On the Limits of Machine Knowledge: Completeness, Recall and Negation in Web-scale Knowledge Bases

Simon Razniewski[1], Hiba Arnaout[1], Shrestha Ghosh[1], Fabian Suchanek[2]
[1]Max Planck Institute for Informatics
[2]Institut Polytechnique de Paris
{srazniew,harnaout,ghoshs}@mpi-inf.mpg.de,fabian@suchanek.name

ABSTRACT

General-purpose knowledge bases (KBs) are an important component of several data-driven applications. Pragmatically constructed from available web sources, these KBs are far from complete, which poses a set of challenges in curation as well as consumption.

In this tutorial we discuss how completeness, recall and negation in DBs and KBs can be represented, extracted, and inferred. We proceed in 5 parts: (i) We introduce the logical foundations of knowledge representation and querying under partial closed-world semantics. (ii) We show how information about recall can be identified in KBs and in text, and (iii) how it can be estimated via statistical patterns. (iv) We show how interesting negative statements can be identified, and (v) how recall can be targeted in a comparative notion.

PVLDB Reference Format:
Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek. On the Limits of Machine Knowledge:
Completeness, Recall and Negation
in Web-scale Knowledge Bases. PVLDB, 14(12): 3175 - 3177, 2021.
doi:10.14778/3476311.3476401

1 MOTIVATION

Web-scale knowledge bases (KBs) like Wikidata [32], DBpedia [2] or Yago [30] are used in applications ranging from question-answering to personal assistants. Pragmatically constructed from web resources, they focus on representing *positive* knowledge, i.e., statements that are true. They do not store *negative* statements. They are also *incomplete*, i.e., they do not contain all true statements in the domain of interest. This means that if a statement is not in the KB, we do not know whether it is false in the real world or just absent.

This poses major challenges for the curation and application of KBs: First, KB curators may want to know where the KB is incomplete, so that they can prioritize their completion efforts. This holds in particular for KBs such as NELL [4], which want to autocomplete themselves. Second, KB applications need to know where the data is incomplete, so as to alert end users of quality issues. For example, a query for "the largest city in Japan" may return the wrong answer if Tokyo happens to be absent in the KB. Similarly, a KB that is used for question-answering in an enterprise setting needs awareness of when a question surpasses its knowledge [22]. This holds in particular for boolean questions, such as "Did Airbus produce this plane", where a "no" could come simply from missing information. Finally, a comprehensive answer to the request to summarize the salient information about an entity should contain also *salient facts that do not apply*.

Traditionally, KB construction and curation has focused on the aspects of provenance and accuracy [21, 33]. Yet recent years have seen a maturing of formalisms for describing recall and negative knowledge [1, 5, 18], as well as a rise of statistical and text-based methods for estimating recall [3, 7, 12–14, 17, 24, 29] and deriving negative statements [1, 13]. Systematizing these approaches, and making them accessible to the general database audience, is the topic of this tutorial. The tutorial will be of interest to theoreticians as well as practitioners. It will inform the audience about the latest advances in completeness assessment and negation, and equip them with a repertoire of methodologies to better represent and assess the recall of specific datasets.

2 DETAILED DESCRIPTION

Length: Half day (3 hours)

Outline

(1) **Introduction (10 minutes):** We outline the gaps in existing web-scale KBs [26], and motivate the importance of capturing information about completeness, recall and salient negations in KBs with several use cases.
(2) **Logical foundations (20 minutes):** We outline the logical framework in which KBs operate, the open world assumption, the partial-closed world assumption (PCWA) [5, 6, 8], the implications this framework has for query answering [25], as well as the formal semantics of completeness assertions, how they can be practically represented in databases [15, 20, 28] and knowledge base [5].
(3) **Cardinalities from KBs and text as ground truth (45 min):** We explain the challenges in obtaining human ground truth, and the role that relation cardinality information plays in recall assessment. In particular, we show how existing cardinality information inside KBs can be identified and used to assess completeness [10], as well as how this information can be extracted from natural language documents [18].

Proceedings of the VLDB Endowment, Vol. 14, No. 12 ISSN 2150-8097.
doi:10.14778/3476311.3476401

(4) **Predictive recall assessment (45 min):** We present three lines of approaches: (i) Supervised machine learning to identify complete or incomplete regions of KBs [7], (ii) unsupervised statistical techniques, such as species sampling techniques from ecology [17, 31], density-based estimators [14] and statistical invariants about number distributions [29], (iii) linguistic theories about human conversations, which tell in which contexts information is likely complete, and in which contexts it is not [24].
(5) **Identifying salient negations (30 min):** We show why explicit negations are needed in open-world settings, and how they can be automatically mined by locally inferring closed-world topics from reference peer entities [1]. We contrast this approach with text extraction based on search engine query logs or Wikipedia text revisions [13], as well as language-modelling-based extraction [27], and outline open issues in terms of ontology modelling.
(6) **Relative recall (30 min):** We finally relax stricter absolute notions of recall, and show how recall can be measured in a relative manner, especially via extrinsic use cases like question answering and entity summarization [12, 23], by comparison with open information extraction or external reference resources [9, 19], and by comparison with other comparable entities inside the KB [3, 11, 16].

Practical sessions Each part will have a practical component, where participants familiarize themselves with sample tools for the respective topic (for example, data quality tools like Recoin (https://www.wikidata.org/wiki/Wikidata:Recoin), ProWD (https://prowd.id), Wikinegata (https://d5demos.mpi-inf.mpg.de/negation), CounQER (https://counqer.mpi-inf.mpg.de/spo)), Cool-WD (cool-wd.inf.unibz.it).

3 TUTORIAL MATERIAL

Attendees will be provided with slides, as well as link collections to relevant tools, code repositories and datasets, under a permissive license.

4 AUDIENCE

We expect a broad audience for this tutorial that can be divided into three groups: (1) knowledge acquisition engineers interested in tracking quality, and focusing extraction efforts; (2) application engineers interested in understanding knowledge quality; (3) formal semantics researchers interested in modeling knowledge beyond the open-world assumption.

Correspondingly, the tutorial will be organized so that participants with varying backgrounds can benefit. Only a basic familiarity with general database modelling (e.g., ER data model) is required, beyond that, all important concepts will be introduced formally, as well as by examples.

5 PRESENTERS

Simon Razniewski (primary contact) - Max Planck Institute for Informatics, srazniew@mpi-inf.mpg.de, http://simonrazniewski.com. Simon Razniewski is a senior researcher at the Max Planck Institute for Informatics in Saarbrücken, Germany, where he heads the Knowledge Base Construction and Quality research area. He has been a driver behind recent research around completeness, recall and negation in KBs, and has ample didactical experience from university teaching, and conference tutorials on commonsense knowledge (e.g., AAAI'21, WSDM'21).

Hiba Arnaout - Max Planck Institute for Informatics, harnaout@mpi-inf.mpg.de, http://people.mpi-inf.mpg.de/~harnaout/. Hiba Arnaout is a PhD student at the Max Planck Insitute for Informatics, in Saarbrücken, Germany. Her primary academic interests include Knowledge Base quality and negation in Knowledge Bases. Hiba has authored an award-wining paper on interesting negative statements in Knowledge Bases, published at AKBC'20, and presented at ISWC'20, as significant Web Semantic related work in a sister-conference.

Shrestha Ghosh - Max Planck Institute for Informatics, ghoshs@mpi-inf.mpg.de, https://people.mpi-inf.mpg.de/~ghoshs/. Shrestha Ghosh is a PhD student at the Max Planck Institute for Informatics in Saarbrücken, Germany. Her primary research is on exploring set information in Knowledge Bases and text to improve recall on count queries. She has published her work in JWS'20, ESWC'20 and presented at the doctoral consortium track of ISWC'20.

Fabian Suchanek - Institut Polytechnique de Paris, fabian@suchanek.name, http://suchanek.name. Fabian Suchanek is a professor at Institut Polytechnique de Paris in France, and the creator of the YAGO knowledge base. He has ample didactical experience, and authored more than 100 publications in the area of knowledge bases (with 12k citations in total), several of these specifically concerning completeness.

6 RELATED EVENTS

This tutorial represents completely novel material and has not been presented anywhere so far. Modified versions of the proposal have also been submitted to two complementary communities, semantic web (ISWC'21), and computational logics (KR'21), but these conferences take place several months after VLDB, and the communities have different foci.

We are not aware of any similar tutorials by other presenters in the past.

7 REQUIREMENTS

A a live internet connection is beneficial to try out examples on web-deployed KBs, but we will show them in the presentation screen too, so recordings can also be watched offline.

REFERENCES

[1] H. Arnaout, S. Razniewski, and G. Weikum. Enriching knowledge bases with interesting negative statements. In *AKBC*, 2020.
[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a web of open data. *ISWC*, 2007.
[3] V. Balaraman, S. Razniewski, and W. Nutt. Recoin: relative completeness in Wikidata. In *Wiki workshop at WWW*, 2018.
[4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.
[5] F. Darari, W. Nutt, G. Pirro, and S. Razniewski. Completeness statements about RDF data sources and their use for query answering. In *ISWC*, 2013.
[6] M. Denecker, Á. Cortés-Calabuig, M. Bruynooghes, and O. Arieli. Towards a logical reconstruction of a theory for locally closed databases. *TODS*, 2008.
[7] L. Galárraga, S. Razniewski, A. Amarilli, and F. M. Suchanek. Predicting completeness in knowledge bases. In *WSDM*, 2017.
[8] L. A. Galárraga, C. Teflioudi, K. Hose, and F. Suchanek. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW*, 2013.

[9] K. Gashteovski, R. Gemulla, B. Kotnis, S. Hertling, and C. Meilicke. On aligning OpenIE extractions with knowledge bases: A case study. In *Eval4NLP*, 2020.

[10] S. Ghosh, S. Razniewski, and G. Weikum. Uncovering hidden semantics of set information in knowledge bases. *JWS*, 2020.

[11] L. C. Gleim, R. Schimassek, D. Hüser, M. Peters, C. Krämer, M. Cochez, and S. Decker. Schematree: Maximum-likelihood property recommendation for wikidata. In *ESWC*, 2020.

[12] A. Hopkinson, A. Gurdasani, D. Palfrey, and A. Mittal. Demand-weighted completeness prediction for a knowledge base. In *NAACL*, 2018.

[13] G. Karagiannis, I. Trummer, S. Jo, S. Khandelwal, X. Wang, and C. Yu. Mining an "anti-knowledge base" from Wikipedia updates with applications to fact checking and beyond. *VLDB*, 2019.

[14] J. Lajus and F. M. Suchanek. Are all people married? determining obligatory attributes in knowledge bases. In *WWW*, 2018.

[15] W. Lang, R. V. Nehme, E. Robinson, and J. F. Naughton. Partial results in database systems. In *SIGMOD*, 2014.

[16] M. Luggen, J. Audiffren, D. Difallah, and P. Cudré-Mauroux. Wiki2prop: A multimodal approach for predicting wikidata properties from wikipedia. In *WWW*, 2021.

[17] M. Luggen, D. Difallah, C. Sarasua, G. Demartini, and P. Cudré-Mauroux. Non-parametric class completeness estimators for collaborative knowledge graphs - the case of Wikidata. In *ISWC*, 2019.

[18] P. Mirza, S. Razniewski, F. Darari, and G. Weikum. Enriching knowledge bases with counting quantifiers. In *ISWC*, 2018.

[19] B. D. Mishra, N. Tandon, and P. Clark. Domain-targeted, high precision knowledge extraction. *TACL*, 2017.

[20] A. Motro. Integrity= validity+completeness. *TODS*, 1989.

[21] H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *SWJ*, 2017.

[22] P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad. In *ACL*, 2018.

[23] S. Razniewski and P. Das. Structured knowledge: Have we made progress? an extrinsic study of KB coverage over 19 years. In *CIKM*, 2020.

[24] S. Razniewski, N. Jain, P. Mirza, and G. Weikum. Coverage of information extraction from sentences and paragraphs. In *EMNLP*, 2019.

[25] R. Reiter. On closed world data bases. In *Readings in artificial intelligence*. 1981.

[26] D. Ringler and H. Paulheim. One knowledge graph to rule them all? analyzing the differences between DBpedia, YAGO, Wikidata & co. In *KI*, 2017.

[27] T. Safavi and D. Koutra. Generating negative commonsense knowledge. *KR2ML*, 2020.

[28] Z. Shang, W. Brackenbury, A. J. Elmore, and M. J. Franklin. Cyadb: a database that covers your ask. *VLDB*, 2018.

[29] A. Soulet, A. Giacometti, B. Markhoff, and F. M. Suchanek. Representativeness of knowledge bases with the generalized Benford's law. In *ISWC*, 2018.

[30] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.

[31] B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar. Crowdsourced enumeration queries. In *ICDE*, 2013.

[32] D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledge base. *CACM*, 2014.

[33] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *SWJ*, 2016.