Instance Matching Benchmark for Spatial Data: A Challenge Proposal to OAEI

Irini Fundulaki[1] and Axel-Cyrille Ngonga-Ngomo[2]

[1] Institute of Computer Science - FORTH, Greece
[2] Institute for Applied Informatics, University of Leipzig, Germany

Keywords: Instance Matching, Benchmarks, Spatial Datasets, Linked Data

1 Introduction

The number of datasets published in the Web of Data as part of the Linked Data Cloud is constantly increasing. The Linked Data paradigm is based on the unconstrained publication of information by different publishers, and the interlinking of Web resources across knowledge bases. In most cases, the cross-dataset links are not explicit in the dataset and must be automatically determined using Instance Matching (IM) tools (also known as record linkage [1], duplicate detection [2] and, entity resolution [3]) amongst others. The large variety of techniques requires their comparative evaluation to determine which one is best suited for a given context. Performing such an assessment generally requires well-defined and widely accepted benchmarks to determine the weak and strong points of the proposed techniques and/or tools.

A number of real and synthetic benchmarks that address different data linking challenges have been proposed for evaluating the performance of such systems. Those include, but are not limited to, IIMB 2012 [4], Sandbox 2012 [4], RDFT 2013 [5], ID-REC 2014 [6], ONTOBI 2010 [7], Author - Task 2015 [8] and Lance 2015 [9] to mention few. A more complete survey can be found in [10].

2 A benchmark for linking geo-spatial entities

So far, only a limited number of link discovery benchmarks target the problem of linking geo-spatial entities e.g., PABench [11]. However, some of the largest knowledge bases on the Linked Open Data Web are geo-spatial knowledge bases (e.g., LinkedGeoData with more than 30 billion triples). Linking spatial resources requires techniques that differ from the classical mostly string-based approaches. In particular, considering the topology of the spatial resources and the topological relations between them is of central importance to systems driven by spatial data.

We believe that due to the large amount of available geo-spatial datasets employed in Linked Data and in several domains, it is critical that benchmarks for geo-spatial link discovery are developed. For OAEI 2017, we propose the introduction of a new challenge for such systems. The benchmark that the systems will use will be based upon the Lance [9] scalable, schema-agnostic benchmark generator extended with appropriate transformations to tackle geo-spatial link

discovery tasks. Lance is able to produce, from an initial ontology, datasets of arbitrary size and complexity. The configuration of the transformations will be derived from real data. More specifically, we will use the widely accepted Linked Data datasets such as GeoNames, LinkedGeoData, and DBpedia for this task. The tasks proposed will focus on the different types of spatial object representations and will be provided with different severity levels for the applied transformations. In these transformations, objects may keep their representation, they may change their geometry, type or attributes, merge with other objects, or can completely disappear. This is a scenario that stems from the heterogeneous datasets (in structure and semantics) used to describe geo-spatial entities. The produced tasks will be used by IM tools that implement string-based as well as topological approaches for identifying matching entities. The IM frameworks will be evaluated for both accuracy (precision, recall and f-measure) and scalability. Furthermore, the results will be made available in both human and machine-readable form for further processing. Since Lance is schema-agnostic, contrary to PABench, it will be used to produce benchmarks for different (source) ontologies to accommodate the different requirements that stem from a variety of applications.

Acknowledgments

This work was supported by a grant from the EU H2020 Framework Programme provided for the project HOBBIT (GA no. 688227).

References

1. C. Li, L. Jin, and S. Mehrotra. Supporting efficient record linkage for large data sets using mapping techniques. In *WWW*, 2006.
2. A. K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate Record Detection: A Survey. *TKDE*, 19(1), 2007.
3. I. Bhattacharya and L. Getoor. *Entity resolution in graphs. Mining Graph Data.* Wiley and Sons, 2006.
4. J. Aguirre, K. Eckert, J. Euzenat, et al. Results of the ontology alignment evaluation initiative 2012. In *OM*, 2012.
5. B. Cuenca Grau, Z. Dragisic, K. Eckert, et al. Results of the ontology alignment evaluation initiative 2013. In *OM*, 2013.
6. Z. Dragisic, K. Eckert, J. Euzenat, et al. Results of the ontology alignment evaluation initiative 2014. In *OM*, 2014.
7. K. Zaiss, S. Conrad, and S. Vater. A Benchmark for Testing Instance-Based Ontology Matching Methods. In *KMIS*, 2010.
8. M. Cheatham, Z. Dragisic, J. Euzenat, et al. Results of the ontology alignment evaluation initiative 2015. In *OM*, 2015.
9. T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, and A. Ngonga-Ngomo. LANCE: Piercing to the Heart of Instance Matching Tools. In *ISWC*, 2015.
10. E. Daskalaki, G. Flouris, I. Fundulaki, and T. Saveta. Instance matching benchmarks in the era of linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2016.
11. B. Berjawi and F. Duchateau and F. Favetta and M. Miquel and R. Laurini. Pabench: Designing a taxonomy and implementing a benchmark for spatial entity matching. In *GeoProcessing*, 2015.