L3 Research Internship Topic in Computer Science

Incremental Schema Inference for the Covid-19 Dataset

Keywords: Big Data, Graph Databases, Schema Inference, Schema Validation, Schema Evolution, Incremental Maintenance, Graph Updates.

Supervision:

- Prof. Angela Bonifati, angela.bonifati@univ-lyon1.fr
  Équipe BD, Laboratoire Liris, Lyon 1 University & INRIA Grenoble Rhône-Alpes, Lyon.

- Dr. Stefania Dumbrava, stefania.dumbrava@ensiie.fr

Setting: The advent of social networks and of highly interconnected data, in a wide variety of areas, ranging from network infrastructures to life science repositories, has led to an increased need to efficiently store and process graph-structured data. To address this, a special type of NoSQL systems, namely graph databases, have been developed. Prominent commercial implementations are Neo4j, TigerGraph, Amazon Neptune, Oracle PGX, etc. The most expressive data model underlying these systems is the property graph one, in which lists of properties can be attached to both the nodes and the edges of a directed, labeled, multi-graph. Due to this rich foundational model and to the fact that graph databases do not impose a rigid schema, graph databases can be employed to analyze complex, heterogeneous data. Recently, the Covid-19 Knowledge Graph ([1]), integrating data from the Genotype Tissue Expression(GTEx) and Covid-19 Disease Map(Cord-19) repositories, as well as from academic article collections, such as ArXiv, BioRxiv, MedRxiv, PubMed, and PubMed Central, has been openly published as a Neo4j graph instance. Given the various sources it integrates and its rapid evolution, the Covid-19 Knowledge Graph is an ideal setting for schema inference and analysis. Indeed, as shown in [2], one can build on MapReduce type inference and on the expressivity of Neo4j’s query language, Cypher, to extract the corresponding schema.

Goal: The goal of this work is to improve the quality of the schema inferred in [2]. To this end, we aim to develop a new inference method, which combines the label and property oriented algorithms described in [2], so as to ensure robustness with respect to both incomplete labeling and to spurious node additions. In addition to this, we set to investigate incremental maintenance approaches that would allow us to preserve the schema validity under updates, inspired by the work done in [3] and [4].
Opportunities:

The selected students will have the opportunity to be involved in the writing of a research article documenting the obtained results.

Requirements:

- Very good programming skills;
- Familiarity with graph databases is not mandatory, but appreciated.

Bibliography:


