A scalable schema aware instance matching benchmark for semantic publishing

Tzanina Saveta[1,2]
jsaveta@ics.forth.gr
[#1] Computer Science Department, University of Crete
[#2] Institute of Computer Science, Foundation for Research & Technology - Hellas

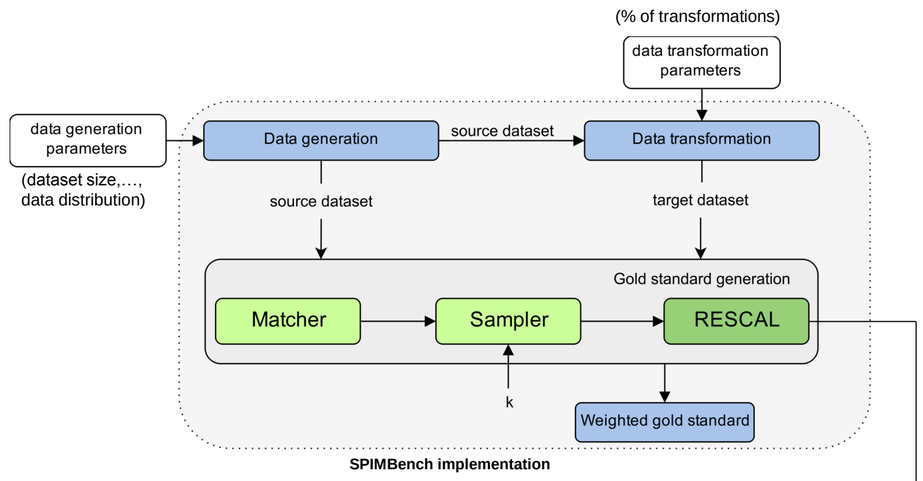
Motivation

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

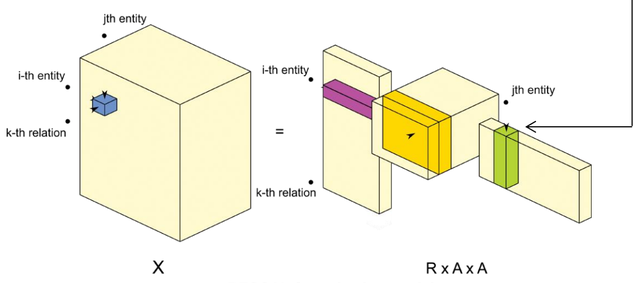
SPIMBench is a benchmark for the Semantic Publishing Domain which takes into consideration RDFS and OWL constructs in order to evaluate instance matching systems.

- Schema aware transformations (logical).
- Standard value and structural transformations.[3,4]
- Multithreaded scalable data generation in order of billion triples.
- Weighted gold standard based on tensor factorization.



SPIMBench implementation

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

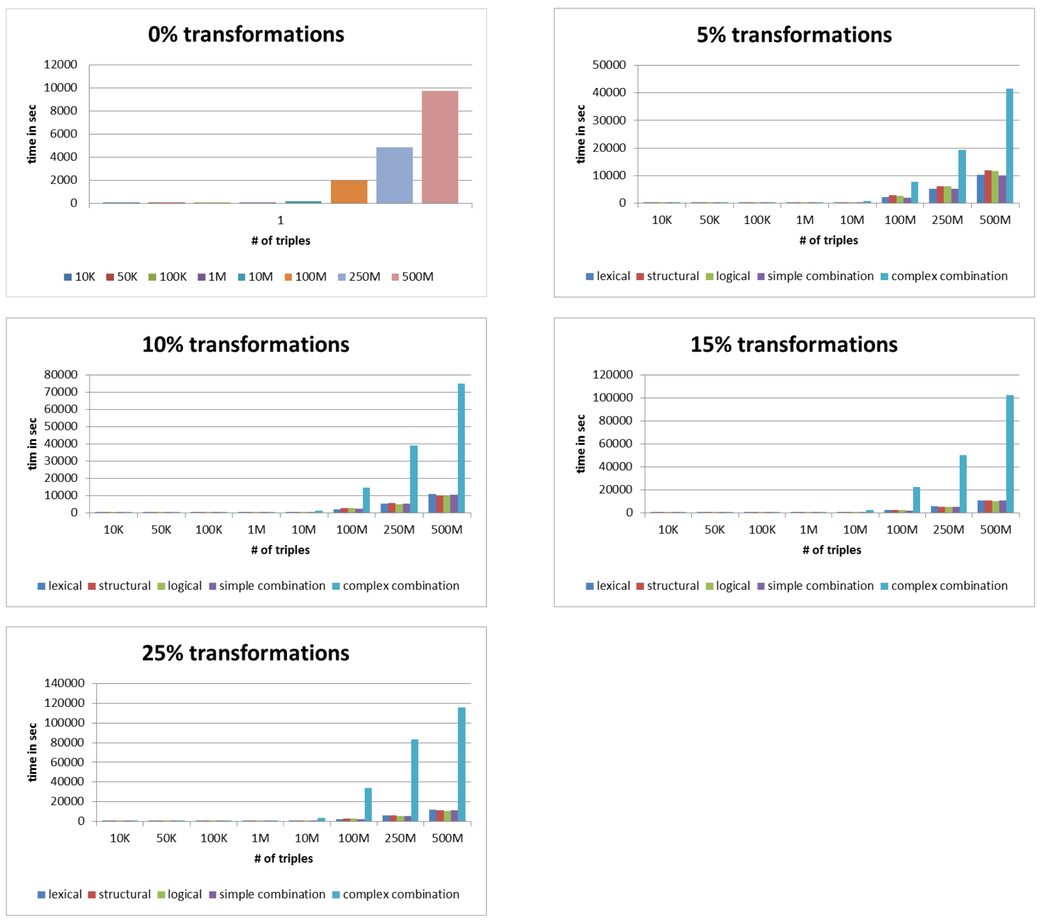


RESCAL factorization model

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.



Scalability experiments for n% of simple combination transformation type

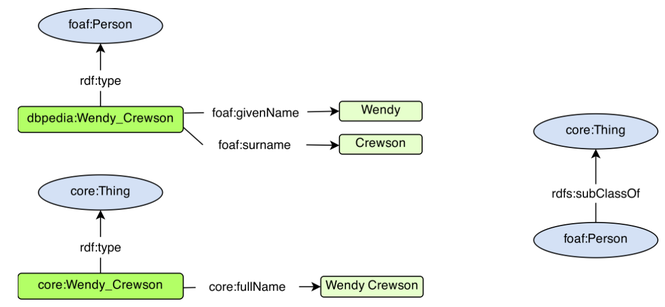
Transformations

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



rdfs:subClassOf and property aggregation transformations

Logical

RDFS/OWL	SD	TD	SCHEMA TRIPLES	GS
owl:sameAs	(u_1, rdf:type, C) (u_2, rdf:type, C)	(u_1', rdf:type, C) (u_2', rdf:type, C) (u_1', owl:sameAs, u_2')		$u_1 \sim u_1$' $u_1 \sim u_2$' $u_2 \sim u_2$' $u_2 \sim u_1$'
owl:differentFrom	(u_1, rdf:type, C)	(u_1', rdf:type, C) (u_1'', rdf:type, C) (u_1', owl:differentFrom, u_1'')		$u_1 \sim u_1$'
owl:equivalentClass	(u_1, rdf:type, C)	(u_1', rdf:type, C')	(C, owl:equivalentClass, C')	$u_1 \sim u_1$'
owl:disjointWith	(u_1, rdf:type, C)	(u_1', rdf:type, C')	(C, owl:disjointWith, C')	
owl:FunctionalProperty	(u_1, rdf:type, C) (u_1, p_1, o_1)	(u_1, rdf:type, C) (u_1, p_1, o_2)	(p_1, rdf:type, owl:FunctionalProperty)	$o_1 \sim o_2$
owl:InverseFunctionalProperty	(u_1, rdf:type, C) (u_1, p_1, o_1)	(u_1', rdf:type, C) (o_1, p_1, u_1')	(p_1, rdf:type, owl:InverseFunctionalProperty)	$u_1 \sim u_1$'
owl:unionOf	(u_1, rdf:type, C)	(u_1', rdf:type, C')	(C', owl:unionOf, {C_0, C_1,..})	$u_1 \sim u_1$'
owl:intersectionOf	(u_1, rdf:type, C)	(u_1', rdf:type, C')	C owl: intersectionOf{C,D,E,F} C' owl: intersectionOf {C,D}	$u_1 \sim u_1$'

Combination of transformations

More than one transformation types per instance.

Simple:
One transformation per triple.

Complex:
Combination of two transformations per triple (Lexical – Structural or Lexical - Logical) based on the data transformation parameters.

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.

References

[1] Maximilian Nickel, and Volker Tresp. Tensor Factorization for Multi-relational Learning. ECML/PKDD 3, volume 8190 of Lecture Notes in Computer Science, page 617-621. Springer, 2013.
[2] I. Fundulaki, E. Daskalaki, G. Flouris and T. Saveta. D4.4.3 Benchmark Design for Instance Matching. Technical report, Linked Data Benchmark Council, 2014. Available at http://ldbc.eu/results/deliverables
[3] A. Ferrara, D. Lorusso, S. Montanelli, and G. Varese. Towards a Benchmark for Instance Matching.In OM, 2008.
[4] A. Ferrara, S. Montanelli, J. Noessner, and H. Stuckenschmidt. Benchmarking Matching Applications on the Semantic Web. In *ESWC*, 2011.

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