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

T. Savață1, E. Daskalaki1, G. Flouris1, I. Fundulaki1, M. Herschel2, A.-C. Ngonga Ngomo3

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

Motivation

The widespread adoption of Semantic Web Technologies and the publication of large interrelated 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
- Scalable data generation in order of billions triples.
- Weighted gold standard based on tensor factorization.

Transformations

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

Structure-based
- Property Addition/Deletion
- Property Aggregation
- Property Extraction

Seminatics-aware
- owl:reverseFunctionalProperty
- owl:transitiveProperty
- owl:inverseFunctionalProperty
- owl:symmetricProperty

Combination of transformations

More than one transformation types per instance.

Simple
- One transformation per triple.

Complex
- Combination of two transformations per triple (value-based and structure-based or value-based and semantics-aware) based on the transformation parameters.

Scalability

Scalability experiments for datasets up to 500M triples with simple combination of transformations.

- 1000 triples ~ 36 entities.
- Data generation is linear to the size of triples.
- Transformation overhead is negligible for value, structure-based, semantics-aware and simple combinations.
- Overhead for logical transformations is higher by one magnitude.

Applicability of SPIMBENCH

We demonstrated the applicability of SPIMBENCH by using it to evaluate LogMap® with different data set sizes and different 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