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.
- Schema aware transformations (logical).
- Standard value and structural transformations.
- Multithreaded scalable data generation in order of billion triples.
- Weighted gold standard based on tensor factorization.

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 lexical, structural, logical and simple combinations.
- Overhead for logical transformations is higher by one magnitude.

Transformations

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

Structural
- Property Addition/Deletion
- Property Aggregation
- Property Extraction
- Combination of transformations

Logical

RDFS/OWL
- owl:sameAs
- owl:differentFrom
- owl:equivalentClass
- owl:inverseFunctionalProperty
- owl:FunctionalProperty
- owl:equivalentClass
- rdfs:subClassOf and property aggregation transformations

SBT
- S
- T
- B
- S
- T
- B
- S
- T
- B

Logics

GS
- $\gamma$
- $\phi$
- $\varphi$

References


Acknowledgment

This work was partially supported by the ongoing FP7 European Project LDBC (Linked Data Benchmark Council) and is done in collaboration with E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, A. Ngonga Ngomo and D. Plexousakis.

The research leading to these results has received funding from the European Community’s Seventh Framework Programme FP7 – 317548