From Big Data to Big Knowledge

Large-Scale Information Extraction Based on Statistical Methods (Invited Talk)

Martin Theobald(✉)

University of Luxembourg, Esch-sur-Alzette, Luxembourg
martin.theobald@uni.lu

Abstract. Today's knowledge bases (KBs) capture facts about the world's entities, their properties, and their semantic relationships in the form of subject-predicate-object (SPO) triples. Domain-oriented KBs, such as DBpedia, Yago, Wikidata or Freebase, capture billions of facts that have been (semi-)automatically extracted from Wikipedia articles. Their commercial counterparts at Google, Bing or Baidu provide back-end support for search engines, online recommendations, and various knowledge-centric services.

This invited talk provides an overview of our recent contributions—and also highlights a number of open research challenges—in the context of extracting, managing, and reasoning with large semantic KBs. Compared to domain-oriented extraction techniques, we aim to acquire facts for a much broader set of predicates. Compared to open-domain extraction methods, the SPO arguments of our facts are canonicalized, thus referring to unique entities with semantically typed predicates. A core part of our work focuses on developing scalable inference techniques for querying an uncertain KB in the form of a probabilistic database. A further, very recent research focus lies also in scaling out these techniques to a distributed setting. Here, we aim to process declarative queries, posed in either SQL or logical query languages such as Datalog, via a proprietary, asynchronous communication protocol based on the Message Passing Interface.

Keywords: Information extraction
Probabilistic databases · Distributed graph databases

1 Information Extraction

The World Wide Web is the most comprehensive—but likely also the most complex—source of information that we have access to today. A vast majority of all information in the Surface Web, i.e., the part of the Web that is publicly accessible either as static pages or in the form of dynamically created contents, in fact consists of unstructured text. This textual information just happens to occasionally be interspersed with semi-structured components such as form fields, lists,

B. Catania et al. (Eds.): SOFSEM 2019, LNCS 11376, pp. 50–53, 2019.
https://doi.org/10.1007/978-3-030-10801-4_5

and tables—or so-called "infoboxes" in Wikipedia. These infoboxes, plus perhaps some more metadata, however, still constitute the main source of information for all of the currently available, Wikipedia-centric knowledge bases such as DBpedia (Auer et al. 2007), YAGO (Suchanek et al. 2007), Freebase (Bollacker et al. 2008), and Wikidata (Vrandeĉić and Krötzsch, 2014). This means that we currently exploit only a very small fraction of the information that is published on the Web for the purpose of *information extraction* (IE) and *knowledge base* (KB) construction.

In a recent series of publications on systems like KORE (Hoffart et al. 2012), AIDA-light (Nguyen et al. 2014), J-NERD (Nguyen et al. 2016), J-REED (Nguyen et al. 2017b) and QKBfly (Nguyen et al. 2017a), we stepwisely investigated the transition from basic, domain-oriented IE tasks like *named-entity recognition* (NER) and *disambiguation* (NED) toward a more extensible, open-domain IE setting, which combines NER and NED with a flexible form of pattern matching of relational paraphrases (Nakashole et al. 2011; Nakashole et al. 2012) into a comprehensive framework for *relation extraction* (RE). A focal point of our efforts thereby lies in combining these IE steps via various forms of joint-inference, rather than by following the more traditional, pipelined architectures for IE and NLP: J-NERD, for example, merges the tasks of NER and NED by performing a form of joint-inference over an underlying probabilistic-graphical model (in the form of a conditional random field), while J-REED and QKBfly further integrate NER and NED with RE and other NLP tasks like pronoun and co-reference resolution (via a semantic graph representation and corresponding graph-densification algorithm). Compared to domain-oriented extraction techniques, we thereby aim to acquire facts for a much broader set of predicates. Compared to open-domain extraction methods, the SPO arguments of our facts are canonicalized, thus referring to unique entities with semantically typed predicates.

2 Probabilistic Databases

Probabilistic databases (PDBs) (Suciu et al. 2011) encompass a plethora of applications for managing uncertain data, ranging from scientific data management, sensor networks, data integration, to information extraction and reasoning with semantic knowledge bases. While classical database approaches benefit from a mature and scalable infrastructure for the management of relational data, PDBs aim to further combine these well-established data management strategies with efficient algorithms for probabilistic inference by exploiting given independence assumptions among database tuples whenever possible. Moreover, PDBs adopt powerful query languages from relational databases, including Relational Algebra, the Structured Query Language (SQL), and logical query languages such as Datalog. The Trio PDB system (Mutsuzaki et al. 2007), which we developed at Stanford University back in 2006, was the first such system that explicitly addressed the integration of data management (using SQL as query language), lineage (aka. "provenance") management via Boolean formulas, and probabilistic inference based on the lineage of query answers. The Trio data model, coined

"*Uncertainty and Lineage Databases*" (ULDBs) (Benjelloun et al. 2008), provides a closed and complete probabilistic extension to the relational data model under all of the common relational (i.e., SQL-based) operations. Our recent research contributions in the domain of PDBs comprise a lifted form of evaluating top-k queries over non-materialized database views (Dylla et al. 2013b), learning of tuple probabilities from user feedback (Dylla et al. 2014), as well as further temporal-probabilistic database extensions (Dylla et al. 2013a; Dylla et al. 2011; Papaioannou et al. 2018).

3 Distributed Graph Databases

The third part of the talk finally takes an in-depth look at the architecture of our TriAD (for "*Triple-Asynchronous-Distributed*") (Gurajada et al. 2014a; Gurajada et al. 2014b) engine, which provides an end-to-end system for the distributed indexing of large RDF collections and the processing of queries formulated in the SPARQL 1.0 standard. TriAD combines a novel form of sharded, main-memory-based index structures with an asynchronous communication protocol based on the Message Passing Interface (MPI). It thus aims to bridge the gap between shared-nothing MapReduce-based RDF engines, on the one hand, and shared-everything native graph engines, on the other hand (see (Abdelaziz et al. 2017) for a recent overview). TriAD is designed to achieve higher parallelism and less synchronization overhead during query executions than MapReduce engines by adding an additional layer of multi-threading for entire execution paths within a query plan that can be executed in parallel. TriAD is the first RDF engine that employs asynchronous join executions, which are coupled with a lightweight join-ahead pruning technique based on graph summarization. Our current work also considers the processing of multi-source, multi-target graph-reachability queries (coined "*Distributed Set Reachability*" (DSR)) (Gurajada and Theobald, 2016b), as they may occur, for example, in the recent "Property Paths" extension of SPARQL 1.1 (Gurajada and Theobald, 2016a).

References

Abdelaziz, I., Harbi, R., Khayyat, Z., Kalnis, P.: A survey and experimental comparison of distributed SPARQL engines for very large RDF data. PVLDB **10**(13), 2049–2060 (2017)

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Ives, Z.: DBpedia: a nucleus for a web of open data. In: ISWC, pp. 11–15 (2007)

Benjelloun, O., Sarma, A.D., Halevy, A.Y., Theobald, M., Widom, J.: Databases with uncertainty and lineage. VLDB J. **17**(2), 243–264 (2008)

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD, pp. 1247–1250 (2008)

Dylla, M., Miliaraki, I., Theobald, M.: A temporal-probabilistic database model for information extraction. PVLDB **6**(14), 1810–1821 (2013a)

Dylla, M., Miliaraki, I., Theobald, M.: Top-k query processing in probabilistic databases with non-materialized views. In: ICDE, pp. 122–133 (2013b)
Dylla, M., Sozio, M., Theobald, M.: Resolving temporal conflicts in inconsistent RDF knowledge bases. In: BTW, pp. 474–493 (2011)
Dylla, M., Theobald, M., Miliaraki, I.: Querying and learning in probabilistic databases. In: Reasoning Web, pp. 313–368 (2014)
Gurajada, S., Seufert, S., Miliaraki, I., Theobald, M.: TriAD: a distributed shared-nothing RDF engine based on asynchronous message passing. In: SIGMOD, pp. 289–300 (2014a)
Gurajada, S., Seufert, S., Miliaraki, I., Theobald, M.: Using graph summarization for join-ahead pruning in a distributed RDF engine. In: SWIM, pp. 41:1–41:4 (2014b)
Gurajada, S., Theobald, M.: Distributed processing of generalized graph-pattern queries in SPARQL 1.1. CoRR, abs/1609.05293 (2016a)
Gurajada, S., Theobald, M.: Distributed set reachability. In: SIGMOD, pp. 1247–1261 (2016b)
Hoffart, J., Seufert, S., Nguyen, D.B., Theobald, M., Weikum, G.: KORE: keyphrase overlap relatedness for entity disambiguation. In: CIKM, pp. 545–554 (2012)
Mutsuzaki, M., et al.: Trio-one: layering uncertainty and lineage on a conventional DBMS. In: CIDR, pp. 269–274 (2007)
Nakashole, N., Theobald, M., Weikum, G.: Scalable knowledge harvesting with high precision and high recall. In: WSDM, pp. 227–236 (2011)
Nakashole, N., Weikum, G., Suchanek, F.M.: PATTY: a taxonomy of relational patterns with semantic types. In: EMNLP-CoNLL, pp. 1135–1145 (2012)
Nguyen, D.B., Abujabal, A., Tran, K., Theobald, M., Weikum, G.: Query-driven on-the-fly knowledge base construction. PVLDB **11**(1), 66–79 (2017a)
Nguyen, D.B., Hoffart, J., Theobald, M., Weikum, G.: AIDA-light: high-throughput named-entity disambiguation. In: LDOW (2014)
Nguyen, D.B., Theobald, M., Weikum, G.: J-NERD: joint named entity recognition and disambiguation with rich linguistic features. TACL **4**, 215–229 (2016)
Nguyen, D.B., Theobald, M., Weikum, G.: J-REED: joint relation extraction and entity disambiguation. In: CIKM, pp. 2227–2230 (2017b)
Papaioannou, K., Theobald, M., Böhlen, M.H.: Supporting set operations in temporal-probabilistic databases. In: ICDE, pp. 1180–1191 (2018)
Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: WWW, pp. 697–706 (2007)
Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic databases. Synth. Lect. Data Manag. **3**(2), 1–180 (2011)
Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Comm. of the ACM **57**(10), 78–85 (2014)