Discovery of Genuine Functional Dependencies from Relational Data with Missing Values [Abstract for INFORSID 2019]

Laure Berti-Equille, Hazar Harmouch, Felix Naumann, Noël Novelli, Thirumuruganathan Saravanan

Discovery of Genuine Functional Dependencies from Relational Data with Missing Values

Laure Berti-Équille [1,2], **Hazar Harmouch** [3], **Felix Naumann** [3], **Noël Novelli** [1], **Saravanan Thirumuruganathan** [4]

1. *Aix Marseille Université, Univ. de Toulon, CNRS, LIS, DIAMS, Marseille, France*
   laure.berti-equille@univ-amu.fr,noel.novelli@lis-lab.fr
2. *ESPACE-DEV/IRD, UMR 228, IRD/UM/UG/UR, Montpellier, France*

3. *Hasso Plattner Institute, University of Potsdam, Germany*
   hazar.harmouch@hpi.de,felix.naumann@hpi.de
4. *QCRI, HBKU, Doha, Qatar*
   sthirumuruganathan@hbku.edu.qa

ABSTRACT. *This article is an extended abstract of our work published at VLDB'2018. The full paper is available at www.vldb.org/pvldb/vol11/p880-berti-equille.pdf.*
Functional dependencies (FDs) play an important role in maintaining data quality in relational databases. They can be used to enforce data consistency and guide data repairs. In this work, we investigate the problem of missing values and its impact on FD discovery. When using existing FD discovery algorithms, some genuine FDs could not be detected precisely due to missing values and some non-genuine FDs can be discovered even though they are caused by missing values depending on the considered semantics for NULL values. We define the notion of genuineness of FDs and propose algorithms to compute the FD genuineness score. This can be used to identify genuine FDs *among the set of all valid dependencies that hold on the data. We evaluate the quality of our method over various real-world and semi-synthetic datasets with extensive experiments. The results show that our method performs well for relatively large FD sets and is able to accurately capture genuine FDs.*

KEYWORDS: *Functional dependencies, missing values, scoring.*

---

1. Context and motivations

Functional dependencies (FDs) are one of the most important types of integrity constraints and have been extensively studied by the DB research community. FDs have a number of applications, such as maintaining data quality in databases, capturing schema semantics, schema normalization, data integration, repairing of data inconsistencies, and data cleaning. An FD $X \rightarrow A$ states that the tuples of attribute set X uniquely determine the value of attribute (set) A. Traditional FDs are typically defined for correct and complete data and there are many efficient algorithms to discover FDs from a given clean dataset. However, many real-world datasets are neither correct nor complete. Traditional FDs often have trouble with incomplete data, such as NULL values, that routinely exist in massive datasets with well-known data error rates that may vary from 20% up to 80%.

2. Genuine FDs discovery

Despite the importance of this problem, very few work has focused on the critical aspects of FD discovery over incomplete data. In our work, we consider three semantics of missing values: (i) all tuples with at least one NULL value are ignored in computing FDs; for each attribute, we substitute all NULL values either by (ii) the same value (all NULL values are considered equal) or (iii) by distinct values (all NULLs are considered distinct). Then, for each NULL semantics, we study the impact on FD discovery. We formally and experimentally show the phenomenon caused by missing values over FD discovery and we formalize the definitions of *genuine*, *ghost*, and *fake* FDs. Furthermore, we study their impact under various NULL semantics and imputation strategies.

Intuitively, given a clean dataset r and a corresponding dirty dataset r' polluted by injecting missing values in r: A *same FD* is a valid FD in r and also in r'. A *ghost FD* is a valid FD in r becomes invalid in r' while a *fake FD* is an invalid FD in r but is valid in r'. A *genuine FD* is a valid exact FD in r and in r'. To estimate FD genuineness score of FDs in a dirty relation, we propose (i) a probabilistic approach using a given imputation technique for estimating the score and we provide an efficient method for enumerating and pruning irrelevant possible worlds, (ii) we propose efficient algorithms to approximate the genuineness score of discovered FDs: the first one is based on possible worlds using *Monte Carlo* sampling and the second methods are based on probabilities per value and per tuple, using respectively the likelihood that the FD $X \rightarrow A$ which holds for the value $V_X \in Dom(X)$ can identify the value V_A. We estimate genuineness with *PerValue* score as the normalization of the sum of *PerValue* over the number of distinct values in X, and with *PerTuple* score as the normalization of the sum of $|V_X, V_A|$ over the number of distinct tuples. We performed extensive experiments of our methods on real-world (Sensors dataset) and semi-synthetic datasets (Abalone, Computer, Glass and Iris datasets) artificially polluted in a controlled experiments and showed the effectiveness and efficiency of our approach.

Acknowledgements

This work is funded by the French National Research Agency, project QualiHealth, ANR-18-CE23-0002 .