

Slovak Web-based Encyclopedia

Patrik GAJDOŠÍK*

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

The Web contains a huge amount of unstructured or semi-structured information of different qualities floating around. In order to search through them and use them effectively, the concept of Semantic Web comes to aid with transforming this chaos into structured machine-readable data. After creating these blocks of knowledge and finding the right relations in between them the cloud of Linked Data is growing, allowing to search for the data in more specific and precise way.

One of the great resources of information on the Web nowadays is Wikipedia. Although it contains huge amounts of data, it comes in an unstructured form. One of the great ontologies that focuses on transforming the information from Wikipedia infoboxes (the boxes usually placed on the right side of the Wikipedia articles) to a structured form is DBpedia [1]. It uses metadata (mappings) to make sense of the data on Wikipedia. However, different mappings are needed for different language mutations of Wikipedia. Mappings for less spread languages such as Slovak are not available. By creating these mappings and using them to extract more information we could enrich the Web of Data with new (even language specific) information.

We propose a method that tries to create the mappings automatically by using mappings of different languages that already exist. The foundations of our method are based on the work in [2] that covers the area of mapping a class of an infobox to a class in the DBpedia ontology. Similar work has also been done in [3] where they add more than 1 million entities with high precision and recall.

In order to automatically create mappings, two part process takes place. The first part is to assign each infobox a class in the DBpedia ontology. Then, the second step is to map individual attributes of the infoboxes to the DBpedia predicates.

For the mapping of infoboxes to classes, our method looks at all the possible classes to which a certain infobox could be mapped and decides which one it will pick for Slovak mapping. To decide the class of one infobox, all of the resources that use it need to be gathered. Then, for each Slovak resource, resources in different language mutations are obtained together with their class that was given to them. One of these classes may or may not be the class to which the Slovak infobox should be mapped.

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

Our method then uses different scoring metrics to find the right mapping. After this, the mapping for Slovak language is generated in a form used by DBpedia and can be used in extraction process of DBpedia.

The mapping of attributes of infoboxes to the predicates in the DBpedia ontology is a harder task to perform. If a value is tied to a Slovak resource through a property (DBpedia extracts raw properties that have the same name as an attribute of the infobox and are in separate namespace), our method looks whether in other language mutations there also exists a property that ties it to the same value. Hopefully, this property will be a part of the DBpedia ontology which is what are we looking for. System of points is put to work to decide the right mapping. Again, after this process, the mapping is generated in the form used by DBpedia.

We evaluate our method by building our own dataset of manually created mappings that cover wide range of different cases that could be problematic for our method such as infoboxes with small amount of articles using them, infoboxes with articles that use much richer hierarchy in different languages, infoboxes with small or big amount of attributes. We then compare results of our method with the built expected results. Except the exact matches we also take into account how close to the correct result our method is.

Since we can bend our method and use it for whichever language mutation, we also evaluate by creating mappings for languages that already have their mappings created and compare the results with the already existing ones.

Extended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 140-145.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

- [1] Lehmann, J., Isele, R., Jakob, M. et al.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2014.
- [2] Aprosio, A.P. et al.: Automatic Mapping of Wikipedia Templates for Fast Deployment of Localised DBpedia Datasets. In: *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies - i-Know '13*, New York, New York, USA, ACM Press, 2013, pp. 1–8.
- [3] Palmero Aprosio, A., Giuliano, C., Lavelli, A.: Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information. In Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S., eds.: *The Semantic Web: Semantics and Big Data*. Volume 7882 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 397–411.