PhD Thesis Proposal

Title : “Missingness-aware and Sample-based Query Processing”

PhD advisors:
Farouk Toumani (ftoumani@isima.fr)
Angela Bonifati (angela.bonifati@univ-lyon1.fr)

Data incompleteness is ubiquitous in real-world applications where data is collected under various circumstances that do not allow to get the necessary missing records. In data management, data incompleteness has interested a large swath of research work [Abi9,Grahne91,Bertossi11,Libkin14], focusing on incomplete databases and the possible worlds semantics.

The problem of incompleteness has also attracted the interest of the statistics community. According to [Rub76], missing data can be classified in three categories: (i) Missing Completely and Random - (MCAR): the missing data are a random subset of data. No relationship between whether a data point is missing and any values in the data set (missing or observed); (ii) Missing at Random — (MAR): The missing data here is affected only by the complete (observed) variables and not by the characteristics of the missing data itself; (iii) Missing Not at Random — (MNAR): the data will be missing based on the missing column itself. All three options are reminiscent of dealing with the problem of data missingness spanning from closed world assumption in databases (in the case of MAR) to open world assumption in databases (MNAR). The latter case is particularly tricky since it deals with missing data that goes beyond the actual content of relational instances.

In the field of statistics, different techniques have been developed to deal with missing data such as deletion (e.g. for MCAR) or data imputation (e.g. for MAR).

The goal of this PhD thesis is to address the problem of query processing over databases affected by the problem of data missingness. As such, it follows a novel research trend aiming at combining databases with statistical approaches:

1. In case of MCAR missing data, the goal is to explore new results [Meliou10,Salimi16] on causality-based explanation together with the notion of tuple responsibility in order to quantify the incompleteness degree of queries answers over incomplete databases, when missing data are ignored by the query evaluation process.
2. In case of MAR missing data, the goal is to investigate query processing in the context of multiple imputations [Rub87,Eub96] of an incomplete database. Given an incomplete database, multiple imputation creates $m$ sets of complete databases where missing values are replaced by imputed values obtained using an imputation process with a
random component. Hence, an incomplete database can be viewed as a finite set of \( m \) complete databases issued from various imputations (in practice, \( m \) usually varies between 5 to 10). A query over an incomplete database can then be translated into \( m \) queries over the associated complete databases, where the final query result is obtained by combining the answers of the \( m \) queries over the complete databases. A question arises on how to formalize and efficiently process queries in this setting.

3. Sample-based query processing [Orr2020] is a new generation of DBMS that can apply query processing techniques to biased sample database instances, as it often occurs in real-world. A third milestone in the thesis will be to tackle the case of MNAR under the closed world assumption. Hence, one objective would be to take into account the incompleteness degrees of database samples and adapt the sample-based query processing techniques accordingly.

References


[Libkin14] Leonid Libkin. Incomplete data: what went wrong, and how to fix it. PODS2014