Exploratory Analytics for RDF Data

Yanlei Diao (Ecole Polytechnique) Ioana Manolescu (INRIA)

Duration: 5-6 months, the starting date is flexible (ideally March 1st, 2016)
Location: INRIA Saclay / Ecole Polytechnique, Palaiseau, France
Keywords: **Databases**, **Data Analytics**, **Semantic Web**

Context: RDF and the Semantic Web

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 [3]. To exploit this wealth of data, the SPARQL query language has been defined [4], and recently enriched to support complex querying using regular path expressions, grouping and aggregation etc.RDF graph structure tends to be complex. In a given graph, different resources may have different sets of properties; a given property may have zero, one or several values; resources may or may not have types. This makes it difficult for users to get acquainted with the content and structure of a large RDF graph.

RDF Analytics

Recent work [2] has proposed a framework for analyzing RDF graphs, in the spirit of relational data warehouse analytics, but redesigned in order to adapt to the specificities of the RDF data model. At the core of this framework are RDF analytical queries, specifying: a set of resources to be analyzed, the set of analysis dimensions, and optionally aggregation functions to be applied for each set of resources having the same dimension values. For instance, one RDF analytical query may count the number of products of a certain type sold in each store and each month of sale (a typical relational analytical query), but it may also compute the average number of distinct properties for each type represented in the RDF graph. The latter query is over the structure of the graph (which can be seen as its schema), and goes well beyond the expressive power of relational analytical queries.

Internship Goal

The purpose of the internship is to devise and implement an exploratory framework for RDF analytics, starting from the concept of RDF analytical queries and extending it significantly. First, we are interested in novel aggregation measures, allowing for instance to select the "the most diverse" resources having the same values along a set of dimensions, or to sort such resource groups according to a metric measuring their respective homogeneity or diversity.

Second, to assist and facilitate the user's effort in discovering information about an RDF graph, a method should be devised for recommending a series of analytical queries to be evaluated over the RDF graph, each being chosen based on the previous one(s) and user feedback. The goal is to help the user identify analytical query result interesting for her, where interest is signaled by the user either explicitly (e.g., classes, properties, or simply keywords of interest) or implicitly (signaling that a certain analytical query is interesting). RDF analytical query refinements correspond to changes in the classifier, measure, and/or aggregation components of the query; simple operations include analytical cube transformations as defined in [1]. The approach should be implemented and its interest and performance validated through experiments.

The internship will take the form of a full-time INRIA employment contract. The intern will be paid 1100 €/month.

Contacts

- Yanlei Diao (yanlei.diao@polytechnique.edu), http://www.lix.polytechnique.fr/~yanlei.diao
- Ioana Manolescu (ioana.manolescu@inria.fr), http://pages.saclay.inria.fr/ioana.manolescu/

References

[1] Elham Akbari Azirani, François Goasdoué, Ioana Manolescu, and Alexandra Roatis. Efficient OLAP Operations for RDF Analytics. Research Report RR-8668, OAK team, Inria Saclay ; INRIA, January 2015. 1

[2] Dario Colazzo, François Goasdoué, Ioana Manolescu, and Alexandra Roatis. RDF Analytics: Lenses over Semantic Graphs. In *23rd International World Wide Web Conference*, Seoul, South Korea, April 2014. 1

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

[4] The World Wide Web Consortium (W3C). SPARQL protocol and RDF query language. http://www.w3.org/TR/rdf-sparql-query. 1