Summary of: The LLUNATIC Data-Cleaning Framework

Vincent J. Maccio

Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada

The Problem

- The known issues of data cleaning
- What does one mean by clean? Repaired?
- If we know something is dirty, how do we clean it?

Example - Data

	SSN	Name	Phone	Conf	Str	City	CC#
t_1	111	M. White	408 – 3334	0.8	Red Ave.	NY	112321
t_2	222	L. Lennon	122 – 1876	0.9	NULL	SF	781658
t_3	222	L. Lennon	000 – 0000	0.0	Fry Dr.	SF	784658

	SSN	Salary	Insur.	Treat	Date
t_4	111	10K	Abx	Dental	10/1/2011
t_5	111	25K	Abx	Cholest.	8/12/2012
t_6	222	30K	Med	Eye Surg.	6/10/2012

	SSN	Name	Phone	Str	City
t_m	222	F. Lennon	122-1876	Sky Dr.	SF

Example - Constraints

- (SSN, NAME) $\rightarrow$ (PHONE, CC#)
- (SSN) $\rightarrow$ (SALARY)
- (INSUR= "Abx")$\rightarrow$ (Treat="Dental")
- If the customer relation agrees with the Master relation on SSN and PHONE, then they must also agree on NAME, STR, and CITY
- The company Abx only accepts customers from SF

Example - Problems

- Repairing the FD, which phone numbers to resolve to isn't clear
- Same problem when resolving SALARY
- When resolving the correct CC# there is no information available to make a "good" decision
- If these dependencies are resolved in a different order, you get different results (different definitions of repaired)

Contributions

The authors identify, address and offer solutions to three problems:

- Missing semantics to deal with these complex constraints
- Missing a formal repair algorithm
- Implementation and scalability of such an algorithm, if it were to exist

Semantics

- Let a schema $\mathcal{S} = \{R_1, ..., R_k\}$, where each table/relation R_i has some arity $n_i \geq 0$
- Given discrete sets NULL and CONSTS, an instance of $\mathcal{S}$, $I = \{I_1, ..., I_k\}$, where forall i, $I_i \subset (\text{NULL} \cup \text{CONSTS})^{n_i}$
- If $\mathcal{T}$ is also a schema and J is an instance of it, then (I, J) is an instance of $(\mathcal{S}, \mathcal{T})$
- Let an Equality Generating Dependency (EGD) be defined as: $\forall \bar{x}(\phi(\bar{x}) \rightarrow x_i = x_j)$

EGD Examples

(SSN, NAME) → (PHONE)

- Cust(**ssn**,**n**,p,s,c,cc),Cust(**ssn**,**n**,p',s',c',cc')→ p=p'

(INSUR= "Abx")→ (Treat="Dental")

- Treat(ssn,s,ins,tr,d), ins=Abx → tr=Dental

If the customer and master tables agree on SSN and PHONE, then they should also agree on NAME

- Cust(**ssn**,n,**p**,s,c,cc),MD(**ssn**,n',**p**,s',c',cc')→ n=n'

LLUNs

Placeholders, the opposite of null

	SSN	Name	Phone	Str	City	CC#
t_1	111	M. White	$408 - 3334$	Red Ave.	NY	112321
t_2	222	F. Lennon	$122 - 1876$	Sky Dr	SF	L_0
t_3	222	F. Lennon	$122 - 1876$	Sky Dr.	SF	L_0

Repairs and Cell Groups

Cell Group

- $g = (v \to \mathcal{C}, \mathcal{C}_s)$, $occ(g) = \mathcal{C}$, $just(g) = \mathcal{C}_s$

Examples

- $(L_0(781658, 784659) \to \{t_2.CC\#, t_3.CC\#\}, by\ \emptyset)$
- $(F.Lennon \to \{t_2.NAME, t_3.NAME\}, by\ \{t_m.NAME\})$
- $(Dental \to \{t_5.TREAT\}, by\ \{t_{c3}.TREAT\})$

Repairs

- $Rep = \{g_1, ..., g_k\}$

The Partial Order

Informative Relation

- $v_1 \lhd v_2$ if v_1 is null and v_2 is not, or if v_2 is a llun and v_1 is not

Partial Order Specification

- $\sqcap$ is a set of assignments of attributes to partial orders
- $\preceq_J^{\sqcap}$ is the partial on cell values
- For two cell values v_1 and v_2; $v_1 \preceq_J^{\sqcap} v_2$ iff $v_1 = v_2$, $v_1 \lhd v_2$, or the corresponding poset values hold under that ordering i.e. $v_1' < v_2'$

The Partial Order - Cell Groups

- $\preceq_\sqcap$ induces a partial order on cell groups (val(g)=lub)
- Given a partial order based on the specification, $g_1 \preceq_\sqcap g_2$ if the following hold: $occ(g_1) \subseteq occ(g_2)$, $just(g_1) \subseteq just(g_2)$, and $val(g_1) \lhd val(g_2)$ or g_1 and g_2 are of the same type

The Partial Order - Cell Groups

- (A) → (B), (A=a)→(B=x), (A=a)→(B=y)
- R(a,1), R(a, 2)

$$(1 \rightarrow \{t_1.B\}, by\ \emptyset) \preceq_{\sqcap}$$
$$(2 \rightarrow \{t_1.B, t_2.B\}, by\ \emptyset) \preceq_{\sqcap}$$
$$(x \rightarrow \{t_1.B, t_2.B\}, by\ \{t_{c1}.x\}) \preceq_{\sqcap}$$
$$(L \rightarrow \{t_1.B, t_2.B\}, by\ \{t_{c1}.x\}, \{t_{c2}.y\})$$

Recall repairs are just set of cell groups, now one can say what repairs are better.

Definitions

Satisfaction after repairs

- A repair Rep is said to satisfy an EGD w.r.t $\preceq_{\sqcap}$ if forall homomorphisms h of $\phi(\bar{x})$, $g_h(x) \preceq_{\sqcap} g_h(x')$ or $g_h(x') \preceq_{\sqcap} g_h(x)$

Solution

- Given a cleaning scenario $\mathcal{CS} = \{(\mathcal{S}, \mathcal{T}), \Sigma, \sqcap\}$ and instance (I, J) a solution is a repair Rep s.t. $J \preceq_{\sqcap}$ Rep, and Rep satisfies Σ

Minimal Solution

- Rep is minimal if there does not exist a solution Rep' s.t. Rep' $\preceq_{\sqcap}$ Rep

Computing the Solution

Chase Algorithm

- Known algorithm for testing implications of data dependencies
- Modifications: chases forward and backwards, uses the partial order, simplifies data into equivalences classes

Chase Example

FD: (SSN)→(SALARY)

	SSN	Salary	Insur.	Treat	Date
t_4	111	10K	Abx	Dental	10/1/2011
t_5	111	25K	Abx	Cholest.	8/12/2012

$$Rep = \{(10K \rightarrow \{t_4.SALARY\}, by\ \emptyset), (25K \rightarrow \{t_5.SALARY\}, by\ \emptyset)\}$$
$$Rep_{f,f} = (25K \rightarrow \{t_5.SALARY, t_4.SALARY\}, by\ \emptyset)$$
$$Rep_{b,f} = (L_1 \rightarrow \{t_4.SALARY\}, by\ \emptyset)$$
$$Rep_{f,b} = (L_2 \rightarrow \{t_5.SALARY\}, by\ \emptyset)$$

Bottom Line

Given a set of EGDs and a dirty data set, this modified chase algorithm terminates, generates a finite set of solutions, and also generates all minimum solutions. What's the catch?

Beating Complexity

Cost Manager

- Maximum size, frequency, and forward only

Delta Database

- Only store what you need to know

Conclusion

- Current methods were not complete enough to specify all dependencies
- Repairing methods were too ad-hoc and no real algorithm existed
- LLUNATIC addresses both of these issues from the ground up and delivers a complete and proven algorithm

The End

Thank you.

Discussion - Andrew Leung

- Q: What is the advantage of the EDG notation described in the paper?
- Q: How to translate the CFD into a formula: insurance company "Abs" only offers dental treatments ("Dental")?

Discussion - Andrew Leung

- Treat(ssn, s, ins, tr, d), ins $=$ “Abs” $\rightarrow$ tr $=$ “Dental”
- Using constant tables,
- Treat(ssn, s, ins, tr, d), Cst(ins, tr’)$\rightarrow$ tr $=$ tr’
- Q: What are some limitations of the LLUNATIC framework?

Discussion - Andrew Leung

- Q: What are some key differences between LLUNATIC and other systems with respect to their data repair model?
- Q: What are some ways to evaluate the data repair in terms of Quality Metrics?
- Q: What are some strengths and weaknesses of the paper?