Blocking for Entity Resolution in the Web of Data: Challenges and Algorithms

Kostas Stefanidis

Abstract In the Web of data, entities are described by interlinked data rather than documents on the Web. In this talk, we focus on entity resolution in the Web of data, i.e., on the problem of identifying descriptions that refer to the same real-world entity within one or across knowledge bases in the Web of data. To reduce the required number of pairwise comparisons among descriptions, methods for entity resolution typically perform a pre-processing step, called blocking, which places similar entity descriptions into blocks and executes comparisons only between descriptions within the same block. The objective of this talk is to present challenges and algorithms for blocking for entity resolution, stemming from the Web openness in describing, by an unbounded number of KBs, a multitude of entity types across domains, as well as the high heterogeneity (semantic and structural) of descriptions, even for the same types of entities.

1 Introduction

Over the past decade, numerous *knowledge bases* (KBs) have been built to power large-scale knowledge sharing, but also an entity-centric Web search, mixing both structured data and text querying (e.g., [10]). These KBs offer comprehensive, machine-readable descriptions of a large variety of real-world entities (e.g., persons, places) published on the Web as *Linked Data* (LD). Traditionally, KBs are manually crafted by a dedicated team of knowledge engineers, such as the pioneering projects Wordnet and Cyc. Today, more and more KBs are built from existing Web content using information extraction tools [5]. Such an automated approach offers an unprecedented opportunity to scale-up KBs construction and leverage existing knowledge published in HTML documents [12].

Kostas Stefanidis
University of Tampere, Finland, e-mail: kostas.stefanidis@uta.fi

Although KBs (e.g., DBpedia [1], Freebase [2]) may be derived from the same data source (e.g., Wikipedia), they may provide multiple descriptions of the same entities. This is mainly due to the different information extraction tools and curation policies [6] employed by KBs, resulting to complementary and sometimes conflicting descriptions. *Entity resolution* (ER) aims to identify descriptions that refer to the same entity within or across KBs [4, 7]. ER is essential in order to improve *interlinking* in the Web of data, even by third-parties[1]. Compared to data warehouses, the new ER challenges stem from the *openness* of the Web of data in describing entities by an unbounded number of KBs, the *semantic and structural diversity* of the descriptions provided across domains even for the same entities, and the *autonomy* of KBs in terms of adopted processes for creating and curating descriptions. In particular:

- The size of the Linking Open Data (LOD) cloud[2], in which nodes are KBs (aka RDF datasets) and edges are links crossing KBs, has roughly doubled between 2011 and 2014 [14], while data interlinking dropped by 30%. In general, the majority of the KBs are sparsely linked, while their popularity in links is heavily skewed[3]. Sparsely interlinked KBs appear in the periphery of the LOD cloud (e.g., Open Food Facts, Bio2RDF), while heavily interlinked ones lie at the center (e.g., DBpedia, GeoNames, FOAF). Encyclopaedic KBs, such as DBpedia, or widely used georeferencing KBs, such as GeoNames, are interlinked with the largest number of KBs both from the LOD center and the periphery.
- The descriptions contained in these KBs present a high degree of semantic and structural diversity, even for the same entity types. The former is due to the frequent creation of new names for entities that have been described in another KB [11], as well as the simultaneous annotation of descriptions with semantic types not necessarily originating from the same vocabulary. The latter is due to the diverse sets of properties used to describe entities both in terms of types and number of occurrences, even within a KB.

The *scale*, *diversity* and *graph structuring* of entity descriptions in the Web of data challenge the way two descriptions can be effectively compared in order to efficiently decide whether they are referring to the same real-world entity. This clearly requires an understanding of the relationships among *somehow similar* entity descriptions that goes beyond duplicate detection without always being able to merge related descriptions in a KB and thus improve its quality. Furthermore, the *very large volume* of entity collections that we need to resolve in the Web of data is prohibitive when examining pairwise all descriptions.

In this context of big Web data, *blocking* is typically used as a pre-processing step for ER to reduce the number of unnecessary comparisons, i.e., comparisons between descriptions that do not match. After blocking, each description can be compared

[1] For instance, the sameas.org service provides co-references of the same entities between different KBs that have been manually collected.

[2] http://lod-cloud.net

[3] http://linkeddata.few.vu.nl/wod_analysis

only to others placed within the same block and thus disregard comparisons between descriptions that are unlikely to be matches. The desiderata of blocking are to place (i) similar descriptions in the same block (*effectiveness*), and (ii) dissimilar descriptions in different blocks (*efficiency*). However, efficiency dictates skipping many comparisons, possibly leading to many missing matches, which in turn implies low effectiveness. This is even more critical in the context of the Web of data, in which we do not know which pieces of the descriptions are the most appropriate to consider for computing the similarities. Thus, the main objective of blocking is to achieve a trade-off between the number of comparisons suggested and the number of missed matches.

Most of the blocking algorithms proposed in the literature (for a survey, refer to [3]) assume both the availability and knowledge of the schema of the input data, i.e., they refer to relational databases. To support a Web-scale resolution of heterogeneous and loosely structured entities across domains, recent blocking algorithms (e.g., [13, 9, 8]) disregard strong assumptions about knowledge of the schema of data and rely on a minimal number of assumptions about how entities match (e.g., when they feature a common token in their description or URI) within or across sources. In this talk, we will focus on the behavior of such blocking algorithms for datasets exhibiting different semantic and structural characteristics. Specifically, we are interested in quantifying the factors that make blocking algorithms take different decisions on whether two descriptions from real LOD sources potentially match or not. Finally, we will investigate typical cases of missed matches of existing blocking algorithms and examine alternative ways for them to be retrieved.

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC*, 2007.
2. K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.
3. P. Christen. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-centric systems and applications. Springer, 2012.
4. V. Christophides, V. Efthymiou, and K. Stefanidis. *Entity Resolution in the Web of Data*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, 2015.
5. P. Cimiano, C. Unger, and J. McCrae. *Ontology-Based Interpretation of Natural Language*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2014.
6. O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. Building, maintaining, and using knowledge bases: a report from the trenches. In *SIGMOD*, 2013.
7. X. L. Dong and D. Srivastava. *Big Data Integration*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2015.
8. V. Efthymiou, G. Papadakis, G. Papastefanatos, K. Stefanidis, and T. Palpanas. Parallel meta-blocking: Realizing scalable entity resolution over large, heterogeneous data. In *IEEE Big Data*, 2015.
9. V. Efthymiou, K. Stefanidis, and V. Christophides. Big data entity resolution: From highly to somehow similar entity descriptions in the web. In *IEEE Big Data*, 2015.

10. A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker. Searching and browsing linked data with SWSE: the semantic web search engine. *J. Web Sem.*, 9(4):365–401, 2011.
11. A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. *Web Semant.*, 14:14–44, 2012.
12. E. H. Hovy, R. Navigli, and S. P. Ponzetto. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artif. Intell.*, 194:2–27, 2013.
13. G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, and W. Nejdl. A blocking framework for entity resolution in highly heterogeneous information spaces. *IEEE Trans. Knowl. Data Eng.*, 25(12):2665–2682, 2013.
14. M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *ISWC*, 2014.