Analyzing Relative Incompleteness of Movie Descriptions in the Web of Data: A Case Study

Wancheng Yuan[1], Elena Demidova[2], Stefan Dietze[2], Xuan Zhou[1]

[1]DEKE Lab, MOE. Renmin University of China. Beijing, China
wancheng.yuan@ruc.edu.cn, zhou.xuan@outlook.com
[2]L3S Research Center and Leibniz University of Hanover, Germany
{demidova, dietze}@L3S.de

1 Introduction and Approach

In the context of Linked Open Data (LOD) [3], datasets are published or updated frequently, constantly changing the landscape of the Linked Data Cloud. In this paper we present a case study, investigating relative incompleteness among subgraphs of three Linked Open Data (LOD) datasets (DBpedia (`dbpedia.org`), Freebase (`www.freebase.com`), LinkedMDB (`www.linkedmdb.com`)) and propose measures for relative data incompleteness in LOD. The study provides insights into the level of accuracy and actual conflicts between different LOD datasets in a particular domain (movies). In addition, we investigate the impact of the neighbourhood size (i.e. path length) under consideration, to better understand the reliability of cross-dataset links.

Fig. 1 presents an example of relative incompleteness in the representation of movie entity "Holy Smoke!" in DBpedia and Freebase. In this example, the difference between the actor sets indicates the "Movie.Actor" property might be incomplete. As we do not know the exact complete set of actors and the noise observed in linked datasets interferes completeness estimation, we call this phenomenon relative incompleteness. If we follow the "Movie.Cinematographer" link in the data graphs of the two datasets, we can observe further relative incompleteness in its "birthPlace" property.

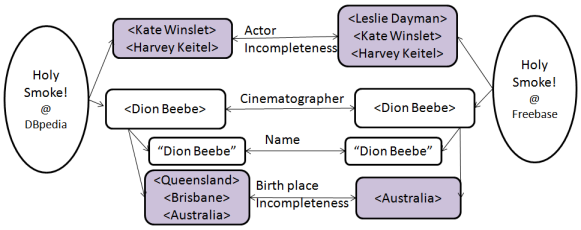

Fig. 1. Representation of the movie "Holy Smoke!" in DBpedia and Freebase.

In this paper we discuss incompleteness related measures that can be obtained by pairwise dataset comparisons exploiting entity co-resolution across

these datasets and apply these measures in a case study. The basis for the proposed measures is the assumption that dataset specific differences in representation of equivalent entities, and in particular the values of multi-value properties, can provide valuable insights into the relative incompleteness of these datasets. To facilitate discovery of relative incompleteness, we assume correct schema mapping and make use of known equivalent entities. In the context of LOD, absolute incompleteness is difficult to judge, as it is difficult to obtain the ground truth of absolute completeness. Therefore, we choose to estimate relative incompleteness of the properties by following paths of limited lengths in the data graphs.

By i^{th}**-Order property**, we mean a property that can be reached from the target entity through a path of length i in the data graph. For instance, "Movie.Actor" is a 1^{st} Order property in Fig. 1, while "Movie.Cinematograher. Name" is a 2^{nd} Order property. Then, we define i^{th}-Order Value Incompleteness as follows:

i^{th}**-Order Value Incompleteness (**D_x**,** D_y**,** P**)** between the pair of datasets D_x, D_y with respect to a i^{th}-Order multi-value property P is the proportion of entities in D_x and D_y having different values in P.

As P is a multi-value property, the difference on P usually indicates that at least one of the datasets does not provide sufficient information on P. In Fig. 1, we observe a 2^{nd} Order Value Incompleteness in the "Movie.Cinematographer .birthPlace" property. To determine equivalent values, we rely on direct value comparisons and identity links.

Considering the LOD cloud as a large interlinked knowledge graph, relative incompleteness of data across different datasets is a crucial and often under-investigated issue. Relative incompleteness can result e.g. from extraction errors, lack of source maintenance [4], imprecise identity links [2] as well as incompatibilities in schemas and their interpretation (as we observed in this study). In the literature, detection and resolution of data inconsistency has been studied in the context data fusion [1]. However, the corresponding methods for the assessment of LOD datasets are underdeveloped. The measures proposed in this paper can help judging relative agreement of datasets on certain properties and thus support source selection. E.g. these statistics can support identification of sources with the highest relative agreement as well as the sources containing complementary information, dependent on the particular scenario.

2 A Case Study

Datasets and Schemas: We used the latest version of three datasets from LOD - LinkedMDB (LMDB), DBpedia and Freebase. The LMDB dataset contains eight concepts about movies, such as *Movie*, *Actor* and *Country*, and more than 200,000 records. The DBpedia and Freebase datasets contain around 150,000 and 1,000,000 movie records respectively. To perform the study, we randomly selected 200 Movie and 200 Actor entities shared between these datasets. To establish the relationship of entities across the three datasets, we obtained the existing interlinking information (i.e., the *owl:sameAs* predicate) of the Movie

entities across all the three datasets as well as the Actor entities in the DBpedia and Freebase. We manually established schema mappings between the Movie and Actor concepts and their properties among the datasets.

Evaluation Results: We computed the 1^{st} and 2^{nd} Order Value Incompleteness for each property in each pair of datasets. Table 1 presents an aggregated 1^{st} and 2^{nd} Order Value Incompleteness results for the Movie and Actor entities. In this result, if there is a single property that is incomplete on an entity, we would count this entity as incomplete. As we can see in Table 1, the relative incompleteness in the DBpedia/Freebase pair reaches 100% in the first order and 89% in the second order, meaning that all the Movie entities in the datasets are affected by incompleteness issues. The overall 1^{st} Order Incompleteness of the Movie entities in the other dataset pairs is also pretty high, e.g., 70% for LMDB/Freebase and 56% for LMDB/DBpedia.

Table 1. Aggregated Incompleteness for Movie and Actor Entities

Datasets	Movie 1^{st} O. Incompleteness	Movie 2^{nd} O. Incompleteness	Actor 1^{st} O. Incompleteness
LMDB/DBpedia	0.56	n/a	n/a
LMDB/Freebase	0.70	n/a	n/a
DBpedia/Freebase	1.00	0.89	0.76

Table 2 presents the details of the evaluation for each property. As we can see in the Table 2, the highest relative incompleteness among all datasets is observed in the DBpedia/Freebase pair on the property Actor, whose incompleteness is 73%. This is because DBpedia tends to include only key people in a movie, whereas Freebase tends include more complete actor lists. For example, for the movie "Stand by Me", DBpedia lists only five actors: Wil Wheaton, Kiefer Sutherland, River Phoenix, Corey Feldman, and Jerry O'Connell. The "starring" property of Freebase includes many more actor names such as Gary Riley, Bradley Gregg, Frances Lee McCain, etc. We also observed that the "starring" property sometimes mixes actor and character names in a movie. For example, for "Stand by Me", it includes characters Teddy Duchamp and Waitress. Regarding the LMDB/Freebase, the incompleteness on the properties Producer, Release Date and Actor are 30%, 26% and 19% respectively. LMDB/DBpedia shows a similar distribution, i.e. 29%, 11% and 15%, on the same properties.

Table 2. 1st Order Value Incompleteness of Selected Movie Properties

Dataset	Release Date	Country	Language	Actor	Director	Writer	Editor	Producer
LMDB/DBpedia	0.11	0.02	0.16	0.15	0.02	n/a	n/a	0.29
LMDB/Freebase	0.26	0.15	0.24	0.19	0.02	n/a	n/a	0.30
DBpedia/Freebase	0.21	0.12	0.25	0.73	0.04	0.25	0.08	0.36

An exemplary evaluation performed on the Actor entities indicates a similar tendency as for the Movie type, with 76% incompleteness in the first order in the DBpedia/Freebase pair. While Actor entities always agree on the names and very often on the birth dates (which makes us think that the existing interlinking of Actor entities was established using these properties), they frequently disagree on the property of birthPlace. This is because the values of property *birthPlace* from DBpedia are much more detailed than those from Freebase. DBpedia typically includes a country name in an address, whereas Freebase does not. For example, the place of birth of the person "Len Wiseman" from DBpedia is "Fremont, California, United States", while that from Freebase is "Fremont, California". As a result, we observe an increased incompleteness in the property *birthPlace*. Interestingly, the property *deathPlace* is much less incomplete, as most actors listed in these databases are still alive (we regard null values as incomparable).

3 Conclusions and Outlook

In this paper we presented measures to automatically evaluate relative incompleteness in linked datasets and applied these measures in a case study. From the experiment performed using three linked datasets in the movie domain we can conclude that incompleteness is a very common phenomenon in these datasets, and its number increases significantly with increasing order, i.e. increase of investigated entity neighbourhood. The main causes of relative incompleteness observed during our experiment are due to different interpretations of properties in the datasets. Our method of classification and identification of quality issues provides not only insights into the level of agreement between datasets but also into the overall quality of datasets. In future work we intend to extend these approaches to infer knowledge about the correctness and agreement of schemas.

Acknowledgments

This work was partially funded by the NSFC Project No. 61272138, ERC under ALEXANDRIA (ERC 339233), the COST Action IC1302 (KEYSTONE) and 973 Program Project of China (Grant Nr.: 2012CB316205).

References

1. X. L. Dong and F. Naumann. Data fusion: resolving data conflicts for integration. *Proc. VLDB Endow.*, 2(2):1654–1655, 2009.
2. H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When owl: sameas isn't the same: An analysis of identity in linked data. In *Proc. of the 9th International Semantic Web Conference, ISWC 2010, Shanghai*, 2010.
3. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136.* Morgan & Claypool, 2011.
4. P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In *Proc. of the 2012 Joint EDBT/ICDT Workshops, Berlin*, 2012.