WebER: Resolving Entities in the Web
(Research Contribution)

Contributor names

Vassilis Christophides
Vassilis Christophides graduated in 1988 from the Electrical engineering at the National Technical University of Athens (NTUA), Greece. He received his DEA in computer science in 1992 from the University PARIS VI, and his Ph.D. in 1996 from the Conservatoire National des Arts et Metiers (CNAM) of Paris, France. He has been scientific coordinator of a number of research projects funded by the European Union, the Greek State and private foundations on the Semantic Web and Digital Preservation at the Institute of Computer Science of FORTH. He recently joined Technicolor, R&I Center in Paris with a leave of absence from the University of Crete. He has published over 130 articles in high-quality international conferences, journals and workshops. He has received the 2004 SIGMOD Test of Time Award and the Best Paper Award at the 2nd and 6th International Semantic Web Conference in 2003 & 2007. He was the General Chair of the joint EDBT/ICDT Conference in 2014 at Athens and he will be the ICDE 2016 Area Chair for "Semi-structured, Web, and Linked Data Management" at Bali, Indonesia.
E-mail: christop@csd.uoc.gr

Kostas Stefanidis
Kostas Stefanidis is a research scientist at ICS-FORTH, Greece. Previously, he worked as a post-doctoral researcher at the IDI Dept. of NTNU in Norway, with a scholarship funded by the ERCIM Marie Curie Network, and the CSE Dept. of CUHK in Hong Kong. He got his PhD in personalized data management from the University of Ioannina, Greece, in 2009. His research interests lie in the intersection of databases, Web and information retrieval, and include personalized and context-aware data management systems, recommender systems, keyword-based search, and information extraction, resolution and integration. Kostas has been involved in several international projects and co-authored more than 35 papers in peer-reviewed conferences and journals, including ACM SIGMOD, IEEE ICDE and ACM TODS. He is the General co-Chair of the Workshop on Exploratory Search in Databases and the Web (ExploreDB), and he will be the Web & Information Chair of SIGMOD/PODS 2016.
E-mail: kstef@ics.forth.gr

Vasilis Efthymiou
Vasilis Efthymiou is a PhD candidate at the University of Crete and a member of the Information Systems Laboratory of the Institute of Computer Science at FORTH. The topic of his PhD research is entity resolution in the Web of data. He got his MSc and BSc degrees from the same university in 2012 and 2010, respectively. He has received undergraduate and postgraduate scholarships from FORTH, working in the areas of Semantic Web, non-monotonic reasoning, and Ambient Intelligence.
E-mail: vefthym@ics.forth.gr

Summary

Entity resolution aims to identify descriptions of the same entity within or across knowledge bases. Our work focuses on designing and developing the WebER platform, capable of resolving entities appearing in the Web. To reduce the required number of comparisons, WebER performs blocking to place similar descriptions into blocks and executes comparisons to identify matches only between descriptions within the same block. It examines ways to exploit in a pay-as-you-go fashion any intermediate results of blocking and matching, iteratively discovering new candidate description pairs for resolution. We also present our preliminary experimental evaluation of the components of our platform.

Extended Abstract

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). 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. Such an automated approach offers an unprecedented opportunity to scale-up KBs construction and leverage existing knowledge published in HTML documents.

Although they may be derived from the same data source (e.g., a Wikipedia entry), KBs (e.g., DBpedia, Freebase, Wikidata) 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. ER is essential in order to improve *interlinking* in the Web of data, even by third-parties and thus improve the level of *analytics* and *sense-making* of the Web data content. 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 georeferencing KBs, such as GeoNames, are interlinked with the largest number of KBs.
- 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 *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. The core task of the ER problem is to decide whether two descriptions match using an adequate similarity function. For specific domains and relatively small number of KBs, such similarity functions can be easily defined eventually using experts' knowledge. In cross-domain and large-scale ER, even deciding which is the most appropriate piece of descriptions for performing comparisons is an open research issue. For example, do we need to care only

for the values of the descriptions, or should we consider any graph structuring of descriptions? What is a reasonable trade-off for assessing similarity between the content-based and structure-based similarity of two descriptions? Moving one step forward, how does schematic information, in terms of employed attribute names and types, affect the degree of similarity of two descriptions?

In our work, we focus on designing and developing the WebER platform that is capable of resolving entity descriptions appearing in the Web of data. 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 resolution algorithm the comparisons only between descriptions within the same block. Its goal is to place as many matching descriptions as possible in common blocks, i.e., identify many matches, and only miss as few matches as possible. To place two descriptions into the same block, different criteria are employed that mostly reflect the content similarity of the descriptions. For further reducing the number of comparisons to be performed by ER, blocking can be accompanied by block post-processing steps. Such steps make sense to be used, when blocking results in missing only few matches, and the whole process is faster than exhaustively performing the comparisons between all descriptions. Clearly, it is not straightforward to attain the best trade-off between pruning many comparisons, while retaining the comparisons between matches, since it is not easy to select, or even construct, the appropriate similarity function to use.

To minimize the number of missed matches, an iterative entity resolution process can exploit in a *pay-as-you-go* fashion any intermediate results of blocking and matching, progressively discovering new candidate description pairs for resolution. Such an iterative process considers similarity evidence provided by entity descriptions placed into the same block or being structurally related in the original entity graph. This way, an iterative approach is more suitable for coping with the *varying data quality* (e.g., incompleteness) and *loose structuring* (e.g., diverse entity graphs) of entity descriptions in the Web of data. In overall, Figure 1 illustrates the general steps involved in the ER process of WebER.

Finally, in this contribution, we provide an experimental evaluation of a large part of the algorithms implemented in WebER and explain the involved tradeoffs for real KBs in the Web of data.

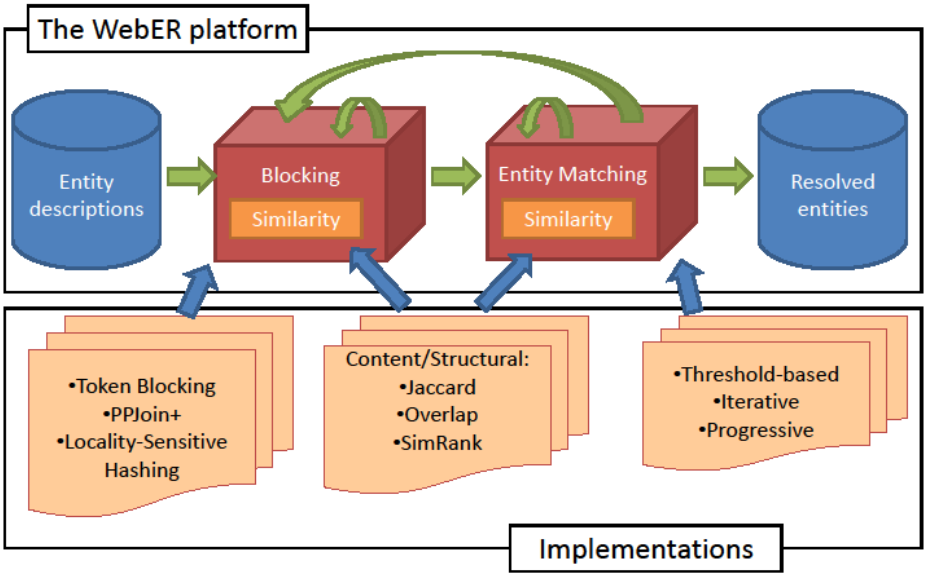

Figure 1: The WebER platform.