SPIMBENCH: A Scalable, Schema-Aware Instance Matching Benchmark for the Semantic Publishing Domain

T. Saviotl, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, A.-C. Ngonga Ngomo

#1 FORTH-ICS, #2 University of Stuttgart, #3 University of Leipzig

Motivation

The widespread adoption of Semantic Web Technologies and the publication of large interlinked RDF datasets and ontologies in the Web has made the integration of data a crucial task. Data linking in this context is essential in order to provide an integrated view of the underlying information; this is achieved by instance and schema matching techniques. To aid the users to choose among the systems that perform such tasks, a number of benchmarks have been developed.

SPIMBENCH Approach

SPIMBENCH is a benchmark for the Semantic Publishing Domain which takes into consideration RDFS and OWL constructs in order to evaluate instance matching systems. SPIMBENCH supports:

- A data generator that extends the one provided by LDBC’s SPB Benchmark.
- Semantics aware transformations.
- Standard value and structure based transformations.[2,3]
- Scalable data generation in order of billion triples.
- Weighted gold standard based on tensor factorization.

Rescal[1] is a tensor factorization for large-scale relational learning from Linked Data, multi-relational data and large multigraphs.

Transformations

- Blank Character Addition/Deletion
- Random Character Addition/Deletion/Modification
- Token Addition/Deletion/Shuffle
- Date Format
- Property Addition/Deletion
- Property Aggregation
- Property Extraction
- Abbreviation
- Synonym/Antonym
- Stem of a Word
- Multilinguality

Structure-based

- Semantic-aware

RDFS/OWL SD TD SCHEMA TRIPLES GS
owl:inverseProperty

Value-based

- $\text{owl:UnionOf} (\text{C}, \text{C}, \text{C})$
- $\text{owl:IntersectionOf} (\text{C}, \text{C}, \text{C})$
- $\text{owl:FunctionalProperty}$
- $\text{owl:InverseFunctionalProperty}$

LogMap responds optimally regarding the precision as it does not find many matches that are not actually a match. On the other hand, fails to find matches when the instance is involved in multiple semantics-aware test cases.

Future Work

- Domain independent instance matching test case generator for Linked Data.
- Definition of more sophisticated metrics that takes into account the difficulty (weight).

Acknowledgments

This work was partially supported by the ongoing FP7 European Project LDBC (Linked Data Benchmark Council).

References