Exploiting Equivalence to Infer Type Subsumption in Linked Graphs

Russa Biswas[1,2], Maria Koutraki[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
firstname.lastname@kit.edu

Abstract. Open Knowledge Graphs (KGs) such as DBpedia and Wikidata have been recognized as the foundations for diverse applications in the field of data mining and information retrieval. Each of these KGs follows a different knowledge organization as well as is based on differently structured ontologies. Moreover, it has been observed that type information are often noisy, incomplete or even incorrect. In general, there is a need for well defined and comparable type information for the entities of the KGs. In this paper, we propose an isomorphism-based approach to infer subsumption relations to RDF type information in Wikidata by exploiting the RDF type information from DBpedia.

Keywords: Knowledge Graph, RDF, Wikidata, DBpedia

1 Introduction

Since the introduction of the Linked Open Data(LOD) cloud, the general purpose KGs like DBpedia, YAGO, Wikidata have been the focal point of research in the field of data mining and information retrieval. Hence, the correctness and completeness of such KGs is of great importance. However, many studies show that information in these KGs often can be noisy, incorrect and incomplete [3,9,7,6]. One way to account for the incompleteness of information in a KG is to harness the complementary information from different KGs.

Nevertheless, the different KGs are following different knowledge organization approaches [8,4,2] and use different underlying ontologies to represent knowledge, where explicit alignments amongst the different ontologies are not always available [5]. Therefore, a direct comparison of the KGs in the content level is a challenging task. For example, in Wikidata, the property `wdt:P31` (instance of)[3] defines what we know as `rdf:type`. However, based on our observations, `wdt:P31` follows different semantics and it differs in its use when compared to `rdf:type` in DBpedia. Thus, by relying only on `wdt:P31` it is not possible to have a direct content-based comparison of the classes of the two KGs.

In this paper, we propose a light-weight isomorphism-based schema matching approach to harmonize two KGs having different underlying schema structure. For this study, we have used the two most popular KGs: DBpedia(English language) and Wikidata. The main aim of this work is to infer type subsumption

[3] https://www.wikidata.org/wiki/Property:P31

relations in Wikidata by leveraging the existing equivalence relations between Wikidata and DBpedia. To this purpose we establish *conditional subsumption relations* between Wikidata properties and `rdf:type`.

2 Type Subsumption

Problem Description - We consider two RDFS[4] KGs, a source K_S and a target K_T, consisting of set of triples $K \subseteq E \times R \times (E \cup L)$, where E is a set of resources referred to as entities, L a set of literals, and R a set of relations. $\{C_{S_i}\}$ and $\{C_{T_j}\}$ is the set of classes in the source and target KG respectively. We assume that the classes and the entities of K_S and K_T are aligned i.e. K_S stores the statement $< C_{S_n}, \texttt{owl:equivalentClass}, C_{T_m} >$ and $< e_S, \texttt{owl:sameAs}, e_T >$.

In this work, we aim for a *conditional subsumption relation alignment*, as the schemas used for KGs vary heavily. Thus only *equivalence* alignments that have merely similar semantics or subsume one another are not enough to map the relations. Following the *relation subsumption* definition in [5] the goal is:

Goal. *For two KGs, a source $K_S \subseteq E_S \times R_S \times (E_S \cup L_S)$ and a target $K_T \subseteq E_T \times R_T \times (E_T \cup L_T)$, and a relation* ***`rdf:type`*** $\in R_S$*, find relations $r_T \in R_T$ s.t. $r_T \subseteq$* ***`rdf:type`****.* The *equivalence* relation between r_S and r_T, can also be expressed as a two-way subsumption relation: $r_S \equiv r_T$, iff $r_S \subseteq r_T$ and $r_T \subseteq r_S$.

Methodology - The aforementioned goal is achieved by exploiting the equivalence relations of classes and instances between the two KGs. The method is described with the help of the illustration in Fig. 1.

Step 1: For each class C_{S_i} in DBpedia, we determine the entities e_S of the class via `rdf:type` relation. Formally: $\forall C_{S_i} \in K_S :< e_S, \texttt{rdf:type}, C_{S_i} >$

Step 2: From the entities e_S, find those with `owl:sameAs` link(s) to corresponding e_T entities in Wikidata. Formally: $\forall e_S \in C_{S_i}, \exists e_T \in K_T :< e_S, \texttt{owl:sameAs}, e_T >$

Step 3: Determine the class C_{T_j} in Wikidata equivalent to DBpedia class, C_{S_i} via the `owl:equivalentClass` relation.
Formally: $\forall C_{S_i} \in K_S, \exists C_{T_j} \in K_T :< C_{S_i}, \texttt{owl:equivalentClass}, C_{T_j} >$

Step 4: For each entity e_{T_j}, check if there is any relation (or relations) r_{T_j}, which connects to C_{T_j}. Formally: $\forall e_T, \exists r_{T_j} \in K_T :< e_T, r_{T_j}, C_{T_j} >$

3 Experimental Evaluation

This section discusses the results of the approach of inferring type subsumption relations in Wikidata leveraging existing mappings to DBpedia. Due to lack of space the full set of results can be found here [1].

For this work, all the experiments were carried out on DBpedia 2016-10 version and Wikidata as of January 11, 2018. Out of the *524* interlinked classes between DBpedia and Wikidata, we conducted experiments on *327* classes, the instances of which are linked via `owl:sameAs`.

[4] https://www.w3.org/RDFS/

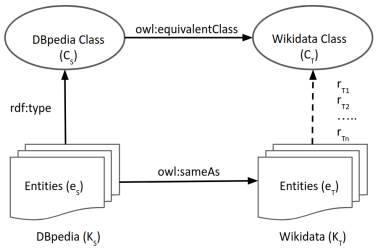

Fig. 1. Isomorphic Approach to Infer Type Subsumption Relations in Wikidata with the help of DBpedia

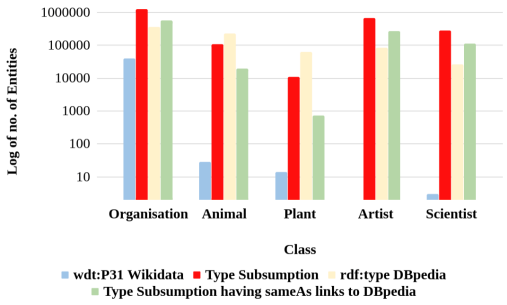

Fig. 2. Comparison of the KGs (best viewed with color print) [1]

Results - The experiments establish the fact that the type information in Wikidata is often implicitly defined and 41 properties, including `wdt:P31`(instance of), hold a subsumption relation with `rdf:type` in DBpedia. Interestingly only the members of about 38% of these Wikidata classes can be accessed via `wdt:P31`. Furthermore, only 58% of the aforementioned 38% of Wikidata classes are using the property `wdt:P31` exclusively to denote the membership in a class. Table 1 shows some Wikidata classes and the properties serving as `rdf:type` ordered by the percentage of the class members which were retrieved via them.

Additionally, it is also interesting to notice that similar classes have similar type subsumption relations. For instance, for the classes in Wikidata denoting different kinds of professions such as, *Artist, Scientist* the property *occupation*(`wdt:P106`) defines the members of the class.

Fig. 2 illustrates a comparison between DBpedia and Wikidata for 5 classes. It is interesting to notice that the number of instances retrieved from Wikidata via the new type subsumption relations (red bar) is much higher than via only `wdt:P31` (blue bar). Hence, more members of the classes can be retrieved using the subsumption relations leading to a strong foundation for the content level comparison of the KGs.

Furthermore, the green bar in the Fig. 2 represents the number of instances of the corresponding Wikidata classes using the type subsumption relations of Table 1, which also have `owl:sameAs` links to DBpedia. For all these classes, it has been observed that the height of the red bar (count of instances with new type subsumption relations) is higher than the green bar (count of instances with new type subsumption relations and `owl:sameAs` links to DBpedia), which reflects that Wikidata potentially contains more information than DBpedia for these classes. Also, it can be inferred that some of these entities in Wikidata are also present in DBpedia but are assigned to some other classes in DBpedia. This however can lead to further research on the correctness of the KG content.

Last, for the classes `dbo:Animal` and `dbo:Plant`, the number of instances in DBpedia(yellow) is higher than the number of instances that possess `owl:sameAs`

Class	Conditional $r_T \equiv$ `rdf:type`
Organisation	instance of (P31) (99.7%), is a list of (P360) (0.2997%), has part (P527) (0.0003%)
Animal	found in taxon(P703) (87%), instance of (P31) (8%), depicts (P180) (3.6%), category combines topics (P971) (1.04%), parent taxon (P171) (0.09%), is a list of (P360) (0.07%), part of (P361) (0.3%)
Artist	occupation (P106) (98.3%), depicts (P180) (1.1%), instance of (P31) (0.4%), is a list of (P360) (0.2%)
Scientist	occupation (P106) (99.99%), instance of (P31) (0.01%)

Table 1. Type Subsumption properties for Wikidata classes [1]

links (green). Thus, some of the instances of these two classes in DBpedia are not instances of the corresponding `owl:equivalentClass` in Wikidata.

4 Conclusion and Future Work

This paper presented an isomorphic approach to infer type subsumption relations in Wikidata with the help of DBpedia. This approach can be extended to any two arbitrary KGs sharing equivalent classes and some equivalent instances. The results obtained in this study can be used as a starting point of further research on discovering potential errors or violations in the content of KGs. Next, we will explore the implicit type information stored in these KGs and contribute towards their completeness by predicting the type information using structural embeddings.

References

1. Directory with all achieved results. https://github.com/ISE-AIFB/Wiki_DB
2. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked Data Quality of DBpedia, Freebase, Opencyc, Wikidata, and YAGO. Semantic Web 9(1), 77–129 (2018)
3. Fleischhacker, D., Paulheim, H., Bryl, V., Völker, J., Bizer, C.: Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection. In: ISWC. pp. 357–372 (2014)
4. Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., Hellmann, S.: Wikidata through the Eyes of DBpedia. CoRR abs/1507.04180 (2015)
5. Koutraki, M., Preda, N., Vodislav, D.: Online Relation Alignment for Linked Datasets. In: ESWC. pp. 152–168 (2017)
6. Melo, A., Paulheim, H., Völker, J.: Type Prediction in RDF Knowledge Bases Using Hierarchical Multilabel Classification. In: WIMS. p. 14 (2016)
7. Paulheim, H., Bizer, C.: Type Inference on Noisy RDF Data. In: ISWC. pp. 510–525 (2013)
8. Ringler, D., Paulheim, H.: One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co. In: KI. pp. 366–372 (2017)
9. Wienand, D., Paulheim, H.: Detecting Incorrect Numerical Data in DBpedia. In: ESWC. pp. 504–518 (2014)