Assessing the Completeness of Entities in Knowledge Bases

Albin Ahmeti[1], Simon Razniewski[2], Axel Polleres[1]

[1]Vienna University of Economics & Business [2]Free University of Bozen-Bolzano

Abstract. While human-created knowledge bases (KBs) such as Wikidata provide usually high-quality data (precision), it is generally hard to understand their completeness. In this paper we propose to assess the relative completeness of entities in knowledge bases, based on comparing the extent of information with other similar entities. We outline building blocks of this approach, and present a prototypical implementation.

1 Introduction

Knowledge bases such as Wikidata, YAGO or DBpedia are becoming increasingly popular as structured sources of data, and are used in a variety of tasks such as structured search, question answering, or entity recognition, even though they are generally highly *incomplete* [8]. In particular, when incomplete KBs are combined with query languages that contain negation such as SPARQL, the result easily yields unsound answers [6]. Understanding how complete KBs are on different aspects is important for KB curators so they know where to focus their efforts, and for consumers to know to which extent they can rely on a KB.

It is difficult to talk about the completeness of KBs because completeness can be investigated on various levels and with varying semantics. While it is relatively easy to understand when a knowledge base is complete for children of Obama (when Malia and Sasha are there), it is not clear what completeness of Obama himself, or of US politicians as a whole, could mean. Previous work on knowledge base completeness has focused on the lowest level, i.e., finding out when a subject is complete for a predicate (like Obama for *child*) [2, 4, 7], whereas more abstract levels have not been investigated so far.

In this paper we propose investigating completeness on the level of entities, i.e., to give statements about how complete entities such as Barack Obama or Portoroz are. We propose to compute these statements by comparison with other, similar entities. More specifically, for a designated entity we check its coverage of frequent properties, computed among similar entities. We have implemented a prototype as *Recoin (Relative Completeness Indicator)* in Wikidata.

2 Background

While general-purpose knowledge bases already find application in a variety of tasks, due to their ill-defined scope (for instance, unlike Wikipedia, Wikidata has

no relevance criteria other that new items should be linked to at least one existing item) and/or ambition to capture as much knowledge as possible, they are in general highly incomplete. In Wikidata, for instance, only 48% of politicians are member of a party, or only 0.02% of people do have a child.

Previous work on assessing KB completeness has focused on the level of subject-predicate pairs. [7] provides a plugin for Wikidata that allows to assert completeness for such pairs directly on the Wikidata website. [2] has used association rule mining for automatically determining complete pairs. [4] used Wikipedia texts to mine the cardinalities of such pairs, using these cardinalities in turn to assess completeness. A recent survey paper, [5], provides a comprehensive overview on the state-of-the-art KB refinement approaches aimed at improving the KB completeness.

For more holistic descriptions of quality, Wikipedia has so-called status indicators (like "Featured article", "Good article"). For Wikidata, such indicators do not yet exist, but their introduction is planned.[1]

3 Relative Completeness Indicators

For basic granularities, such as children of Obama, as discussed in [2, 4, 7], boolean completeness annotations generally suffice. In contrast, on the entity level, given that Wikidata contains over 2700 properties, of which 101 are used at least 1000 times for the class *human*, containing further ill-defined properties such as *medical condition*, *notable work* and *participant of*, it is clear that boolean statements such as "Data about Obama is complete", or "Data about Trump is incomplete", are not meaningful.

To allow statements for entities, we thus propose to define a relative completeness measure. More specifically, we propose to compare the extent of information about an entity with the extent of information that is available for other, similar entities. For instance, in assessing the completeness of Obama, we would compare the information available about him with that available for other politicians, while when assessing the completeness of Slovenia, we would compare with other countries.

There are three crucial components to this approach, (i) the definition of similar entities, (ii) the way how the extent of information is compared among similar entities, and (iii) the way how explanations are provided.

(i) For similarity, classes are a natural baseline, and class-like properties such as *occupation* allow a further refinement. Semantic similarity measures [9] could provide even better way to find similar entities.
(ii) Baselines for comparison could be counts of facts or properties, while better results can be expected if the relevance or importance of properties and facts is taken into account [1].
(iii) The way explanations are generated depends highly on the choices made for (i) and (ii), and will in turn impact usability for knowledge base authors

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

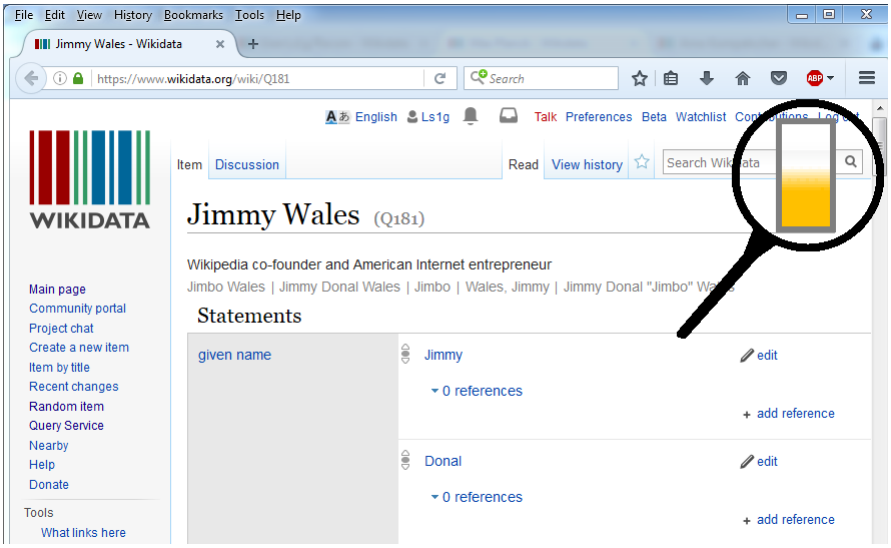

Fig. 1. Recoin core module on the Wikidata page of Jimmy Wales.

and users. We may expect a tradeoff between accuracy and complexity, i.e., more complex choices may lead to more accurate assertions, which however are harder to explain, thus not necessarily increasing usability.

4 Wikidata Implementation

We have implemented a relative completeness indicator called *Recoin* in Wikidata.[2] It is provided as user script, i.e., logged in Wikimedia users can enable it in a user configuration file. It consists of two components. The core component, which adds a relative completeness indicator to the status indicator section of Wikidata articles, is shown in Fig. 1. The indicator is a color-coded progress bar, which can show 5 levels of completeness, ranging from "very detailed" to "very basic". An explanation module adds information about the relevant missing properties, based on which the completeness level is calculated. Further details about the architecture are on the tools website. It is currently available on the Wikidata pages of all *humans* that have a profession. Internally, the completeness level is computed as follows:

1. Each entity is compared with the set of all entities that have at least one profession in common.
2. For that set, the 50 most frequent properties are computed. The completeness level is then computed using fixed thresholds, i.e., if the entity has more than 40 out of these 50 properties, completeness is on the highest level, if it has between 30 and 40 of these properties, second highest level, and so on.

[2] https://www.wikidata.org/wiki/User:Ls1g/Recoin

3. As explanation, the properties absent wrt. the comparison set are shown along with their frequency in the comparison set.

The tool was made available to the Wikidata community on 15th of November, 2016. An expansion to all *humans* and other classes of entities are planned.

5 Evaluation and Future Work

Some completeness levels computed by Recoin are for Obama 4 (detailed), for Trump 3 (fair), for Jimmy Wales 3 (fair), or for Dijkstra 2 (basic). While many levels appear reasonable (more popular entities are more complete, less popular ones less), others can only be understood using the explanations. The comparably low level for Jimmy Wales, for instance, is based on the fact that he misses properties such as *member of political party*, *position held* and *father*, which in the comparison set, exist for 10%, 8% and 6% of entities.

To further evaluate the levels computed by Recoin, in a crowdsourcing experiment, we compared a three-level scheme with levels that human annotators would give. Using 20 entities and 7 opinions per entity, we found that Recoin agreed in 60% of cases with the majority opinion, while in 25% it was off by one level, and in 15% off by two levels.

As future work, we aim to evaluate how methods based on semantic similarity can provide more meaningful sets of entities for comparison, and how relevance and importance of properties can be taken into account when comparing entities. More specifically, we aim to investigate [3], which uses statistical analysis of predicate-value pairs in order to find similar entities.

Acknowledgment We thank Werner Nutt for comments, and Fariz Darari for technical help. This work has been partially supported by the project "TaDaQua", funded by the Free University of Bozen-Bolzano.

References

1. A. Dessi and M. Atzori. A machine-learning approach to ranking RDF properties. *Future Generation Comp. Syst.*, 54:366–377, 2016.
2. L. Galárraga, S. Razniewski, A. Amarilli, and F. M. Suchanek. Predicting completeness in knowledge bases. *WSDM*, 2017.
3. A. Hogan, A. Polleres, J. Umbrich, and A. Zimmermann. Some entities are more equal than others: statistical methods to consolidate linked data. In *NeFoRS*, 2010.
4. P. Mirza, S. Razniewski, and W. Nutt. Expanding Wikidata's parenthood information by 178%, or how to mine relation cardinalities. *ISWC Posters & Demos*, 2016.
5. H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, 2017.
6. A. Polleres, C. Feier, and A. Harth. *Rules with Contextually Scoped Negation.* 2006.
7. R. E. Prasojo, F. Darari, S. Razniewski, and W. Nutt. Managing and consuming completeness information for wikidata using COOL-WD. *COLD*, 2016.
8. S. Razniewski, F. M. Suchanek, and W. Nutt. But what do we actually know. *AKBC*, 2016.
9. M. A. Rodríguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *TKDE*, 15(2):442–456, 2003.