Entity Resolution in the Web of Data

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

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

Melanie Herschel
Université Paris Sud / Inria
melanie.herschel@lri.fr

Vassilis Christophides
R&I Center, Technicolor
vassilis.christophides@technicolor.com

ABSTRACT

This tutorial provides an overview of the key research results in the area of entity resolution that are relevant to addressing the new challenges in entity resolution posed by the Web of data, in which real world entities are described by interlinked data rather than documents. Since such descriptions are usually partial, overlapping and sometimes evolving, entity resolution emerges as a central problem both to increase dataset linking but also to search the Web of data for entities and their relations.

Categories and Subject Descriptors

H.2 [**DATABASE MANAGEMENT**]: General

Keywords

Web of Data, Entity Resolution

1. MOTIVATION & TUTORIAL OUTLINE

Nowadays, a vast and rapidly increasing quantity of government, scientific, corporate, and crowd-sourced data is published on the Web. The emerging *Web of data* realizes the vision of a *data-driven decision-making* and is expected to play a catalyst role in the way structured information is exploited across-domains in a large scale. According to the Linked Data paradigm, real world entities are described on the Web by interlinked data rather than documents. Data publishers are thus encouraged to describe entities using W3C standards (e.g., RDF) by naming them with HTTP URIs, so that people can access their descriptions on the Web, as well as by including links to other URIs, so that people can discover more entities, such as persons, things, places and artifacts.

Due to the open and decentralized nature of the Web, real world entities are usually described in multiple datasets using different URIs in a *partial*, *overlapping* and sometimes *evolving* way. Recognizing descriptions of the same real-world entities across, and sometimes within, data sources emerges as a central problem in the context of the Web of data. Addressing this problem, referred to as *entity resolution*, is a prerequisite to various applications, namely, semantic search in terms of entities and their relations on top of the Web of text, interlinking entity descriptions in autonomous sources to strengthen the Web of data, and supporting deep reasoning using related ontologies to create the Web of knowledge.

Data describing entities are made available in the Web under different formats (e.g., tabular, tree or graph) of varying *structuredness*. Traditional entity resolution techniques, for instance, used for merging customer databases or library catalogues, are not suited for the Web of data, due to *high heterogeneity* (i.e., different properties are used to describe the same kind of entity in different domains) and *non-regularity* in data structuring (i.e., even within the same domain, properties describing the same kind of entity significantly vary in terms of occurrences and types). Typically, an entity described in knowledge bases, such as Yago or Freebase, is declared to be instance of several *semantic types*, i.e., classes. The description of such an entity may employ properties from different vocabularies, resulting in quite different *structural types* even for descriptions of same type, e.g., person or place. Thus, matching these loosely structured entities is one of the major challenges of entity resolution that arise for the Web of data and that have not been addressed by entity resolution in relational databases. Even when Web data are accompanied with ontologies, these are more simple vocabularies (of classes or properties) rather than schemas imposing extensive structural and semantic data constraints. In fact, there exist 366 distinct vocabulary spaces in the LOD cloud, while only Freebase describes 25M entities using 4,000 properties. Also, given the large scale of the Web of data (>60B of triples), entity resolution in this context calls for efficient parallel and distributed techniques.

After a brief overview of entity resolution and its unique challenges in the Web of data, we present the general solution space. We then explore three directions in detail. First, we discuss iterative methods increasing the number of matched entities. We then present blocking techniques that reduce entity resolution runtime. Finally, special emphasis will be given to MapReduce techniques for entity resolution, since they appear to be a good potential for coping with the large scale of the Web of data, as well as to recent approaches on entity resolution with the help of crowdsourcing. We conclude the tutorial with open research questions.

2. ACKNOWLEDGMENTS

This work was partially supported by LODGOV and IdeaGarden (FP7-318552).

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577263.