Relation Discovery on the DBpedia Semantic Web

Obey Liu (obey.liu@ensimag.imag.fr)

May 2009
1. The Semantic Web
   - The current Web and the Semantic Web
   - Semantic data representation
   - Relation discovery

2. The DBpedia Dataset
   - The DBpedia project
   - Relation discovery
   - Problematic paths

3. DBpedia algorithms and heuristics
   - Multi-directional Breadth-First Search with Priority Queue
   - Formalizing islands of data
   - Experiments with distance metrics

4. Conclusions
   - Conclusions
   - Future work
   - Acknowledgments
Most of the current Web is destined to *human* consumption:
- Poorly structured information, even if created from structured data
- No software understandable semantics

We would like to store and process the meaning of data:
- Store and universally present information in a software processable way
- Use formal techniques and algorithms to exploit this structured information
Hi, my name is Dr. Eric Miller and my mail is em@w3.org:

```xml
<contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
  <contact:fullName>Eric Miller</contact:fullName>
  <contact:mailbox rdf:resource="mailto:em@w3.org"/>
  <contact:personalTitle>Dr.</contact:personalTitle>
</contact:Person>
```
Semantic data representation

Explicit metadata representation

**Figure:** Explicit metadata representation in a graph format
Getting from here to here

Figure: Discovering relations with the Semantic Web (SCARLET)
Getting from here to here

Figure: Finding a path on the Semantic Web (SCARLET)
The DBpedia dataset

- Multi-domain ontology derived from Wikipedia
- Data from English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish and Norwegian Wikipedia
- 2.6 million “things” with 274 million “facts”, including:
  - 213,000 persons
  - 328,000 places
  - 57,000 music albums
  - 36,000 films
  - 20,000 companies
A nucleus of Semantic Data

Figure: DBpedia in the Semantic Web
Relation discovery

A real world graph

**Figure:** Finding a path on the DBpedia Semantic Web
Figure: A path on the DBpedia Semantic Web
Better paths

- Prune useless paths during discovery
- Define a quantitative measure of path length
- Do it quick on a large dataset
Figure: Disambiguation node
Redirections

RDF nodes pointing to another node.

- Misspellings:
  - “GOOGLE” redirects to “Google”
  - “E. coli” redirects to “Escherichia coli”

- Sub-topics and small topics in broader contexts:
  - “Distributed denial of service” redirects to “Denial of service”
  - “Blackberry 8820’” redirects to “List of BlackBerry products”
Figure: Wikilinks
Multi-directional Breadth-First Search with Priority Queue

Figure: Searching from multiple sources with weighted edges
Figure: An example of category apex
A hierarchy of categories

Figure: The hierarchy of categories between “flower” and “cucumber”.
Example of weights

- Some weights:
  - $+\infty$, source/destination is Disambiguates
  - $+10$, relation is Wikilink
  - $+5$, relation is a subsumption between two categories
  - $+40$, source/destination is year

- Objective: maybe longer but “common-sense” good paths

- Possibilities: finer-grained matchers, bayesian learning...
Conclusions

- A very varied range of relations, some useful, some not
- DBpedia leverageable to create rudimentary ontologies from word sets
- Reusable Python tools to execute relations searches
- Multiple and vast DBpedia datasets still areas to study for semantic exploration algorithms
Future work

- More datasets: other languages, OWL ontology (RDF with logic relations)
- Better heuristics: more pattern matchers, PageRank style algorithms, Bayesian networks
Acknowledgments

TER 2009, supervised by Jérôme Euzenat