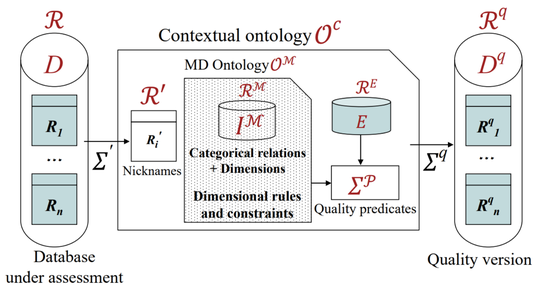
Data and Dimension Repair for Multidimensional Ontologies by Record Deletion

McMaster University

Enamul Haque, Mostafa Milani, Fei Chiang – Data Science Research Laboratory

Introduction

Data quality assessment of data cleaning are context dependent activities, where context can be modeled as logic-based ontologies comprising dimensions, quality predicates and the data itself.



Our Research:

- Inconsistent Ontological Multidimensional Data Model
- Reasoning for inconsistency (data, constraints or dimension)
- Weight based framework for data cleaning based on Answer Set Programming

Problem Definition

A MD ontology is composed of dimensional schema & instances, category predicates, child-parent predicates and categorical predicates. Instance has the complete extensions of predicates. It also contains basic application-independent constraints, dimensional rules and dimensional constraints.

A MD ontology is inconsistent if there is no instance which satisfies all these constraints.

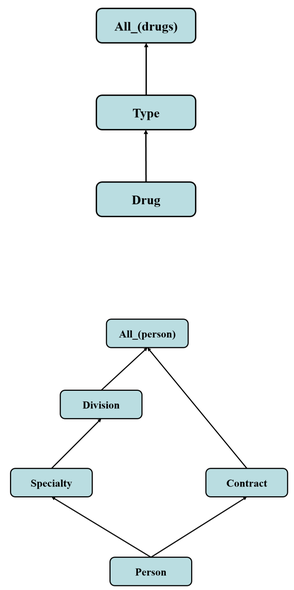
A repair for the MD ontology is the possible set of instances that satisfies all the constraints over it.

We are proposing a repair by record deletion from the data or the dimensions.

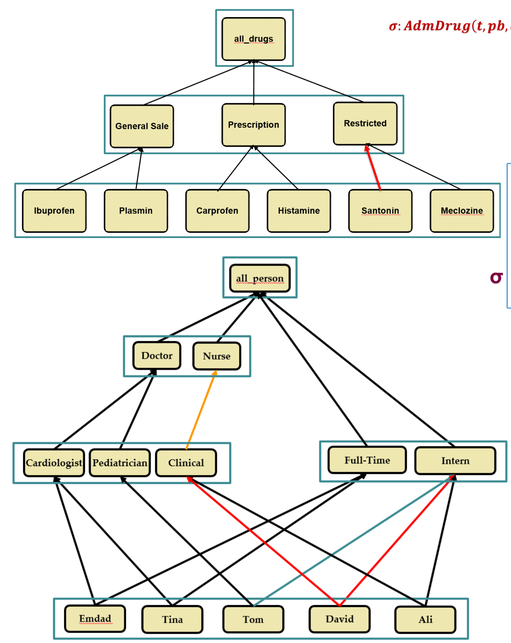
Here the complexity lies within, number of dimensions involved, the level of the dimension structure needs fixation, and the tuples which may be generated from other tables.

Example

Dimensional Schema



Dimensional Instance



Relational Data Tables

$\sigma: AdmDrug(t, pb, d; p, ag), PersonSpec(pb, sp), DrugType(d, dt) \rightarrow \exists am: Bills(t, sp, dt; p, am)$

AdmDrug(t,pb,d;p,ag)

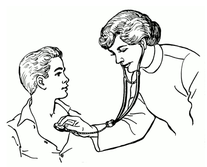
Time	Prescribed By	Drug	Patient	Age
28-Mar-18	Emdad	Ibuprofen	Rafi	80
12-Feb-18	Tom	Plasmin	Anika	2
14-Feb-18	Tom	Santonin	Ruby	1
16-Dec-17	David	Santonin	Harry	15

σ

Bills(t,sp,dt;p,am)

Day	Specialization	Drug Type	Patient	Amount
28-Mar-18	Cardiologist	General Sale	Rafi	200
12-Feb-18	Pediatrician	General Sale	Anika	150
14-Feb-18	Pediatrician	Restricted	Ruby	60
16-Dec-17	Clinical	Restricted	Harry	?

Dimensional rule: If any patient is given a drug s/he should be billed for that amount.

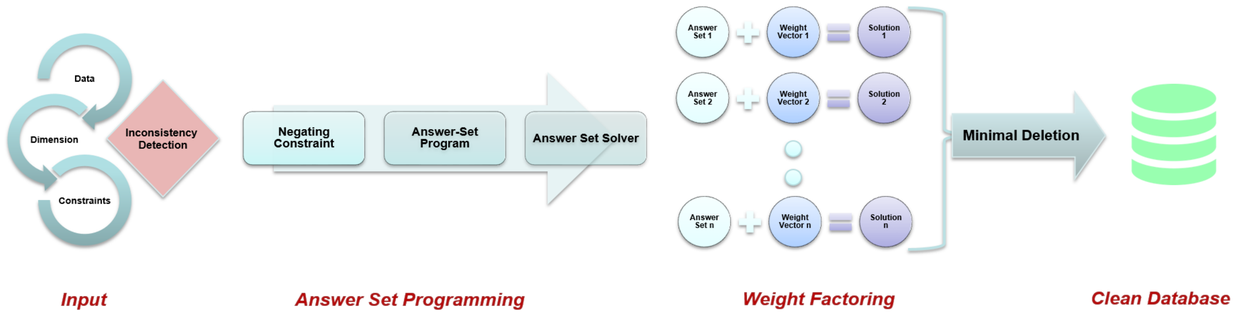


"Restricted drugs must be prescribed by a full-time doctor"

$\kappa_1: AdmDrug(t, sp, Restricted; p, ag), SpecDiv(sp, div), PersonSpec(pb, sp), PersonContract(pb, con) \rightarrow div = Doctor$

$\kappa_2: AdmDrug(t, sp, Restricted; p, ag), SpecDiv(sp, div), PersonSpec(pb, sp), PersonContract(pb, con) \rightarrow con = FullTime$

Proposed Solution



Example Output

We introduce weights for each edge in the dimension as well as for each relation. For example, the predicates can be written in the following way (with corresponding weights):

$\mu_1 = DrugType(Santonin, Restrcited)$	$w_1 = 3$
$\mu_2 = PersonContract(Tom, Intern)$	$w_2 = 2$
$\mu_3 = PersonContract(David, Intern)$	$w_3 = 2$
$\mu_4 = PersonSpec(David, Clinical)$	$w_4 = 5$
$\mu_5 = SpecDiv(Clinical, Nurse)$	$w_5 = 7$
$\mu_6 = AdmDrug(14Feb, Tom, Santonin, Ruby, 1)$	$w_6 = 6$
$\mu_7 = AdmDrug(16Dec, David, Santonin, Harry, 15)$	$w_7 = 9$

ASP Engine will search for the models which satisfy the negation of the constraint and generate the answer sets which should be deleted. The cost from weights:

$M_1 = \{\mu_1\}$	$Cm_1 = (3) = 3$
$M_2 = \{\mu_2, \mu_3, \mu_4\}$	$Cm_2 = (2 + 2 + 5) = 9$
$M_3 = \{\mu_2, \mu_3, \mu_5\}$	$Cm_3 = (2 + 2 + 7) = 11$
$M_4 = \{\mu_6, \mu_7\}$	$Cm_4 = (6 + 9) = 15$

We propose that weights are provided by the user. Based on the weight here we get the minimum cost for the models is 3. That means the cost for model-1 (M_1) is minimum. This suggests that we can delete record:

$$\mu_1: DrugType(Santonin, Restrcited)$$

Hence, *Santonin* is no longer a restricted drug and all the constraints are satisfied after this record deletion.

References

[1] Ahmetaj, S., Ortiz, M. and Simkus, M. Polynomial Datalog Rewritings for Expressive Description Logics with Closed Predicates. Proc. IJCAI 2016, pp. 878-885.

[2] Ahmetaj, S., Ortiz, M. and Simkus, M.: Polynomial Datalog Rewritings for Ontology Mediated Queries with Closed Predicates. Proc. AMW 2016. CEUR 1644.

[3] Bertossi, L. and Milani, M. Ontological Multidimensional Data Models and Contextual Data Quality. Journal submission, 2017. Corr ArXiv paper cs.DB/1704.00115.

[4] Bertossi, L. Database Repairing and Consistent Query Answering. Morgan & Claypool, Synthesis Lectures on Data Management, 2011.

[5] Bertossi, L., Bravo, L. and Caniupan, M. Consistent Query Answering in Data Warehouses. Proc. AMW 2009. CEUR 450.

[6] A. Cali, G. Gottlob, and T. Lukasiewicz. Datalog: A Unied Approach to Ontologies and Integrity Constraints. Proc. ICDT 2009, pp. 14-30.

[7] A. Cali, G. Gottlob, and A. Pieris. Towards more Expressive Ontology Languages: The Query Answering Problem. Articial Intelligence, 2012, 193:87-128.

[8] Hurtado, C. and Mendelzon, A. OLAP Dimension Constraints. Proc. PODS 2002.