Harvesting Knowledge from Web Data and Text

Hady Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, Gerhard Weikum

Harvesting Knowledge from Web Data and Text

Tutorial Proposal for CIKM 2010 (1/2 Day)

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

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

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

1 Overview and Motivation (to be used as abstract)

The Web bears the potential of being the world's greatest encyclopedic source, but we are far from fully exploiting this potential. Valuable scientific and cultural content is interspersed with a huge amount of noisy, low-quality, unstructured text and media. The proliferation of knowledge-sharing communities like Wikipedia and the advances in automated information extraction from Web pages give rise to an unprecedented opportunity: *Can we systematically harvest facts from the Web and compile them into a comprehensive machine-readable knowledge base?* Such a knowledge base would contain not only the world's entities, but also their semantic properties, and their relationships with each other. Imagine a "Structured Wikipedia" that has the same scale and richness as Wikipedia itself, but offers a precise and concise representation of knowledge, e.g., in the RDF format. This would enable expressive and highly precise querying, e.g., in the SPARQL language (or appropriate extensions), with additional capabilities for informative ranking of query results.

The benefits from solving the above challenge would be enormous. Potential applications include

1) a formalized *machine-readable encyclopedia* that can be queried with high precision like a semantic database;
2) a key asset for *disambiguating entities* by supporting fast and accurate mappings of textual phrases onto named entities in the knowledge base;
3) an enabler for entity-relationship-oriented *semantic search* on the Web, for detecting entities and relations in Web pages and reasoning about them in expressive (probabilistic) logics;
4) a backbone for *natural-language question answering* that would aid in dealing with entities and their relationships in answering who/where/when/ etc. questions;
5) a key asset for *machine translation* (e.g., English to German) and interpretation of spoken dialogs, where world knowledge provides essential context for disambiguation;
6) a *catalyst for acquisition of further knowledge* and largely automated maintenance and growth of the knowledge base.

While these application areas cover a broad, partly AI-flavored ground, the most notable one from a database perspective is semantic search: finally bringing DB methodology to Web search! For example, users (or tools on behalf of users) would be able to formulate queries about succulents that grow both in Africa and America, politicians who are also scientists or are married to singers, or flu medication that can be taken by people with high blood pressure. The search engine would return precise and concise answers: lists of entities or entity pairs (depending on the question structure), for example, Angela Merkel, Benjamin Franklin, etc., or Nicolas Sarkozy for the questions about scientists. This would be a quantum leap over today's search where answers are embedded if not buried in lots of result pages, and the human users would have to read them to extract entities and connect them to other entities. In this sense, the envisioned large-scale *knowledge harvesting* [42] from Web sources may also be viewed as *machine reading* [13].

2 Target Audience, Aims, and Organization of the Tutorial

The tutorial is aimed towards a broad audience of researchers from the DB, IR, and KM communities, especially those interested in 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.

The tutorial is organized into three main parts, with ample opportunity for questions and discussion:

- Part 1 (30-45 minutes) 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 (60-90 minutes) gives an overview of different methodological paradigms and concrete techniques for automatically constructing such knowledge bases from Web sources and maintaining them with high quality;
- Part 3 (30-45 minutes) discusses querying knowledge bases for entity-relationship-oriented facts and ranking the results in a principled, informative manner.

The tutorial uses material from an invited tutorial presented at PODS 2010 [43]. However, the PODS tutorial was limited to a 60-minute session. The proposed tutorial deepens the PODS tutorial by going into greater details on the construction of knowledge bases, and broadens it by adding new materials on applications of the knowledge bases, as well as on searching for knowledge and ranking the answers.

3 Content of the Tutorial

3.1 Current Large-Scale Knowledge Bases

This part of the tutorial introduces the audience to large-scale knowledge bases through an overview of several major publicly-available knowledge bases. Universal knowledge bases have been an AI objective since the pioneering work on *Cyc* [27] and the early enthusiasm about the original roadmap for the Semantic Web [38]. However, there are many more favorable assets available today, as evidenced by the fair number of ongoing endeavors, both in academia and industrial research. These include *freebase.com*, *trueknowledge.com*, DBpedia [4], KnowItAll [21], TextRunner [45], Kylin/KOG [44], Omnivore [10], ReadTheWeb [13], the *sig.ma* engine [41], as well as our own YAGO project [39]. In addition to general knowledge bases, further services are being pursued for specific communities, such as DBLife [17] for database research, or MedIE [28] for the biomedical domain.

Available Resources. Endeavors to build large knowledge bases generally exploit three different kinds of resources, namely: *manual* human efforts, knowledge-sharing communities such as *Wikipedia*, or the entire *Web*. The pioneering work *Cyc* is an example of a knowledge base constructed by hand. Cyc has been followed by the SUMO ontology [30] and WordNet [22]. Most knowledge bases build upon Wikipedia, which contains a wealth of semi-structured information, such as a category system, attribute-value pairs contained in the infoboxes, inter-language links, and change logs. Knowledge bases such as DBpedia, YAGO, TrueKnowledge, and Freebase have relied on Wikipedia as a main resource. The Linking Open Data project [7] aims to interconnect many of them.Other projects aim more broadly at gathering and structuring information from the entire Web. The *sig.ma* engine, for example, taps on "triplified" RDF data on the Web.

Knowledge Representation Model. Most current knowledge bases use the RDF framework proposed by the W3C to represent their data. An RDF knowledge base can be seen as a graph, in which the nodes are *entities* (such as persons, companies, cities, products, etc.), and the edges are *relationships* (such as "livesIn", "hasPopulation" or "hasCEO"). Some nodes represent *classes*, which are arranged in a partial order, ranging from broad (e.g., "person") to more specific (e.g., "German politician"). RDF also provides a scheme of worldwide unique identifiers for the entities, thus enabling cross-ontology linking.

Applications. We review three groups of applications, which—though by no means being exhaustive—give the audience a flavor of how the knowledge is employed in various scenarios. The first group of applications make use of the *linguistic* component of the knowledge. For instance, the data can serve as a lexicon (distinguishing English from non-English words) or thesaurus (detecting semantic similarity of terms), which have found uses in sentiment analysis and document classification. The second group make use of the knowledge about *entities and relationships* for tasks such as word sense disambiguation [9], information extraction [19], and logical reasoning. The last group treat the knowledge bases as a form of *semantic database*, and surface relevant information queried from the database for human consumption. This can either happen implicitly, e.g., map annotations [5], or explicitly, through semantic search, as discussed later in this tutorial.

3.2 Extracting Knowledge

Extracting large-scale knowledge bases from Web data can be split into three major tasks: 1) detecting and disambiguating entities in their given context, 2) detecting binary relationships among entities (e.g., RDF-style *facts*), and 3) filtering and detecting inconsistencies among facts which may involve also reasoning about

higher-arity relationships (e.g., when combining binary facts also with additional properties such as time annotations). The boundaries between these tasks are not necessarily static. Several recent machine learning approaches pursue joint inferencing approaches which are able to combine tasks such as record segmentation and entity resolution (i.e., disambiguation), which had previously been studied only in isolation [37, 35].

Entities and classes. In the first level of knowledge harvesting, we are interested in collecting as many individual entities – persons, companies, cities, products, etc. – as possible and organizing them into semantic classes (types) such as artists, scientists, molecular biologists, singers, guitar players, movies, etc. A key asset for this is WordNet [22], a hand-crafted collection of more than 100,000 semantic classes along with fairly comprehensive subclass/superclass relations.

Wikipedia. The English version of Wikipedia contains more than 3 million articles, most of which correspond one-to-one to individual entities. A major breakthrough in extracting entities and classes from Wikipedia are YAGO (Yet Another Great Ontology) [39], the parallel and independent work on WikiTaxonomy [32], and the follow-up work on KOG (Kylin Ontology Generator) [44]. YAGO, for example, initializes its class system by importing all WordNet classes and their hyponymy/hypernymy (subclass/superclass) relations. Yago constitutes a major part of DBpedia [4].

Online dictionaries. Of course, Wikipedia is not complete.For many domains, explicit dictionaries are available, e.g., *imdb.com* for movies, *librarything.com* for books, and many more. Further sources for mining taxonomic relations include "social tags" from online communities such as *del.icio.us*, *citeulike.org*, or *flickr.com*, and also tags assigned to blog postings and news articles. Informally compiled directories such as *dmoz.org* are potentially valuable, although their directory structure is not based on taxonomic relations.

Entity disambiguation. When mapping a category to a WordNet class based on a clever name-matching method, a thorny issue that we have so far disregarded is the ambiguity of names. For example, should we map the category "drivers" to the WordNet sense "driver: the operator of a motor vehicle" or to the sense "driver: device driver for a computer" of this polysemous word? This raises the general issue of *entity resolution*, also known as entity reconciliation or record linkage. Given a string, perhaps with a textual context, or a record with a few fields, what is the most likely target for mapping the string onto an individual entity or semantic class?

Pattern-based fact extraction. Seminal work by Brin [8] was centered around the following observation on the *duality of facts and patterns*: if we knew enough facts for a relation (e.g., instances of married couples) we could automatically find textual patterns and distill the best ones, and if we knew good patterns, we could automatically find more facts. This iterative process between fact and pattern harvesting is powerful but difficult to tune (regarding thresholds, weighting parameters, etc.) and susceptible to drifting away from its target. A series of improvements led to a variety of projects, most notably, Snowball [1], KnowItAll [21], Text2Onto [16], and TextRunner [45].

Wrappers and wrapper induction. Dynamic Web pages often are generated from a database-backed content management system, which makes it possible to construct or automatically infer *wrappers* for fact extraction from HTML headings, tables, lists, form fields, and other semistructured elements. To this end, powerful languages for extraction scripts have been developed, and methods for learning structure from examples have been successfully applied (see, e.g., [26]). The latter is also known as *wrapper induction*. Some approaches employed ML techniques like Conditional Random Fields (CRFs), Hidden Markov Models (HMMs) or classifiers [34], but the general rationale has been to arrive at a set of good extraction rules that could be applied in a deterministic manner. Systems of this kind include Rapier [12], the W4F toolkit [33], and Lixto [14].

Declarative extraction. More recent work on rule-based fact gathering is based on DB-style *declarative* IE (see [18] for several overview articles). SystemT [25] has developed a declarative language, coined AQL, for fact extraction tasks, along with an algebra and query rewriting rules. A very nice showcase is the (at least largely) automated construction and maintenance of the DBLife community portal (*dblife.cs.wisc.edu*), which is based on the Cimple tool suite [17].

Statistical Relational Learning. The field of statistical relational learning (SRL) has gained strong interest in both the AI and DB communities. Within the SRL family, Markov Logic Networks (MLN) are probably the most versatile approach in combining first-order logic rules and probabilistic graphical models. Kylin/KOG [44] is an interesting application of MLNs and a suite of other learning techniques, which aims to infer also "missing infobox values" in Wikipedia. In the ReadTheWeb project [13], semi-supervised learning ensembles have been combined with constraints for extracting entities and facts from a huge Web corpus. StatSnowball [47] is another powerful machinery that makes intensive use of MLNs and other machine learning techniques.

Leveraging existing knowledge. Parallel work that has found new ways of combining pattern-based harvesting with consistency reasoning is the SOFIE methodology [40], which was developed to enable automatic growth of YAGO while retaining the high level of near-human quality. SOFIE maps all ingredients – known facts from the knowledge base, new fact hypotheses, patterns, constraints, and possible entity disambiguations

– into a set of weighted clauses, where the weights are derived from the automatically gathered statistical evidence. SOFIE exploits YAGO and its rigorous entity typing for both accuracy and efficiency.

Open information extraction. Recent work by [11] addresses the goal of *open information extraction* (Open IE). The TextRunner system [45] aims at extracting all meaningful relations from Web pages (coined "*assertions*"), rather than a predefined set of canonical relations. Here, entities are not yet necessarily disambiguated , and relations are not canonicalized: all verbal phrases found in natural language sentences may constitute a valid relation type between two or more entities.

Higher-arity relations and temporal reasoning. So far we have simplified our knowledge-harvesting setting by assuming that facts are time-invariant. This is appropriate for some relation types (e.g., birth dates), but inappropriate for evolving facts. Extracting the validity time of facts involves detecting explicit temporal expressions such as dates as well as implicit expressions in the form of adverbial phrases such as "last Monday", "next week", or "years ago", which entails difficult issues also for reasoning about interrelated time points or intervals. Initial work on these issues includes [2, 46].

3.3 Querying and Ranking

With this new availability of large-scale semantic knowledge, a new brand of semantic-search and knowledge-discovery engines have begun to emerge. Representatives include Wolfram Alpha which computes knowledge answers from a set of hand-crafted databases, Google Squared which arranges search results in a tabular form with entities and attributes, EntityCube which provides dynamically gathered facts about named entities, *opencalais.com* which provides services to superimpose structure on documents or Web pages, or *kosmix.com* which uses a large ontology for categorizing questions and identifying entities that are related to the user's input.

Moreover, whenever queries return many results, we need ranking. For example, a query about *politicians who are also scientists* can easily yield hundreds of persons. Even the more specific query about *French politicians who are married to singers* may overwhelm the users with possible answers. A meaningful ranking should consider the two fundamental dimensions *informativeness* and *confidence*.

Informativeness. Users prefer prominent entities and salient facts as answers. For example, the first query above should return politicians such as Benjamin Franklin (who made scientific discoveries), Paul Wolfowitz (a mathematician by training), or the German chancellor Angela Merkel (who has a doctoral degree in physical chemistry). The second query should prefer an answer like Nicolas Sarkozy over the mayor of a small provincial town. This ranking criterion calls for appropriate statistical models about entities and relationships.

Confidence. We need to consider the strength or certainty in believing that the result facts are indeed correct. This is largely determined at the time when facts are harvested and placed in the knowledge base, and it can be based on aggregating different sub-criteria. First, the extraction methods can assign an *accuracy* weight to each fact that based on the empirically assessed goodness of the extractor and the extraction target.and the total number of *witnesses* for the given fact.Second, the *provenance* of the facts should be assessed by considering the *authenticity and authority* of the sources from which facts are derived. PageRank-style link-graph-based models come to mind for authority ranking, but more advanced models of *trustworthiness* and entity-oriented rather than page-oriented importance are needed.

Entity-based ranking. State-of-the-art ranking models in IR are based on *statistical language models*, LM's for short. Recently, extended LMs have been developed for entity ranking in the context of expert finding in enterprises and Wikipedia-based retrieval and recommendation tasks [29, 31, 36]. For ranking entity-search results e to a keyword query q, one needs to compute $P[q|e]$. As an entity cannot be directly compared to query words, one considers the words in a Web page d that occur in a proximity window around the position from which e was extracted. Additional sophistication is needed for considering also an entity's attributes [29]. An alternative paradigm for entity ranking is to generalize PageRank-style link-analysis methods (see, e.g., [15, 23]) to graphs that connect entities rather than Web pages. This line of models is useful, but appears to be more of an ad-hoc flavor compared to the principled LM approaches.

Fact-based ranking. The models discussed above are limited to entities – the nodes in an entity-relationship graph. In contrast, general knowledge search needs to consider also the role of relations – the edges in the graph – for answering more expressive classes of queries. Ranking for structured queries has been intensively investigated for XML [3] and in the context of keyword search on relational graphs [6]. However, these approaches do not carry over to graph-structured, largely schema-less RDF data collections. What we need for RDF knowledge ranking is a generalization of entity LM's that considers relationships (RDF properties) as first-class citizens. Recent work on the NAGA search engine [24] has addressed these issues and developed a full-fledged LM for ranking the results of extended SPARQL queries [20]. Efficiently evaluating the LM-based scores at query run-time in order to return the top-k best answers however still is an open issue.

4 About the Speakers

Hady W. Lauw is a researcher at Institute for Infocomm Research in Singapore. 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.

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.

Fabian Suchanek was a visiting researcher at Microsoft Research Silicon Valley and is now a postdoc at INRIA Saclay in Paris. 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 250 citations, and YAGO is used in many major knowledge-base projects around the world (including DBpedia).

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.

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 300 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.

References

1. E. Agichtein, L. Gravano, J. Pavel, V. Sokolova, and A. Voskoboynik. Snowball: a prototype system for extracting relations from large text collections. In *SIGMOD*, 2001.
2. O. Alonso, M. Gertz, and R. A. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *CIKM*, 2009.
3. S. Amer-Yahia and M. Lalmas. XML search: languages, INEX and scoring. *SIGMOD Rec.*, 35(4), 2006.
4. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *ISWC*, 2007.
5. C. Becker and C. Bizer. DBpedia mobile: A location-enabled linked data browser. In *Linking Open Data Workshop*, 2008.
6. G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword Searching and Browsing in Databases using BANKS. In *ICDE*, 2002.

7. C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the Web. In *WWW*, 2008.
8. S. Brin. Extracting patterns and relations from the world wide web. In *WebDB*, 1998.
9. R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, 2006.
10. M. J. Cafarella. Extracting and querying a comprehensive web database. In *CIDR*, 2009.
11. M. J. Cafarella, A. Y. Halevy, and N. Khoussainova. Data integration for the relational web. *PVLDB*, 2(1), 2009.
12. M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In *AAAI/IAAI*, 1999.
13. A. Carlson, J. Betteridge, R. C. Wang, E. R. H. Jr., and T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *WSDM*, 2010.
14. J. Carme, M. Ceresna, O. Frölich, G. Gottlob, T. Hassan, M. Herzog, W. Holzinger, and B. Krüpl. The Lixto project: Exploring new frontiers of web data extraction. In *BNCOD*, 2006.
15. S. Chakrabarti. Dynamic personalized PageRank in entity-relation graphs. In *WWW*, 2007.
16. P. Cimiano and J. Völker. Text2Onto - a framework for ontology learning and data-driven change discovery. In *NLDB*, 2005.
17. P. DeRose, W. Shen, F. C. 0002, Y. Lee, D. Burdick, A. Doan, and R. Ramakrishnan. DBLife: A community information management platform for the database research community. In *CIDR*, 2007.
18. A. Doan, L. Gravano, R. Ramakrishnan, and S. V. (Editors). Special issue on information extraction. *SIGMOD Record*, 37(4), 2008.
19. A. Doan, N. F. Noy, and A. Y. Halevy. Special issue on semantic integration. *SIGMOD Record*, 33(4), 2004.
20. S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum. Language-model-based ranking for queries on RDF-graphs. In *CIKM*, 2009.
21. O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in KnowItAll. In *WWW*, 2004.
22. C. Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, 1998.
23. V. Hristidis, H. Hwang, and Y. Papakonstantinou. Authority-based keyword search in databases. *ACM Trans. Database Syst.*, 33(1), 2008.
24. G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *ICDE*, 2008.
25. R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, and H. Zhu. SystemT: a system for declarative information extraction. *SIGMOD Record*, 37(4), 2008.
26. N. Kushmerick. Wrapper induction: efficiency and expressiveness. *Artif. Intell.*, 118(1-2), 2000.
27. D. B. Lenat. CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11), 1995.
28. Y. Miyao, T. Ohta, K. Masuda, Y. Tsuruoka, K. Yoshida, T. Ninomiya, and J. Tsujii. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *COLING-ACL*, 2006.
29. Z. Nie, Y. Ma, S. Shi, J.-R. Wen, and W.-Y. Ma. Web object retrieval. In *WWW*, 2007.
30. I. Niles and A. Pease. Towards a standard upper ontology. In *Formal Ontology in Information Systems*, 2001.
31. D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *ICTAI*, 2006.
32. S. P. Ponzetto and M. Strube. WikiTaxonomy: A large scale knowledge resource. In *ECAI*, 2008.
33. A. Sahuguet and F. Azavant. Building intelligent web applications using lightweight wrappers. *Data Knowl. Eng.*, 36(3), 2001.
34. S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3), 2008.
35. S. Sarawagi and W. W. Cohen. Semi-Markov conditional random fields for information extraction. In *NIPS*, 2004.
36. P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR*, 2008.
37. P. Singla and P. Domingos. Entity resolution with Markov Logic. In *ICDM*, 2006.
38. S. Staab and R. Studer, editors. *Handbook on Ontologies (2nd Edition)*. Springer, 2009.
39. F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A large ontology from Wikipedia and WordNet. *J. Web Sem.*, 6(3), 2008.
40. F. M. Suchanek, M. Sozio, and G. Weikum. SOFIE: a self-organizing framework for information extraction. In *WWW*, 2009.
41. G. Tummarello. SIG.MA: Live views on the web of data. In *WWW*, 2010.
42. G. Weikum. Harvesting, searching, and ranking knowledge on the web. In *WSDM*, 2009.
43. G. Weikum and M. Theobald. From information to knowledge: harvesting entities and relationships from web sources. In *PODS*, 2010.
44. F. Wu and D. S. Weld. Automatically refining the Wikipedia infobox ontology. In *WWW*, 2008.
45. A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, and S. Soderland. TextRunner: Open information extraction on the web. In *HLT-NAACL*, 2007.
46. Q. Zhang, F. M. Suchanek, L. Yue, and G. Weikum. TOB: Timely ontologies for business relations. In *WebDB*, 2008.
47. J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen. StatSnowball: a statistical approach to extracting entity relationships. In *WWW*, 2009.