Knowledge Bases for Web Content Analytics

[Extended Abstract]

Johannes Hoffart
Max Planck Institute for Informatics
Saarbrücken, Germany
jhoffart@mpi-inf.mpg.de

Nicoleta Preda
University of Versailles
Versailles, France
preda@prism.uvsq.fr

Fabian M. Suchanek
Télécom ParisTech
Paris, France
suchanek@enst.fr

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

Keywords

Knowledge Bases, Ontologies, Web, Information Extraction

1. TUTORIAL DESCRIPTION

The proliferation of knowledge-sharing communities such as Wikipedia and the progress in scalable information extraction from Web and text sources has enabled the automatic construction of very large knowledge bases (KBs). Recent endeavors of this kind include academic research projects such as DBpedia, KnowItAll, Probase, ReadTheWeb, and YAGO, as well as industrial ones such as Freebase, the Google Knowledge Graph, Amazon's Evi, Microsoft's Satori, and related efforts at Bloomberg, Walmart, and others. These projects provide automatically constructed KBs of facts about named entities, their semantic classes, and their mutual relationships. They usually contain millions of entities and hundreds of millions of facts about them. Such world knowledge in turn enables cognitive applications and knowledge-centric services like disambiguating natural-language text, entity linking, deep question answering, and semantic search and analytics over entities and relations in Web and enterprise data. Prominent examples of how knowledge bases can be harnessed include the Google Knowledge Cards and the IBM Watson question answering system.

This tutorial presents state-of-the-art methods, recent advances, research opportunities, and open challenges in the field of knowledge harvesting and its applications. Particular emphasis will be on the twofold role of KBs for big-data analytics: using scalable distributed algorithms for harvesting knowledge from Web and text sources, and leveraging entity-centric knowledge for deeper interpretation of and better intelligence with big data.

In particular, the tutorial will cover a wide spectrum of methods for automatically constructing large KBs, for extending them, and for harnessing them in applications like text annotation, disambiguation, and entity linking. Participants will obtain an in-depth understanding of state-of-the-art KBs, how they are built and maintained, how knowledge harvesting can utilize scalable algorithms, and how knowledge can contribute to big-data analytics. Finally, as the relevant literature is widely dispersed across different communities like Web and Data Mining (WSDM, WWW, and KDD), Artificial Intelligence (IJCAI and AAAI), Natural Language Processing (ACL and EMNLP), Semantic Web (ISWC), Information Retrieval (SIGIR and CIKM), and Data Management (SIGMOD, VLDB, and ICDE), the tutorial also serves as a guided tour on the latest research in these venues and aims to offer a unifying big picture. An extensive bibliography as of 2013 is given in [1].

2. OUTLINE

Constructing knowledge bases. The first part introduces the knowledge representation of the Semantic Web. It will then discuss the automated construction of knowledge bases, as well as the acquisition of commonsense knowledge.
Linking knowledge. The second part discusses techniques that link entities across different sources: named entity recognition and disambiguation, record linking, semantic relatedness of entities, knowledge fusion, and the linking of Web services to KBs.
Web analytics. The last part highlights applications of semantic knowledge for the analysis of and knowledge discovery in Web contents. We will focus on the mining of logical rules in knowledge, entity-based search and analytics, and event prediction.

3. REFERENCES

[1] F. M. Suchanek and G. Weikum. Knowledge Harvesting in the Big-Data Era. SIGMOD 2013.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741984.