Discovering Identity Links between Entities on the Semantic Web

Ondrej PROKSA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
ondrej.proksa@gmail.com

In our times, few millions of specific pages increase on the World Wide Web daily, from which Linked Data can be obtained. Linked Data appear on the Web in different form and goal of gathering structured data is a better possibility of device processing.

Currently, for the purpose of connecting entities between different datasets, the relationship of identity (also known as owl:sameAs) is widely used. However, it turns out, that in some cases the claim, that two entities are identical, is incorrect [3]. The problem might be caused when we have people who create the relationship with their namesakes or there are different entities that are connected with relationship of identity, but their properties are not fully identical. In the case, when two entities are identical, it is possible to create new relationships, which enrich Linked Open Data cloud with relationships across the datasets.

Our main goal is to discover relationships of identity between entities in order to create new connections between existing data sources and datasets in the Linked Data Cloud. Our goal is to propose a method, that will automatically search for connections and owl:sameAs relationships using graph algorithms and specific rules. We use similarities from sub-graphs of properties and classes for determination of identity between two entities. In the case that entities are identical, it is possible to enrich new relationships, which are across the datasets in sub-graphs.

Our proposed method is universal to any particular domain, because it uses sub-graph of properties and classes for entities comparison. Method is applicable on any particular graph, in which similarity relationships between the entities exist. It is possible to use the method in the search for identical organizations among the public government data as well as in discovering duplicate authors with data from digital libraries and also for connecting a new dataset to the cloud of Linked Open Data.

Because the Linked Data datasets use various ontologies to describe their content, there is a problem of ontology diversity, which could cause that identical entities are not connected using owl:sameAs. The method performs ontology alignment [4] on

* Supervisor: Michal Holub, Institute of Informatics and Software Engineering

Spring 2014 PeWe Workshop, March 21, 2014, pp. 35–36.

each of the found graph patterns. Finally, it aggregates similar ontology classes and properties. The method was evaluated on 4 datasets form the Linked Open Data cloud and using this method the authors were able to discover new, missing relationships between the datasets.

We propose a method for finding similarities between entities in a graph. We use this similarity to determine whether two entities are identical or not. This determines whether they should be linked using owl:sameAs relationship. We based our method on a hypothesis that the matching of entities is reflected in the similarity between the sub-graphs composed of classes and properties of the individual entities. This approach was also explored in the ontology matching problem [1].

The similarity between entities (sig. SGN) depends on the similarity of their properties, graph distance between entities and graph distance between neighbouring entities. We define the total similarity as a sum of similarities of its individual components:

- The similarity of properties between entities
- Distance between the entities
- Average distance between adjacent entities

The resulting values of the similarity are from the interval <0, 1>. SGN = 1.0 means that the two entities are 100% similar (i.e. identical), where as SGN = 0.0 means that the two entities are not similar at all (their similarity is 0%).

We have developed a prototype and evaluated it on a dataset from domain digital libraries. We analyzed the new dataset Annota [2], that was created using the principles of linked data. We tried to find a duplicities authors comparing our method. We have proposed two possibilities find candidates for same authors.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 167-172.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Aumueller, D., Do, H.H., Massmann, S., Rahm, E.: Schema and Ontology Matching with COMA++. *Proc. of the 2005 ACM SIGMOD International Conf. on Management of Data*. SIGMOD '05, New York, NY, USA, ACM, 2005.

[2] Bieliková, M., Ševcech, J., Holub, M., Móro, M.: Annota - poznámkovanie dokumentov v prostredí digitálnych knižníc. *Proc. of the Annual Database Conf.*.

[3] Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameAs isn't the same: an analysis of identity in linked data. *Proc. of the 9th international semantic web conference on The semantic web* - Volume Part I. ISWC'10, Berlin, Heidelberg, Springer-Verlag, 2010.

[4] Zhao, L., Ichise, R.: Graph-based Ontology Analysis in the Linked Open Data. *Proceedings of the 8th International Conference on Semantic Systems*. I-SEMANTICS '12, New York, NY, USA, ACM, 2012.