Semantic Knowledge Bases from Web Sources

Tutorial Proposal for IJCAI 2011 (1 day)

Fabian M. Suchanek[1], Martin Theobald[2], Gerhard Weikum[2], Hady W. Lauw[3], and Ralf Schenkel[4]

[1] INRIA Saclay, Paris
[2] Max Planck Institute Informatics, Saarbrücken
[3] Institute for Infocomm Research, Singapore
[4] Saarland University, Saarbrücken

Keywords: information extraction, knowledge harvesting, machine reading, RDF knowledge bases, ranking

1 Short Description

Recent advances in scalable information extraction from Web sources has enabled the automatic construction of large knowledge bases, such as DBpedia, EntityCube, KnowItAll, ReadTheWeb, and YAGO-NAGA. In this tutorial, we explain how these knowledge bases are organized, how they are constructed, how they can be maintained and extended in structure and contents, and how they serve as enabling assets for semantic applications.

2 Longer Description

The advent of knowledge-sharing communities such as Wikipedia and the progress in scalable information extraction from Web sources has enabled the automatic construction of large knowledge bases. Recent endeavors of this kind include academic research projects such as DBpedia, EntityCube, KnowItAll, ReadTheWeb, and YAGO-NAGA, as well as industrial ones such as Freebase and Trueknowledge. These projects provide automatically-constructed, large and rich knowledge bases of facts about named entities, their semantic classes, and their mutual relations. This 1-day tutorial discusses a) the content, organization, and potential of these Web-induced knowledge bases, b) state-of-the-art methods for constructing them from semistructured and textual Web sources, c) recent approaches to maintaining and extending them, which includes introducing a temporal dimension of knowledge, d) use cases of knowledge bases, including semantic search, reasoning for question answering, and entity linking and disambiguation. It also points out open problems and research opportunities on this spectrum of issues.

3 Target Audience

We believe that this tutorial would be of interest to a broad audience at IJCAI, because it brings together different, though related, topics of research. The tutorial bridges the areas of data and text mining, knowledge extraction, knowledge-based search, and uncertain data management. It aims at providing valuable knowledge about available data assets, as well as basic methods for knowledge base construction and querying to researchers working on knowledge discovery, semantic search on Web and enterprise sources, or coping with automatically extracted facts as a major use case for uncertain data management. In addition, it summarizes the state of the art, and points out research opportunities to those who are specifically interested in bringing Web mining and Web search to a more database-oriented, valued-added level of gathering, organizing, searching, and ranking entities and relations from Web sources.

This tutorial uses some material from an invited tutorial presented at PODS 2010 [2] and from a tutorial presented at CIKM 2010 [5].

4 Outline

The tutorial is organized into four main parts:

- Part 1 (2 hours) explains the content and organization of the largest ones of the publicly available knowledge bases, and their value in a variety of application use cases;
- Part 2 (3 hours) gives an overview of different methodological paradigms and concrete techniques for automatically constructing such knowledge bases from Web sources with high quality. This includes pattern-based, learning-based, and reasoning-based techniques.
- Part 3 (1 hour) discusses the maintenance of knowledge bases along the temporal dimension, and approaches to capturing time-variant facts.
- Part 4 (1 hour) discusses further extensions of knowledge bases such as multilingual and multimodal knowledge or common-sense properties, and advanced use cases like entity linking and disambiguation.

5 About the Speakers

Fabian Suchanek is a postdoctoral researcher at INRIA Saclay in Paris. He spent one year as a visiting researcher at Microsoft Research Silicon Valley in 2009. Fabian obtained his doctoral degree from Saarland University in 2008. In his dissertation, Fabian developed methods for the automatic construction and maintenance of a large knowledge base, YAGO. For his thesis, he received the ACM SIGMOD Dissertation Award Honorable Mention. The original YAGO paper at the WWW Conference in 2007 has received more than 300 citations, and YAGO is used in many major knowledge-base projects around the world (including DBpedia). Fabian has published a dozen papers in top-tier conferences (WWW, SIGMOD, ICDE, SIGIR), inter alia two tutorials and two survey articles. He is teaching classes on the Semantic Web, Natural Language Processing and Information Extraction at the École nationale supérieure des Télécommunications in Paris/France.

Fabian M. Suchanek
INRIA, building G, office 116
4, rue Jacques Monod
91893 Orsay Cedex
France

fabian@suchanek.name
phone: +33 01 72 92 59 11
fax: +49 3212-1078865
http://suchanek.name

Martin Theobald is a Senior Researcher at the Max-Planck Institute for Informatics. He obtained a doctoral degree in computer science from Saarland University in 2006, and spent two years as a post-doc at Stanford University where he worked on the Trio probabilistic database system. Martin received an ACM SIGMOD dissertation award honorable mention in 2006 for his work on the TopX search engine for efficient ranked retrieval of semistructured XML data. Martin has been teaching courses and seminars in the area of databases and information systems at Stanford and Saarland University. He has published at top-level conferences and journals, including SIGMOD, SIGIR, VLDB, ICDE, PKDD and the VLDB Journal, and he presented two recent tutorials on information extraction at PODS 2010 and CIKM 2010.

Martin Theobald
Max-Planck-Institut für Informatik,
Room 407
Campus E1 4
66123 Saarbrücken
Germany

martin.theobald@mpi-inf.mpg.de
phone: +49 681 9325 507
Fax: +49 681 9325 599
http://www.mpi-inf.mpg.de/∼mtb/

Gerhard Weikum Gerhard Weikum is a Scientific Director at the Max-Planck Institute for Informatics, where he is leading the research group on databases and information systems. Earlier he held positions at Saarland University in Germany, ETH Zurich in Switzerland, MCC in Austin, and he was a visiting senior researcher at Microsoft Research in Redmond. His recent working areas include peer-to-peer information systems, the integration of database-systems and information-retrieval methods, and information extraction for building and maintaining large-scale knowledge bases. Gerhard has co-authored more than 150 publications, including a comprehensive textbook on transactional concurrency control and recovery. He received the VLDB 2002 ten-year award for his work on self-tuning databases, and he is an ACM Fellow. He is a member of the German Academy of Science and Engineering and a member of the German Council of Science and Humanities. Gerhard has served on the editorial boards of various journals including ACM TODS and the new CACM, and as program committee chair for conferences like ICDE 2000, SIGMOD 2004, CIDR 2007, and ICDE 2010. From 2004 to 2009 he was president of the VLDB Endowment.

Gerhard Weikum has co-authored a number of papers on knowledge management and information extraction, which appeared in conferences such as: WWW 2007, ICDE 08, WWW 2009, CIKM 2009, WSDM 2010, PODS 2010, SIGMOD 2010, ACL 2010, CIKM 2010, and more. He has given an invited keynote on knowledge harvesting at the ACM WSDM 2009 Conference[5]; and he was an invited speaker at the International Workshop on Automated Knowledge Base Construction[6].

Gerhard Weikum has been a university professor since 1990. He has given major courses at ETH Zurich and Saarland University, across a variety of topics including: information retrieval and data mining, database

[5] http://wsdm2009.org/
[6] http://akbc.xrce.xerox.com/

systems, performance analysis, introduction to theoretical computer science, and more. He has co-authored a 900-pages textbook on transactional information systems. He has given tutorials at conferences like SIGMOD, VLDB, and ICDE and at EDBT summer schools, on topics such as foundations of automated database tuning, semantic information retrieval, and others. Recently, he has presented an invited tutorial (1 hour) at the ACM PODS 2010 conference on knowledge harvesting [2].

Gerhard Weikum
Max-Planck-Institut für Informatik,
Room 400
Campus E1 4
66123 Saarbrücken
Germany

weikum@mpi-inf.mpg.de
phone: +49 681 9325 500
Fax: +49 681 9325 599
http://www.mpi-inf.mpg.de/∼weikum

Hady W. Lauw is a researcher at the Institute for Infocomm Research in Singapore. He is also an adjunct assistant professor at the School of Information Systems, Singapore Management University. Previously, he was a postdoctoral researcher at Microsoft Research Silicon Valley, working on mining user-generated content and social networks to improve search. He earned a doctorate degree in computer science at Nanyang Technological University in 2008 on a A*STAR graduate fellowship. He has published papers in top-tier conferences (KDD, ICDE, SDM) and journals (TKDE, TOIS), as well as a tutorial at CIKM'10 [5].

Hady Lauw
1 Fusionopolis Way
#21-01 Connexis (South Tower)
Singapore 138632

hady@hadylauw.com
phone: (+65) 6408-2304
http://www.hadylauw.com/

Ralf Schenkel is a research group leader at Saarland University and an associated senior researcher at the Max-Planck Institute for Informatics. The focus of his work has been on efficient retrieval algorithms for text and XML data, graph indexing, and search in social networks. Within the context of the WisNetGrid project, he is coordinating the efforts on knowledge extraction and knowledge-based search in D-Grid, the German Grid infrastructure.

Ralf Schenkel
Campus E1.7, Room 3.07
Universität des Saarlandes
66123 Saarbrücken
Germany

schenkel@mmci.uni-saarland.de
phone: +49 681 302 - 70 798
Fax: +49 681 302 - 70 155
http://people.mmci.uni-saarland.de/∼schenkel/

References

1. A. Doan et al. (Eds.): Special Issue on Managing Information Extraction
   In *ACM SIGMOD Record 37(4)*, December 2008
2. G. Weikum, M. Theobald: From information to knowledge: Harvesting entities and relationships from Web sources
   Tutorial at *PODS 2010*
3. International Workshop on Automated Knowledge Base Construction
   http://akbc.xrce.xerox.com/
4. Linked Data
   http://linkeddata.org/
5. H. Lauw, R. Schenkel, F. Suchanek, M. Theobald, G. Weikum: Harvesting Knowledge from Web Data and Text
   Tutorial at *CIKM 2010*