Entity Resolution in the Web of Data

Synthesis Lectures on the Semantic Web

Editor

Ying Ding, ***Indiana University***
Paul Groth, ***Elsevier Labs***

Synthesis Lectures on the Semantic Web: Theory and Application is edited by Ying Ding of Indiana University and Paul Groth of Elsevier Labs. Whether you call it the Semantic Web, Linked Data, or Web 3.0, a new generation of Web technologies is offering major advances in the evolution of the World Wide Web. As the first generation of this technology transitions out of the laboratory, new research is exploring how the growing Web of Data will change our world. While topics such as ontology-building and logics remain vital, new areas such as the use of semantics in Web search, the linking and use of open data on the Web, and future applications that will be supported by these technologies are becoming important research areas in their own right. Whether they be scientists, engineers or practitioners, Web users increasingly need to understand not just the new technologies of the Semantic Web, but to understand the principles by which those technologies work, and the best practices for assembling systems that integrate the different languages, resources, and functionalities that will be important in keeping the Web the rapidly expanding, and constantly changing, information space that has changed our lives.
Topics to be included:

- Semantic Web Principles from linked-data to ontology design
- Key Semantic Web technologies and algorithms
- Semantic Search and language technologies
- The Emerging "Web of Data" and its use in industry, government and university applications
- Trust, Social networking and collaboration technologies for the Semantic Web
- The economics of Semantic Web application adoption and use
- Publishing and Science on the Semantic Web
- Semantic Web in health care and life sciences

Entity Resolution in the Web of Data

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis
2015

Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description
Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter
2015

Semantic Mining of Social Networks
Jie Tang and Juanzi Li
2015

Social Semantic Web Mining
Tope Omitola, Sebastián A. Ríos, and John G. Breslin
2015

Semantic Breakthrough in Drug Discovery
Bin Chen, Huijun Wang, Ying Ding, and David Wild
2014

Semantics in Mobile Sensing
Zhixian Yan and Dipanjan Chakraborty
2014

Provenance: An Introduction to PROV
Luc Moreau and Paul Groth
2013

Resource-Oriented Architecture Patterns for Webs of Data
Brian Sletten
2013

Aaron Swartz's A Programmable Web: An Unfinished Work
Aaron Swartz
2013

Incentive-Centric Semantic Web Application Engineering
Elena Simperl, Roberta Cuel, and Martin Stein
2013

Publishing and Using Cultural Heritage Linked Data on the Semantic Web
Eero Hyvönen
2012

VIVO: A Semantic Approach to Scholarly Networking and Discovery
Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding
2012

Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer
2011

Entity Resolution in the Web of Data

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis

ISBN: 978-3-031-79467-4 paperback
ISBN: 978-3-031-79468-1 ebook

DOI 10.1007/978-3-031-79468-1

A Publication in the Springer series
SYNTHESIS LECTURES ON THE SEMANTIC WEB

Lecture #13
Series Editors: Ying Ding, *Indiana University*
Paul Groth, *Elsevier Labs*
Founding Editor Emeritus: James Hendler, *Rensselaer Polytechnic Institute*
Series ISSN
Print 2160-4711 Electronic 2160-472X

Entity Resolution in the Web of Data

Vassilis Christophides
University of Crete, Greece
INRIA, France

Vasilis Efthymiou
University of Crete, Greece
ICS-FORTH, Greece

Kostas Stefanidis
ICS-FORTH, Greece

SYNTHESIS LECTURES ON THE SEMANTIC WEB #13

ABSTRACT

In recent years, several knowledge bases have been built to enable large-scale knowledge sharing, but also an entity-centric Web search, mixing both structured data and text querying. These knowledge bases offer machine-readable descriptions of real-world entities, e.g., persons, places, published on the Web as Linked Data. However, due to the different information extraction tools and curation policies employed by knowledge bases, multiple, *complementary* and sometimes *conflicting* descriptions of the same real-world entities may be provided. Entity resolution aims to identify different descriptions that refer to the same entity appearing either within or across knowledge bases.

The objective of this book is to present the *new entity resolution challenges* stemming from the *openness* of the Web of data in describing entities by an unbounded number of knowledge bases, the *semantic and structural diversity* of the descriptions provided across domains even for the same real-world entities, as well as the *autonomy* of knowledge bases in terms of adopted processes for creating and curating entity descriptions. The *scale*, *diversity*, and *graph structuring* of entity descriptions in the Web of data essentially challenge how two descriptions can be effectively compared for similarity, but also how resolution algorithms can efficiently avoid examining pairwise all descriptions.

The book covers a wide spectrum of entity resolution issues at the Web scale, including basic concepts and data structures, main resolution tasks and workflows, as well as state-of-the-art algorithmic techniques and experimental trade-offs.

KEYWORDS

entity resolution, Web of data

Contents

List of Figures

List of Tables

Preface

Over the past decade, numerous knowledge bases (KBs) have been built to power a new generation of Web applications that provide *entity-centric search* and *recommendation services*. 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). Even when derived from the same data source (e.g., a Wikipedia entry), KBs such as DBpedia, YAGO2, or Freebase may provide multiple, non-identical descriptions for the same real-world entities. This is due to the different information extraction tools and curation policies employed by KBs, resulting in *complementary* and sometimes *conflicting* entity descriptions. Entity resolution (ER) aims to identify different descriptions that refer to the same real-world entity, and emerges as a central data-processing task for an *entity-centric organization* of Web data. ER is needed to enrich interlinking of data elements describing entities, even by third parties, so that the Web of data can be accessed by machines as a *global data space* using standard languages, such as SPARQL. ER can also facilitate an automated KB construction by integrating entity descriptions from legacy KBs with Web content published as HTML documents.

ER has attracted significant attention from many researchers in information systems, database, and machine-learning communities. The objective of this lecture is to present the *new ER challenges* 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. The *scale*, *diversity*, and *graph structuring* of entity descriptions published according to the LD paradigm challenge the core ER tasks, namely, (i) how descriptions can be *effectively compared for similarity* and (ii) how resolution algorithms can *efficiently filter the candidate pairs* of descriptions that need to be compared.

In a multi-type and large-scale entity resolution, we need to examine whether two entity descriptions are *somehow* (or near) *similar* without resorting to domain-specific similarity functions and/or mapping rules. Furthermore, the resolution of some entity descriptions might influence the resolution of other neighborhood descriptions. This setting clearly goes beyond deduplication (or record linkage) of collections of descriptions, usually referring to a single entity type, that slightly differ only in their attribute values. It essentially requires leveraging similarity of descriptions both on their *content* and *structure*. It also forces us to revisit traditional ER workflows consisting of separate *indexing* (for pruning the number of candidate pairs) and *matching* (for resolving entity descriptions) phases.

This Synthesis lecture is intended to provide a starting point for researchers, students, and developers who are interested in a global view of the ER problem in the Web of data. Throughout the lecture, we present the basic concepts and resolution workflows, as well as state-of-the-art

indexing and matching techniques. We additionally survey new ER execution strategies (such as parallel/distributed and progressive strategies) to resolve, under specific efficiency or effectiveness constraints, very large collections of entity descriptions, eventually arriving in streams. We made an effort to define in a self-contained way the similarity measures and data structures involved in various algorithms along with representative examples. We finally provide an experimental evaluation of a large part of the presented techniques and explain the involved trade-offs for real KBs in the Web of data. Much of the material presented in this lecture has been used in graduate courses taught at the University of Crete, as well as in two recent tutorials at CIKM'13 and WWW'14.

Since ER is a specialized problem of Data Integration, our Synthesis lecture provides complementary material with other books in this research area. [Doan et al., 2012] focuses on models, languages, and architectures for Data Integration systems, as well as on techniques for rewriting and processing queries on top. It also covers machine learning techniques for inferring mappings/matchings between heterogeneous relational and Web data. [Dong and Srivastava, 2015] stresses the Data Integration challenges in the Big Data era. In particular, it details how well-known ER algorithms can benefit from parallel and distributed implementations, aiming to reduce the overall execution time of the entire ER process. The book also considers schema alignment, as well as techniques for linking text snippets with embedded attributes, to structured records. Record Linkage techniques and Deduplication techniques for traditional Data Warehouse settings have been the subject of numerous surveys and books, such as [Naumann and Herschel, 2010] and [Christen, 2012]. Finally readers are referred to [Abiteboul et al., 2011] for a comprehensive overview of languages and technologies involved in Web Data Management.

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis
June 2015

Acknowledgments

Several people provided valuable support during the preparation of this book, without whose help the project could not have been satisfactorily and timely completed. We warmly thank Ying Ding and Paul Groth for inviting us to write this book and Michael Morgan for managing the entire publication process. Special thanks also go to Christian Bizer for constructive feedback during the review process of the book. We would also like to acknowledge our many collaborators who have influenced our thoughts and our understanding of this research area over the years, and the following projects for their support in our research efforts: EU FP7-ICT-2011-9 DIACHRON (Managing the Evolution and Preservation of the Data Web), EU FP7-PEOPLE-2013-IRSES SemData (Semantic Data Management), EU FP7-ICT-318552 IdeaGarden (An Interactive Learning Environment Fostering Creativity), and LoDGoV (Generate, Manage, Preserve, Share, and Protect Resources in the Web of Data) of the Research Programme ARISTEIA (EXCELLENCE), GSRT, Ministry of Education, Greece, and the European Regional Development Fund. Finally, we would like to thank the ~okeanos GRNET cloud service that is used in our experimental evaluation.

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis
June 2015