Take What You Need: Efficiently Querying Semantic Web Data

Dario Colazzo, François Goasdoué and Ioana Manolescu
Oak *team, Inria Saclay / Database team, Université Paris Sud*
dario.colazzo@lri.fr, francois.goasdoue@lri.fr, ioana.manolescu@inria.fr

The "Web of Data" vision behind the initial World Wide Web project has found its most recent incarnation through the Semantic Web. More and more data sources are being exported or produced as *triples*, using the Resource Description Format (or RDF, in short) model standardized by the W3C [1]. To exploit this wealth of data, the SPARQL query language has been defined [2]; subsequently, novel techniques and algorithms have been proposed for the processing of SPARQL queries, based on indexing [3], efficient join processing [4], and optimized stores [5], to name a few.

The Oak team brings together Inria and University of Paris Sud faculty working on large-scale management of complex data. The team has worked extensively on efficient algorithms for query processing on Web data. A class of techniques of particular interest are based on *data projection*: the idea is to prune (restrict) a data set before querying it, to only a subset of the data which is strictly needed by the query evaluation. Pruning out irrelevant data limits the memory needs of the query processor and typically speeds up query evaluation [6].

The goal of the internship is to devise techniques for *efficient processing of SPARQL queries on RDF data, based on the principle of projection*. As is generally the case for projection-based optimization, the core of the work consists of analyzing the query and possibly other sources of information about the data (e.g., such as expressed for instance through a schema) in order to identify the relevant data subset; the focus is on identifying the tightest possible subset, while ensuring that all necessary data is preserved.

A further subtle point specific to the RDF setting is due to the presence of *implicit information*: an RDF database consists of both explicitly declared data, and of data *implicitly* present into the database, due to semantic constraints that may hold on it. Implicit data requires *reformulating* the query before evaluating it, in order to guarantee that the query results due to the implicit information are correctly accounted for [7]. In some cases, reformulated queries may be syntactically very large, while exhibiting numerous repeated sub-expressions; the nature of reformulated queries must be taken into account when devising RDF projection techniques. A prototype implementation validating the proposed technique on top of existing SPARQL query processors is also expected.

We are currently seeking a Master (or PhD) intern with good background in databases, Web techniques, and/or knowledge representation to join us on this work.
The position is paid by Inria. Our offices are located in the Paris Sud/Inria Saclay building in Gif-sur-Yvette, south of Paris. The position can start at any time, however interested applicants should be available for at least three consecutive months.

How to apply / further questions Send to Ioana Manolescu (`Ioana.Manolescu@inria.fr`) your CV, highlighting the courses taken in databases, and the name of one or two professors that can serve as references.
For more information, see:

- Oak team web site: `http://team.inria.fr/oak`
- Some of our data management projects for the Semantic Web: `http://tripleo.saclay.inria.fr`
- Inria Saclay web site: `http://www.inria.fr/en/centre/saclay`
- LRI, CS lab of Université Paris Sud: `http://www.lri.fr/index_en.php`

References

[1] The World Wide Web Consortium (W3C). Resource description framework. `http://www.w3.org/RDF`.

[2] The World Wide Web Consortium (W3C). SPARQL protocol and RDF query language. `http://www.w3.org/TR/rdf-sparql-query`.

[3] Cathrin Weiss, Panagiotis Karras, and Abraham Bernstein. Hexastore: sextuple indexing for Semantic Web data management. *Proceedings of the VLDB Endowment (PVLDB)*, 1(1), 2008.

[4] Thomas Neumann and Gerhard Weikum. The RDF-3X engine for scalable management of RDF data. *VLDB Journal*, 2010.

[5] François Goasdoué, Konstantinos Karanasos, Julien Leblay, and Ioana Manolescu. View Selection in Semantic Web Databases. *Proceedings of the VLDB Endowment (PVLDB)*, 5(2), October 2011.

[6] Nicole Bidoit, Dario Colazzo, and Federico Ulliana. Type-Based Detection of XML Query-Update Independence. In *Proceedings of the VLDB Endowment (PVLDB)*, 2012.

[7] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Web Data Management and Distribution*. Cambridge University Press, 2011.