Evaluating Semantic Technology: Towards a Decisional Framework

Roberto Armenise, Daniele Caso, and Cosimo Birtolo

Poste Italiane S.p.A.
TI - RS - Centro Ricerca e Sviluppo - 80133 Napoli, Italy
{armenis5,casodan2,birtoloc}@posteitaliane.it

Abstract. Time required to access data and query knowledge base is one of the most important parameter in designing information system. When the size increases, the complexity of an ontology makes reasoning and querying processes less efficient and very time consuming. Although performances are crucial, other features, such as the type of license, the availability of a related community support or the ease of adoption of a particular technology are often key elements in a decision-process of an industrial designer. This paper proposes an evaluation of main semantic technologies in terms of different metrics at varying ontology sizes. The evaluation aims at building a concrete framework for supporting industrial designers of ontology-based software systems to make proper decisions, taking into account the scale of their knowledge base and the main metrics.

Keywords: benchmark, ontology, OWL, semantic, reasoning.

1 Related Works

In spite of their usefulness for automatic reasoning in complex software systems, ontology-based technologies still have a limited adoption in industrial contexts where performance plays a crucial role among non-functional attributes. Performance is considered the real Achilles' heel of ontologies, especially when their size grows. Indeed, the larger the ontology size, the higher the complexity that makes less efficient both reasoning and querying activities. Nevertheless, ontology-based semantic technologies could have a wider diffusion if their performances would be related to the application contexts where they are proposed.

In literature, many benchmarks have been developed in order to test performance of semantic technologies and different metrics arise. A widely used benchmark for comparing the performance, completeness and soundness of OWL reasoning engines is the Lehigh University Benchmark (LUBM), proposed by Guo et al. [1,2]. LUBM consists of a university domain ontology, customizable and repeatable synthetic data, a set of test queries, and several performance metrics developed to facilitate the evaluation of Semantic Web repositories and to evaluate the performance of those repositories over a large data set. Garcia et al. [3] denote the importance of evaluation of quality parameters as a guide to

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 677–680, 2012.

choose an ontological tool. They considered usability, defined as the measure of the ease of use of the ontology tools including aesthetic factors and compliance with usability and accessibility guidelines. In the current state-of-the-art, several implementations of semantic technologies are available either from opensource communities or from commercial vendors. Among them, the main opensource semantic projects can be represented by: Apache Jena, openRDF Sesame and Minerva; while in the commercial offer we can find: Oracle semantic technology, openLink Virtuoso and Ontotex OWLIM.

2 Comparing Semantic Technologies

The aim of this work is to build a complete and exhaustive decisional framework able to guide designers of industrial information systems into the huge plethora of semantic technologies. For this reason, we decide to evaluate the leading semantic technologies (choosing among commercial and opensource offer) on the basis of a set of qualitative and performance metrics. In particular, we tested

The adopted benchmark model is univ-bench LUBM [1,2]. We use Data Generator tool (UBA tool) in order to create three different ontologies: (i) LUBM(1,0) consisting of 15 departments and 1 university, (ii) LUBM(5,0) consisting of 5 universities and 93 departments, and (iii) LUBM(10,0) with 10 universities and 189 departments. In order to evaluate the inference capability and scalability of ontology systems under test, we use a set of 13 queries, defined by Guo et al.[1] and we take into account the following metrics: (i) **Specifications** which is a set of qualitative non-functional characteristic supplied by each technology, (ii) **Load Time** which measures the elapsed time to load the ontology model into storage repository, (iii) **Query Response Time** which is the time elapsed to issue the query, to obtain the result set and to traverse that set sequentially, and (iv) **Query Completeness** which is the degree of completeness of each query answer and is evaluated by the ratio between the number of the provided answers and the expected ones.

Results: We investigate[1] the different technologies by means of three directions: (i) Specifications, (ii) Load Time, (iii) Query Response Time and Completeness. Collecting together results of the qualitative and quantitative evaluations, we will define a decisional tool that gives a general overview of each technology.

Specifications are introduced to investigate some important qualitative features non-related to the performance requirements. We take into account: (i) the type of license, (ii) the documentation supporting the deployment phase, (iii) the existence of a support community, (iv) the ease of installation and (v) the ease of use of the technology itself. We express a rating for each metric according to our experience in adopting, installing and testing these technologies. The ratings

[1] Experiments were carried out on Intel Xeon E5650 QuadCore 2.67GHz machine with 32 GB of RAM running Windows Server 2003 R2 Enterprise x64 Edition Service Pack 2 with Java 1.6 update 26 64-bit. We performed all the execution with Java VM's heap space set between 2GB and 24GB.

are integer numbers on a -1("poor")-to-1("good") scale. Collecting the scores of different feature, we obtain that: Jena reaches a total score of 4, Sesame 3, Minerva −3, Oracle 2 (good in Documentation and Ease of use), Virtuoso 2 (good in Ease of installation and Ease of use), and OWLIM 3.

Investigating load time for the different semantic technologies, we prove that the larger the dataset size, the higher the load time. This kind of analysis lets us to make a choice of one technology rather than another one depending on the frequency of loading tasks expected in our software design.

In order to investigate query response time and completeness, we define a fitness function of a technology x: $fitness(x) = \frac{1}{Q} \cdot \sum_{q=1}^{Q} \sqrt{\left(1 - \frac{t_q(x)}{T_q}\right) \cdot C_q(x)}$ where q is the query, $t_q(x)$ is the average time (10 different runs) requested by the technology x to answer the query q, T_q is the maximum time requested for the execution of a target query q, $C_q(x)$ is the completeness of the answer provided by the technology x after the question q, and Q is the number of queries. Investigating the fitness value per technology, we can state that Minerva and OWLIM outperform in the three case studies. Indeed, these two technologies guarantee a completeness equal to 100% thus entails that they are able to provide all the expected answers in a shorter response time.

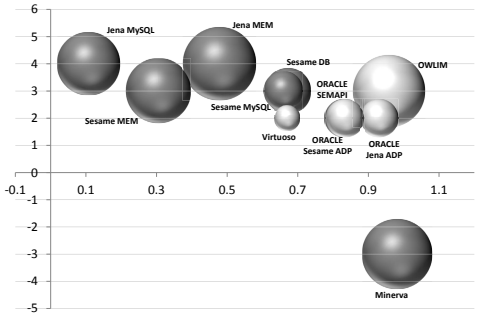

Fig. 1. Decisional framework: Investigating fitness (x axis), load time (dimensions of the bubbles), qualitative features (y axis) and license type (opensource (dark-grey bubbles), proprietary (light-grey bubbles))

The various analysis performed are able to support the choice of a semantic technology. Indeed, industrial designers of ontology-based software systems have to take into account different requirements that arise from the application they need to design.

In order to provide a decisional framework, Fig. 1 depicts our findings[2]. Each bubble refers to a semantic technology. Load time is represented by the dimension d of the bubbles (where $d = 1 - \frac{loadTime}{maxAllowedTime}$). In detail, the largest

[2] *MEM* refers to memory-based approach; while *DB* and *MySQL* are two different database-based approaches; *SEMAPI*, *Sesame ADP* and *Jena ADP* are different adapters of Oracle technology.

bubbles represent the technologies with the shortest loading time. X axis represents the average fitness value, while y axis represents the sum of qualitative ratings ranging in the [-4;4] interval. Fig. 1 is able to support industrial designers of ontology-based software systems in the selection process of a semantic technology. The proper choice is strongly related to the expected performance requirements, available budgets in terms of licence costs and seniority of available developers. Indeed, qualitative features which include well-supported technology and extremely comprehensive, detailed and up to date documentation can be a key element in some decisional processes. For instance, if a designer is looking for a higher level of performance and good qualitative features, he can choose OWLIM; instead, if qualitative features are not so crucial but he needs higher performances, he can choose Minerva.

3 Conclusion and Future Work

In this paper we investigate different knowledge management technologies: (i) Jena, (ii) Oracle with different adapters, (iii) SESAME, (iv) Virtuoso, (v) Minerva and (vi) OWLIM. Performance and completeness are key metrics in some decisional processes, but qualitative features such as the license, the provided documentation, the support of related communities are very important at the same time. Moreover, the easiness in installation and use play a relevant role too. Experimental results investigated different directions: (i) Specifications, (ii) Load Time, (iii) Query Response Time and Completeness. The different criteria address the problem of the selection of the proper technology according to desired characteristic of the target application. As next steps, we plan to investigate semantic technology behavior increasing dataset size and examining the relative strengths and weaknesses of these different technologies when different ontology-based software systems have to be designed.

Acknowledgments. This work was partially supported by MIUR under the MODERN Project PON01-01949. The authors would like to thank Eugenio Zimeo, whose contributions made this work possible.

References

1. Guo, Y., Pan, Z., Heflin, J.: 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)
2. Guo, Y., Qasem, A., Pan, Z., Heflin, J.: A requirements driven framework for benchmarking semantic web knowledge base systems. IEEE Transactions on Knowledge and Data Engineering 19(2), 297–309 (2007)
3. García-Castro, R., Gómez-Pérez, A.: Guidelines for Benchmarking the Performance of Ontology Management APIs. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 277–292. Springer, Heidelberg (2005)