Properties of real-world RDF Datasets and SPARQL Queries

François Picalusa and Stijn Vansummeren
Université Libre de Bruxelles
CoDE - Web and Information Technology

October 26, 2010

RDF is a widely used schema-less graph data model, with datasets hundreds of millions triples large. It is accompanied by the SPARQL query language for querying RDF data based on graph pattern matching.

In this talk we discuss the results of some experimental analysis of both the RDF datasets available in practice, and the SPARQL workloads used to query them. We believe that such analysis may prove useful in defining more efficient query processing algorithms. Besides improvements to the query processors themselves, knowledge of the data and queries is also important for the generation of synthetic data and query workloads; the design of efficient physical storage schemes; and the design of indexes.

Other than general statistics on the size of the RDF data, we discuss the underlying topology in terms of node degrees; clustering coefficient; and diameter. On all of these aspects, RDF data shows strong similarities with the existing complex-networks models considered e.g. for Web Graphs.

With respect to the properties of SPARQL queries we discuss how each feature of the language is used in the workloads, as well as the topology of the queries themselves. In particular, most of the queries are conjunctive and acyclic, in the sense of classical relational database theory.