

Harvesting, Searching, and Ranking Knowledge on the Web

[Invited Talk]

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

ABSTRACT

There are major trends to advance the functionality of search engines to a more expressive semantic level (e.g., [2, 4, 6, 7, 8, 9, 13, 14, 18]). This is enabled by employing large-scale information extraction [1, 11, 20] of entities and relationships from semistructured as well as natural-language Web sources. In addition, harnessing Semantic-Web-style ontologies [22] and reaching into Deep-Web sources [16] can contribute towards a grand vision of turning the Web into a comprehensive knowledge base that can be efficiently searched with high precision.

This talk presents ongoing research towards this objective, with emphasis on our work on the YAGO knowledge base [23, 24] and the NAGA search engine [14] but also covering related projects. YAGO is a large collection of entities and relational facts that are harvested from Wikipedia and WordNet with high accuracy and reconciled into a consistent RDF-style “semantic” graph. For further growing YAGO from Web sources while retaining its high quality, pattern-based extraction is combined with logic-based consistency checking in a unified framework [25]. NAGA provides graph-template-based search over this data, with powerful ranking capabilities based on a statistical language model for graphs. Advanced queries and the need for ranking approximate matches pose efficiency and scalability challenges that are addressed by algorithmic and indexing techniques [15, 17].

YAGO is publicly available and has been imported into various other knowledge-management projects including DBpedia. YAGO shares many of its goals and methodologies with parallel projects along related lines. These include Avatar [19], Cimple/DBlife [10, 21], DBpedia [3], Know-ItAll/TextRunner [12, 5], Kylin/KOG [26, 27], and the Libra technology [18, 28] (and more). Together they form an exciting trend towards providing comprehensive knowledge bases with semantic search capabilities.

Categories and Subject Descriptors

H.1. [Information Systems]: Models and Principles

General Terms

Algorithms, Design

Keywords

Information extraction, knowledge management, information retrieval, scalability

Biography

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. Weikum has co-authored more than 150 publications, including a comprehensive textbook on transactional concurrency control and recovery. He received several best paper awards including the VLDB 2002 ten-year award, and he is an ACM Fellow. He has served on the editorial boards of various journals and book series, including ACM TODS, the Springer LNCS series, and the new CACM, and as program committee chair for international conferences like ICDE 2000, ACM SIGMOD 2004, and CIDR 2007. He is currently the president of the VLDB Endowment.

1. REFERENCES

- [1] Eugene Agichtein: Scaling Information Extraction to Large Document Collections. IEEE Data Eng. Bull. 28(4), 2005
- [2] Kemafor Anyanwu, Angela Maduko, Amit P. Sheth: SemRank: Ranking Complex Relationship Search Results on the Semantic Web. WWW 2005
- [3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary G. Ives: DBpedia: A Nucleus for a Web of Open Data. ISWC/ASWC 2007
- [4] Ricardo A. Baeza-Yates, Massimiliano Ciaramita, Peter Mika, Hugo Zaragoza: Towards Semantic Search. NLDB 2008

- [5] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, Oren Etzioni: Open Information Extraction from the Web. IJCAI 2007
- [6] Holger Bast, Alexandru Chitea, Fabian M. Suchanek, Ingmar Weber: ESTER: Efficient search on Text, Entities, and Relations. SIGIR 2007
- [7] Michael J. Cafarella: Extracting and Querying a Comprehensive Web Database. CIDR 2009
- [8] Soumen Chakrabarti: Breaking Through the Syntax Barrier: Searching with Entities and Relations. ECML 2004
- [9] Tao Cheng, Xifeng Yan, Kevin Chen-Chuan Chang: EntityRank: Searching Entities Directly and Holistically. VLDB 2007
- [10] Pedro DeRose, Warren Shen, Fei Chen, AnHai Doan, Raghu Ramakrishnan: Building Structured Web Community Portals: A Top-Down, Compositional, and Incremental Approach. VLDB 2007
- [11] AnHai Doan, Luis Gravano, Raghu Ramakrishnan, Shivakumar Vaithyanathan (Editors): Special Issue on Information Extraction, SIGMOD Record 37(4), December 2008
- [12] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates: Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artif. Intell. 165(1), 2005
- [13] Jens Graupmann, Ralf Schenkel, Gerhard Weikum: The SphereSearch Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents. VLDB 2005
- [14] Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Maya Ramanath, Gerhard Weikum: NAGA: Searching and Ranking Knowledge. ICDE 2008
- [15] Gjergji Kasneci, Maya Ramanath, Mauro Sozio, Fabian M. Suchanek, Gerhard Weikum: STAR: Steiner Tree Approximation in Relationship-Graphs. ICDE 2009
- [16] Jayant Madhavan, Loredana Afanasiev, Lyublena Antova, Alon Halevy: Harnessing the Deep Web: Present and Future. CIDR 2009
- [17] Thomas Neumann, Gerhard Weikum. RDF-3X: a RISC-style Engine for RDF. PVLDB 1(1), 2008
- [18] Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, Wei-Ying Ma: Web Object Retrieval. WWW 2007
- [19] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, Shivakumar Vaithyanathan: An Algebraic Approach to Rule-Based Information Extraction. ICDE 2008
- [20] Sunita Sarawagi: Information Extraction. Foundations and Trends in Databases 2(1), 2008
- [21] Warren Shen, AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan: Declarative Information Extraction Using Datalog with Embedded Extraction Predicates. VLDB 2007
- [22] Steffen Staab, Rudi Studer: Handbook on Ontologies, 2nd Edition. Springer 2008
- [23] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum: YAGO: a Core of Semantic Knowledge. WWW 2007
- [24] Fabian Suchanek, Gjergji Kasneci, Gerhard Weikum: YAGO: A Large Ontology from Wikipedia and WordNet. Journal of Web Semantics 6(39), 2008
- [25] Fabian Suchanek, Mauro Sozio, Gerhard Weikum: SOFIE: a Self-Organizing Framework for Information Extraction. Technical Report MPI-I-2008-5-004, 2008
- [26] Fei Wu, Daniel S. Weld: Autonomously Semantifying Wikipedia. CIKM 2007
- [27] Fei Wu, Daniel S. Weld: Automatically Refining the wikipedia Infobox Ontology. WWW 2008
- [28] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, Wei-Ying Ma: Simultaneous Record Detection and Attribute Labeling in Web Data Extraction. KDD 2006