# QOCO-R:Query and Rules Oriented Data Cleaning

Ahmad Assadi Tova Milo Slava Novgorodov

## ABSTRACT

Fill Content

## 1. INTRODUCTION

Fill intro

## 2. RULES

The rules in QOCO-R are a database assertions that delivered to the algorithm by the Oracle. In order to get an effective rules language it must enable expressing all popular relational database assertions. As in [1, 2] The embedded dependencies include all of the naturally-occurring constraints on relational databases. They are a first order logic formulas of the form:

$$\forall \mathtt{x_1,...,x_n}\phi(\mathtt{x_1,...,x_n}) \rightarrow \exists \mathtt{z_1,...,z_m}\psi(\mathtt{x_1,...,x_n,z_1,...,z_m})$$

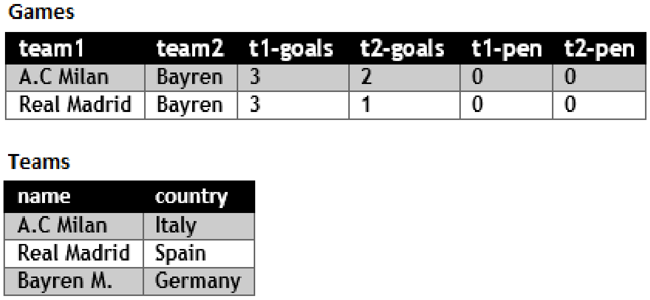
where the left hand side (LHS) of the implication, $\phi$, is a conjunction of relational formulas over variables $\bar{x}$, and the right hand side (RHS) of the implication, $\psi$, is a conjunction of relational or equality formulas over variables $\bar{x}$ and $\bar{z}$. The embedded dependencies is comprised of tuple-generating dependencies (tgds) of the form:

$$\forall \mathtt{x_1,...,x_n}\phi(\mathtt{x_1,...,x_n}) \rightarrow \exists \mathtt{z_1,...,z_m}\mathtt{R}(\mathtt{x_1,...,x_n,z_1,...,z_m})$$

and equality generating dependencies (egds) of the form:

$$\forall \mathtt{x_1,...,x_n}\phi(\mathtt{x_1,...,x_n}) \rightarrow \mathtt{x_i} = \mathtt{x_j}$$

In tgds the RHS contains only relational formulas and in egds the RHS contains only equality formulas. Given a particular combination of tuples satisfying the constraint of the LHS, tgds expresses an assertion about the existence of a tuple in the instance on the RHS,and egds expresses an assertion about an equality between two variables.
As we mentioned above our rules language should enable expressing tgds and egds, therefor the rules language consists of two sets of rules:



Games

| team1 | team2 | t1-goals | t2-goals | t1-pen | t2-pen |
|---|---|---|---|---|---|
| A.C Milan | Bayren | 3 | 2 | 0 | 0 |
| Real Madrid | Bayren | 3 | 1 | 0 | 0 |

Teams

| name | country |
|---|---|
| A.C Milan | Italy |
| Real Madrid | Spain |
| Bayren M. | Germany |

**Figure 1: Portion of a football league database.**

1. Tuple-generating rules (tgrs). They have the same form as tgds but the LHS may contain also constraints on variables (not only relational formulas), a constraint is a boolean expression of the form $v\, OP\, w$ where $v, w \in \bar{x} \cup \mathcal{C}$ and $OP$ is a boolean operation that defined on the variables domain, for example if $v$ and $w$ value should be a number then $OP$ should be $=, \neq, \leq, \geq, ...$ For

2. Condition-generating rules (cgrs). They have the same form as egds but both the LHS and the RHS could contain a conjunction of constraints.

Example 1. *Consider the database in Figure 1 which shows portions of two relations of a football league. The Games relation describes the results of a match between two teams, it stores the teams name, goals score and penalties score. The Teams relation describes a football team, it stores the team name and country. This database must satisfy the facts: (i) If a game ends with a draw then the penalties stage must determine the winner (ii) The team name uniquely determines it's country (iii) team1 column in Games relation should be included in name column in the Teams relation. Those facts are equal to the following rules:*

- $\forall \bar{\mathtt{x}}\, \mathtt{Games}(\mathtt{x_1,x_2,x_3,x_4,x_5,x_6}) \wedge \mathtt{x_3} = \mathtt{x_4} \rightarrow \mathtt{x_5} \geq 0 \wedge \mathtt{x_6} \geq 0 \wedge \mathtt{x_5} \neq \mathtt{x_6}$
- $\forall \bar{\mathtt{x}}\, \mathtt{Teams}(\mathtt{x_1,x_2}) \wedge \mathtt{Teams}(\mathtt{x_3,x_4}) \wedge \mathtt{x_1} = \mathtt{x_3} \rightarrow \mathtt{x_2} = \mathtt{x_4}$
- $\forall \bar{\mathtt{x}}\, \mathtt{Games}(\mathtt{x_1,x_2,x_3,x_4,x_5,x_6}) \rightarrow \exists \bar{\mathtt{z}}\, \mathtt{Teams}(\mathtt{x_1,z_1})$

*respectively where the first two are cgrs and the third is a tgr.*

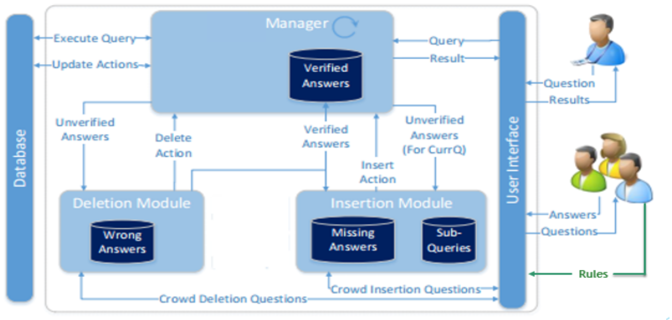


**Figure 2: QOCO-R framework architecture.**

## 3. SYSTEM ARCHITECTURE

Query and Rules Oriented Data Cleaning (QOCO-R) System comprises of 3 major blocks: Manager, Deletion module and insertion module, as shown in Figure 2. The QOCO-R's input is a relational database $D$ where $D$ can contain invalid or missing data. The system has a user interface for enabling interaction with the crowd (oracles) and the users. As in QOCO, through the UI the user can specify two target actions: (i) remove a wrong answer from $Q(D)$ or, (2) add a missing answer to $Q(D)$

## 4. CONCLUSIONS

Content

## 5. ACKNOWLEDGMENTS

Content

## 6. REFERENCES

[1] A. Deutsch, L.Popa, and V. Tannen. *Query reformulation with constraints. SIGMOD Record, 35(1), 2006.*

[2] R. Fagin, P. Kolaitis, R. J. Miller, , and L. Popa. *Data exchange: Semantics and query answering. Theoretical Computer Science, 336, 2005.*

## APPENDIX

Content