Storage and Reasoning Systems Evaluation Campaign 2010

Mikalai Yatskevich[1], Ian Horrocks[1], Francisco Martin-Recuerda[2], and Giorgos Stoilos[1]

[1] Oxford University Computing Laboratory,
Wolfson Building, Parks Road,
Oxford OX1 3QD, UK
mikalai.yatskevich@comlab.ox.ac.uk
ian.horrocks@comlab.ox.ac.uk
giorgos.stoilos@comlab.ox.ac.uk
[2] Universidad Politcnica de Madrid,
Campus de Montegancedo,
sn 28660 Boadilla del Monte, Spain
fmartinrecuerda@fi.upm.es

1 Evaluation Criteria

According to the ISO-IEC 9126-1 standard [5], interoperability is a software functionality sub-characteristic defined as "the capability of the software product to interact with one or more specified systems". In order to interact with other systems a DLBS must conform to the standard input formats and must be able to perform standard inference services. In our setting, the standard input format is the OWL 2 language. We evaluate the standard inference services:

- Class satisfiability;
- Ontology satisfiability;
- Classification;
- Logical entailment.

The last two are defined in the OWL 2 Conformance document[3], while the first two are extremely common tasks during ontology development, and are de facto standard tasks for DLBSs.

The performance criterion relates to the efficiency software characteristic from ISO-IEC 9126-1. Efficiency is defined as "the capability of the software to provide appropriate performance, relative to the amount of resources used, under stated conditions". We take a DLBS's performance as its ability to efficiently perform the standard inference services. We will not consider the scalability criterion for the Storage and Reasoning Systems Evaluation Campaign 2010 because suitable test data is not currently available. The reason for this is the fact that while hand crafted ontologies can be tailored to provide interesting tests, at least for particular systems it is very difficult to create hand crafted

[3] http://www.w3.org/TR/2009/PR-owl2-conformance-20090922/

Proceedings of the International Workshop on Evaluation of Semantic Technologies (IWEST 2010). Shanghai, China. November 8, 2010.

ontologies that are resistant to various optimizations used in modern systems. Furthermore, hand crafted ontologies are rather artificial since their potential models often are restricted to those having a very regular structure. Synthetic DL formulas may be randomly generated [4, 7, 9]. Thus, no correct answer is known for them in advance. There have been extensive research on random ABox generation in recent years [2, 6]. These works are tailored to query answering scalability evaluation. Real-world ontologies provide a way to assess the kind of performance that DLBSs are likely to exhibit in end-user applications, and this is by far the most common kind of evaluation found in recent literature. However, it is not a trivial task to create a good scalability test involving real-world ontologies. To the best of our knowledge, no such tests are available at the moment. The particular problems are parametrization and uniformity of the input.

2 Evaluation Metrics

The evaluation must provide informative data with respect to DLBS interoperability. We use the number of tests passed by a DLBS without parsing errors is a metric of a system's conformance to the relevant syntax standard. The number of inference tests passed by a DLBS is a metric of a system's ability to perform the standard inference services. An inference test is counted as passed if the system result coincides with a "gold standard". In practice, the "gold standard" is either produced by a human expert or computed. In the latter case, the results of several systems should coincide in order to minimize the influence of implementation errors. Moreover, systems used to generate the "gold standard" should be believed to be sound and complete, and should be known to produce correct results on a wide range of inputs.

The evaluation must also provide informative data with respect to DLBS performance. The performance of a system is measured as the time the system needs to perform a given inference task. We also record task loading time to assess the amount of preprocessing used in a given system. It is difficult to separate the inference time from loading time given that some systems perform a great deal of reasoning and caching at load time, while others only perform reasoning in response to inference tasks. Thus, we account for both times reflecting the diversity in DLBSs behavior.

3 Evaluation Process

We evaluate both interoperability and performance for the standard inference services. Our evaluation infrastructure requires that systems either conform to the OWL API 3 [3]. The output of an evaluation is the evaluation status. The evaluation status is one of the following TRUE, FALSE, ERROR, UNKNOWN. TRUE is returned if ontology and ontology class are satisfiable and in the case the entailment holds. FALSE is returned if ontology and ontology class are unsatisfiable and in the case the entailment does not hold. ERROR indicates a

parsing error. UNKNOWN is returned if a system is unable to determine an evaluation results.

4 Testing Data

Our collected data set contains most of the ontologies that are well established and widely used for testing DBLS's inference services. More precisely, it contains:

- The ontologies from the Gardiner evaluation suite. This suite now contains over 300 ontologies of varying expressivity and size. The test suite was originally created specifically for the purpose of evaluating the performance of ontology satisfiability and classification of DLBSs [1]. It has since been maintained and extended by the Oxford University Knowledge Representation and Reasoning group[4], and has been used in various other evaluations (e.g., [8]).
- Various versions of the GALEN ontology [10]. The GALEN ontology is a large and complex biomedical ontology which has proven to be notoriously difficult for DL systems to classify, even for modern highly optimized ones. For this reason several "weakened" versions of GALEN have been produced by system developers in order to provide a subset of GALEN which some reasoners are able to classify.
- Various ontologies that have been created in EU funded projects, such as SEMINTEC, VICODI and AEO.

We use the OWL 2 conformance document as a guideline for conformance testing data. In particular, we aim at semantic entailment and non-entailment conformance tests. 148 entailment tests and 10 non-entailment tests from the OWL 2 test cases repository[5] are used for evaluating a DLBS's conformance.

5 Conclusion

We have provided a general framework for evaluating advanced reasoning systems. We also described Storage and Reasoning Systems Evaluation Campaign 2010 evaluation design, including criteria and metrics definitions, test data, test workflows and API methods to be implemented by evaluation participants. Furthermore, we described the ontologies to be used as testing data in the evaluation campaign.

References

1. T. Gardiner, I. Horrocks, and D. Tsarkov. Automated benchmarking of description logic reasoners. In *Proc. of the 2006 Description Logic Workshop (DL 2006)*, volume 189, 2006.

[4] http://web.comlab.ox.ac.uk/activities/knowledge/index.html
[5] http://owl.semanticweb.org/page/OWL_2_Test_Cases

2. Y. Guo, Z. Pan, and J. Heflin. Lubm: A benchmark for owl knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):158 – 182, 2005.
3. M. Horridge and S. Bechhofer. The OWL API: A Java API for Working with OWL 2 Ontologies. In Rinke Hoekstra and Peter F. Patel-Schneider, editors, *OWLED*, volume 529 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
4. I. Horrocks, P. F. Patel-Schneider, and R. Sebastiani. An analysis of empirical testing for modal decision procedures. *Logic Journal of the IGPL*, 8(3):293–323, 2000.
5. ISO/IEC. *ISO/IEC 9126-1. Software Engineering – Product Quality – Part 1: Quality model.* 2001.
6. L. Ma, Y. Yang, Z. Qiu, G. T. Xie, Y. Pan, and S. Liu. Towards a complete OWL ontology benchmark. In *ESWC*, pages 125–139, 2006.
7. F. Massacci and F. M. Donini. Design and results of tancs-2000 non-classical (modal) systems comparison. In *TABLEAUX '00: Proceedings of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods*, pages 52–56, London, UK, 2000. Springer-Verlag.
8. B. Motik, R. Shearer, and I. Horrocks. Hypertableau reasoning for description logics. *J. of Artificial Intelligence Research*, 2009. To appear.
9. P. F. Patel-Schneider and R. Sebastiani. A new general method to generate random modal formulae for testing decision procedures. *J. Artif. Intell. Res. (JAIR)*, 18:351–389, 2003.
10. A. L. Rector and J. Rogers. Ontological and practical issues in using a description logic to represent medical concept systems: Experience from galen. In *Reasoning Web*, pages 197–231, 2006.
11. D. Tsarkov, I. Horrocks, and P. F. Patel-Schneider. Optimizing terminological reasoning for expressive description logics. *J. Autom. Reasoning*, 39(3):277–316, 2007.
12. T. D. Wang, B. Parsia, and J. Hendler. A survey of the web ontology landscape. In *Proceedings of the International Semantic Web Conference, ISWC*, 2006.