Data Quality: State of the Art

What is Data Quality?

- Data Quality (DQ) is a perception of data’s fitness to serve its purpose in a given context.
- DQ problems cost U.S. businesses around $600 billion annually (TDWI Journal).
- DQ problems are characterized by usually independent dimensions, such as completeness, correctness, consistency, accuracy, timeliness... (no silver bullet for all DQ problems).

What is Data Completeness?

- Is all necessary data present?
- Completeness means correctness to which data are of sufficient breadth, depth, and scope for the task at hand (e.g., query answering). A database might be incomplete in general but still sufficiently complete for the task at hand.
- Application domains includes areas like Data Warehousing, data generated by business processes (workflow), etc.

Research Motivation

- Managers see the data throughout Reports/Dashboards
- Data quality is a major concern for decision support data (according to interviews with IT experts in companies and School Administration)
- Diffuse market of enterprise solutions: very expensive, not sufficiently comprehensive, non-standardized solutions

Research Goals

- Take advantage of widely present meta-information that accompanies data (e.g., business processes workflows, ETL processes, master data management systems, etc.) to assess data quality aspects.
- Track back the causes of bad data quality and propose fixes.
- Design and implement algorithms that automate the proposed solutions.

MAGIK at Work: School Information System

Scenario

The school administration (e.g., school personnel or school administrative process) provides Completeness Statements (meta-information) that holds over the existing database (that is in general incomplete).

School Schema

```
pupil(name, level, code) ... pupils
class(level, code, dept) ... every class belongs to a department
language(name, language) ... pupils attend language courses
```

Plain Reasoning

Statement 1: We are complete for all pupils.
```
TABLE: pupil(Name,Level,Class) WHERE:
```
Query 1: Who are the pupils at the 2nd level?
```
SELECT p.name
FROM pupil AS p
WHERE p.level='2'
```

Statement 2: We are complete for all pupils in the class ‘la’.
```
TABLE: pupil(Name,Level,Class) WHERE: Level=1 AND Class='a'
```

Formalization of the Problem

Provided meta-information (called Table Completeness (TC) statements) that some parts of the existing database (called available database - D) is complete can we guarantee (deduce) that a query answer is same (called Query Completeness (QC)) as if the query was evaluated over the complete database (called ideal database - D)?

- To express partial completeness of database we use table completeness (TC) statements [A. Levy '96].
- For example, (TC1) we are complete for science pupils, we express using the notation:
  `TABLE: pupil(name,Level,Code) (-- complete table) WHERE: Level=1 AND Code='sc'` (== condition)
- Alternatively, we can express this using datalog notation:
  `pupil(N, L, C) :- pupil(N, L, C), class(N, L, C)`. (== condition)

Implementation

Encoding of the Problem in Logic Programming (Answer Set Programming)

System Architecture

Summary and Publications

- The first realized system that can reason about query completeness based on partially complete database (reasoning about TC-QC entailment)
- We gone beyond original TC-QC problem, and we investigate the impact of Schema Constraints, like Foreign keys and Finite Domains, on TC-QC entailment.
- We developed a component for explanations and suggestions, that in the case the query is not complete indicates which parts of database are incomplete w.r.t. the query.

DEMO-PAPER: MAGIK: Managing Completeness of Data
Ognjen Savkovic, Mirza Paramita, Sergey Paramonov, and Werner Nutt
Proc. of the 21st ACM Int. Conf. on Information and Knowledge Management (CIKM 2012)

MAGIK: Managing Completeness of Data
http://magik-demo.inf.unibz.it/public-version/

8th Reasoning Web Summer School (Reasoning Web 2012)
September 3 – 8, 2012 Vienna, Austria joint work with Sergey Paramonov, Mirza Paramita, and Werner Nutt

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