Noisy Semantic Data Processing in Seoul Road Sign Management System

Zhisheng Huang[1], Jun Fang[2], Stanley Park[3], and Tony Lee[3]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands
[2] School of Automation, Northwestern Polytechnical University, China.
[3] Saltlux Inc., Seoul, Korea

Abstract. The Seoul Road Sign Management (RSM) is a system which provides the semantic integration of LOD's Linked Geo Data and Open Street Map with Korean POI data set. That is an attempt to develop intelligent road sign management system based on the LarKC platform. The RSM data set contains over 1.1 billion triples of semantic data. However, significant amount of the RSM data are noisy (e.g., inconsistent, partial, or erroneous). We have facilitated the RSM system with the capability of processing and reasoning with noisy semantic data, so that the RSM system is robust enough to return intended answers in spite of the poor quality of the semantic data.

1 Introduction

With ever growing scales of semantic data, the noises in those large scale data sets are increasing significantly. We have observed this trend in the Seoul Road Sign Management (RSM)[4], a system which provides the semantic integration of LOD's Linked Geo Data and Open Street Map with Korean POI data set. That is an attempt to develop intelligent road sign management system based on the LarKC platform[4][2].

Effective management of road signs requires processing the directions that are given on each road sign together with a large amount of urban-related information. The Open Street Map (OSM) creates free editable maps of the world. Wikipedia provides many POI (Point of Interest) descriptions. In the LarKC project, we have investigated a data integration solution for urban information that provides a basis for intelligent road sign management. This solution supports data modeling and the integration of massive amounts of linked geo-data, POI data, and road sign data, as well as scalable querying and reasoning.

The existing RSM contains over 1.1 billion triples of semantic data. The large scale of street map data were converted from the Web automatically or semi-automatically by running some script programs, or created manually. That results in the poor quality of the data. Actually, within selected ranges on the Seoul street maps, we have found that average 10 percent of the street data are

[4] http://www.larkc.eu

noisy. Namely, they are either inconsistent, duplicated, or have other types of errors.

We have facilitated the RSM system with the capability of processing and reasoning with noisy data by developing various methods of noisy data checking and processing with heuristic rules. In this demo, we will show that this extended RSM system is robust enough to return meaningful answers in spite of the poor quality of the semantic data.

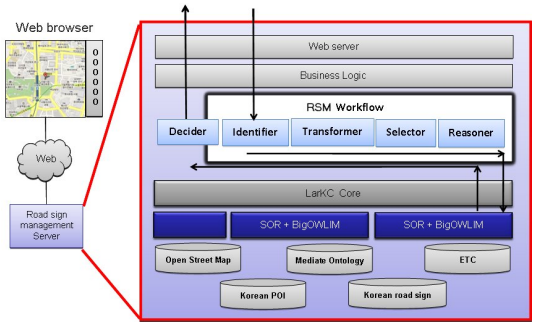

Fig. 1. RSM System

2 Noises in the RSM Semantic Data Sets

The architecture of the RSM system is shown as Figure 1. In the RSM system, users use the web interface to post the operation requirements to the RSM server. Those operation requirements include validation checking on selected road signs, noisy data checking, finding a path (without or with noisy data processing, etc.). The RSM server sends the SPARQL queries to the SPARQL end point, which is launched by the RSM workflow on the LarKC platform. The SOR+BIGOWLIM is located at the data layer of the LarKC platform to maintain the RSM data sets which include the Open Street Map data, the RSM ontologies, Seoul Road Sign Data, and POI data, etc. In this demo, we will show that how the LarKC platform is used for massive semantic data processing and reasoning in RSM.

We have observed the following kinds of noises in the RSM system. 1) Partial data. Figure 2(a) shows an example of partial or uncertain data in which there are two links are crossed each other, however, there is no any junction with those two links. It is unclear whether or not the junction is missing in this scenario, or there is a bridge on the links such that there is no any junction with those two links. 2) Inconsistent Data. Figure 2(b) shows such an example in which a junction appears on a POI (e.g. a building). 3) Duplicated Data. Figure 2(c) shows the case in which the same ID is assigned to two different nodes. 4) Erroneous Data. Figure 2(d) shows an example of noisy links in which there are

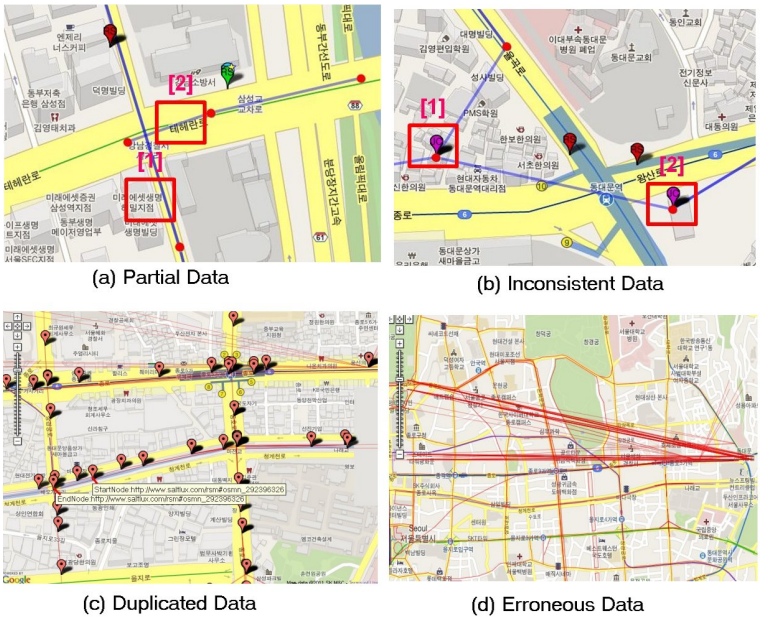

Fig. 2. Noises in RSM

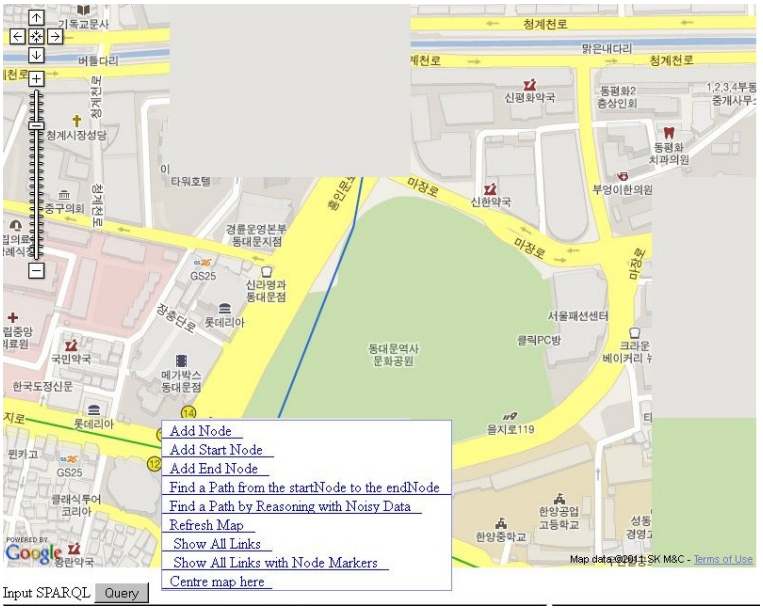

Fig. 3. Screenshot of the RSM Interface

many meaningless links appear on the map. Those links are not connected with any road.

3 Processing and Reasoning with Noisy Data

We introduce heuristic rules to deal with those domain-specific noises. Those heuristic rules can be either provided by domain experts as default rules for noisy data processing or by users who detect a special noise, or obtained by machine learning methods. For each case of known noises in the RSM, we have created the corresponding heuristic rules for noisy data processing and reasoning. Those heuristic rules are used for reasoning with noisy data, which is similar with the methods of reasoning with inconsistent ontologies[3]. Given a reasoning query, the system selects relevant part of data, then launches the noise checking procedure to detect and remove the noises in the data. After obtaining a clean data, the system uses the standard reasoners to reason with the clean data and return the answers.

This demo will show various methods of noisy data processing in RSM. Furthermore, it will show different results of queries with and without noisy data processing. The screenshot of the RSM interface is shown in Figure 3. In an experiment in which 40 node pairs are selected randomly for finding paths between the pairs, we found that the average rate of inconsistent links, missing links and duplicated nodes are 7.8%, 72.1% and 3.5%. In such noisy environment, we have detected that the system fails to find any path in almost half cases (47.5%) without noisy data processing. On the contrary, the system can find the path in almost all the cases (95%) with the support of noisy data processing[5].

4 Conclusion

This demo will show how the LarKC platform can be used for massive semantic data processing and reasoning in RSM, and why the noisy data processing is so useful to obtain intended answers.

Acknowledgement: The work reported in this paper was partially supported by the EU-funded LarKC project.

References

1. Zhisheng Huang at el. D4.7.3 - final evaluation and revision of reason plug-ins, Sept 2011. Available from: http://www.larkc.eu/deliverables/.
2. Dieter Fensel et al. Towards larkc: A platform for web-scale reasoning. In *Proceedings of the International Conference on Semantic Computing*, pages 524–529, 2008.
3. Z. Huang, F. van Harmelen, and A. ten Teije. Reasoning with inconsistent ontologies. In *Proceedings of IJCAI'05*, pages 454–459, 2005.
4. Tony Lee, Stanley Park, Zhisheng Huang, and Emanuele Della Valle. Toward seoul road sign management on the larkc platform. In *Proceedings of ISWC2010*, 2010.

[5] The details of the methods and the evaluations can be found in [1].