2018-00758 - Expressive and Efficient Analytics for RDF Graphs

Level of qualifications required: Graduate degree or equivalent
Function: PhD Position

About the research centre or Inria department

Located at the heart of the main national research and higher education cluster, member of the Université Paris Saclay, a major actor in the French Investments for the Future Programme (Idex, LabEx, IRT, EquipeX) and partner of the main establishments present on the plateau, the centre is particularly active in three major areas: data and knowledge; safety, security and reliability; modelling, simulation and optimisation (with priority given to energy).

The 450 researchers and engineers from Inria and its partners who work in the research centre’s 31 teams, the 100 research support staff members, the high-level equipment at their disposal (image walls, high-performance computing clusters, sensor networks), and the privileged relationships with prestigious industrial partners, all make Inria Saclay Île-de-France a key research centre in the local landscape and one that is oriented towards Europe and the world.

Context

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. To exploit this wealth of data, the SPARQL query language has been defined, and recently enriched with regular path expressions, grouping and aggregation etc. RDF graph structures tend to be complex: each resource may have a different sets of properties; a property of a resource may have one or several values; resources may or may not have types. This makes it difficult for users to exploit RDF graph through SPARQL queries.

Recent work (Colazzo et al., WWW 2014) has proposed a framework for analyzing RDF graphs, redefining core data warehouse concepts to adapt them to the specificities of the RDF data model. This framework is based on RDF analytical schemas and analytical queries. Building upon it, (Diao et al, ISWC 2017) makes a first step toward automatically selecting interesting RDF analytical queries, based on statistic properties of the query results.

Assignments

The purpose of this PhD study is to develop a formal model and scalable algorithms for rich, efficient analytics on RDF graphs.

Main activities

We outline three main research objectives below.

1. A first objective is to revisit the analytical schema and analytical query framework and extend it to support more elaborate aggregation measures, allowing for instance to select 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.

2. The second objective consists of devising system-aided exploratory techniques for discovering interesting data in an RDF graph, where “interesting” is defined based on user feedback on a series of exploratory queries. One possible method, building on prior work from a relational setting, can 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 all analytical results interesting for her, which is eventually characterized by a precise, complex analytical query over an RDF graph. Here, the user interest can be signaled either explicitly (e.g., classes, properties, or simply keywords of interest to the user) or implicitly (signaling that a certain answer of an exploratory query is interesting). RDF analytical query refinements correspond to changes in the classifier, measure, and/or aggregation components of the query. The main challenge in designing...
such system-aided exploratory techniques is to capture the user interest quickly through a small set of exploratory queries and/or a small set of query answers with positive user feedback.

3. Efficient algorithms should be devised for: (i) eliciting user's interest in an RDF graph through a series of queries; the output of this process may be for instance an analytical schema; (ii) supporting the query-driven interaction with the user seeking to discover the data fragments most interesting to her; this can be seen as an assistance in formulating (extended) analytical queries. This is a query-intensive process, during which user waiting time should be kept to a minimum; special physical execution operators (or analytical execution paradigms) may be devised to maximize performance; (iii) distributing computations in a massively parallel framework inspired e.g., from Scalla or Spark, given that the large and growing volumes of RDF graphs increasingly require parallel processing.

Skills
Interesting candidates should demonstrate very good knowledge of algorithms and databases, as well as experience with significant-size programming projects. Knowledge of data mining and/or data exploration would be a plus.

Benefits package
- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Flexible working hours
- Sports facilities

Remuneration
1.982 euros brut/month for 2 first year, then 2.085 euros brut/month for the 3rd year