Minoan ER: Progressive Entity Resolution in the Web of Data

Vasilis Efthymiou, Kostas Stefanidis, Vassilis Christophides

Poster Paper

Minoan ER: Progressive Entity Resolution in the Web of Data

Vasilis Efthymiou
Univ. of Crete &
ICS-FORTH
vefthym@ics.forth.gr

Kostas Stefanidis
ICS-FORTH
kstef@ics.forth.gr

Vassilis Christophides
Univ. of Crete &
INRIA Paris-Rocquencourt
Vassilis.Christophides@inria.fr

ABSTRACT

Entity resolution aims to identify descriptions of the same entity within or across knowledge bases. In this work, we present the Minoan ER platform for resolving entities described by linked data in the Web (e.g., in RDF). To reduce the required number of comparisons, Minoan ER performs blocking to place similar descriptions into blocks and executes comparisons to identify matches only between descriptions within the same block. Moreover, it explores in a pay-as-you-go fashion any intermediate results of matching to obtain similarity evidence of entity neighbors and discover new candidate description pairs for resolution.

1. DESCRIPTION

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. These KBs offer comprehensive, machine-readable descriptions of a large variety of real-world entities (e.g., persons, places, products, events) published on the Web as Linked Data (LD). Although KBs (e.g., DBpedia, Freebase) may be derived from the same data source (e.g., a Wikipedia entry), they may provide multiple, non-identical descriptions of the same real-world entities. This is mainly due to the different information extraction tools and curation policies employed by KBs, resulting to complementary and sometimes conflicting entity descriptions. Entity resolution (ER) aims to identify descriptions that refer to the same real-world entity appearing either within or across KBs [2, 3]. 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 real-world entities, as well as the autonomy of KBs in terms of adopted processes for creating and curating entity descriptions. In particular:

- The number of KBs (aka RDF datasets) in the Linking Open Data (LOD) cloud has roughly tripled between 2011 and 2014 (from 295 to 1014), while KBs interlinking dropped by 30%. The main reason is that with more KBs available, it becomes more difficult for data publishers to identify relations between the data they publish and the data already published. Thus, the majority of KBs are sparsely linked, while their popularity in links is heavily skewed. 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). Encyclopaedic KBs, such as DBpedia, or widely used geo-referencing KBs, such as GeoNames, are interlinked with the largest number of KBs [6].
- The descriptions contained in these KBs present a high degree of semantic and structural diversity, even for the same entity types. Despite the Linked Data principles, multiple names (e.g., URIs) can be used to refer to the same real-world entity. The majority (58.24%) of the 649 vocabularies currently used by KBs are proprietary, i.e., they are used by only one KB, while diverse sets of properties are commonly used to describe the entities both in terms of types and number of occurrences even in the same KB. Only YAGO contains 350K different types of entities, while Google's Knowledge Graph contains 35K properties, used to describe 600M entities.

The two core ER problems, namely how can we (a) effectively compute similarity of entity descriptions and (b) efficiently resolve sets of entities within or across sources, are challenged by the large scale (both in terms of the number of sources and entity descriptions), the high diversity (both in terms of number of entity types and properties) and the importance of relationships among entity descriptions (not committing to a particular schema defined in advance). In particular, in addition to *highly similar* descriptions that feature many common tokens in values of semantically related attributes, typically met in the center of the LOD cloud and heavily interlinked mostly using owl:sameAs predicates, we are encountering *somehow similar* descriptions with significantly fewer common tokens in attributes not always semantically related, that appear usually in the periphery of the LOD cloud and are sparsely interlinked with various kinds of predicates. Plainly, the coming up of highly and somehow similar semi-structured entity descriptions requires solutions that go beyond those applicable to duplicate detection. A promising area of research in this respect is cross-domain similarity search and mining [8, 7], aiming to exploit similarity of objects described by different modalities (i.e., text,

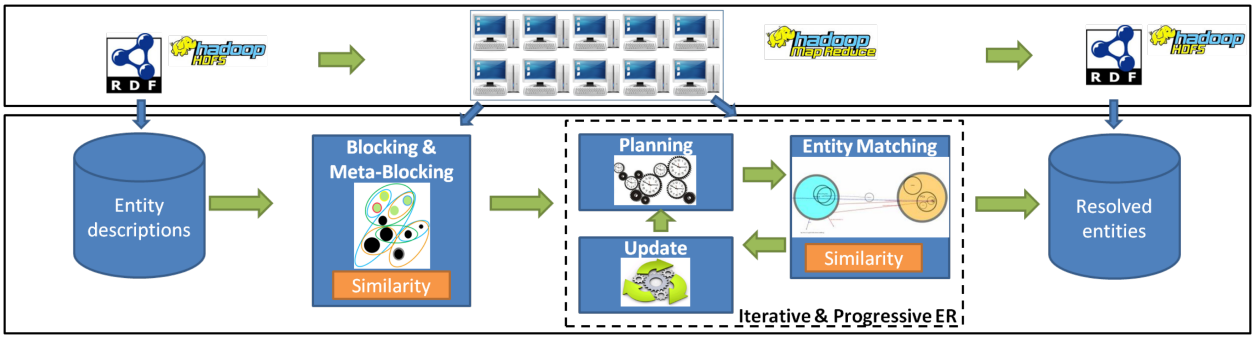

Figure 1: The Minoan ER Framework.

image) and contexts (i.e., facets) and support research by analogy. Such techniques could be also beneficial for matching highly heterogeneous entity descriptions and thus support ER at the Web scale.

We present in this poster the Minoan ER platform for resolving entities described by linked data in the Web (e.g., in RDF). Figure 1 illustrates the general steps involved in our process.

Blocking and Meta-blocking in Minoan ER: We use *blocking* as a pre-processing step for ER to reduce the number of required comparisons. Specifically, blocking places similar entity descriptions into blocks, leaving to the entity matching algorithm the comparisons only between descriptions within the same block.

Typically, token-based blocking algorithms place highly similar descriptions (having many common tokens) in many common blocks; intuitively, the more common blocks two descriptions share, the more likely it is that they match. This leads to many repeated comparisons between the same pairs of descriptions. To overcome this problem, we accompany blocking with *meta-blocking*, which prunes such repeated comparisons. Moreover, meta-blocking aims at discarding comparisons between descriptions that share few common blocks and are thus less likely to match. In Minoan ER, to support a Web-scale resolution of heterogeneous and loosely structured entities across domains, we use algorithms for blocking and meta-blocking that disregard strong assumptions about knowledge of the data schema and rely on a minimal number of assumptions about how entities match (e.g., when they feature a common token in their descriptions or URIs) within or across sources. For doing so, we exploit the parallel processing power of a computer cluster via Hadoop MapReduce, as presented in [5, 4].

Progressive Entity Matching in Minoan ER: Blocking approaches in the Web of data, especially when handling somehow similar descriptions appearing in the periphery of the LOD cloud, may miss highly heterogeneous matching descriptions featuring few common tokens [5]. To overcome that, we focus on exploiting the partial matching results as a similarity evidence for their neighbor (i.e., linked) descriptions. Since this inherently iterative process entails an additional overhead, we are interested in maximizing its benefit, given a computational cost budget. So, we need to estimate which part of the graph is the most promising to explore in the next iteration, in a *progressive* way.

In this respect, Minoan ER focuses on extending the typical ER workflow with a *scheduling phase*, which is responsible for selecting which pairs of descriptions, that have resulted from blocking, will be compared in the entity matching phase and in what order. The goal of this new phase is to favor more promising comparisons, i.e., those that are more likely to increase the targeted benefit. This way, those comparisons are executed before less promising ones and thus, higher benefit is provided early on in the process. The *update phase* propagates the results of matching, such that a new scheduling phase will promote the comparison of pairs that were influenced by the previous matches. This iterative process continues until the cost budget is consumed.

In contrast to existing works in progressive relational ER (e.g., [1]), which consider the quantity of entity pairs resolved, as the benefit of ER, we explore different aspects of data quality, improved through ER. In particular, we are interested in characterizing the quality of the resolved pairs, with respect to the number of descriptions resolved, corresponding to the same real-world entity (targeting attribute completeness), the number of real-world entities resolved (targeting entity coverage), and the number of real-world entity graphs resolved (targeting relationship completeness).

Acknowledgements: This work was partially supported by the EU H2020 PARTHENOS (#654119), FP7 DIACHRON (#601043) and FP7 SemData (#612551) projects.

2. REFERENCES

[1] Y. Altowim, D. V. Kalashnikov, and S. Mehrotra. Progressive approach to relational entity resolution. *PVLDB*, 7(11):999–1010, 2014.

[2] 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.

[3] X. L. Dong and D. Srivastava. *Big Data Integration*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2015.

[4] 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.

[5] 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.

[6] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *ISWC*, pages 245–260, 2014.

[7] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.*, 30(6):154, 2011.

[8] Y. Zhen, P. Rai, H. Zha, and L. Carin. Cross-modal similarity learning via pairs, preferences, and active supervision. In *AAAI*, pages 3203–3209, 2015.