Quality-driven Query Answering and refinement

Ousmane Issa

Institut Supérieur d'Informatique, de Modélisation et de leurs Applications
Université Clermont Auvergne

Objectives

1. Investigate how data quality indicators can be used in order to improve the quality of query answers,
2. Ability to reason about quality-ranked query answers,
3. Design and implemente quality-aware query evaluation algorithm that enables the usage of quality indicators to compute context-dependant and quality-aware query answers,
4. Effective approach to automatically rewrite the query by taking into account the quality indicators of the involved data.

Introduction

- Conclusions are done with extracted data from stored data. From these conclusions, decisions are made. If stored data contain erroneous data, extrated data from them can also contain error. So, wrong decision can made from conclusions that are from these extracted data. To deal with this problem, some quality indicators are introduced. Given a query and a stored data called database, extracted data is answers of this query over database. Having a database that contains error and a query, our main goal is to compute answers of query over this database with a quantity that measures level of errors arise in answers. In litterature, there are some related works but they are about error quantification in stored data. Using this quantity, when one takes decision with answers of query so one has idea about the risk of result.

Overview of the problem

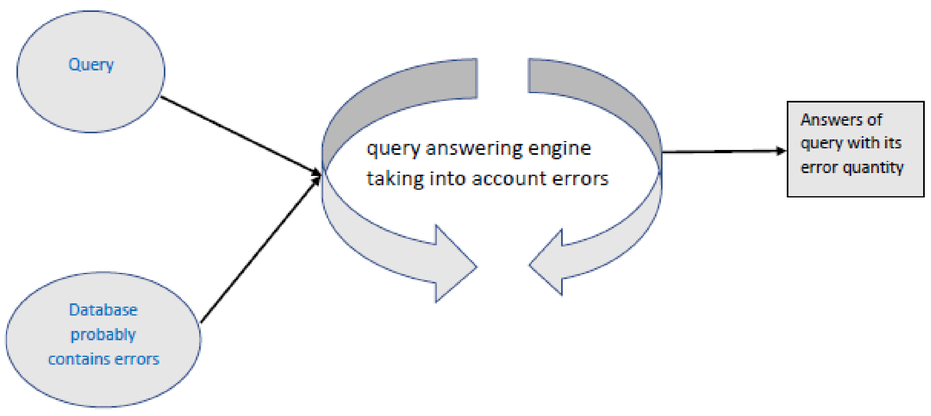

Figure 1:Overiew of our the problem

- Some interesting questions are:
  1. How to identify these erronous data from database ?
  2. How to quantify errors of answers of queries ?
  3. What means this error quantity for answers of queries ?
  4. How to compute this quantity, once it is defined ?
  5. depending on the context, what are the underlying filtering problems ?

First contribution, quantify inconsistency

- First, we are interested in inconsistency,
- Inconsistency arise when some integrity constraints defined over database are violated.
- These constraints make it possible to give meaning to the data.
- Example of integrity constraint is that two students can not have the same school number or two person can not have the same social security number.
- An other example of integrity constraint is any disease is diagnosed before its surgical operation

Related works

- Measure of inconsistency in database [1, 2]. In case of [1], inconsistency is measured as the number of contraductions that can be arise in database when integrity constrains are applied over it. In [2], inconsistency measure is the number of cases that violate constraints.
- Other related works are works about consistent query answering. For more details, reader can see [3, 4, 5, 6].

Formalization of the problem

- Let I, IC and q respectivelly a database, a set of integrity constraints and a query.

1. Identify set of tuples in I that violate IC, it is denoted $IncT(I, IC)$
2. Compute answers of q over I, denoted $q(I)$, and during this processing compute for each tuple $t \in q(I)$ its set of set tuples in I from whitch it is derived, denoted $prov(I, q, t)$
3. All tuples in answers of query with their inconsistency degree, denoted $q^{IC}(I)$ is the following

$$q^{IC}(I) = \{< t, m >:\ t \in q(I) \text{ and } m = Min_{E \in prov(I,q,t)}(|E \cap IncT(I, IC)|)\}$$ with $||$ the cardinal function.

Running example

Diagnosis (P)

Id	PatientID	DiseaseRef	Date
t1	P01	D1	2
t2	P02	D7	4
t3	P03	D7	10
t4	P04	D1	3
t5	P04	D4	8
t6	P02	D4	7
t7	P01	D2	5
t8	P08	D7	4
t9	P10	D1	5

Vaccination (R)

Id	PatientID	DiseaseRef	Date
t10	P01	D2	3
t11	P02	D4	5
t12	P03	D4	6
t13	P04	D4	7
t14	P02	D5	10
t15	P08	D5	7
t16	P10	D3	8

Surgery (S)

Id	PatientID	DiseaseRef	Date
t17	P01	D1	1
t18	P02	D7	3
t19	P03	D7	9
t20	P04	D1	4
t21	P08	D7	5

Integrity Constraints

(1) not (P (x, y, z) and S(x, y, u) and z > u)
(2) not (P(x, 'D2', y) and R(x, 'D2', z) and y>z)
(3) not (P(x, 'D4', y) and R(x, 'D4', z) and y>z)

Query

Q(y, u) ← P(x, y, z), R(x, u, v), S(x, y, z')

Tuples in answers of query and their provenance tuples

Surgery Disease	Vaccinated Disease	Provenance tuples
D1	D2	[t1, t10, t17]
D7	D4	[t2, t18, t11] or [t3, t12, t19]
D1	D4	[t4, t13, t20]
D7	D5	[t2, t18, t14] or [t8, t15, t21]

Tuples in answers of query and their inconsistency degree

Surgery Disease	Vaccinated Disease	Inconsistency degree
D1	D2	3
D7	D4	2
D1	D4	1
D7	D5	0

Figure 2:Running example of answers of query computing with inconsistency degree of tuples in answers

Conclusion

- My thesis work is about quality of query answering over database that contains errors
- My first contribution is definition of a new measure of inconsistency of query answers.

References

[1] John Grant and Anthony Hunter.
Measuring inconsistency in knowledgebases.
Journal of Intelligent Information Systems, 2009.

[2] Hendrik Decker and Davide Martinenghi.
Modeling, measuring and monitoring the quality of information.
In *ER '09 Proceedings of the ER 2009 Workshops (CoMoL, ETheCoM, FP-UML, MOST-ONISW, QoIS, RIGiM, SeCoGIS) on Advances in Conceptual Modeling*, 2009.

[3] Leopoldo Bertossi.
Database Repairing and Consistent Query Answering.
2011.

[4] Marcelo Arenas, Leopoldo Bertossi, and Jan Chomicki.
Consistent query answers in inconsistent databases.
In *the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 68–79. ACM, 1999.

[5] Dany Maslowski and Jef Wijsen.
A dichotomy in the complexity of counting database repairs.
Journal of Computer and System Sciences, 2013.

[6] JEF WIJSEN.
Database repairing using updates.
ACM Transactions on Database Systems (TODS), 2005.

[7] Todd J. Green, Grigoris Karvounarakis, and Val Tannen.
Provenance semirings.
In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40. ACM, 2007.

Contact Information

- Email: issaousmane7@gmail.com
- Phone: +33767128651