The Sindice-2011 Dataset for Entity-Oriented Search in the Web of Data

S. Campinas, D. Ceccarelli, T.E. Perry, R. Debru, K. Balog, G. Tummarello

Introduction

Web of Data

Data Collection

Statistics

Tools

Conclusion

Towards Entity-Oriented Search

Retrieving entities? people, organisation, products, locations

Emphasis of the research community as well as the commercial sector on the entity-oriented search

International Benchmarking campaigns foster research in that direction

TREC Expert finding and Related Entity finding tasks

INEX Entity Ranking track

SemSearch Entity Ranking and Lists search task

The Web of Data fits the need of entity-oriented search:

- Data naturally centered around entities
- Provides a semi-structured description of the data

To foster research and development in that direction:

Sindice-2011 Entity-oriented dataset representative of the entities found on the Web of Data

SIREn Entity-oriented search infrastructure and tools for experiments

In need for research material: the Web of Data

The Web of Data fits the need of entity-oriented search:

- Data naturally centered around entities
- Provides a semi-structured description of the data

To foster research and development in that direction:

Sindice-2011 Entity-oriented dataset representative of the entities found on the Web of Data

SIREn Entity-oriented search infrastructure and tools for experiments

The Sindice-2011 Dataset

Sindice crawls the Web for documents containing semantic markups since 2009

Sindice-2011 is a up-to-date snapshot of the Web of Data

Large coverage of domains: e-commerce, social networks, publications, ...

Highly heterogeneous collection: many ontologies and predicates as well as data formats are used
The Sindice-2011 dataset provides two views on the entities:
- **Sindice-DE**: A Document-Entity view where an entity is presented within a context, i.e., the document.
- **Sindice-ED**: An Entity-Document view providing a global description of an entity across documents.
We provide statistics about the Sindice-2011.

We organize statistics in three groups:
- **data heterogeneity**
- **entity size**
- **term distribution**

### Data heterogeneity

- 230 million distinct documents
- 270,404 distinct second-level domains
- 1.7 billion entities

### Description

<table>
<thead>
<tr>
<th>Description</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>230,643,642</td>
</tr>
<tr>
<td>Domains</td>
<td>6,286,468</td>
</tr>
<tr>
<td>Second-level domain</td>
<td>270,404</td>
</tr>
<tr>
<td>Entities</td>
<td>1,738,089,611</td>
</tr>
<tr>
<td>Statements</td>
<td>11,163,347,487</td>
</tr>
<tr>
<td>Bytes</td>
<td>1,274,254,991,725</td>
</tr>
<tr>
<td>Literals</td>
<td>3,968,223,334 (35.55%)</td>
</tr>
<tr>
<td>URIs (as objects)</td>
<td>5,461,354,581 (48.92%)</td>
</tr>
<tr>
<td>Blank nodes (as objects)</td>
<td>1,732,798,072 (15.08%)</td>
</tr>
<tr>
<td>Unique ontologies</td>
<td>863,952</td>
</tr>
<tr>
<td>Unique predicate URIs</td>
<td>5,386,273</td>
</tr>
<tr>
<td>Unique literal words</td>
<td>116,130,896</td>
</tr>
<tr>
<td>Unique URIs (as objects)</td>
<td>498,189,991</td>
</tr>
</tbody>
</table>

### Sindice-2011 Statistics

- 11 billion of statements
- 1.7 billion entities
- 11 billion of statements
Data heterogeneity

- 230 million distinct documents
- 270,404 distinct second-level domains
- 1.7 billion entities
- 11 billion of statements
- using more than 600 thousand ontologies

Data heterogeneity - Data Formats

- Half (142 million) of the documents are containing RDFa markups.
- Half (130 million) of the documents are containing Microformats markups.
- A third (87 million) are containing plain RDF.
- All standardised data formats are well covered.

Data heterogeneity - Predicate Distribution

- Frequency of use of predicate URIs follows a power-law distribution
- A large number of predicates (96%) are used only once
- Distribution tail is sparse: a few predicates are used in a large proportion of documents

Entity Size

- Sindice-DE
  - 25% of entities have three or four triples
  - 75% have ten or fewer triples
  - Average entity size is 5.32 triples
- Sindice-ED
  - 20% of entities have three or four triples
  - 75% have ten or fewer triples.
  - Average entity size is 14.19 triples

Coverage

- We checked all the URIs in the relevance assessment files from the 2010 and 2011 editions of the Semantic Search Challenge and from the ELC task of the 2010 TREC Entity track
- The Sindice-2011 Dataset covers 99% of these URIs
- The missing ones are either:
  - not exist anymore on the Web
  - synthetic URIs that were generated to replace blank nodes with URIs in the BTC-2009 dataset

Semantic Information Retrieval Engine

- Field-based techniques (e.g., Lucene) inefficient with highly heterogeneous structured data
- SIREn, an IR search engine for the Web of Data
  - Based on XML IR technique
  - Built for dealing with highly heterogeneous structured data sources
  - No limit in the number of attributes
  - Extended support for structured query
  - http://siren.sindice.com
Search infrastructure and set of tools developed on top of SIREn
- Provide out of the box functionalities to process, index and search Sindice-2011 dataset
- One can develop and test new entity retrieval model
  - Two-step retrieval process
  - TF-IDF ranking baseline provided, but possibilities to implement its own ranking
  - Query expansion techniques that combines full-text and structured queries

> https://github.com/rdelbru/trec-entity-tool

Sindice-2011, a dataset to foster research on entity-oriented search
- Accurate reflection of the current Web of Data
- A corpus of 1.7 Billion entities
- A search infrastructure tool based on SIREn, the Semantic Information Retrieval Search Engine, in order to ease development and research

Thank You!

Questions?

Thank You!