Review of the Doctoral Thesis by RNDr. Tomáš Knap

Title of the thesis: Towards Trustworthy Linked Data Integration and Consumption

Author: RNDr. Tomáš Knap

Reviewer: Ing. Radim Bača, Ph.D.

The work "Towards Trustworthy Linked Data Integration and Consumption" consists of three major topics: (1) introduction of a data fusion algorithm, (2) introduction of a provenance model for the web of data, and (3) definition of a trust model in SoSIReČR. Every topic begins with an extensive description of state-of-the-art approaches accompanied by many examples which help a reader to better understand the main problems. An interesting output of the thesis is that the ideas are implemented in a tool called ODCleanStore which can be used for storing and querying linked data.

Major negative comments and questions:

- Sometimes it is not easy to recognize whether the information contained in the thesis is new or is just preliminaries taken from a previous work. For example, Sections 3.1.3 and 3.1.4 describe transformer classes implementing non-trivial functions related to data quality. I was not able to recognize whether it is a novel approach or not even though there is a paragraph called "Related Work".
- Some parts of Section 5 describing W3P look like a definition of a standard. It seems that the W3P model has a better expressiveness when compared to other models, but under special circumstances an extension of W3P would be necessary. Defining standards is a commendable work, however, is it a research problem? Personally, I am not quite sure, however, I am not an expert in this area. On the other hand, the acceptance of the author's paper (concerning the W3P model) by a prestigious research journal is a good proof that it is an interesting research work.
- Section 6 defines a concept of trust for SoSIReČR. Section 6.3 begins with the definition of several trust beliefs. The author selects several of them based on a questionnaire and evaluates the selection. Nine pages (136-145) are dedicated to a definition and a selection of trust beliefs. My first question is: did you ask in the questionnaire if there is some trust belief missing in the preselected list of five beliefs? Clearly, correspondents considered the preselected beliefs important, but what about the missing ingredients of trust? Did they have an opportunity to answer such question?
- The result is a selection of three trust beliefs, where the thesis says: "The list of sources for the beliefs is not complete …". My second question is: how do you recognize that the list of sources is complete? If I build my trust model on totally different sources and different trust beliefs, how would you disproof my concept? Is there any possibility to compare different trust models? This seems to me to be rather a philosophical disputation with a little connection to computer science.
- Section 6 describes trust too generally and it ends when the problem starts to be really interesting from the computer science perspective. For example, the estimation of trust beliefs

mostly depends on explicit values from users (Section 6.3.5). I feel that some statistical methods for automatic values extraction could be employed which would help to fill missing values from users.

Positive comments:

- The whole thesis is very nicely written with many examples and logical flow of ideas. I really like the example situations which help a lot to understand different concepts.
- Having a journal paper with several external citations in the time of the thesis submission is something that I appreciate. It is a clear indication that this work is interesting for a research community. Even though I have many comments to the thesis (as I stated above) I think that it is an interesting work that I definitely recommend to be defended.
- The topic of the thesis is very up-to-date. Linked data represent a hot topic and, recently, there has been an enormous research effort in this area. It is clear that this work contributes to this effort.
- It is great that the presented ideas are not "only on paper", but there is a tool ODCleanStore which materializes most of the ideas in the thesis. The result is highly practical and it can improve the first experience with linked data for many users.

Minor comments:

p.30, 2nd para - I recommend not to split the URL address using a hyphen, it seems that the URL address is http://db-pedia.org

p.34, Def 2.6 – Definition 2.6 is not properly ended. The sentence “Thus, every quad …” is not a part of the definition.

p.34, Def 2.8 – “Suppose a **named** graph …”

p.34 – It is not quite clear why some definitions are introduced. For example, I was not able to find the usage of the *nodeIn* and *graphIn* functions in the whole thesis.

p. 36, 1st para – It is hard to understand the beginning of the 1st paragraph in Section 2.2.3. I would say that it is mainly by the fact that the term ‘RDF class’ is not defined.

p.36, last para – Property dc:title in not present in Listing 3 even though the following text references it.

p.88 – „…, such **se** where-provenance …“

p.146, explicit source – There is the sentence „…negative honesty relation (u,v,d) – (u,v,d) \in Eph …“. I was really hard for me to realize that it is a hyphen, not a minus.

Conclusion:

The author shows that he has an excellent writing style with a perfect analytical skill of the state of the art. I strongly recommend the thesis to be defended.

In Ostrava, April 30, 2013

Radim Bača

Department of Computer Science

FEECS, VŠB - Technical University of Ostrava

17. listopadu 15, 708 00 Ostrava-Poruba