Lab II

Entity Retrieval

Daan Odijk
University of Amsterdam

Overview of Lab II

- Entity Retrieval toolkits and webservices

- Hands-on Entity Retrieval using webservices
  - Building an Entity Retrieval system using webservices

Public Toolkits and Web Services for Entity Linking

- YAGO
- Freebase
- EARS
- Sindice & SIREn
- DBpedia

YAGO

- Accuracy manually evaluated
  - Confirmed accuracy of 95%
  - Relation is annotated with its confidence value.
- Anchored in Time and Space
- Thematic domains (e.g. "music" or "science")
- Includes the WordNet class hierarchy

J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis Kelham, G. de Melo, G. Weikum. **YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages**. *WWW 2011.*
F. M. Suchanek, G. Kasneci, G. Weikum. **YAGO - A Core of Semantic Knowledge.** *WWW 2007*.

Freebase

- Initially seeded from high-quality open data
- Now composed mainly by community
- Harvested from many sources
  - Wikipedia, MusicBrainz, and others.

- Acquired by Google in 2010
  - Basis of Google Knowledge Graph

EARS

- Entity and Association Retrieval System
    - Developed in context of expertise retrieval
    - Open source, built on top of Lemur in C++
        - Not actively maintained
- Entity-topic association finding models
    - Suited for other tasks, e.g. blog distillation
    - Focuses on two entity-related tasks:
        - Finding entities:
            - "Which entities are associated with topic X?"
        - Profiling entities:
            - "What topics is an entity associated with?"

K. Balog. **People Search in the Enterprise.** *PhD thesis, University of Amsterdam, June 2008*.

Sindice/SIREn

- Handling of semi-structured data
  - Efficient, large scale
  - Typically based on DBMS backends
  - Apache Lucene plugin for semi-structured search
- Search engine features: top-k query processing, real time updates, full text search, distributed indexes over shards, etc.
- Open source

R. Delbru, N. Toupikov, M. Catasta and G. Tummarello. **A Node Indexing Scheme for Web Entity Retrieval**. ESWC'10.

DBpedia

- Extract structured information from Wikipedia
- Crowd-sourced community effort
- Open source
  - Written in Scala, Java and VSP
  - Virtuoso Universal Server Operating system

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann: **DBpedia – A Crystallization Point for the Web of Data**. JWS, 2009.

Sense of Scale

- YAGO: 10 million entities and 120 million facts.
- Freebase: 37 million topics, 1,998 types, and more than 30,000 properties
- DBpedia: 3.77 million things
  - 2.35 million classified in Ontology, including:
    - 764,000 persons, 573,000 places,
    - 333,000 creative works, 192,000 organizations,
    - 202,000 species and 5,500 diseases.
  - 111 languages, together 20.8 million things

Overview of Lab II

- Entity Retrieval toolkits and webservices

- Hands-on Entity Retrieval using webservices
  - Building an Entity Retrieval system using webservices