Data Quality and Data Completeness

What is Data Quality?
- Data is of a high quality if it is fit for intended uses
- Data quality (DQ) has different aspects: correctness, completeness, accuracy, etc.
- Little work has been done on data completeness

What is Data Completeness?
- A database is complete for a domain if it contains all facts that are true in the domain
- In practice no DB is complete, but a database can be sufficiently complete for a given query e.g., IMDb database does not contain all movies but it contains all movies by Charlie Chaplin

MAGIK at Work: School Database

Scenario
The school administration provides completeness statements that describe which parts of the database are complete

School director creates a statistical report about the pupils in the school
Teacher at the school
System administrator maintains the database

School Database

pupil(name, level, code) ...
... a pupil belongs to a class of certain level and code

learns(name, language) ...
... a pupil learns a language

Plain Reasoning

Statement 1: The school database is complete for all pupils.

Query: Who are the pupils at the 1st level? Can I trust the query answer?

Reasoning under Finite Domain Constraints (FDCs)

In our school every pupil can be in class ‘a’ or class ‘b’. That is a finite domain constraint.

Statement 2: Wait, we made a mistake. The database is only complete for pupils of class ‘a’.

Query: Who are the pupils at the 2nd level? Can I trust the query answer?

Reasoning under Foreign Keys (FKs)

I defined the FDC on table class, not on pupil, so every class[code] is in {a,b}.

Wait, I designed the database! Because statement 2 guarantees completeness for a specific part of the data asked by the query. Other parts might be incomplete.

Statement 3: Now, we are complete for all pupils at class ‘b’. Who are the pupils at the 1st level? Can I trust the query answer?

Implementation

Query is Complete if Query is Complete wrt Canonical Ideal Database

Input

DB (pupil), TDC, FK, TDC

Output

Q(ideal database, Q(ideal database, pupil))

Table Completeness (TC) Statements

pupil(name, level, code)

Table Completeness (TC) Statements

pupil(name, level, code, branch)

Table Completeness (TC) Statements

pupil(name, language)

Query Completeness

Q(ideal (complete) database) = Q(incomplete database)

Formalization of the Problem

Data Quality and Data Completeness

Meta-information about completeness
- Completeness cannot be checked by inspecting the database
- One cannot see what is missing
- We need information about the database completeness state – meta-information
- Often meta-information about the data completeness is available

Information about partial completeness can come from:
- Business Processes that manipulate the data
- Humans assertions (e.g., school administration)
- Origin of the data (data provenance)

System Architecture

Web based application written in JAVA with a GUI that allows one to:
- Create of database schema and to extract a scheme from the database catalog
- Create/modify/add TC-statements, FKs, FDs and Queries
- Four SQL select-project-join queries, possibly with DISTINCT, COUNT(*) and GROUP BY
- Encodes the completeness reasoning into Logic Programming (JESS+SAT Programming), which is executed by DLV engine

MAGIK: Managing Completeness of Data

Ognjen Savkovic
Free University of Bozen-Bolzano, Italy
savkovic@inf.unibz.it

joint work with Sergey Paramonov, Mirza Paramita, and Werner Nutt