Instance Matching Benchmarks in the Era of Linked Data

Tzanina Saveta,
Evangelia Daskalaki, Irini Fundulaki, Giorgos Flouris
Institute of Computer Science – FORTH, Greece
Linked Data – The LOD Cloud

Same entity can be described in different sources

*Adapted from Suchanek & Weikum tutorial@SIGMOD 2013
Instance Matching: the cornerstone for Linked Data

Problem: How can we automatically recognize multiple mentions of the same entity across or within sources?

Solution: Instance Matching

Open/Social data

Data Integration

Data Acquisition

Data Evolution
Instance Matching in the LOD

Sets RDF/OWL triples

Many sources to match

Rich semantics

Value Structure
Logical variations

*Adapted from Suchanek & Weikum tutorial@SIGMOD 2013
Instance Matching Techniques and Tools

• People *interconnect* their dataset with existing ones.
  – These links are often *manually curated* (or semi-automatically generated).

• Size and number of datasets is *huge*, so it is vital to automatically detect additional links that *is making the graph more dense*

• Instance matching research has led to the development of a number of systems.

1. How to compare these?
2. How can we assess their performance?
3. How can we push the systems to get better?

*The systems must be benchmarked!*
Ingredients of an Instance Matching Benchmark

• It is organized in test cases each addressing different kind of requirements

• Datasets
  – The raw material of the benchmarks. These are the source and the target dataset that will be matched together to find the instances that refer to the same real world entity

• Gold Standard (Ground Truth / Reference Alignment)
  – The “correct answer sheet” used to judge the completeness and soundness of the instance matching algorithms

• Metrics/Key Performance Indicators
  – The performance metric(s) that determine the systems behavior and performance
Datasets

Distinguish between Real and Synthetic datasets:

– Real Datasets:
  • Realistic conditions for heterogeneity problems
  • Realistic distributions
  • Error prone Reference Alignment

– Synthetic Datasets:
  • Fully controlled test conditions
  • Accurate Gold Standards
  • Unrealistic distributions
Data Variations in Datasets

Variations

Value
- Name style abbreviation
- Typographical errors
- Change format (date/gender/number)
- Synonym Change
- Multilingualism

Structural
- Change property depth
- Delete/Add property
- Split property values
- Transformation of object to data type property
- Transformation of data to object type property

Logical
- Delete/Modify Class Assertions
- Invert property assertions
- Change property hierarchy
- Assert disjoint classes
Gold Standards and Metrics

- **Metrics/Key Performance Indicators**
  - Precision/Recall/F-measure

- **Recall** \( r = \frac{TP}{TP + FN} \)
- **Precision** \( p = \frac{TP}{TP + FP} \)
- **F-measure** \( f = \frac{2 \times p \times r}{p + r} \)
# Design Criteria for Instance Matching Benchmarks

<table>
<thead>
<tr>
<th><strong>Systematic Procedure</strong></th>
<th><em>matching tasks are reproducible and the execution is comparable</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Availability</strong></td>
<td><em>related to the availability of the benchmark in time</em></td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td><em>precise evaluation rules and high quality ontologies</em></td>
</tr>
<tr>
<td><strong>Equity</strong></td>
<td><em>no system privileged during the evaluation process</em></td>
</tr>
<tr>
<td><strong>Dissemination</strong></td>
<td><em>how many systems have used this benchmark to be evaluated with</em></td>
</tr>
<tr>
<td><strong>Volume</strong></td>
<td><em>how many instances did the datasets contain</em></td>
</tr>
<tr>
<td><strong>Gold Standard</strong></td>
<td><em>existence of gold standard and it’s accuracy</em></td>
</tr>
</tbody>
</table>
Benchmarking

- Instance matching techniques have, until recently, been benchmarked in an ad-hoc way.
- There does not exist a standard way of benchmarking the performance of the systems, when it comes to Linked Data.
- On the other hand, Instance Matching benchmarks have been mainly driven forward by the Ontology Alignment Evaluation Initiative (OAEI):
  - organizes annual campaign for ontology matching since 2005
  - hosts independent benchmarks
- In 2009, OAEI introduced the Instance Matching (IM) Track:
  - focuses on the evaluation of different instance matching techniques and tools for Linked Data.
Benchmark Frameworks

- Benchmark Generators
  - SWING [FMN+11]
  - SPIMBENCH [SDF+15]
  - LANCE [SDFF+15]

- Synthetic Benchmarks
  - OAEI IIMB 2009 [EFH+09]
  - OAEI IIMB 2010 [EFM+10]
  - OAEI Persons-Restaurants 2010 [EFM+10]
  - OAEI IIMB 2011 [EHH+11]
  - Sandbox 2012 [AEE+12]
  - OAEI IIMB 2012 [AEE+12]
  - OAEI RDFT 2013 [GDE+13]
  - SPIMBENCH 2015 [CDE+15]
  - ONTOBI 2010 [Z10]
  - ID-REC Task 2014 [DEE14]
  - Author Task 2015 [CDE+15]

- Real Benchmarks
  - ARS (OAEI 2009) [EFH+09]
  - DI (OAEI 2010) [EFM+10]
  - DI-NYT (OAEI 2011) [EHH+11]
Comparison of synthetic Benchmarks

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic Procedure</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Quality</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Equity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Availability</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Volume</td>
<td>~200</td>
<td>~1400</td>
<td>~860</td>
<td>~4000</td>
<td>~375</td>
<td>~1500</td>
<td>~430</td>
<td>~2650</td>
<td>~100000</td>
<td>~8500</td>
<td>~13700</td>
</tr>
<tr>
<td>Dissemination</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Reference Alignment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Value variations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Structural variations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Logical variations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multilinguality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blind Evaluations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-n Mappings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Comparison of real benchmarks

<table>
<thead>
<tr>
<th></th>
<th>ARS</th>
<th>DI 2010</th>
<th>DI 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systematic Procedure</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Equity</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Availability</strong></td>
<td>✔</td>
<td>✔</td>
<td>-</td>
</tr>
<tr>
<td><strong>Volume</strong></td>
<td>~1M</td>
<td>~6000</td>
<td>?</td>
</tr>
<tr>
<td><strong>Dissemination</strong></td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Reference Alignment</strong></td>
<td>✔</td>
<td>✔</td>
<td>+</td>
</tr>
<tr>
<td><strong>Value variations</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Structural variations</strong></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td><strong>Logical variations</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Multilinguality</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Blind Evaluations</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Wrapping Up

Which benchmarks included multilingual datasets?

- OAEI RDFT 2013 (French-English)
- ID-REC 2014 (English-Italian)
- Author Task (English–Italian)
Wrapping Up: Benchmarks

Which benchmarks included value variations into the test cases?

- OAEI IIMB 2009
- OAEI IIMB 2010
- OAEI Persons-Restaurants 2010
- ONTOBI
- OAEI IIMB 2011
- Sandbox 2012
- OAEI IIMB 2012
- OAEI RDFT 2013
- ID-REC 2014
- SPIMBENCH 2015
- Author Task 2015
- ARS
- DI 2010
- DI 2011
Wrapping up: Benchmarks

Which benchmarks included structural variations into the test cases?

- OAEI IIMB 2009
- OAEI IIMB 2010
- OAEI Persons-Restaurants 2010
- ONTOBI
- OAEI IIMB 2011
- OAEI IIMB 2012
- OAEI RDFT 2013
- ID-REC 2014
- SPIMBENCH 2015
- Author Task 2015
- ARS
- DI 2010
- DI 2011
Wrapping up: Benchmarks

Which benchmarks included logical variations into the test cases?

- OAEI IIMB 2009
- OAEI IIMB 2010
- OAEI IIMB 2011
- OAEI IIMB 2012
- SPIMBENCH 2015
Wrapping up: Benchmarks

Which benchmarks included combination of the variations into the test cases?

- IIMB 2009
- IIMB 2010
- IIMB 2011
- IIMB 2012
- RDFT 2013
- ID-REC 2014
- SPIMBENCH 2015
- Author Task 2015
Wrapping up: Benchmarks

Which benchmarks are more voluminous?

- SPIMBENCH 2015
- ARS
- DI 2011
Wrapping up: Benchmarks

Which benchmarks included both combination of the variations and was voluminous at the same time?

SPIMBENCH 2015
Open Issues

• **Issue 1:**
  Only one benchmark that tackles both, **combination** of variations and **scalability** issues

• **Issue 2:**
  Not enough IM benchmark using the **full expressiveness** of RDF/OWL language
Wrapping Up: Systems for Benchmarks

Outcomes as far as systems are concerned:

• Systems can handle the value variations, the structural variation, and the simple logical variations separately

• More work needed for complex variations (combination of value, structural, and logical)

• More work needed for structural variations

• Enhancement of systems to cope with the clustering of the mappings (1-n mappings)
Conclusion

- Many instance matching benchmarks have been proposed
- Each of them answering to some of the needs of instance matching systems
- It is time now to start creating benchmarks that will "show the way to the future" and make them better
This work was supported by grants from the EU H2020 Framework Programme provided for the project HOBBIT (GA no. 688227).