Knowledge Harvesting in the Big-Data Era

Fabian Suchanek
Max Planck Institute for Informatics
D-66123 Saarbruecken, Germany
suchanek@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
D-66123 Saarbruecken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

The proliferation of knowledge-sharing communities such as Wikipedia and the progress in scalable information extraction from Web and text sources have enabled the automatic construction of very large knowledge bases. Endeavors of this kind include projects such as DBpedia, Freebase, KnowItAll, ReadTheWeb, and YAGO. These projects provide automatically constructed knowledge bases of facts about named entities, their semantic classes, and their mutual relationships. They 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, semantic search for entities and relations in Web and enterprise data, and entity-oriented analytics over unstructured contents. Prominent examples of how knowledge bases can be harnessed include the Google Knowledge Graph 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.

Categories and Subject Descriptors

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

Keywords

Big Data, Information Extraction, Knowledge Base, Ontology, Entity Recognition, Web Contents

1. MOTIVATION AND OVERVIEW

1.1 Knowledge Bases

Knowledge harvesting from Web and text sources has become a major research avenue in the last five years. It is the core methodology for the automatic construction of large knowledge bases [2, 3, 51], going beyond manually compiled knowledge collections like Cyc [63], WordNet [33], and a variety of ontologies [102]. Salient projects with publicly available resources include KnowItAll [30, 6, 31], ConceptNet [100], DBpedia [5], Freebase [11], NELL [16], WikiTaxonomy [89], and YAGO [103, 46, 8]. Commercial interest has been strongly growing, with evidence by projects like the Google Knowledge Graph, the EntityCube/Renlifang project at Microsoft Research [84], and the use of public knowledge bases for type coercion in IBM's Watson project [52].

These knowledge bases contain many millions of entities, organized in hundreds to hundred thousands of semantic classes, and hundred millions of relational facts between entities. All this is typically represented in the form of RDF-style subject-predicate-object (SPO) triples. Moreover, knowledge resources can be semantically interlinked via owl:sameAs triples at the entity level, contributing to the Web of Linked Open Data (LOD) [45].

Large knowledge bases are typically built by mining and distilling information from sources like Wikipedia which offer high-quality semi-structured elements (infoboxes, categories, tables, lists), but many projects also tap into extracting knowledge from arbitrary Web pages and natural-language texts. Despite great advances in these regards, there are still many challenges regarding the scale of the methodology and the scope and depth of the harvested knowledge:

- covering more entities beyond Wikipedia and discovering newly emerging entities,
- increasing the number of facts about entities and extracting more interesting relationship types in an open manner,
- capturing the temporal scope of relational facts,
- tapping into multilingual inputs such as Wikipedia editions in many different languages,
- extending fact-oriented knowledge bases with commonsense knowledge and (soft) rules,
- detecting and disambiguating entity mentions in natural-language text and other unstructured contents, and
- large-scale sameAs linkage across many knowledge and data sources.

1.2 Enabling Intelligent Applications

Knowledge bases are a key asset that enables and contributes to intelligent computer behavior. Application areas along these lines include the following:

- Semantic search and question answering: Machine-readable encyclopediae are a rich source of answering expert-level questions in a precise and concise manner. Moreover, interpreting users' information needs in terms of entities and relation-

SIGMOD'13, June 22–27, 2013, New York, New York, USA.

ships yields strong features for informative ranking of search results and entity-level recommendations over Web and enterprise data.

- Deep interpretation of natural language: Both written and spoken language are full of ambiguities. Knowledge is the key to mapping surface phrases to their proper meanings, so that machines interpret language as fluently as humans. As user-generated social-media contents is abundant and human-computer interaction is more and more based on smartphones, coping with text, speech, and gestures will become crucial.
- Machine reading at scale: The deluge of online contents overwhelms users. Users wish to obtain overviews of the salient entities and relationships for a week of news, a month of scientific articles, a year of political speeches, or a century of essays on a specific topic.
- Reasoning and smart assistants: Rich sets of facts and rules from a knowledge base enable computers to perform logical inferences in application contexts.
- Big-Data analytics over uncertain contents: Daily news, social media, scholarly publications, and other Web contents are the raw inputs for analytics to obtain insights on business, politics, health, and more. Knowledge bases are key to discovering and tracking entities and relationships and thus making sense of noisy contents.

1.3 Scope and Structure of the Tutorial

This tutorial gives an overview on knowledge harvesting and discusses hot topics in this field, pointing out research opportunities and open challenges. As the relevant literature is widely dispersed across different communities, we also venture into the neighboring fields of Web Mining, Artificial Intelligence, Natural Language Processing, Semantic Web, and Data Management. The presentation is structured according to the following sections and subsections.

2. KNOWLEDGE BASE CONSTRUCTION

2.1 Knowledge Bases in the Big-Data Era

Many Big-Data applications need to tap unstructured data. News, social media, web sites, and enterprise sources produce huge amounts of valuable contents in the form of text and speech. Key to making sense of this contents is to identify the entities that are referred to and the relationships between entities. This allows linking unstructured contents with structured data, for value-added analytics. Knowledge bases are a key asset for lifting unstructured contents into entity-relationship form and making the connection to structured data. We give an overview of several large and publicly available knowledge bases, and outline how they can support Big-Data applications.

2.2 Harvesting of Entities and Classes

Every entity in a knowledge base (such as `Steve_Jobs`) belongs to one or more 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 discuss two groups of approaches to harvest information on classes and their instances: i) Wikipedia-based approaches and ii) Web-based approaches using set expansion and other techniques. Relevant work in the first group includes [89, 90, 103, 124]. Relevant work in the second group includes [4, 21, 44, 57, 88, 106, 115, 125].

3. HARVESTING FACTS AT WEB SCALE

3.1 Harvesting Relational Facts

Relational facts express relationships between two entities, for example, the following facts about Steve Jobs:
`Steve_Jobs founded Apple_Inc.`,
`Steve_Jobs was_Board_Member_of Walt_Disney_Company`,
`Steve_Jobs died_on 5-Oct-2011`,
`Steve_Jobs died_of Pancreas_Cancer`,
`Steve_Jobs has_Friend Joan_Baez`, and more.

There is a large spectrum of methods to extract such facts from Web data, tapping both semistructured sources like Wikipedia infoboxes, lists, and tables, and natural-language text sources like Wikipedia full-text articles, news and social media. We 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 MLN's), and logical consistency reasoning (e.g., weighted MaxSat or ILP solvers). We also discuss to what extent these approaches scale to handle big data.

Overviews of information extraction methods for knowledge base population are given in [26, 96, 123]. For specific state-of-the-art methods, see the following original papers and references given there: [1, 10, 13, 15, 16, 17, 18, 30, 32, 36, 40, 46, 49, 58, 60, 69, 70, 76, 85, 86, 87, 94, 104, 112, 128]. For foundations of statistical learning methods used in this context, see [27, 39, 55].

3.2 Open Information Extraction

In contrast to approaches that operate on a predefined list of relations and a huge, but fixed set of 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. For example, in addition to capturing the pre-specified hasWonPrize relation, we aim to automatically learn that nominatedForPrize is also an interesting relation expressed by natural-language patterns such as "candidate for ... prize" or "expected to win ... prize". We discuss recent methods that follow this Open IE direction [6, 12, 24, 31, 41, 56, 72, 75, 77, 83, 116, 126]. Some methods along these lines make clever use of Big-Data techniques like frequent sequence mining and map-reduce computation.

3.3 Temporal, Multilingual, Commonsense, and Visual Knowledge

In this part, we venture beyond entity-relationship facts and describe approaches that attach meta-information to facts. This concerns the temporal or spatial context of a fact [38, 61, 68, 107, 108, 113, 117, 118, 119], or describes entities in multiple languages [22, 23, 78, 81]. Along the temporal dimension, we would like to capture the timepoints of events and the timespans during which certain relationships hold, for example:
`Steve_Jobs Chairman_of Apple_Inc. @[1976,1985]`,
`Steve_Jobs CEO_of Apple_Inc. @[Sep-1997,Aug-2011]`,
`Pixar acquired_by Walt_Disney_Company @5-May-2006`.

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 (see, e.g., [37, 62, 71, 99, 109, 114]). For example, snakes can crawl and hiss, but they cannot fly or sing. An example for a (soft) commonsense rule is that the husband of a mother is the father of her child (husband at the time of the child's birth). Here again, state-of-the-art methods use techniques that scale out to handle Big-Data inputs.

Finally, another dimension of knowledge is to associate entities and classes with visual data: images and videos [25, 95, 110, 111].

4. KNOWLEDGE FOR BIG DATA

When analytic tasks tap into text or Web data, it is crucial to identify entities (people, places, products, etc.) in the input for proper grouping and aggregation. 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). Kowledge about entities is an invaluable asset here.

4.1 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"). Entity mentions can be discovered by named-entity recognition (NER) methods, usually based on CRF's [35] or other probabilistic graphical models and/or using dictionary of surface forms [101]. Some methods infer semantic types for mentions, e.g., telling that "the Apple founder" is a person, or in a fine-grained manner, an entrepreneur (see, e.g., [66, 67, 127] and references there).

Nevertheless, entity mentions are just noun phrases and still ambiguous. Mapping mentions 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 [14, 19, 20, 28, 34, 43, 47, 48, 59, 74, 93]. Although these principles are well understood, NED remains an active research area towards improving robustness, scalability, and coverage.

The NED problem also arises in structured but schema-less data like HTML tables in Web pages [65]. NED is a special case of the general word-sense disambiguation problem [80], which considers also general nouns (concepts that are not entities, e.g., rugby or peace), verbal phrases, adjectives, etc. Finally note that NED is not the same as co-reference resolution [91, 97]. The latter aims to find equivalence classes of surface forms (e.g., "Michelle" and the "First Lady of America" are the same entity), but without mapping to an entity catalog.

4.2 Entity Linkage

We see more and more structured data on the Web, in the form of (HTML) tables, microdata embedded in Web pages (using, e.g., the `schema.org` vocabulary), and Linked Open Data. Even when entities are explicitly marked in these kinds of data, 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) [29, 54, 79]. For knowledge bases and Linked Open Data, it is of particular interest because of the need for generating and maintaining owl:sameAs linkage across knowledge resources. We give an overview of approaches to this end, covering statistical learning approaches (e.g., [7, 42, 92, 98]) and graph algorithms (see, e.g., [9, 50, 53, 64, 73, 82, 105, 120, 121, 122] and further references given there).

5. PRESENTERS' BIOGRAPHIES

Fabian M. Suchanek is the leader of the Otto Hahn Research Group "Ontologies" at the Max Planck Institute for Informatics in Germany. He obtained his PhD from Saarland University in 2008, and was a postdoc at Microsoft Research Search Labs in Silicon Valley (in the group of Rakesh Agrawal) and in the WebDam team at INRIA Saclay in France (in the group of Serge Abiteboul). Fabian is the main architect of the YAGO ontology, one of the largest public knowledge bases.

Gerhard Weikum is a Scientific Director at the Max Planck Institute for Informatics in Saarbruecken, Germany, where he is leading the department on databases and information systems. He co-authored a comprehensive textbook on transactional systems, received the VLDB 10-Year Award for his work on automatic DB tuning, and is one of the creators of the YAGO knowledge base. Gerhard is an ACM Fellow, a member of the German Academy of Science and Engineering, and a recipient of a Google Focused Research Award and an ACM SIGMOD Contributions Award.

6. REFERENCES

[1] E. Agichtein, L. Gravano: Snowball: Extracting Relations from Large Plain-Text Collections. ACM DL 2000

[2] AKBC 2010: First Int. Workshop on Automated Knowledge Base Construction, Grenoble, 2010, `http://akbc.xrce.xerox.com/`

[3] AKBC-WEKEX 2012: The Knowledge Extraction Workshop at NAACL-HLT, 2012,`http://akbcwekex2012.wordpress.com/`

[4] E. Alfonseca, M. Pasca, E. Robledo-Arnuncio: Acquisition of Instance Attributes via Labeled and Related Instances. SIGIR 2010

[5] S. Auer, C. Bizer, et al.: DBpedia: A Nucleus for a Web of Open Data. ISWC 2007

[6] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni: Open Information Extraction from the Web. IJCAI 2007

[7] I.Bhattacharya, L. Getoor: Collective Entity Resolution in Relational Data. TKDD 1(1), 2007

[8] J. Biega et al.: Inside YAGO2s: a Transparent Information Extraction Architecture. WWW 2013

[9] C. Böhm et al.: LINDA: Distributed Web-of-Data-Scale Entity Matching. CIKM 2012

[10] P. Bohannon et al.: Automatic Web-Scale Information Extraction. SIGMOD 2012

[11] K.D. Bollacker et al.: Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge. SIGMOD 2008

[12] D. Bollegala, Y. Matsuo, M. Ishizuka: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web. WWW 2010

[13] S. Brin: Extracting Patterns and Relations from the World Wide Web. WebDB 1998

[14] R.C. Bunescu, M. Pasca: Using Encyclopedic Knowledge for Named Entity Disambiguation. EACL 2006

[15] M.J. Cafarella: Extracting and Querying a Comprehensive Web Database. CIDR 2009

[16] A. Carlson et al.: Toward an Architecture for Never-Ending Language Learning. AAAI 2010

[17] L. Chiticariu et al.: SystemT: An Algebraic Approach to Declarative Information Extraction. ACL 2010

[18] P. Cimiano, J. Völker: Text2Onto. NLDB 2005

[19] M. Cornolti, P. Ferragina, M. Ciaramita: A Framework for Benchmarking Entity-Annotation Systems. WWW 2013

[20] S. Cucerzan: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP 2007

[21] B.B. Dalvi, W.W. Cohen, J. Callan: WebSets: Extracting Sets of Entities from the Web using Unsupervised Information Extraction. WSDM 2012

[22] G. de Melo, G. Weikum: Towards a Universal Wordnet by Learning from Combined Evidence. CIKM 2009

[23] G. de Melo, G. Weikum: MENTA: Inducing Multilingual Taxonomies from Wikipedia. CIKM 2010

[24] L. Del Corro, R. Gemulla: ClausIE: Clause-Based Open Informtion Extraction. WWW 2013

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. CVPR 2009

[26] A. Doan et al. (Eds.): Special Issue on Managing Information Extraction, SIGMOD Record 37(4),

[27] P. Domingos, D. Lowd: Markov Logic: An Interface Layer for Artificial Intelligence. Morgan & Claypool 2009

[28] M. Dredze et al.: Entity Disambiguation for Knowledge Base Population. COLING 2010

[29] A.K. Elmagarmid, P.G. Ipeirotis, V.S. Verykios: Duplicate Record Detection: A Survey. IEEE TKDE 19(1), 2007

[30] O. Etzioni et al.: Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artif. Intell. 165(1), 2005

[31] A. Fader. S. Soderland, O. Etzioni: Identifying Relations for Open Information Extraction, EMNLP 2011

[32] Y. Fang, K. Chang: Searching Patterns for Relation Extraction over the Web: Rediscovering the Pattern-Relation Duality. WSDM 2011

[33] C. Fellbaum, G. Miller (Eds.): WordNet: An Electronic Lexical Database, MIT Press, 1998

[34] P. Ferragina, U. Scaiella: TAGME: On-the-Fly Annotation of Short Text Fragments. CIKM 2010

[35] J.R. Finkel, T. Grenager, C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. ACL 2005

[36] T. Furche et al.: DIADEM: Domain-centric, Intelligent, Automated Data Extraction Methodology. WWW 2012

[37] L. Galarraga, C. Teflioudi, K. Hose, F. Suchanek: AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases. WWW 2013

[38] G. Garrido et al.: Temporally Anchored Relation Extraction. ACL 2012

[39] L. Getoor, B. Taskar (Eds.): Introduction to Statistical Relational Learning. MIT Press 2007

[40] G. Gottlob et al.: The Lixto Data Extraction Project - Back and Forth between Theory and Practice. PODS 2004

[41] R. Gupta, S. Sarawagi: Joint Training for Open-Domain Extraction on the Web: Exploiting Overlap when Supervision is Limited. WSDM 2011

[42] R. Hall, C.A. Sutton, A. McCallum: Unsupervised Deduplication using Cross-Field Dependencies. KDD 2008

[43] X. Han, L. Sun, J. Zhao: Collective Entity Linking in Web Text: a Graph-based Method. SIGIR 2011

[44] M.A Hearst: Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992

[45] T. Heath, C. Bizer: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 2011

[46] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum: YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia, Artif. Intell. 194, 2013

[47] J. Hoffart, M. A. Yosef, I. Bordino, et al.: Robust Disambiguation of Named Entities in Text. EMNLP 2011

[48] J. Hoffart et al.: KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. CIKM 2012

[49] R. Hoffmann, C. Zhang, D.S. Weld: Learning 5000 Relational Extractors. ACL 2010

[50] A.Hogan et al.: Scalable and Distributed Methods for Entity Matching. J. Web Sem. 10, 2012

[51] E. Hovy, R. Navigli, S.P. Ponzetto: Collaboratively Built Semi-Structured Content and Artificial Intelligence: the Story So Far, Artif. Intell. 194, 2013

[52] IBM Journal of Research and Development 56(3/4), Special Issue on "This is Watson", 2012

[53] H. Köpcke et al.: Evaluation of Entity Resolution Approaches on Real-World Match Problems. PVLDB 2010

[54] H. Köpcke, E. Rahm: Frameworks for entity matching: A comparison. Data Knowl. Eng. 69(2), 2010

[55] D. Koller, N. Friedman: Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009

[56] S.K. Kondreddi, P. Triantafillou, G. Weikum: HIGGINS: Knowledge Acquisition meets the Crowds. WWW 2013

[57] Z. Kozareva, E.H. Hovy: A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web. EMNLP 2010

[58] S. Krause, H. Li, H. Uszkoreit, F. Xu: Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web. ISWC 2012

[59] S. Kulkarni et al.: Collective Annotation of Wikipedia Entities in Web Text. KDD 2009

[60] N. Kushmerick, D.S. Weld, R.B. Doorenbos: Wrapper Induction for Information Extraction. IJCAI 1997

[61] E. Kuzey, G. Weikum: Extraction of temporal facts and events from Wikipedia. TempWeb 2012

[62] N. Lao, T.M. Mitchell, W.W. Cohen: Random Walk Inference and Learning in A Large Scale Knowledge Base. EMNLP 2011

[63] D.B. Lenat: CYC: A Large-Scale Investment in Knowledge Infrastructure. CACM 38(11), 1995

[64] J.Li, J.Tang, Y.Li, Q.Luo: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. TKDE 21(8), 2009

[65] G.Limaye et al: Annotating and Searching Web Tables Using Entities, Types and Relationships. PVLDB 2010

[66] T. Lin et al.: No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. EMNLP 2012

[67] X. Ling, D.S. Weld: Fine-Grained Entity Recognition. AAAI 2012

[68] X. Ling, D.S. Weld: Temporal Information Extraction. AAAI 2010

[69] A. Machanavajjhala et al.: Collective extraction from heterogeneous web lists. WSDM 2011

[70] B. Marthi, B. Milch, S. Russell, First-Order Probabilistic Models for Information Extraction. IJCAI 2003

[71] C. Matuszek et al.: Searching for Common Sense: Populating Cyc from the Web. AAAI 2005

[72] Mausam, M. Schmitz, S. Soderland, et al.: Open Language Learning for Information Extraction. EMNLP 2012

[73] S. Melnik, H. Garcia-Molina, E. Rahm: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. ICDE 2002

[74] D.N. Milne, I.H. Witten: Learning to link with wikipedia. CIKM 2008

[75] T. Mohamed, E.R. Hruschka, T.M. Mitchell: Discovering Relations between Noun Categories. EMNLP 2011

[76] N. Nakashole, M. Theobald, G. Weikum: Scalable Knowledge Harvesting with High Precision and High Recall. WSDM 2011
[77] N. Nakashole, G. Weikum, F. Suchanek: PATTY: A Taxonomy of Relational Patterns with Semantic Types. EMNLP 2012
[78] V. Nastase et al.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. LREC 2010
[79] F. Naumann, M. Herschel: An Introduction to Duplicate Detection. Morgan & Claypool, 2010
[80] R. Navigli: Word Sense Disambiguation: a Survey. ACM Comput. Surv. 41(2), 2009
[81] R. Navigli, S. Ponzetto: BabelNet: Building a Very Large Multilingual Semantic Network. ACL 2010
[82] T. Nguyen et al.: Multilingual Schema Matching for Wikipedia Infoboxes. PVLDB 2012
[83] M. Nickel, V. Tresp, H.-P. Kriegel: Factorizing YAGO: Scalable Machine Learning for Linked Data. WWW 2012
[84] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, W.Y. Ma: Web Object Retrieval. WWW 2007
[85] F. Niu, C. Re, A. Doan, et al.: Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS, VLDB 2011
[86] F. Niu et al.: DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference, VLDS Workshop 2012
[87] M. Palmer, D. Gildea, N. Xue: Semantic Role Labeling Morgan & Claypool 2010
[88] M. Pasca: Ranking Class Labels Using Query Sessions. ACL 2011
[89] S.P. Ponzetto, M. Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007
[90] S.P. Ponzetto, M. Strube: Taxonomy induction based on a collaboratively built knowledge repository. Artif. Intell. 175(9-10), 2011
[91] A. Rahman, V. Ng: Coreference Resolution with World Knowledge. ACL 2011
[92] V. Rastogi, N. Dalvi, M. Garofalakis: Large-Scale Collective Entity Matching. PVLDB 2011
[93] L. Ratinov et al.: Local and Global Algorithms for Disambiguation to Wikipedia. ACL 2011
[94] S. Riedel, L. Yao, A. McCallum: Modeling Relations and their Mentions without Labeled Text. ECML 2010
[95] M. Rohrbach et al.: What Helps Where - and Why? Semantic Relatedness for Knowledge Transfer. CVPR 2010
[96] S. Sarawagi: Information Extraction. Foundations & Trends in Databases 1(3), 2008.
[97] S. Singh, A. Subramanya, F.C.N. Pereira, A. McCallum: Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models. ACL 2011
[98] P. Singla, P. Domingos: Entity Resolution with Markov Logic. ICDM 2006
[99] R.Speer, C.Havasi, H.Surana: Using Verbosity: Common Sense Data from Games with a Purpose. FLAIRS 2010
[100] R. Speer, C. Havasi: Representing General Relational Knowledge in ConceptNet 5, LREC 2012
[101] V.I. Spitkovsky, A.X. Chang: A Cross-Lingual Dictionary for English Wikipedia Concepts. LREC 2012
[102] S. Staab, R. Studer: Handbook on Ontologies, Springer, 2009
[103] F.M. Suchanek, G. Kasneci, G. Weikum: YAGO: a Core of Semantic Knowledge. WWW 2007
[104] F.M. Suchanek, M. Sozio, G. Weikum: SOFIE: a Self-Organizing Framework for Information Extraction. WWW 2009
[105] F. Suchanek et al.: PARIS: Probabilistic Alignment of Relations, Instances, and Schema. PVLDB 2012
[106] P.P. Talukdar, F. Pereira: Experiments in Graph-Based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL 2010
[107] P.P. Talukdar, D.T. Wijaya, T. Mitchell: Coupled temporal scoping of relational facts. WSDM 2012
[108] P.P. Talukdar, D. Wijaya, T. Mitchell: Acquiring Temporal Constraints between Relations. CIKM 2012
[109] N. Tandon, G. de Melo, G. Weikum: Deriving a Web-Scale Common Sense Fact Database. AAAI 2011
[110] B. Taneva et al.: Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity. WSDM 2010
[111] B. Taneva et al.: Finding Images of Difficult Entities in the Long Tail. CIKM 2011
[112] P. Venetis, A. Halevy, J. Madhavan, et al.: Recovering Semantics of Tables on the Web. PVLDB 2011
[113] M. Verhagen et al.: Automating Temporal Annotation with TARSQI. ACL 2005
[114] J. Völker, P. Hitzler, P. Cimiano: Acquisition of OWL DL Axioms from Lexical Resources. ESWC 2007
[115] R. Wang, W.W. Cohen: Language-independent Set Expansion of Named Entities using the Web. ICDM 2007
[116] C. Wang, J. Fan, A. Kalyanpur, D. Gondek: Relation Extraction with Relation Topics. EMNLP 2011
[117] Y. Wang et al.: Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. EDBT 2010
[118] Y. Wang et al.: Harvesting Facts from Textual Web Sources by Constrained Label Propagation. CIKM 2011
[119] Y. Wang, M. Dylla, M. Spaniol, G. Weikum: Coupling Label Propagation and Constraints for Temporal Fact Extraction. ACL 2012
[120] J. Wang, T. Kraska, M. Franklin, J. Feng: CrowdER: Crowdsourcing Entity Resolution. PVLDB 2012
[121] Z. Wang, J. Li, Z. Wang, J. Tang: Cross-lingual knowledge linking across wiki knowledge bases. WWW 2012
[122] S.E. Whang, H. Garcia-Molina: Joint Entity Resolution. ICDE 2012
[123] G. Weikum, M. Theobald: From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. PODS 2010
[124] F. Wu, D.S. Weld: Automatically Refining the Wikipedia Infobox Ontology. WWW 2008
[125] W. Wu, H. Li, H. Wang, K.Q. Zhu: Probase: a Probabilistic Taxonomy for Text Understanding. SIGMOD 2012
[126] L. Yao, S. Riedel, A. McCallum: Unsupervised Relation Discovery with Sense Disambiguation. ACL 2012
[127] M.A. Yosef et al.: HYENA: Hierarchical Type Classification for Entity Names. COLING 2012
[128] J. Zhu et al.: StatSnowball: a Statistical Approach to Extracting Entity Relationships. WWW 2009