Discovering Links between Entities on the Web of Data

Ondrej Proksa*

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

A few million unique websites appear on the Web every day. Information on them is usually published in an unstructured format. Linked Data is structured data which contains entities and relationships between them, which are available on the Web. Some datasets are made via automatized processing of freely available data. These are useful for personalization, web search or for knowledge deduction. One of the main problems is the conversion from various unstructured datasets to a uniform format and the linking of the data to existing datasets.

The ontologies behind Linked Data sources, however, remain unlinked. They describes an extensional approach to generate alignments between these ontologies [1]. They present an extension of the YAGO knowledge base with focus on temporal and spatial knowledge. It contains nearly 10 million entities and events, as well as 80 million facts representing general world knowledge [2]. The goal is to automatically construct and maintain a comprehensive knowledge base of facts about named entities, their semantic classes, and their mutual relations as well as temporal contexts, with high precision and high recall [3].

In this work we analyze the issue of mining structured data from various sources available on the Web and the issue of linking the mined data in order to create a domain knowledge base. We analyze various approaches to automatized dataset creation, gathering information about named entities and linking of the entities and integration of new datasets with the existing ones. We propose a method to automatically process chosen sources of unstructured data and create a structured knowledge base, which is based on the Linked Data principles.

The designed method is experimentally evaluated on data from chosen domain by implementing a software prototype, which uses the knowledge base for a chosen problem from the field of Web personalization - search, navigation, recommendation based on relationships between entities. We validate the created knowledge base by comparing it to other existing knowledge bases.

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 105–106.

We divide our work into these parts:

1. Creating structured data – selection of data source, discovering facts about entities
2. Creating a dataset – identifying relationships, elimination of duplicate entities, linking entities in created dataset, linking dataset with selected existing datasets
3. Verification – evaluation of facts about entities, automatic answering of search queries

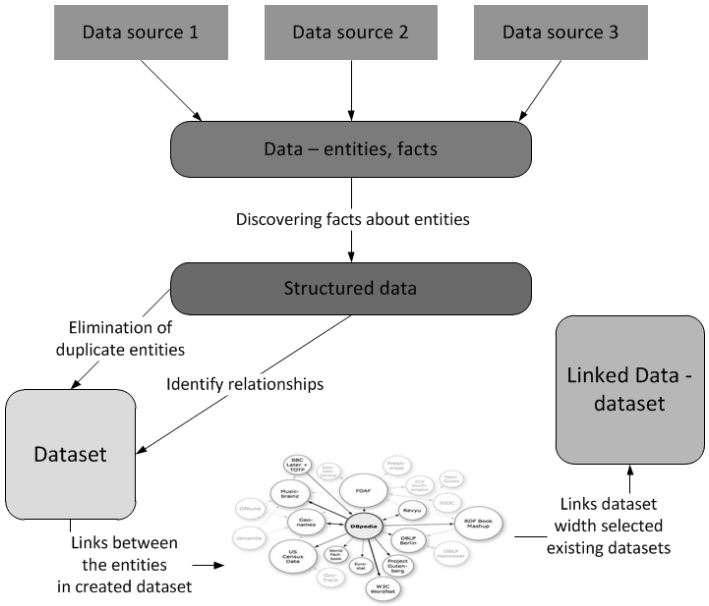

Figure 1. Our work divided into parts

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

References

[1] Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite. 2010. Linking and building ontologies of linked data. In *Proc. of the 9th Int. Semantic Web Conf. on The Semantic Web - Volume Part I* (ISWC'10), Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, and Lei Zhang (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 598-614.

[2] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proc. of the 20th Int. Conf. Companion on World Wide Web* (WWW '11). ACM, New York, NY, USA, 229-232.

[3] Gerhard Weikum and Martin Theobald. 2010. From information to knowledge: harvesting entities and relationships from web sources. In *Proc. of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (PODS '10). ACM, New York, NY, USA, 65-76.