Efficient Querying for the Semantic Web

François Goasdoué and Ioana Manolescu
fg@lri.fr, ioana.manolescu@inria.fr

April 2, 2010

We witness an increasing convergence of efforts in the areas of data management and Semantic Web. This is due partly to the recent interest of the database community in the management of unstructured semantic Web data (such as RDF) and on the other hand, to the increasing availability of models, tools and techniques for handling knowledge (around the RDF-S and OWL technologies, as well as on the fundamental side around suitable Description Logics models for such technologies).

Research carried in the Leo group (http://leo.saclay.inria.fr) has focused on the efficient querying of Web data described by distributed knowledge. These works have lead in particular to building the SomeWhere distributed inference platform and the ViP2P distributed RDF and XML data management platform, both validated by large-scale experiments on several hundreds of peers. This research work is related to the ANR projects WebContent (ended in 2009) and CODEX (2009-2012).

The subject we propose seeks to further the understanding of the possible interplay between query processing and Semantic Web-style reasoning. We consider in particular the efficient querying of RDF data, since RDF is by now widely accepted as a standard for Web data. Several lines of work are envisioned in this context :

1. **Automated selection of materialized views** RDF data lacks structure and state-of-the-art storage models are quite primitive. When a query workload is known, we seek to exploit it by adding a set of data access support structures (under the form of indices or materialized views) in order to improve the efficiency of query answering. Knowledge about the RDF data sources, under the form of an RDF-S schema, can also be used as an input to the view selection process.
2. **Large-scale distributed RDF data management** Large volumes of RDF data require distributed platforms to handle them. We envision two distribution models. The first one is in the context of structured peer-to-peer networks, which could be realized as an extension to the ViP2P platform (http://vip2p.saclay.inria.fr). While some algorithms have been previously proposed for disseminating and querying RDF in such contexts, their efficiency has not been experimentally validated and further optimizations could be brought. The second context is a cloud computing environment, where data is distributed strictly according to the application needs and not produced independently by different sites as in the peer-to-peer context.
3. **OLAP processing for RDF** In a recent proposal, we planned to work on devising a model, languages, and practical algorithms for analytical processing on RDF. Challenges to be addressed in this area concern finding methods for efficiently computing OLAP-style aggregates on this data.

The candidate should have solid background on data and knowledge management systems, on the practical and formal side. The project is likely to involve significant system prototyping, to be carried in collaboration with engineers, researchers, and other students.