Checking Query Completeness over Incomplete Data

Simon Razniewski
Free University of Bozen-Bolzano
Dominikanerplatz 3
39100 Bozen, Italy
simon.razniewski@stud-inf.unibz.it

Werner Nutt
Free University of Bozen-Bolzano
Dominikanerplatz 3
39100 Bozen, Italy
nutt@inf.unibz.it

We study the problem of *query completeness* (QC) over partially complete databases. Query completeness is an important goal of data quality assurance. Incomplete databases can occur in many contexts, especially whenever many users are supposed to insert data into the same database. Although the overall data can often not be guaranteed to be complete, some information may exist about completeness of parts of the data, e.g., completeness of parts of tables. Given a query, the question arises whether that information implies that the query returns the same set of answers over the available data as it would over the complete database.

Motro [6] formalized *partially complete databases* as pairs of instances $\mathcal{D} = (D^i, D^a)$ with $D^a \subseteq D^i$, where D^i stands for the *ideal* data, which hold in the world, and D^a for the *available* data, which are present in the database. Additionally, Motro defined that a query Q is *complete* over $\mathcal{D}$ if $Q(D^a) = Q(D^i)$. He studied when completeness of queries $Q_1, \ldots, Q_n$ entails completeness of another query Q (QC-QC entailment) and found that rewritability of Q in terms of $Q_1, \ldots, Q_n$ is a sufficient condition.

A formalism for making statements about the completeness of parts of relations was introduced by Halevy [5]. Using *table completeness* (TC) statements, which he called *local completeness* statements, one can say that all tuples of the ideal relation R^i, satisfying a condition involving selections and semijoins with other ideal relations, are already present in the available relation R^a. In this style, one could state that the *film* relation of a film database contains all films produced between 2000 and 2005 starring Johnny Depp. Halevy then studied TC-QC entailment and gave a reduction to a variant of the problem of queries independent of updates [2]. However, for the class of instances resulting from this reduction, no general solution is known.

Denecker et al. [1] investigated TC-QC entailment with respect to a database instance. They showed the problem has coNP data complexity and is coNP-hard for some first order queries and TC statements and developed methods for approximating the certain and possible answers of queries over partially complete databases.

Recently, Fan and Geerts studied the problem of query completeness in the presence of master data [3, 4].

In our work, we review the query completeness (QC) statements introduced by Motro and the table completeness (TC) statements of Halevy. We argue that from a practical point of view, TC statements are a natural way to assert database completeness, while QC statements are the goal of completeness inferences.

We focus on *conjunctive queries* and compare TC and QC statements with respect to their expressivity. We define *canonical TC statements* for a query Q and show that Q is complete over every partial database satisfying the the canonical statements for Q. Furthermore, the canonical TC statements exactly characterize query completeness in the case of projection-free queries or queries under multiset semantics. For queries and statements without comparisons, the canonical TC statements are the weakest TC statements that entail query completeness. Consequently, TC-QC entailment can be reduced to TC-TC entailment in these cases, which itself, as we show, has a natural translation to query containment.

We present a decision procedure for TC-QC entailment with respect to a concrete database and show that this problem is Π_2^P-complete already for conjunctive queries without comparisons and without repeated predicates. In contrast to [1], data complexity is polynomial for conjunctive queries.

For the problem of QC-QC entailment raised by Motro, we show the strong connection to the problem of *query determinacy* [8], a problem for which decidability is yet open for conjunctive queries.

Finally, we discuss practical issues regarding on which grounds table completeness guarantees can be given and how schema constraints can alleviate the process of formulating such statements.

The complete paper has appeared as a technical report [7].

LID 2011 March 25, 2011, Uppsala, Sweden

1. REFERENCES

[1] M. Denecker, A. Cortés-Calabuig, M. Bruynooghe, and O. Arieli. Towards a logical reconstruction of a theory for locally closed databases. *TODS*, 35(3), 2010.

[2] Ch. Elkan. Independence of logic database queries and updates. In *Proc. PODS*, pages 154–160, 1990.

[3] W. Fan and F. Geerts. Relative information completeness. In *Proc. PODS*, pages 97–106, 2009.

[4] W. Fan and F. Geerts. Capturing missing tuples and missing values. In *Proc. PODS*, pages 169–178, 2010.

[5] A.Y. Levy. Obtaining complete answers from incomplete databases. In *Proc. VLDB*, pages 402–412, 1996.

[6] A. Motro. Integrity = Validity + Completeness. *ACM TODS*, 14(4):480–502, 1989.

[7] S. Razniewski and W. Nutt. Checking query completeness over incomplete data. Technical Report KRDB11-2, KRDB Research Centre, Free University of Bozen-Bolzano, 2011.

[8] L. Segoufin and V. Vianu. Views and queries: Determinacy and rewriting. In *Proc. PODS*, pages 49–60, 2005.