The Call for Recall

PI: Simon Razniewski
Team: Werner Nutt, Fariz Darari, Radityo Eko Prasojo
External Partner: Lydia Pintscher, Wikimedia Foundation

April 25, 2016

Abstract

General-purpose knowledge bases (KBs) such as YAGO, Wikidata or the Google Knowledge Graph usually contain facts that have a high precision. In contrast, little is known about the recall of such KBs, and anecdotal evidence indicates that knowledge bases have a low recall on many topics. This project aims to develop techniques to evaluate and improve the recall of knowledge bases. We aim to use focus on text extraction in this approach, and to compare the result with a complementary approach using rule mining. In cooperation with the Wikimedia Foundation, the output from both techniques shall lead to an extension for Wikidata which uses the retrieved recall information in order to help users in viewing, adding and managing recall information.

1 Introduction

General-purpose knowledge bases (KBs) such as Wikidata [20], the Google Knowledge Vault [3], NELL [9], or YAGO [18] aim to collect as much factual information about the world as possible. They store information about entities (such as Ötzi, South Tyrol, or Messner), and information about relationships between these entities (such as the fact that Ötzi died in South Tyrol, and that Messner climbed Mount Everest). These pieces of information typically take the form of triples, as in ⟨*Messner, climbed, Mt. Everest*⟩. KBs find applications in question answering, automated translation, or information retrieval.

The quality of a KB can be measured along several dimensions. A prominent one is the size. Today's KBs can contain millions, if not billions of triples. Another criterion is precision, i.e., the proportion of triples that are correct. YAGO, e.g., was manually evaluated on a sample, and was shown to have a precision of 95%. In this project, we propose to look at a third criterion for quality that has so far been largely neglected: recall, i.e., the proportion of facts of the real world that are covered by the KB. For some topics, today's KBs show very good recall values. For example,

- 160 out of 199 Nobel laureates in Physics are in DBpedia;

- 2 out of 2 children of Obama are in Wikidata;
- 36 out of 48 movies by Tarantino are shown in the Google Knowledge Graph.

On some other topics, today's KBs are nearly completely incomplete:

- DBpedia contains currently only 0.03% of all living persons.
- According to YAGO, the average number of children per person is 0.02.
- The Google Knowledge Graph contains a predicate called "Points of Interest" for countries. Since this predicate is subjective, it is not even clear how to measure its recall.
- In Biperpedia, 67k different attributes can be expressed [5], for which each recall is a question.

Previous research [17] has shown that between 69% and 99% of instances in popular KBs lack at least one property that other entities in the same class have. This gives us a hint of how incomplete KBs really are.
The problem is not just that KBs do not contain missing triples, but also that they do not know how many are missing, or whether some are missing at all. There exist approaches to manually create and maintain recall information, such as the no-values in Wikidata [20] or the completeness statements presented in COOL-WD [2]. However, to date, there is no support based on automated methods to create such information. In the project TQTK - "The Quest to Know" [13], in collaboration with Paris TelecomTech University, we are investigating how rule mining can be used to infer recall information from existing recall information. One of the preliminary conclusions from that project is that rule-mining performs well in observing general patterns, but the quality of the results is limited by the large number of cases in the real world that cannot be covered by rules. In the present project proposal, the goal is to extract recall information for such cases from unstructured web documents. In collaboration with Wikimedia Germany, which oversees the initial development of Wikidata[1], the results of this project shall lead to an extension for Wikidata.

2 Problem

Not knowing about the recall of knowledge bases is an problem from several perspectives:

- Philosophical perspective: We do not know what we actually know, and what we don't.
- Data collection perspective: KB contributors and engineers do not know where to focus their effort. If they knew that 39 Nobel laureates in Physics are missing, they could focus on tracing and adding the missing ones.

[1] https://en.wikipedia.org/wiki/Wikidata

- KB debugging perspective: One does not know when too much data is added. If there is reason to believe that Obama has two children, but a KB contains three, this could be highlighted.
- Rule learning perspective: KBs are often used for rule induction in order to learn new patterns and facts about the real world. But in order to evaluate learned rules, negative information is needed, which is usually not contained in KBs, but could be inferred from completeness information. Distant supervision, a popular pattern-based technique for automated knowledge base construction, faces the same challenge [8, 15, 19].
- Data consumption perspective: Consumers do not know whether a query really retrieves all answers. Also, results of aggregate queries (such as the average number of children per person) and queries with negation cannot be trusted.

We formulate the conceptual problem as follows:

Problem 1. We do not know, what we know and what we do not know.

Consequently, we cannot assess the limits of the knowledge of knowledge bases, which corresponds to the phrase of the "unknown unknowns". To phrase the problem more technically, we observe the following.

Problem 2. There are no tools to assist in the creation of information about what we know, or what we do not know.

We next discuss the state of the art in KBs and in techniques that could be used to assess their recall.

3 State Of The Art

Traditionally, knowledge bases focus on providing high-precision information. As already mentioned, YAGO [6], for instance, only contains facts with at least 95% confidence. Other knowledge bases [9, 3] contain facts with varying precision, and lowering the precision requirement naturally increases the recall, but nevertheless, does not help towards a qualitative understanding of recall.

There exists some largely anecdotal evaluations of the recall of KBs, with [17] for instance reporting that between 69% and 99% of instances in popular KBs lack at least one property that other entities in the same class have, or [3] reporting that 71% of people in Freebase have no known place of birth, and 75% have no known nationality. A commonly used proxy for comparing the recall of KBs is to count triples, objects or relations, as shown e.g. in Fig. 3. This is however a very coarse methods, especially if KBs are not aligned, and does not give an absolute understanding of the recall of an individual KB. Similarly, work such as [11], which explicitly mentions recall, refers to the recall relative to the information extracted by other systems.

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	271M
DeepDive [32]	4	2.7M	34	7M[a]
NELL [8]	271	5.19M	306	0.435M[b]
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M[c]
Freebase [4]	1,500	40M	35,000	637M[d]
Knowledge Graph (KG)	1,500	570M	35,000	18,000M[e]

Table 1: Couting information is often used as a proxy for relative recall comparisons. Table taken from [3].

While the only exact way to evaluate the absolute recall of a KB would be the comparison with a gold-standard dataset, we sketched some ideas that could help users in creating recall information in [14]. For manually constructed KBs such as Wikidata, it is expected that only manually created recall information can meet the quality standards of the KBs, and indeed, the Wikidata community manually listings that explain where information is still missing, for instance a list of people having no birth date[2]. Also, it is possible to add no-value statements to Wikidata, for instance it is stated that Elisabeth I has no children, which implies that the set of children for Elisabeth I that is contained in Wikidata is complete. An extension for Wikidata, that allows to manually add and manage recall information is described in [2].

For automatically created KBs, or to assist editors of automatically created KBs, we have proposed to estimate recall by extracting information from documents, use mark-and-recapture techniques, or to infer recall information from other information [14].

Extracting recall information might be doable in parallel to automated knowledge base construction from document corpora. One might look for phrases such as *"John has no children"*, or *"Alice married only once"*, where the latter, in combination with a spouse for Alice in the KB, allows to infer that the list of spouses of Alice is complete.

Mark-and-recapture techniques [16], stemming from the domain of ecology, allow to estimate the size of a population based on samples, even in cases where the samples are not independent. They might be also useful to estimate the size of a set of entities, which, in combination with counting the number of entities already in the KB, would allow to estimate the recall.

Some evidence that might be useful for learning recall, as suggested in [14], is recall information about other parts of the data (if parents are complete, then also children are more likely complete), popularity of entities (children of US presidents are likely to be complete), similarity to entities with known recall, or data changes over time.

Evidence that worked well for creating counterexamples for rule-mining so far was the *partial-closed-world assumption* (PCA) [4, 3], which states that for any subject-predicate-pair, for which at least one object is in the database, all objects

[2] `https://www.wikidata.org/wiki/Wikidata:Database_reports/top_missing_properties_by_number_of_sitelinks/P569`

are in the database. For instance, if a KB contains a child for John, then the PCA implies that the KB contains all children of John.

On the database side, work has been done on combining recall information about parts of databases to recall annotations for query results, but so far only for boolean information [10, 7, 12] (stating whether a certain topic is complete or not). More sophisticated techniques are needed to combine recall information that comes with confidences, maybe building upon techniques from probabilistic databases [1].

4 Objectives

Our objectives are threefold:

1. To develop a techniques to mine recall information from documents.
2. To compare the results of text extraction with those of rule mining as produced by the TQTK project [13].
3. To implement an extension for Wikidata that allows supports creating, viewing, and managing statements.

We expect text mining and rule mining to be complementary techniques, as text extraction can make use of unstructured knowledge that is not accessible to rule mining, while rule mining can make use of statistical information about correlations, general patterns and similar, which is not accessible to text extraction. Thus in Objective (3) we expect to use both the results from Objective 1, and also the results from the TQTK project, wherever, based on the results from Objective 2, the one performs better than the other.

5 Work Plan

We define four work packages. Package 1 is required for the evaluation in Packages 2 and 3, while Package 4 would use the best results identified in Package 3 in order to implement a tool for managing recall information in a real scenario.

5.1 Gold-standard Dataset

In this package, a human annotated dataset for the evaluation of the techniques developed in Packages 2 and 3 is created. Based on previous experience from the TQTK project, we would use the Crowdflower platform, where for 1000€, we would be able to annotate about 20.000 data items with 3 opinions and a correctness of about 95% per item. We will also reuse the gold-standard dataset from TQTK, however will need to label new items that are chosen to represent various textual phrases possibly indicating completeness, which was not a focus of the gold-standard dataset of TQTK.

5.2 Text Extraction

In this package, recall information shall be extracted from texts such as Wikipedia. We intend to focus on three kinds of recall information:

1. Novalues (e.g., "Merkel has no children")
2. Complete mentions (e.g., "The children of the Obamas are Malia and Sasha")
3. Cardinalities (e.g., "The Obamas have two children").

Novalues would tell us that the number of objects for a certain subject-predicate pair, here children of Merkel, is empty, and thus the recall is 100%. *Complete mentions*, in combination with entity disambiguation techniques, would allow to identify whether all mentioned objects (here Malia and Sasha) appear in the mentioned relation (here child) of the subject. Cardinalities would allow to compare the number of objects that should be in the KB with the number that is actually present, thus allowing to quantify the recall.

Technically, we would start with simple regular-expression matching techniques, based on manually defined patterns, and use tools such as Dexter or Stanford CoreNLP to extract named entities, relying on Anaphora resolution using dcoreref where needed. In a second step, we would evaluate whether *distant supervision* techniques can be used to automatically identify phrases that describe any of the three abovementioned types of recall information. Distant supervision is a technique for automated knowledge base construction from texts, based on a set of seed fact [8, 15, 19]. For instance, knowing that Malia is Obama's child, distant supervision would look for all contexts in which Malia and Obama cooccurr, and infer that common contexts such as *"X together with his daughter Y"* indicate that the fact $\langle X, hasChild, Y \rangle$ holds.

We would start with Wikipedia as document corpus, as Wikipedia is one of the best sources for knowledge extraction techniques, and as there are prepared dumps available.

5.3 Merging with Rule Mining

In Package 3, we would merge the recall information resulting from text extraction with the one obtaining using rule mining in the TQTK project. Rule mining is a popular technique for inferring additional knowledge in KBs. Concerning recall information, rule mining is able to learn rules such as *"if children are complete, then also parents are complete with a higher probability"*, or *"if someone is a US president, then pets are complete"*. As identified in the TQTK project, rule mining is very good at detecting general patterns (*"most popes die in Rome"*), but has an insufficient coverage of less standard topics, for instance, there is no common rule indicating how many children a person might have. We thus need to combine rule mining and text extraction in order to achieve sufficiently good recall information.

5.4 Integrated Wikidata Extension

The goal of this package is to implement a tool that uses the results from this project and TQTK in a real-world setting, in order to promote our results, and to make them available to the community. We intend to implement an extension for Wikidata because (a) Wikidata has a big community that is open to new technical ideas, (b) it is technically easy to develop extensions for Wikidata, and (c) as we have contacts to a Wikidata community manager, Lydia Pintscher, that has could help us in this process.

6 Dissemination and Continuation

On the scientific side, we plan to disseminate the results of this project at premier conferences in the area of semantic web and data management, such as ISWC, SIGMOD, VLDB or CIKM.

To attract external funding, we plan to apply for an engagement grant from the Wikimedia foundation that would further help us in developing the Wikidata extension[3]. Also, we consider applying for for a Google Faculty Award[4], as techniques for recall estimation are very relevant for Google's structured data efforts.

7 Budget

We intent to hire one research assistant (Cocopro) for six months for leading the text extraction effort and the data annotation (15000€). Furthermore, we plan to spend 1000€ for the data annotation, which from previous experience would allow us to label 20000 data items. Finally, we intend to use 4000€ for collaboration expenses, e.g. to invite or visit researchers in the field, and to present our work at Wikimedia community conferences or at international scientific conferences.

8 Delineation

Simon Razniewski is currently PI of two other projects funded by the Free University of Bozen-Bolzano, TaDaQua and TQTK.

The project *TQTK* is closely related to the present proposal, as it focuses also on the problem of estimating the recall of KBs. The technical approach in TQTK is based on rule mining. Based on the observed limitations of rule-mining, we expect text extraction and rule mining to be complementary techniques, as text extraction can make use of unstructured knowledge that is not accessible to rule mining, while rule mining can make use of statistical information

[3] https://meta.wikimedia.org/wiki/Grants:IEG

[4] http://research.google.com/research-outreach.html#/research-outreach/faculty-engagement/faculty-research-awards

about correlations, general patterns and similar, which is not accessible to text extraction.

The project *TaDaQua* has an unrelated goal, the estimation of the completeness of complex data objects in databases in the presence of optional information.

9 Team

Simon Razniewski is an assistant professor at the Free University of Bozen-Bolzano. In his PhD research, for which he has received the best PhD student award in 2013, he has investigated the reasoning about completeness information in various contexts such as relational databases, business processes, spatial databases and the semantic web. His work has been published at VLDB, CIKM, BPM, SIGSPATIAL and SIGMOD, and he maintains industrial collaborations with AT&T Labs-Research and Google Research.

Werner Nutt is a full professor at the Free University of Bozen-Bolzano, working on formal aspects of data quality and process governance. In the past, he has been prominently involved in the field of description logic knowledge bases.

Fariz Darari is a third year PhD student at the Free University of Bozen-Bolzano that contributes experience in semantic web and data completeness.

Radityo Eko Prasojo is a first year PhD student at the Free University of Bozen-Bolzano, that contributes experience in information extraction from Wikipedia.

References

[1] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB Journal*, 2007.

[2] F. Darari, S. Razniewski, R. Prasojo, and W. Nutt. Enabling fine-grained RDF data completeness assessment. *ICWE*, 2016.

[3] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.

[4] L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases . In *WWW*, 2013.

[5] R. Gupta, A. Y. Halevy, X. Wang, S. E. Whang, and F. Wu. Biperpedia: An ontology for search applications. *PVLDB*, 2014.

[6] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal*, 2013.

[7] A. Y. Levy. Obtaining complete answers from incomplete databases. In *VLDB*, 1996.

[8] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*, pages 777–782, 2013.

[9] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *AAAI*, 2015.

[10] A. Motro. Integrity = Validity + Completeness. *TODS*, 1989.

[11] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 227–236. ACM, 2011.

[12] S. Razniewski, F. Korn, W. Nutt, and D. Srivastava. Identifying the extent of completeness of query answers over partially complete databases. In *SIGMOD*, 2015.

[13] S. Razniewski and W. Nutt. The quest to know (TQTK). *Unibz RTD Call*, 2015.

[14] Simon Razniewski, Fabian M. Suchanek, and Werner Nutt. But what do we actually know? *AKBC*, 2016.

[15] Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A survey of noise reduction methods for distant supervision. In *AKBC*, pages 73–78. ACM, 2013.

[16] P. Spoor, M. Airey, C. Bennett, J. Greensill, and R. Williams. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ*, 1996.

[17] F. M. Suchanek, D. Gross-Amblard, and S. Abiteboul. Watermarking for Ontologies . In *ISWC*, 2011.

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

[19] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *ACL*, pages 721–729. ACL, 2012.

[20] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.