LA TROBE UNIVERSITY

Dept. of Computer Science & Computer Engineering
School of Engineering & Mathematical Sciences
Faculty of Science, Technology & Engineering

Victoria 3086 Australia
T +61 3 9479 1107
F 03 9479 3060
E info@cs.latrobe.edu.au
latrobe.edu.au/cs

Melbourne, 2 May 2013

Reviewer's Report

Thesis: Towards Trustworthy Linked Data Integration and Consumption
Author: RNDr. Tomas Knap

In this thesis, the author presents his research in the area of Linked Data Management. The motivations of the work are the increasing amount of Linked Data provided by various data providers, the maturity of various Linked Data standards and the increasing usage of Linked Data consumption by data consumers.

The author identifies seven problems that exist in the current Linked Data communities, and that motivate his research. The problems identified are data linkage, data cleansing and transformation, data integration, data provenance, data quality, trustworthiness of agents and trustworthy Linked Data consumption. The author tackles the problems into four research stages, each of which is interconnecting in solving the problems. This report will address the four research stage and how the author undertakes his research in the area.

Development of Linked Data Management Tool (ODCleanStore)

The author's first contribution is the improvement of a Linked Data Management tool called ODCleanStore, which was initially developed by the research group in the Faculty of Mathematics and Physics at Charles University in Prague. The authors ***extends the capabilities of ODCleanStore*** by providing mechanisms for better data linkage, improving data quality, supporting provenance and enhancing trustworthiness of the data.

The author explains the original components of the tool and identifies the work that he contributes in two modules namely Data Processing Module and Query Execution Modules. The former enables automated cleansing, linking and quality assessments of Linked Data. The latter produces data in the formats that can be customised to the data consumers' requirements. The author also provides an early prototype of Linked Data browser that enables data consumers to browse the integrated data.

This part of the thesis has significance importance for the research area, because it has made ODCleanStore as the first Linked Data Management tool that has addressed the seven research problems that the current Linked Data communities are facing.

With maturity of ODCleanStore, it can be used to cleanse, link, integrate public procurement data from Czech national portal of public contracts and can be extended to include European portals of public contracts. Not only for data on public contracts, can the tool also be used for various governmental projects in national and European level.

The tool has also gained interest from other institutions, currently it is further developed with collaboration with research groups in Germany and Italy.

Data Fusion Methodology for Linked Data

The author's second contribution is ***proposing customizable data fusion algorithm*** that is used to integrate RDF data as well as generate quality score for the integrated data. The definition, algorithm, step-by-step process and experimentations in ODCleanStore are clearly presented.

ABN 64 804 735 113
CRICOS Provider 00115M

When it is applied on the ODCleanStore, the proposed data fusion algorithm offers two benefits. First, it improves the quality of the integrated datasets by providing mechanism to solve conflicts among the original datasets. Second, it enables customization of the data fusion algorithm by the data consumers, so that the integrated result really suits the need of the data consumers.

This part of the thesis has significance importance for the research area because the author proposes what to be the first data fusion algorithm available for Linked Data Management. The proposed algorithm is complemented with the ability to determine the quality score of the integrated data and the customizable feature by the data consumers.

There are two aspects that the author can clarify to improve the section:

- The author claims that the data fusion algorithm provides the provenance metadata and thus, contributes to the trustworthy Linked Data consumption (page 167). However, this claim is not clearly explained in chapter 4.
- For completeness, in section 4.3, the author is recommended to discuss s-conflicting quads in terms of definition and algorithm. How complex will it be to consider this type of conflicts to the current algorithm, which only considers o-conflicting quads?

Data Provenance Methodology for Linked Data

The author's third contribution is ***proposing provenance model for Web data*** named W3P. Prior to proposing the model, the author identifies a set of requirements for data provenance for the Web. Each component of the model is formally explained, and the author also demonstrates how to build the model and to apply it in a case study.

The author applies the proposed W3P model into ODCleanStore during data filtering process. It has answered the problems of the provenance of Linked Data as well as the trustworthiness of data consumption. In addition, the proposed model can also be used by data consumers to express their data consumption requirements.

This part of the thesis has significance importance for the research area because the author proposes one of earlier model for Wed data provenance, almost at the same time W3C works on the problem through its W3C Provenance Group.

W3P model is built over core Linked Data Standards and has an ability to reuse other vocabularies. Hence, it can easily be adopted by any tool and application that uses the same Linked Data standards. In addition, W3P model is applicable for any applications that use Web data and thus, not only used for Linked Data Project nor can only be implemented in OCDCleanStore tool. Finally, stated as a future work, W3P model can be aligned with PROV-O, the standard for Web data provenance model proposed by W3C.

Computing Trust in Social Trust Beliefs Network

The author's fourth contribution is ***proposing a trust model formalization*** applied for a national social network project called SOSIReCR. The author identifies relevant trusting beliefs for the project and enables the calculation of the beliefs using trust metrics.

This work has significance importance in improving the trustworthy property of SOSIReCR, which was identified as one major problem in the current state of the network.

While the work in formalization of trust model seems to have contribution, there is no explicit explanation on how the author applies this model in the Linked Data project and how it will be integrated with three previous contributions. There is no justification or demonstration on how ODCleanStore can use the proposed trust model to address two of the problems namely trustworthiness of agents and trustworthy Linked Data consumption. It is recommended that the author clearly explains how this last component of the thesis is integrated with the rest of the contributions.

Conclusion

This thesis contributes to the Linked Data Management, in particular the development of a system for Linked Data integration and consumption, the proposal of data fusion methodology, the proposal of a data provenance model and the proposal of a trust model. Most of the works have been published by the author. Despite minor problem in the last section/last contribution, the examiner believes that **the author of the thesis has proven to have ability for creative scientific work.** The reviewer thus recommends the thesis to PhD defense.

Dr Eric Pardede
Department of Computer Science and Computer Engineering
La Trobe University, Melbourne AUSTRALIA 3083
Email: E.Pardede@latrobe.edu.au
Ph: +61 3 9479 3459 – Fax: +61 3 9479 3060