ML-Based Knowledge Graph Curation: Current Solutions and Challenges

Laure Berti-Equille [1,2]
[1]ESPACE-DEV/IRD, UMR 228, IRD/UM/UG/UR, Montpellier, France
[2]Aix Marseille Université, Université de Toulon, CNRS, LIS, DIAMS, Marseille, France
laure.berti@ird.fr

ABSTRACT

With the success of machine learning (ML) techniques, ML has already proved a tremendous potential to impact the foundations, algorithms, and models of several data management tasks, such as error detection, data quality assessment, data cleaning, and data integration. In Knowledge Graphs, part of the data preparation and cleaning processes, such as data linking, identity disambiguation, or missing value inference and completion could be automated by making a ML model "learn" and predict the matches routinely with different degrees of supervision. This talk will survey the recent trends of applying machine learning solutions to improve and facilitate Knowledge Graph curation and enrichment, as one of the most critical tasks impacting Web search and query-answering. Finally, the talk will discuss the next research challenges in the convergence of machine learning and management of Knowledge Graph evolution and preservation.

CCS CONCEPTS

• **Information systems → Data cleaning**; **Semantic web description languages**; *Graph-based database models*; • **Computing methodologies → Machine learning approaches**.

KEYWORDS

Knowledge base curation; knowledge graph completion; entity disambiguation

ACM Reference Format:
Laure Berti-Equille [1,2]. 2019. ML-Based Knowledge Graph Curation: Current Solutions and Challenges. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316522

1 INTRODUCTION

Over the past few years, massive amounts of world knowledge have been accumulated in Web knowledge bases, both commercial and openly available such as Freebase, DBpedia, and YAGO, among the most prominent ones. These KBs contain millions of entities (e.g., people, places, or organizations) and millions of related facts expressed in RDF (Resource Description Framework) in the form of triples such as < `head entity, relation, tail entity`> indicating the relation between two entities.

Knowledge graphs are often constructed from semi-structured knowledge as Wikipedia or harvested from the Web with a combination of statistical and NLP methods. However, Web Data used for knowledge base construction is noisy, unreliable, highly imbalanced, heterogeneous, and evolve over time [1]. The result are large-scale knowledge graphs that try to make a good trade-off between completeness and correctness [3].

Not surprisingly, Web knowledge graphs suffer from a wide range of anomalies: they are greatly incomplete (i.e., with missing entities and links), some entities are ambiguous, and identity relations between resources that refer to the same real world entity are erroneous.

Building high-quality knowledge bases critically depends on the data curation technologies, which are rapidly evolving with the recent advances in representation learning and machine learning. These advances are now leveraged for knowledge base completion, refinement, entity linking, and entity disambiguation.

In this talk, we will survey the recent trend of applying machine learning solutions to improve knowledge graph curation tasks and establish new paradigms to sharpen knowledge graph error detection and cleaning, as an attempt to fill the gap in knowledge curation science [4]. We will discuss the advantages and limitations of these techniques and their extensions to address the current challenges, not only in extracting knowledge from both structured and unstructured data, across a large variety of domains, and in multiple languages, but also in maintaining high quality in evolving knowledge repositories.

2 SPEAKER'S BIO

Laure Berti-Équille is a Research Director at IRD, the French research institute for sustainable development, currently leading the research group DIAMS (Data Integration, Analysis, and Management as Services, http://diams.lis-lab.fr at Aix-Marseille University. Before, she was a full professor at Aix-Marseille University (AMU), senior scientist at Qatar Computing Research Institute (Hamad Bin Khalifa University, Qatar), Associate Professor (with tenure) at University of Rennes 1 (France), and visiting researcher at AT&T Labs Research (USA) as a recipient of the prestigious European Marie Curie Outgoing Fellowship. Her interests are at the intersection of large-scale data analytics, and statistical machine learning with a focus on data quality, anomaly detection, and truth discovery, with more than 80 publications and three monographs. She initiated the very first workshop editions on information and data quality in information systems (IQIS 2005) and quality in databases (QDB 2009 and 2016) in conjunction with SIGMOD and VLDB respectively,

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316522

and co-organized the first French workshops on Data and Knowledge Quality in conjunction with EGC (Extraction et Gestion de Connaissances) in 2005, 2006, 2010, and 2011. She has received various grants from the French Agency for National Research (ANR), the French National Research Council (CNRS), and the European Union.

REFERENCES

[1] Berti-Equille L., Scannapieco M. (2016). Quality of Web Data (Chapter). In the 2nd Edition of the book Data Quality: Concepts, Methodologies and Techniques, Springer, 2016

[2] Zaveri A., Maurino A., Berti-Equille L. (2014). Web Data Quality: Current State and New Challenges, Int. J. Semant. Web Inf. Syst.,10(2):1552-6283, IGI Global.

[3] Paulheim H. (2017) Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8 3 489-508.

[4] Paritosh P. (2018). The Missing Science of Knowledge Curation (Improving incentives for large-scale knowledge curation). In Companion of The Web Conference 2018, Lyon, France.