Recently, the W3C Linking Open Data effort has boosted the publication and interlinkage of larger amounts of RDF/S datasets on the Semantic Web (SW). Various ontologies and knowledge bases with millions of RDF/S triplets from Wikipedia and other sources have been created and are available online. In addition, numerous data sources in e-science are published nowadays as RDF/S graphs, most notably in the area of life sciences to facilitate community annotation and interlinkage of both scientific and scholarly data of interest. Finally, Web 2.0 platforms are considering RDF/S as a non-proprietary exchange format for the construction of information mashups. It is clear that the increasing number and size of the available SW datasets presents today a real challenge for Semantic Web systems in order to cope with scalability and performance concerns. We must build systems that process terabytes of data, have response times on the order of seconds or less, and rely on reasoning to solve problems not easily solved before. If the Semantic Web cannot rise to the challenge of the available data, it may be dismissed as a broken technology.

For the purpose of our discussion, we consider Semantic Web systems to be those systems that consume and/or produce information encoded in RDF(S) or OWL. When we apply the term scalability to these systems, we mean that the systems must be capable of executing these tasks with a performance level that is acceptable to users. Defining adequate performance targets is one of the major challenges for the development of SW systems that integrate reasoning with manipulation of large scale RDF/OWL ontologies and their instances. In particular, many traditional reasoning algorithms raise serious scalability concerns because they do not rely on secondary storage and appropriate indexing techniques.

In this special issue, we have included four articles (out of 11 submissions) covering a wide range of techniques for benchmarking or enhancing the scalability of Semantic Web systems.

In “The Berlin SPARQL Benchmark,” Christian Bizer and Alexander Schultz propose a new benchmark focused on evaluating the SPARQL performance on knowledge base systems. Unlike prior Semantic Web benchmarks such as the LUBM (Guo, Pan, & Heflin, 2005) and UOBM (Ma, Yang, Qiu, Xie, Pan, & Liu, 2006) this work focuses less on reasoning capabilities and more on efficient support of SPARQL features such as OPTIONAL, ORDER BY, UNION, REGEX, CONSTRUCT and the ability to have variables in the predicate position. In addition, they compare the performance of three popular RDF stores to two SPARQL-to-SQL rewriters. The main conclusion drawn from the conducted experiments is that none of the benchmarked systems was superior within the single client use case for all queries and all dataset sizes. Another observation is that
the overall performance (100M triple, single client, all queries) of the fastest SPARQL-to-SQL rewriter with the fastest relational database exhibits an overhead for query rewriting of 106%. This is an indicator that there is still room for improving the rewriting algorithms. Finally, the authors clearly identify the need for more robust SPARQL optimizers.

“Learning of OWL Class Descriptions on Very Large Knowledge Bases” by Sebastian Hellmann, Jens Lehmann and Sören Auer introduces a technique for learning summaries of a set of instances. These summaries could be used for many purposes, including analysis of instance data, improving data quality and supporting navigation through large knowledge bases. This article is included here because it applies machine learning techniques to very large knowledge bases such as DBpedia. This scaling is enabled by a pre-processing step that selects a subset of the knowledge base that is likely to be relevant to the task.

In “Scalable Authoritative OWL Reasoning for the Web,” Aidan Hogan, Andreas Harth, and Axel Polleres devise an algorithm for scalable reasoning that is complete for a fragment of OWL. Scalability is achieved by using file scan and sorting as the basic primitives. One unique emphasis of this article is that the algorithm ignores axioms that they consider non-authoritative.

Finally, “Enabling Scalable Semantic Reasoning for Mobile Services” by Luke Steller, Shonali Krishnaswamy and Mohamed Gaber considers the problem of semantic reasoning on mobile devices such as smart phones and PDAs. Standard Tableaux reasoning cannot be executed due to need for mobile device applications to be very computationally cost efficient and resource aware. The authors show that their system is faster than RacerPro and Pellet on three different ontologies, and is comparable to Fact++. Furthermore, when resources become low, it is capable of reducing result accuracy in order to proceed with the limited resources. Although this article does not deal with data sets of the size considered in the first three articles, we consider it relevant to scalability because it addresses both the ubiquity of data in the future of the Semantic Web and also scaling techniques to fit on mobile devices.

We hope that these articles will stimulate your thinking on the problem of scalability in the Semantic Web, and we encourage you to contemplate your own solutions to the challenges that lay before us.

Vassilis Christophides and Jeff Heflin would like to thank the following reviewers of this special issue: Jan Paralic, Bijan Parsia, Peter Patel-Schneider, Jerome Euzenat, Raúl García Castro, Irini Fundulaki, Natasha Noy, Giorgos Flouris, Jan Wielemaker, Yannis Tzitzikas, Li Ding, Grigoris Antoniou, and Achille Fokoue

Vassilis Christophides and Jeff Heflin

REFERENCES


ENDNOTES

1. esw.w3.org/topic/SwoeIG/TaskForces/CommunityProjects/LinkingOpenData
2. www.w3.org/TR/wordnet-rdf, www.dbpedia.org
5. simile.mit.edu/wiki/RDFizers
Vassilis Christophides studied electrical engineering at the National Technical University of Athens (NTUA), Greece. He received his DEA in computer science from the University PARIS VI and his PhD from the Conservatoire National des Arts et Metiers (CNAM) of Paris, France. He is an associate professor at the Department of Computer Science, University of Crete, and affiliated researcher at the Information Systems and Software Technology Laboratory of the Institute of Computer Science Foundation for Research and Technology-Hellas (FORTH-ICS). His main research interests include Semantic Web information management systems, semistructured data models and query languages as well as description and composition languages for e-services. He has published over 70 articles in international conferences and journals and has served on numerous conferences program committees (ACM SIGMOD, VLDB, EDBT, WWW, ISWC, ICWE, ICWS, ECDL). He has received the 2004 SIGMOD Test of Time Award and the Best Paper Award at the 2nd and 6th International Semantic Web Conference in 2003 and 2007.

Heflin is an associate professor in the department of Computer Science and Engineering at Lehigh University. His specific research interests include establishing semantic interoperability between heterogeneous information systems, scalable ontology reasoning, and developing formal theories of distributed ontology systems. He has been involved in the design of many important Semantic Web languages, including SHOE, DAML+OIL, and OWL. In addition to serving on numerous program committees, he was the demonstrations chair at the second International Semantic Web conference, the chair of the special track on AI and the Web at FLAIRS 2004, a guest editor for IEEE Internet Computing, and is on the editorial board of the Journal of Web Semantics. Dr. Heflin received his BS in computer science from the College of William and Mary. He received his M.S. and PhD in computer science from the University of Maryland.