From data to knowledge – profiling and interlinking Web datasets

Stefan Dietze
L3S Research Center
Introduction

Research areas

- Web science, Information Retrieval, Semantic Web & Linked Data, data & knowledge integration (mapping, classification, interlinking)
- Application domains: education/TEL, Web archiving, ...

Some projects

- KNOWeSCAPE
- EUMSSI
- arc&mem
- LINKEDUP
- keystone
- DURAAK
- DURABLE ARCHITECTURAL KNOWLEDGE

Recent work on Linked Data exploration/discovery/search

- Entity interlinking & dataset interlinking recommendation
- Dataset profiling
- Data consistency & conflicts

- See also: http://purl.org/dietze
Hm, really?

- "HTTP-accessibility" (SPARQL, URI-dereferencing)
- "Structure" & "Semantics" (=> shared/linked vocabularies)
  - "Interlinked"
  - "Persistent"

...why are there so few datasets actually used?

- Date reuse and in-links focused on trusted "reference graphs" such as DBpedia, Freebase etc
- Long tail of LD datasets which are neither reused nor linked to (LOD Cloud alone 300+ datasets, 50 bn triples)
- Explanations?
Linked data is more diverse than we think

Accessibility of datasets?
- Less than 50% of all SPARQL endpoints actually responsive at given point of time
- “THE” SPARQL protocol?
  No, but many variants & subsets
- ...

“Semantics”, links, quality?
- ...data consistency? [? Yuan2014 ?]
- ...data accuracy (eg DBpedia)? [Paulheim2013]
- ...vocabulary reuse? [D’AquinWebSci13]
- ...schema compliance (RDFS, schemas) [HoganJWS2012]
Too many/diverse datasets, too little knowledge

- Which datasets are useful & trustworthy for case XY (e.g., “learning about the solar system“) ? Which topics are covered?
- Types: which datasets describe statistics, videos, slides, publications etc?
- Currentness, dynamics, accessability/reliability, data quantity & quality?
Data curation, linking and dataset profiling

BBC Programme
<po:Programme>
<po:Series>Wonders of the Solar System</po:Series>
<po:Actor>Brian Cox</po:Actor>
</po:Programme>

Yovisto Video
<yov:Video>
<dc:title>Pluto & the Dwarf Planets</dc:title>
</yov:Video>

Schema mappings [WebSci13]
Entity disambiguation & linking [ESWC13]
Topic profile extraction [WWW13, ESCW14]

Dataset Catalog/Registry

Contains

Stefan Dietze 31/07/14
How standards proliferate:
(See: A/C chargers, character encodings, instant messaging, etc.)

Situation:
There are 14 competing standards.

14?! Ridiculous!
We need to develop one universal standard that covers everyone’s use cases.

Yeah!

Situation:
There are 15 competing standards.

schemas & vocabularies
Schema assessment and mapping

Co-occurrence of data types
(in 146 datasets: 144 Vocabularies, 588 highly overlapping types, 719 Properties)

Co-occurrence after mapping into most frequent schemas
(201 frequent types mapped into 79 classes)

LinkedUp Data Catalog

- RDF (VoID) dataset catalog: browse & query distributed datasets
- Live information about endpoint accessibility
- Federated queries using type mappings

http://datahub.io/group/linked-education
http://data.linkededucation.org/linkedup/catalog/

Stefan Dietze
31/07/14
Dataset interlinking recommendation

Candidate datasets for interlinking?

Approach

- Given dataset $t$, ranking datasets from $D$ according to probability score $(d_j, t)$ to contain linking candidates (entities)

  - Features:
    - Vocabulary overlap
    - Existing links (SNA)
    - Linking candidates likely if datasets share common (a) schema elements, or (b) links (friend of a friend)

Conclusions

- Roughly 60% MAP for both approaches
- Future work: quantity of links, extraction of experimental data from datasets...

---


---

Rank

<table>
<thead>
<tr>
<th>Rank</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DBLP</td>
</tr>
<tr>
<td>2</td>
<td>ACM</td>
</tr>
<tr>
<td>3</td>
<td>OAI</td>
</tr>
<tr>
<td>4</td>
<td>CiteSeer</td>
</tr>
<tr>
<td>5</td>
<td>IBM</td>
</tr>
<tr>
<td>6</td>
<td>Roma</td>
</tr>
<tr>
<td>7</td>
<td>IEEE</td>
</tr>
<tr>
<td>8</td>
<td>Ulm</td>
</tr>
<tr>
<td>9</td>
<td>Pisa</td>
</tr>
</tbody>
</table>
Dataset & entity linking: semantics of resources/datasets?


Slideset

<sioc:Item 2139393292>
<title>Planetary motion & gravity</title>
...
</sioc:Item 2139393292>

Programme

<po:Programme519215>
<po:Series>Wonders of the Solar System</po:Series>
<po:Episode>Emp. of the Sun</po:Episode>
<po:Actor>Brian Cox</po:Actor>
</po:Programme519215>

Video

<yoyo:Video 8748720>
<dc:title>Pluto & the Dwarf Planets</dc:title>
...
</yoyo:Video 8748720>

Topics/categories addressed?
Relatedness of resources/entities? (types, semantics)
Disambiguation/linking using background knowledge „Semantic relatedness“ of resources?

Programme
<po:Programme519215>
<po:Series>Wonders of the Solar System</po:Series>
<po:Episode>Emp. of the Sun</po:Episode>
<po:Actor>Brian Cox</po:Actor>
</po:Programme519215>

Video
<y0:Video 8748720>
<dc:title>Pluto & the Dwarf Planets</dc:title>
...</y0:Video 8748720>

Slideset
<sioc:Item 2139393292>
<title>Planetary motion & gravity</title>
...</sioc:Item 2139393292>

BBC

DBpedia

db:Astronomy

db:Astronomical Objects

db:Pluto (Dwarf Planet)

db:Sun
Entity linking: semantic relatedness

http://purl.org/vol/ns/
http://purl.org/vol/doc/

- Computation of connectivity scores between resources/entities
  - Method: combination of a
    - (i) semantic (graph-based) connectivity score (SCS) with
    - (ii) a Web co-occurrence-based measure (CBM) (similar to NGD)
  - For (i): adaptation of Katz-Index from SNA for (linked) data graphs (considering path number and path lengths of transversal properties)

SCS = 0.32
CBM = 0.24
Entity linking: evaluation

- Evaluation based on USA Today News items (80,000 entity pairs)
- Manually created gold standard (1000 entity pairs)
- Baseline: Explicit Semantic Analysis (ESA)

=> CBM/SCS: „relatedness“; ESA: „similarity“

<table>
<thead>
<tr>
<th></th>
<th>CBM</th>
<th>SCS</th>
<th>ESA</th>
<th>CBM+SCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.32</td>
<td>0.34</td>
<td>0.16</td>
<td>0.34</td>
</tr>
<tr>
<td>Recall (GS)</td>
<td>0.81</td>
<td>0.78</td>
<td>0.23</td>
<td>0.90</td>
</tr>
<tr>
<td>Recall</td>
<td>0.52</td>
<td>0.51</td>
<td>0.15</td>
<td>0.58</td>
</tr>
<tr>
<td>F1 (GS)</td>
<td>0.46</td>
<td>0.47</td>
<td>0.19</td>
<td>0.50</td>
</tr>
<tr>
<td>F1</td>
<td>0.40</td>
<td>0.41</td>
<td>0.15</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Dataset profiling: what’s the data about?

- Extracting representative metadata ("topic profile") for datasets
- Ranking of most representative (DBpedia) categories (= topics); applied to all responsive LOD datasets
- Scalability vs representativeness: sampling & ranking for good scalability/accuracy balance
Dataset profiling: approach

1. **Sampling of resource instances** (random sampling, weighted sampling, resource centrality sampling)

2. **Entity and topic extraction** (NER via DBpedia Spotlight, category mapping and expansion)

3. **Normalisation and ranking** (using graphical models such as PageRank with Priors, HITS with Priors and K-Step Markov)

=> Result: weighted dataset-topic profile graph
Dataset profiling: exploring LOD datasets/topics

- Automatic extraction of dataset “topics” [ESWC2014] => RDF/VoiD dataset profiles
- Visualisation & exploration of dataset-topic graph (datasets, topics, relationships)
- Includes all (responsive) datasets of LOD Cloud
Dataset profiling: results evaluation

Datasets & Ground Truth
- Yovisto, Oxpoints, LAK Dataset, Semantic Web Dogfood
- Crowd-sourced topic indicators from datasets (keywords, tags)
- Manual mapping to entities & category extraction (ranking according to frequency)

Baselines
- 1) LDA, 2) tf/idf (applied to entire datasets)
- Topic extraction according to our approach, weighting/ranking based on term weight

Measure
- NDCG @ rank /
- Performance (time/NDCG) for different sampling strategies/sizes etc

NDCG (averaged over all datasets).
What have these categories in common?

- dbp:Category:People_from_London
- dbp:Category:Web_Services
- dbp:Category:Buzzwords
- dbp:Category:Royal_Medal_winners
- dbp:Category:1955_births
- dbp:Category:Unitarian_Universalists
- dbp:Category:HTTP
- dbp:Category:World_Wide_Web
Diversity of category profile for a single paper

Type-specific exploration of dataset categories

- Type specific views on datasets/categories
  - “Document” (foaf:document)
  - “Person” (foaf:person)
  - “Course” (aaiso:course)

- Currently applied to datasets in LinkedUp Catalog only (as schema mappings already available here)
LinkedUp Challenge: Linking Web Data (for Education)

Series of Open Data Competitions to promote applications which exploit Linked Open Data

- May – September 2013
- October 2013 – May 2014
- May 2014 – October 2014

http://www.linkedup-challenge.org/

- “Vici” open just now
- Final events at ISWC2014
- Submission: 5 September

http://www.linkedup-project.eu/
Conclusions & future work

Summary

- Increasing amounts of data => require knowledge about nature and relationships of datasets
- **Profiling**: scalable methods for extracting dataset metadata
- **Interlinking**: connectivity of entities or datasets

Future work – LD evolution, preservation, consistency

- In RDF graphs (eg LOD Cloud), „all“ nodes are connected
- LD preservation: which datasets to preserve (entity „neighbourhood“)? => semantic relatedness
- Link correctness in evolving LD: investigating impact of changes on link correctness
- Application: informed preservation and enrichment strategies
Thank you!

Acknowledgements

- Besnik Fetahu (L3S)
- Elena Demidova (L3S)
- Bernardo Pereira Nunes (PUC Rio)
- Marco Casanova (PUC Rio)
- Luiz Andre Paes Leme (PUC Rio)
- Giseli Lopes (PUC Rio)
- Davide Taibi (CNR, IT)
- Mathieu d’Aquin (Open University, UK)
- and many more...

WWW

See also (general)

- http://linkedup-project.eu
- http://linkededucation.org
- http://data.l3s.de
- http://purl.org/dietze

See also (data)

- http://data.linkededucation.org
- http://data.linkededucation.org/linkedup/catalog/
- http://lak.linkededucation.org