Reasoning with Noisy Semantic Data

Qiu Ji[1], Zhiqiang Gao[1,⋆], and Zhisheng Huang[2]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China
{jiqiu,zqgao}@seu.edu.cn

[2] Department of Mathematics and Computer Science, Vrije University Amsterdam
huang@cs.vu.nl

1 Problem Statement

Based on URIs, HTTP and RDF, the Linked Data project [3] aims to expose, share and connect related data from diverse sources on the Semantic Web. Linked Open Data (LOD) is a community effort to apply the Linked Data principles to data published under open licenses. With this effort, a large number of LOD datasets have been gathered in the LOD cloud, such as DBpedia, Freebase and FOAF profiles. These datasets are connected by links such as owl:sameAs. LOD has gained rapidly progressed and is still growing constantly. Until May 2009, there are 4.7 billion RDF triples and around 142 million RDF links [3]. After that, the total has been increased to 16 billion triples in March 2010 and another 14 billion triples have been published by the AIFB according to [17].

With the ever growing LOD datasets, one problem naturally arises, that is, the generation of the data may introduce noise, thus hinders the application of the data in practice. To make the Linked Data more useful, it is important to propose approaches for dealing with noise within the data. In [6], the authors classify noise in Linked Data into three main categories: accessibility[1] and derefencability[2] w.r.t. URI/HTTP, syntax errors, and noise[3] and inconsistency[4] w.r.t. reasoning. In our work, we focus on dealing with the third category of noise, namely noise and inconsistency w.r.t. reasoning, but may also consider other categories of noise. We further consider one more noise in the logical level, that is, the logical inconsistency caused by ontology mapping.

2 State of the Art

In [6], a comprehensive study of various kinds of noise in Linked Data have been conducted over 149,057 URIs concluding 54,836 valid RDF/XML document. In

⋆ Corresponding author.

[1] The problem of accessibility here means some of the retrieved documents have no structured data or contain misreported content-types.

[2] Dereferencing means providing information about a resource lookup of its URI using HTTP.

[3] The noise can be atypical use of OWL/RDFS vocabularies, or use of undefined classes and properties, and so on.

[4] Inconsistency here means logical inconsistency in OWL ontologies.

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 497–502, 2011.

[1], an approach was proposed to detect accidental syntactic errors or vocabulary misuse and then apply patches to produce OWL DL ontologies. As owl:sameAs has been heavily used in LOD to connect different data resources, it becomes more and more important to use it correctly. That is, any two URI references connected by owl:sameAs should be the same thing. But in reality, the correctness can not be ensured. Therefore, the authors in [5] explored the origins of this situation and developed an Similarity Ontology by systematizing various theoretically-motivated distinctions which are 'kind of close' to owl:sameAs.

Compared with other kinds of noise in Linked Data, there have been much work on dealing with logical contradictions in OWL ontologies (see [2] for a survey). Given an inconsistent ontology, one can either use an inconsistency-tolerant approach to reasoning with it (e.g. [7]) or repair it (e.g. [16]). To provide some additional information to deal with inconsistency, some researchers have proposed to measure inconsistency in an OWL ontology [10]. Logical inconsistency can also occur when mappings among ontologies are established [11].

In our work, we will mainly focus on noise w.r.t. reasoning in Linked Data. We will enhance the state of the art in the following aspects. First, we will consider learning expressive ontologies from noisy LOD datasets and provide methods to measure the noise. Second, we will enhance existing approaches to handle the inconsistency in learned ontologies by using some patterns and defining novel inconsistency-tolerant approaches. Third, we will propose measures to evaluate inconsistent mappings and novel methods to repair mappings. We explain each of the three aspects in detail in the following section.

3 Proposed Approach and Methodology

3.1 Statistical Reasoning with Noisy Linked Data

Developing an ontology is not an easy task and often introduces noise and incompleteness. This happens in LOD datasets as well. The ontologies in LOD datasets are generally inexpressive and may contain a lot of noise. For example, one of the most popular ontology, DBpedia ontology[5], is claimed as a shallow ontology. The TBox of this ontology mainly includes a class hierarchy.

To deal with the incomplete ontologies, we plan to use statistical relational learning(SRL) techniques to learn expressive ontologies from LOD datasets (more details can be seen in [19]).

Before learning ontologies, we will propose methods to measure the noise of a data, which can be defined according to the data quality assessment [13] in database area. For instance, an objective metric can be defined as the degree to which misused vocabularies from all vocabularies in an ontology. Such kind of measures provides a reference to decide whether a dataset needs to be cleaned or not. If it is necessary to do cleaning, we could apply various cleaning strategies to correct or remove the noise. For example, we can correct the misused

[5] `http://wiki.dbpedia.org/Ontology`

vocabularies manually with the help of an ontology editor like Protege[6]. After ontology learning, we can define the measure of ontology incompleteness which is the degree to which axioms are missing from the ontology.

3.2 Handling OWL Inconsistency in Linked Data

After learning expressive ontologies from LOD datasets and linking them with other datasets, we may confront the problem of inconsistency[7] handling. According to [6], it may be quite difficult to deal with this kind of noise. Due to the large scale of the data, it is hard to apply existing approaches for reasoning with inconsistent OWL ontologies to deal with OWL inconsistency in Linked Data.

In [6], the authors suggested that, to handle inconsistency in Linked Data, inconsistent data may be pre-processed with those triples causing inconsistencies dropped according to some heuristic measures. We fully agree with this proposal. One measure that we will consider is the inconsistency measure defined by four-valued semantics (see [10]). In the open philosophy of the Web, it may be not desirable to completely repair inconsistent ontologies. One reason, as suggested in [6], is that contradiction could be considered as a 'healthy' symptom of different opinion. Therefore, when we repair inconsistency in OWL ontologies in Linked Data, our goal is not to result in fully consistent ontologies, but to reduce the inconsistency degrees of those ontologies. After that, we can apply some inconsistency-tolerant approaches to reasoning with those inconsistent ontologies. To partially repair an inconsistent ontology, we will apply some patterns to efficiently detect the sources of inconsistency, such as patterns given in [18]. To provide inconsistency-tolerant reasoning services, we will further develop the idea of using selection functions [7] to reasoning with inconsistent ontologies. The idea is to propose specific selection functions for specific ontology languages.

3.3 Mapping Repair and Evaluation

As reported in [8], LOD datasets are well connected by RDF links on the instance level. But on the schema level, the ontologies are loosely linked. It is interesting to consider aligning these ontologies based on the plentiful resources of LOD datasets. A few such approaches have been proposed in [8,12].

With the mappings generated, we may confront the problem of dealing with inconsistency caused by mappings and ontologies if we interpret mappings with OWL semantics. We will first consider evaluating the inconsistent mappings[8] by defining a nonstandard reasoner. We will then consider mapping repair based on work in [14,11]. For example, we can apply some patterns to efficiently detect problematic correspondences in the mappings.

[6] `http://protege.stanford.edu/`

[7] A data is inconsistent iff it has no model.

[8] An inconsistent mapping means no concepts in $O_1 \cup O_2$ are interpreted as empty but there is such a concept in the union of O_1, O_2 connected by $\mathcal{M}$.

3.4 Evaluation

To evaluate our work, we will implement our proposed approaches and do evaluation over LOD datasets. Based on our previously developed system RaDON[9], which is a tool to repair and diagnose ontology networks, we will develop a system for reasoning with noisy Linked Data.

4 Results

We have studied repair and diagnosis in ontology networks and developed a tool, called RaDON, to deal with logical contradictions in ontology networks (see [9]). The functionalities provided by RaDON have been implemented by extending the capabilities of existing reasoners. Specifically, the functionalities include debugging and repairing an inconsistent ontology or mapping, and coping with inconsistency based on a paraconsistency-based algorithm.

In [15], we proposed possibilistic extension of OWL to deal with inconsistency and uncertainty in OWL ontologies. Some novel inference services have been defined and algorithms for implementing these inference services were given. We have implemented these algorithms and provided evaluations for their efficiency.

For an inconsistent mapping, the semantic precision and recall defined in [4] meet the trivialization problems. To resolve such kind of problems, we define the meaningfulness of an answer given by an inconsistency reasoner: Given two ontologies O_1 and O_2 and a mapping $\mathcal{M}$ between them, for a correspondence $c = \langle e, e', r, \alpha \rangle$, an answer provided by an inconsistency reasoner is meaningful iff the following condition holds: $\Sigma \Vdash t(c) \Rightarrow (\exists \Sigma' \sqsubseteq \Sigma)(\Sigma' \not\models e \sqsubseteq \bot \ \ and\ \Sigma' \not\models e' \sqsubseteq \bot \ \ and\ \Sigma' \models t(c))$. Here, e and e' are atomic concepts. r is a semantic relation like equivalence and α is a confidence value. t is a translation function to transfer a correspondence to a DL axiom. Σ is the union of O_1, O_2 and a set of axioms obtained by translating all correspondences in $\mathcal{M}$ to DL axioms. An inconsistency reasoner is regarded as meaningful iff all of the answers are meaningful. Based on this definition, we can redefine semantic measures in [4].

5 Conclusions

Reasoning with noisy Linked Data is a quite challenging and interesting work. In our work, we mainly consider the following work: (1) We will propose methods for measuring noisy LOD datasets like incompleteness and clean the noise in these datasets if necessary. Based on the plentiful LOD datasets, we will propose methods for learning expressive ontologies using SRL techniques. (2) To deal with logical inconsistency, we propose to partially repair an inconsistent ontology by considering some patterns to achieve good scalability for LOD datasets. Then we plan to apply some novel inconsistency-tolerant reasoning strategies like defining specific selection functions for specific ontology languages. (3) We will propose methods for evaluating inconsistent mappings and methods to repair inconsistent mappings by applying some patterns.

[9] http://neon-toolkit.org/wiki/RaDON

Acknowledgements

This paper is sponsored by NSCF 60873153 and 60803061.

References

1. Bechhofer, S., Volz, R.: Patching syntax in OWL ontologies. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 668–682. Springer, Heidelberg (2004)
2. Bell, D., Qi, G., Liu, W.: Approaches to inconsistency handling in description-logic based ontologies. In: Chung, S., Herrero, P. (eds.) OTM-WS 2007, Part II. LNCS, vol. 4806, pp. 1303–1311. Springer, Heidelberg (2007)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. In: IJSWIS, pp. 1–22 (2009)
4. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: IJCAI, Hyderabad, India, pp. 348–353 (2007)
5. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
6. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: LDOW, Raleigh, NC, USA (2010)
7. Huang, Z., van Harmelen, F., ten Teije, A.: Reasoning with inconsistent ontologies. In: IJCAI, pp. 454–459. Morgan Kaufmann, San Francisco (2005)
8. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 402–417. Springer, Heidelberg (2010)
9. Ji, Q., Haase, P., Qi, G., Hitzler, P., Stadtmüller, S.: RaDON — repair and diagnosis in ontology networks. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 863–867. Springer, Heidelberg (2009)
10. Ma, Y., Qi, G., Hitzler, P.: Computing inconsistency measure based on paraconsistent semantics. Journal of Logic and Computation (2010)
11. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Repairing ontology mappings. In: AAAI, pp. 1408–1413 (2007)
12. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Linking and building ontologies of linked data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 598–614. Springer, Heidelberg (2010)
13. Pipino, L., Lee, Y.W., Wang, R.Y.: Data quality assessment. ACM Commun. 45(4), 211–218 (2002)
14. Qi, G., Ji, Q., Haase, P.: A conflict-based operator for mapping revision. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 521–536. Springer, Heidelberg (2009)
15. Qi, G., Ji, Q., Pan, J.Z., Du, J.: Extending description logics with uncertainty reasoning in possibilistic logic. International Journal of Intelligent System (to appear, 2011)

16. Schlobach, S., Huang, Z., Cornet, R., van Harmelen, F.: Debugging incoherent terminologies. J. Autom. Reasoning 39(3), 317–349 (2007)
17. Vrandecic, D., Krotzsch, M., Rudolph, S., Losch, U.: Leveraging non-lexical knowledge for the linked open data web. Review of AF Transactions, 18–27 (2010)
18. Wang, H., Horridge, M., Rector, A.L., Drummond, N., Seidenberg, J.: Debugging OWL-DL ontologies: A heuristic approach. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 745–757. Springer, Heidelberg (2005)
19. Zhu, M., Gao, Z.: SRL based ontology learning from linked open data. In: ESWC PhD Symposium,Crete, Greece (to appear, 2011)