Data Science At QCRI

10.5339/qfarc.2014.ITOP1068

Divy Agrawal; Laure Berti; Hosam Hammady; Prasenjit Mitra; Mourad Ouzzani; Paolo Papotti; Jorge Quiane Ruiz; Nan Tang; Yin Ye; Si Yin; Mohamed Zaki

Corresponding author:
lberti@qf.org.qa
Qatar Computing Research Institute, Doha, Qatar

Abstract

The Data Analytics group at QCRI has embarked on an ambitious endeavor to become a premiere world-class research group in Data Science by tackling diverse research topics related to information extraction, data quality, data profiling, data integration, and data mining.

We will present our ongoing projects to overcome different challenges encountered in Big Data Curation, Big Data Fusion, and Big Data Analytics.

(1) Big Data Curation: Due to complex processing and transformation layers, data errors proliferate rapidly and sometimes in an uncontrolled manner, thus compromising the value of information and impacting data analysis and decision making. While data quality problems can have crippling effects and no end-to-end off-the-shelf solutions to (semi-)automate error detection and correction existed, we built a commodity platform, NADEEF that can be easily customized and deployed to solve application-specific data quality problems. This project implements techniques that exploit several facets of data curation including:

- assisting users in the semi-automatic discovery of data quality rules;
- involving users (and crowds) in the cleaning process with simple and effective questions;
- and unifying logic-based methods, such as declarative data quality rules, together with quantitative statistical cleaning methods.

Moreover, implementation of error detection and cleaning algorithms has been revisited to work on top of distributed processing platforms such as Hadoop and Spark.

(2) Big Data Fusion: When data is combined from multiple sources, it is difficult to assure its veracity and it is common to find inconsistencies. We have developed tools and systems that tackle this problem from two perspectives:

(a) In order to find the true value for two or more conflicting ones, we automatically compute the reliability (accuracy) of the sources and dependencies among them, such as who is copying from whom. Such information allows much higher precision than simple majority voting and ultimately leads to values that are closer to the truth.

(b) Given an observed problem over the integrated view of the data, we compute explanations for it over the sources. For example, given erroneous values in the integrated data, we can explain which source is making mistakes.

(3) Big Data Analytics: Data analysis tasks typically employ complex algorithmic computations that are hard/tedious to express in current data processing platforms. To cope with this problem, we are developing Rheem, a data processing framework that provides an abstraction on top of current data processing platforms. This abstraction allows users to focus only on the logics of their applications and developers to provide ad-hoc implementations (optimizations) over existing data processing platforms. We have already created two different applications using Rheem, namely data repair and data mining. Both have shown benefits in terms of expressivity of the Rheem abstraction as well as in terms of query performance through ad-hoc optimizations. Additionally, we have developed a set of scalable data profiling techniques to understand relevant properties of big datasets in order to be able to improve data quality and query performance.