions
Future Work

Table: 2. TALPGeoIR results at GeoCLEF 2007.

Run	IR System	AvgP.	R-Prec.	Recall (%)
TD1	Terrier	0.2711	0.2847	91.23%
TD2	Terrier & GeoKB	**0.2850**	**0.3170**	90.30%
TDN1	Terrier	0.2625	0.2526	**93.23%**
TDN2	Terrier & GeoKB	0.2754	0.2895	90.46%
TDN3	Terrier & GeoKB	0.2787	0.2890	92.61%

Conclusions

TALPGeoIR
Daniel Ferrés
Introduction
System Overview
Geographical Resources
Geographical Thesaurus
Collection Pre-processing
Shape Files Toolbox
Document Retrieval
Thematic IR
Geographical IR
Document Filtering
Experiments
Results
Conclusions
Future Work

- Geographical Knowledge improved standard IR.
- The approach with Terrier and the GeoKB was slightly better in terms of MAP than the one with Terrier alone.
- the BorderFiltering approach applied without Relevance Feedback improved slightly the results in MAP and Recall.
- Good results at GeoCLEF 2007.

Future Work

TALPGeoIR
Daniel Ferrés
Introduction
System Overview
Geographical Resources
Geographical Thesaurus
Collection Pre-processing
Shape Files Toolbox
Document Retrieval
Thematic IR
Geographical IR
Document Filtering
Experiments
Results
Conclusions
Future Work

- A precision-oriented toponym resolution (disambiguation) algorithm
- Experiments with the Divergence From Randomness schema.
- Improvement of the Shape Files toolbox and the Border Filtering algorithm.

Thanks!

TALPGeoIR

Daniel Ferrés

Introduction

System Overview

Geographical Resources

Geographical Thesaurus

Collection Pre-processing

Shape Files Toolbox

Document Retrieval

Thematic IR

Geographical IR

Document Filtering

Experiments

Results

Conclusions

Future Work

Thanks for your attention!

Questions?