Knowledge Bases in the Age of Big Data Analytics

Fabian M. Suchanek
Telecom ParisTech University
46 Rue Barrault
F-75634 Paris Cedex 13, France
fabian@suchanek.name

Gerhard Weikum
Max Planck Institute for Informatics
Campus E1.4
D-66123 Saarbruecken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

This tutorial gives an overview on state-of-the-art methods for the automatic construction of large knowledge bases and harnessing them for data and text analytics. It covers both big-data methods for building knowledge bases and knowledge bases being assets for big-data applications. The tutorial also points out challenges and research opportunities.

1. MOTIVATION AND SCOPE

Comphrehensive machine-readable knowledge bases (KB's) have been pursued since the seminal projects Cyc [19, 20] and WordNet [12]. In contrast to these manually created KB's, great advances have recently been made on automating the building and curation of large KB's [1, 16], using information extraction (IE) techniques and harnessing high-quality Web sources like Wikipedia. Prominent endeavors of this kind include academic research projects such as DBpedia [3], KnowItAll [10], NELL [5] and YAGO [29], as well as industrial ones such as Freebase. These projects provide automatically constructed KB's of facts about named entities, their semantic classes, and their mutual relationships. They contain millions of entities and billions of facts about them. Moreover, several KB's are interlinked at the entity level, forming the backbone of the Web of Linked Data [14]. Such world knowledge in turn enables cognitive applications and knowledge-centric services like disambiguating natural-language text, entity linking, text summarization, deep question answering, and semantic search and analytics over entities and relations in Web and enterprise data (e.g., [2, 6, 8, 13]). Prominent examples of how KB's can be harnessed include the Google Knowledge Graph [27] and the IBM Watson question answering system [17].

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 is on the twofold role of KB's 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. The following sections outline the structure of the tutorial. An extensive bibliography on this theme is given in [30].

Proceedings of the VLDB Endowment, Vol. 7, No. 13

2. BUILDING KNOWLEDGE BASES

Digital Knowledge: Today's KB's represent their data mostly in RDF-style SPO (subject-predicate-object) triples. We introduce this data model and the most salient KB projects, which include KnowItAll [10, 11], BabelNet [22], ConceptNet [28], DBpedia [3, 18], DeepDive [24], Freebase [4], ImageNet [7], NELL [5], Wikidata [31], WikiNet [21], WikiTaxonomy [26], and YAGO [29, 15]. We briefly discuss industrial projects like the Google Knowledge Graph and related work at Google [9, 13, 25], the EntityCube and Probase projects at Microsoft Research [23, 32], and IBM's Watson project [17].

Harvesting Knowledge on Entities and Classes: Every entity in a KB (e.g., Steve Jobs) belongs to one or multiple classes (e.g., computer pioneer, entrepreneur). These classes are organized into a taxonomy, where more special classes are subsumed by more general classes (e.g., person). We discuss two families of methods to harvest such information: Wikipedia-based approaches that analyze the category system, and Web-based approaches that use techniques like set expansion.

3. HARVESTING FACTS AT WEB SCALE

Harvesting Relational Facts: Relational facts express properties of and relationships between entities. There is a large spectrum of methods to extract such facts from Web documents. 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.

Open Information Extraction: Alternatively to methods that operate on a pre-specified set of relations and entities, open information extraction harvests arbitrary SPO 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 and Multilingual Knowledge: Properly interpreting entities and facts in a KB often requires additional meta-information like entity names in different languages and the temporal scope of facts. We discuss tech-

niques for tapping multilingual Web sources, and we cover techniques for extracting temporal expressions and for inferring the timepoints of events and timespans during which certain facts hold.

Commonsense Knowledge: Current KB's focus on facts about entities. However, there is an orthogonal dimension of commonsense that machines should acquire, too. This includes relations between concepts (e.g., mouthpiece partOf clarinet, clarinet hasShape cylindrical), properties of concepts that every child knows but are not obvious to a computer (e.g., the statement that apples can be red, green, juicy, sweet, sour, but not fast or funny), and also commonsense rules (e.g., the assertion that the father of a mother's child is usually the husband or partner of the mother as of the child's birth). We discuss methods for acquiring such kinds of commonsense knowledge.

4. KNOWLEDGE FOR BIG DATA

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 is a key asset.

Named Entity Disambiguation: With text or tables as input, 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 KB 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 Linkage: Even when entities are explicitly marked in (semi-) structured data, the problem arises to tell whether two entities are the same or not. This is a variant of the record-linkage problem (aka. entity matching, entity resolution, entity de-duplication). For KB's and Linked Open Data, the goal is to generate and maintain owl:sameAs information across knowledge resources at large scale. We give an overview of approaches to this end, covering statistical learning approaches and graph algorithms.

5. REFERENCES

[1] 3rd Workshop on Automated Knowledge Base Construction, 2013, http://www.akbc.ws/

[2] E. Alfonseca et al.: HEADY: News Headline Abstraction through Event Pattern Clustering. ACL 2013

[3] S. Auer et al.: DBpedia: A Nucleus for a Web of Open Data. ISWC 2007

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

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

[6] T. Cheng et al.: Data Services for E-tailers Leveraging Web Search Engine Assets. ICDE 2013

[7] J. Deng et al.: ImageNet: A Large-scale Hierarchical Image Database. CVPR 2009

[8] O. Deshpande et al.: Building, Maintaining, and Using Knowledge Bases: a Report from the Trenches. SIGMOD 2013

[9] X. Dong et al.: Knowledge Vault: a Web-scale Approach to Probabilistic Knowledge Fusion. KDD 2014

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

[11] A. Fader et al.: Identifying Relations for Open Information Extraction, EMNLP 2011

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

[13] R. Gupta et al.: Biperpedia: An Ontology for Search Applications. PVLDB 7(7), 2014

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

[15] J. Hoffart et al.: YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artif. Intell. 194, 2013

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

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

[18] J. Lehmann et al.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 2014

[19] D.B. Lenat, R. V. Guha: Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, 1990

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

[21] V. Nastase, M. Strube: Transforming Wikipedia into a large scale multilingual concept network. Artif. Intell. 194, 2013

[22] R. Navigli, S.P. Ponzetto: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. Artif. Intell. 193, 2012

[23] Z. Nie, J.-R. Wen, W.-Y. Ma: Statistical Entity Extraction From the Web. Proc. of the IEEE 100(9), 2012

[24] F. Niu et al.: DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. VLDS 2012

[25] M. Pasca: Acquisition of Open-domain Classes via Intersective Semantics. WWW 2014

[26] S.P. Ponzetto, M. Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007

[27] A. Singhal: Introducing the Knowledge Graph: Things, not Strings. Google Blog, May 16, 2012

[28] R. Speer, C. Havasi: Representing General Relational Knowledge in ConceptNet 5, LREC 2012

[29] F.M. Suchanek, G. Kasneci, G. Weikum: YAGO: a Core of Semantic Knowledge. WWW 2007

[30] F.M. Suchanek, G. Weikum: Knowledge Harvesting in the Big-Data Era. SIGMOD 2013

[31] D. Vrandecic, M. Krötzsch: Wikidata: a Free Collaborative Knowledge Base. CACM 57, 2014

[32] W. Wu et al.: Probase: a Probabilistic Taxonomy for Text Understanding. SIGMOD 2012