Towards semantic-flavored queries for GIR systems: RENOIR at the GikiP pilot task.

Nuno Cardoso - XLDB
University of Lisbon, Faculty of Sciences, LaSIGE
ncardoso@xldb.di.fc.ul.pt
xldb.di.fc.ul.pt/wiki/Nuno_Cardoso

9th Workshop of CLEF 2008 - 17-19th September, 2008 - Århus, Denmark
Motivation

Research focus: **Geographic IR (GReaSE Project)**.

- Reason over the geographic domain.
- Handle geographic queries.
- Reformulate queries according to subject and geographic criteria.
- Ground geographic evidence from documents, assign scopes to them.
Motivation

“Biography of (Denzel) Washington”

“Attacks in Washington (D.C.)”

“Hotels in (lake) Washington”

“Studying in Washington (College)”
GIR Framework

GIR components (used at GeoCLEF 2008):

- **REMBRANDT** – named-entity recognition (NER) software for all kinds of entities.
- **QuerCol** – Query reformulator for non-geographic terms (BRF) and geographic terms (ontology-driven).
- **MG4J™** extended with multiple term indexes from **REMBRANDT's** annotations.

MG4J™ is developed by the University of Milan. [http://mg4j.dsi.unimi.it](http://mg4j.dsi.unimi.it)
Objectives

• Go beyond query strings.
  “Hotels `LOCAL:{Washington}`” vs “biography `PERSON:{Washington}`”? “`PERSON:? president of ORG:{USA}`”? 

• Use REMBRANDT's annotations to enrich QuerCol actions for a semantically-flavored query reformulation.

• MG4J's “biased” document retrieval can favour a given subject or geographic criteria.
Renoir

- Executes query procedures instead of query strings.
Query Procedures (QP)

QPs are pipeline of actions (automatic, manual and supervised) that express GikiP topics.

- Four types of QP actions: Retrieval, Annotation, Filtering, Mapping.

- Ex: topic #12, “Places where Goethe lived”:
  2. Annotate Goethe's page with REMBRANDT.
  3. Filter geographic named entities.
Query Procedures (cont.)

GikiP topic #7: “African capitals with a population of two million inhabitants or more”.

1. MANUAL: Choose Wikipedia category as a start;
2. MANUAL: Select the category “Capitals of Africa”;
3. AUTOMATIC: Retrieve Wikipedia pages w/category;
4. AUTOMATIC: Annotate documents w/REMBRANDT;
5. AUTOMATIC: Filter named entities VALUE/QUANTITY;
6. MANUAL: Evaluate named entity > 2.000.000;
7. SUPERVISED: Map documents to Wikipedia.
Lessons learnt

- Some (obvious) over-simplifications and human intervention, but aiming to a fully-automatic system soon.
- We need to extract more data from Wikipedia (ex: infoboxes) or use other resources (Dbpedia.org?)
- Use an geo-ontology for RENOIR's geographic reasoning, as in QuerCol (ex: 'how is Venice related to Italy?')
Lessons learnt (cont.)

Wikipedia looks promising, but it may be tricky:

- **Choice of categories** ("German composers", not "Composers of Germany")...
- "*Wars on Greece soil*"... is not so straightforward to answer with only categories...
- **Language differences**: "Australia" refers to a country (PT) or a continent (EN)...  
- **Categories now have sub-categories**... more depth levels to help classification, or to be lost?
Future work

QuerCol's query reformulation:

- Use semantic labels on query strings
  "Riots in LOCAL:{Turkey}" vs "Cooking a Turkey"

- Reason over the Wikipedia extracted knowledge
  (ex: "Hotels in African capitals")

- Detect types of queries (informational, navigational, transactional, geographical, non-geographical).

MG4J:

- Support "semantic indexes" and term weights
- Scale up to index Web collections
Future work

**REMBRANDT:**

- Detect the context of named-entities (ex: 'Brussels' as a city, or as the UE headquarters”).
- Better Wikipedia mining procedures → better NER quality → more useful semantic layer.

**RENOIR:**

- Make it fully automatic.
Thank you.

Towards semantic-flavored queries for GIR systems: RENOIR at the GikiP pilot task.

Nuno Cardoso - XLDB
University of Lisbon, Faculty of Sciences, LaSIGE
xldb.di.fc.ul.pt
ncardoso@xldb.di.fc.ul.pt
xldb.di.fc.ul.pt/wiki/Nuno_Cardoso
xldb.di.fc.ul.pt/Renoir

9th Workshop of CLEF 2008 - 17-19th September, 2008 - Århus, Denmark