Preprocessing Linked Data in order to Answer Natural Language Queries

Peter MACKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
pmacko@outlook.com

Nowadays, the Web contains a lot of webpages providing information intended to be processed by humans. Many of these pages need data storage with dynamical data loading and for these reasons they use object-relational databases. But what they don't have is fully linked content. The Semantic Web is based on a different concept. In the world of sematic datasets there are many semantic databases which are linked together. Therefore, we can use these databases for answering more complex queries than using traditional keyword-based search engines. The easiest way for this user is to ask for information in natural language. Remember, how many times you have written a sentence like "How to do something". This type of query is now rare on the Web.

There already has been some research done in the field of querying data using natural language on classical databases [1]. In field of semantic databases there is too some methods like [2, 3].

Pre-processing is a key part of the natural language interface, as we mentioned earlier, therefore it is also in the method we propose. We scan the whole dataset and create two lexicons: Classes and properties lexicon, Values lexicon.

The first lexicon is based on structural part of our dataset. It consists of the names, labels etc. of all classes and properties. Next, all structural parts are decorated with synonyms from WordNet, which allows us to formulate query using different words than the ones used in the dataset. We call these alternative names *descriptors* and we provide ranking based on their source.

The second lexicon, using of which is completely new in our approach and none of the examined methods uses it, consists of property values that were obtained during the pre-processing phase. When the user types a value to his query, this lexicon can navigate us to an object type, which contains this value.

One of the main points of our method is processing of transformation to onto-dictionary which is shown in Figure 1. In this part we use preprocessed lexicons for transformations. Then we use transformation rules to convert user request to SPARQL

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 19–20.

language. In this process we identified modifiers in user query and add them to SPARQL request.

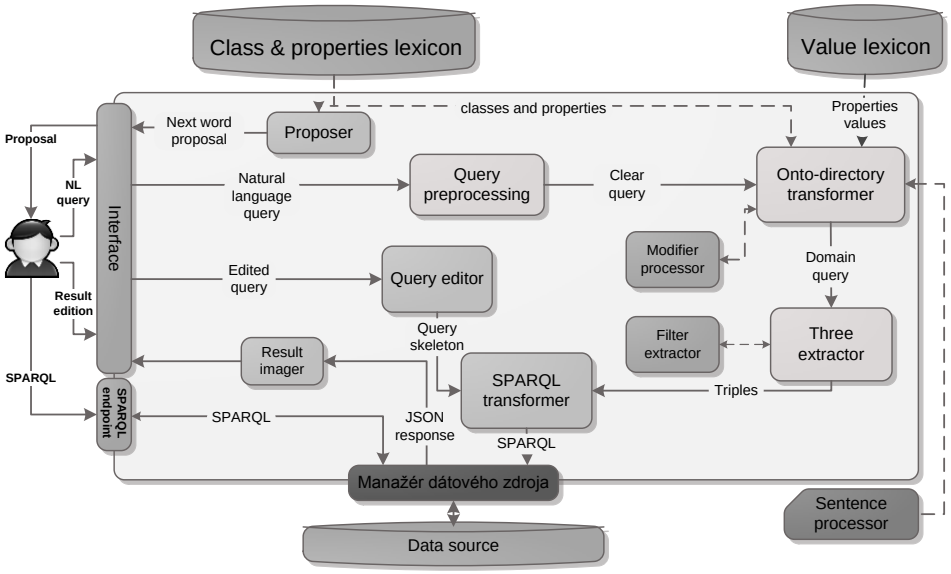

Figure 1 Query processor schema

Next, we plan to evaluate our method with experiment using Annota Firefox extension. We will enhance the ACM digital library web site with our own search engine. We fill our dataset with data produced by Annota which currently has metadata from various digital libraries (ACM, IEEE, etc.) and store them in an ontological dataset.

Amended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 131-136.

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

References

[1] Owda, M., Bandar, Z., Crockett, K.: Conversation-Based Natural Language Interface to Relational Databases. In: *Proc. of 2007 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology - Workshops*. IEEE Computer Society (2007) , pp. 363-367.

[2] Valencia-García, R. et al.: OWLPath: An OWL Ontology-Guided Query Editor. In: *Systems, Man and Cybernetics, Part A: Systems and Humans*. IEEE Systems, Man, and Cybernetics Society (2011), pp. 121-136.

[3] Wang, C., Xiong, M., Qi Z., Yong Y.: PANTO: A Portable Natural Language Interface to Ontologies. In: *4TH ESWC, INNSBRUCK*. Innsbruck, Springer-Verlag (2007), pp. 473-487.