Proposal for WWW 2015 Tutorial on "Knowledge Bases for Web Content Analytics"

Presenters:

Johannes Hoffart
Max Planck Institute for Informatics, Germany
Homepage: http://www.mpi-inf.mpg.de/~jhoffart
E-Mail: jhoffart@mpi-inf.mpg.de

Nicoleta Preda
University of Versailles, France
Homepage: http://preda.fr
E-Mail: nicoleta@preda.fr

Fabian Suchanek
Télécom ParisTech University, France
Homepage: http://suchanek.name
E-Mail: fabian@suchanek.name

Gerhard Weikum
Max Planck Institute for Informatics, Germany
Homepage: http://www.mpi-inf.mpg.de/~weikum/
E-Mail: weikum@mpi-inf.mpg.de

Duration: 3 hours. No additional equipment is required.

Topic and Relevance:

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. 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 Vault, and Amazon's Evi. These projects provide automatically constructed knowledge bases 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 along this avenue of knowledge harvesting and its applications. Particular emphasis will be on the twofold role of knowledge bases 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.

Description:

The tutorial will cover a wide spectrum of methods for automatically constructed large knowledge bases, for extending them, and for harnessing them in intelligent applications like text annotation, disambiguation, and entity linking. Participants will obtain an in-depth understanding of state-of-the-art knowledge bases, how they are built and maintained, how knowledge harvesting can utilize scale-out algorithms, and how knowledge can contribute to big-data analytics. In addition, the tutorial will discuss hot topics in knowledge base construction, pointing out research opportunities and open challenges. Finally, as the relevant literature is widely dispersed across different communities like Web Mining (WSDM, WWW, and KDD), Artificial Intelligence (IJCAI and AAAI), Natural Language Processing (ACL and EMNLP), Semantic Web (ISWC), 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.

A tentative outline of the tutorial looks as follows.

Part 1: Knowledge Bases and their Automatic Construction (60 minutes)

Knowledge Bases in the Big Data Era.

Today's knowledge bases represent their data mostly in RDF or in triple form. We will introduce this data format briefly. We will then go on to discuss the most salient knowledge base construction projects, which include KnowItAll, ConceptNet, DBpedia, Freebase, NELL, WikiTaxonomy, and YAGO. We will briefly cover industrial projects like the Google Knowledge Graph, the EntityCube project and the Probase project at Microsoft Research, and IBM's Watson project.

Harvesting Knowledge on Entities and Classes.

Every entity in a knowledge base (such as `Steve_Jobs`) belongs to one or multiple classes (such as `computer_pioneer`). These classes are organized into a taxonomy, where more special classes are subsumed by more general classes (such as `person`). We will discuss two groups of approaches to harvest such information: Wikipedia-based approaches (such as YAGO and WikiTaxonomy), and Web-based approaches that use set-expansion.

Part 2: Harvesting Facts at Web Scale (60 minutes)

Harvesting Relational Facts.

Relational facts express relationships between two entities. There is a large spectrum of methods to extract such facts from Web documents. We will give an overview on methods from pattern matching (e.g., regular expressions), computational linguistics (e.g., dependency parsing), statistical learning (e.g., factor graphs and MLNs), and logical consistency reasoning (e.g., weighted MaxSat or ILP solvers). We will also discuss to what extent these approaches scale to handle big data.

Open Information Extraction.

In contrast to approaches that operate on a predefined list of relations and entities, open IE harvests arbitrary subject-predicate-object triples from natural language documents. It aggressively taps into noun phrases as entity candidates and verbal phrases as prototypic patterns for relations. We discuss recent methods that follow this direction. Some methods along these lines make clever use of big-data techniques like frequent sequence mining and map-reduce computation.

Temporal, Multilingual, Dynamic, and Commonsense Knowledge.

In this part, we venture beyond simple factoids and describe approaches that attach meta-information to their facts. This meta information can concern the time or location of a fact, or describe entities in multiple languages. We also discuss a dimension that complements factual knowledge by commonsense knowledge: properties and rules that every child knows but are hard to acquire by a computer. Here again, state-of-the-art methods use techniques that scale out to cope with huge inputs. Finally, we also discuss the integration of dynamic knowledge, i.e., knowledge obtained from Web services. This includes the schema matching and the query evaluation for integrating data dynamically into the knowledge base.

Part 3: Knowledge for Big-Data Analytics (60 minutes)

When analytic tasks tap into text or Web data, it is often crucial to identify entities (people, places, products, etc.) in the input for proper grouping and other purposes. An example application could aim to track and compare two entities in social media over an extended timespan (e.g., the Apple iPhone vs. Samsung Galaxy families). In this context, knowledge about entities can be a valuable asset.

Named-Entity Disambiguation.

When extracting knowledge from text or tables, entities are first seen only in surface form: by names (e.g., “Jobs”) or phrases (e.g., “the Apple founder”). Such entity mentions are often ambiguous; mapping them to canonicalized entities registered in a knowledge base is the task of named-entity disambiguation (NED). State-of-the-art NED methods combine context similarity between the surroundings of a mention and salient phrases associated with an entity, with coherence measures for two or more entities co-occurring together. Although these principles are well understood, NED remains an active research area towards improving robustness, scalability, and coverage.

Entity Search.

More recently, entity annotations have been used as basis to improve end-user oriented tasks such as information retrieval. Here, regular keyword search is augmented by entities, allowing users to specify their information need more accurately, increasing both precision (by resolving ambiguous names) and recall (additional names for entities are expanded). Using the knowledge base taxonomies, users can also search by classes, opening up the possibility to search for sets of entities without actually knowing the content.

Entity Linkage.

Even when entities are explicitly marked in structured or semi-structured data (e.g., RDF triples), the problem arises to tell whether two entities are the same or not. This is a variant of the classical record-linkage problem (aka. entity matching, entity resolution, entity de-duplication). For knowledge bases and Linked Open Data, it is of particular interest because of the need for generating and maintaining owl:sameAs information across knowledge resources. We give an overview of approaches to this end, covering statistical learning approaches and graph algorithms.

Audience and Prerequisites:

The tutorial will be useful for doctoral students and faculty interested in a wide spectrum of topics: knowledge bases, text mining, information extraction, semantic annotation of Web data with entities and relations, linked open data, semantic search and question answering, big data analytics on Web contents and social media, and more. The tutorial should also appeal to industrial researchers and practitioners working on semantic technologies for Web, social-media, or enterprise contents, including all kinds of applications where sense-making from text or semi-structured data is an issue. Prior knowledge on natural language processing or statistical learning is not required; the tutorial will introduce these techniques as they are needed.

Previous Editions:

Prior versions of this tutorial have been given at SIGMOD 2013 in New York City (3 hours[1]), ICDE 2013 in Brisbane (90 minutes[2]), and VLDB 2014 in Hangzhou (3 hours[3]). Each was attended by 20 to 25 participants. The one at VLDB was attended by ca. 50 participants. Given the size of the Web research community and the high number of parallel tracks at each of the flagship conferences, we would be confident to attract 20 to 40 attendees who have not yet been at any of the previous tutorials. Also, as the topic of knowledge bases for analytics is a hot issue, the content of the tutorial is constantly being updated. The present proposal puts particular emphasis on dynamic Web data and entity search. One of the authors has also presented related material at two summer schools: for 3 hours at the EDBT 2011 Summer School in St. Petersburg, and over 2.5 days at the VLDB School in Kunming, China, in July 2012.

[1] http://resources.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial

[2] http://resources.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

[3] http://resources.mpi-inf.mpg.de/yago-naga/vldb2014-tutorial/