Efficient Query Containment Checking Using Logical Reasoning Engines

Sergey Paramonov

Technical University of Vienna

EMCL Workshop 2012 Vienna
Historical perspective

- Query completeness problem has roots in the development of school system in Bolzano.

- Central school database is needed for administration, final grades, statistical reports etc.

- Teachers and administrators have only local records.
Settings

• People involved:
  • the KRDB group in Bolzano
  • the KBS group in Vienna

• Bolzano: developed theory of query completeness

• Vienna: developed a powerful disjunctive datalog engine (DLV)

• shortcoming of current theory lack of implementations

• Our goal: put theory into practise.
Motivation for Query Completeness

- When does query completeness matter?
- in data integration
- if several people, institutions independently contribute data
- some data are final and others provisional
Query Completeness

• What does it mean for a query to be complete?

• Intuitively it captures in the answer all tuples.

• Could you imagine that EMCL administration is missing you personal record?

• Now we can verify that everything is in the right place!¹

¹”Beware! I have only proved it correct, not tried it.” Donald Knuth
Definition (Partial Database)

A partial database is a pair $D = (D^i, D^a)$ of two instances,
- the ideal database $D^i$
- the available database $D^a$

such that $D^a \subseteq D^i$

Intuition:
- $D^i$ reflects real world, what is really true
- $D^a$ reflects data we physically store

Note (We make validity assumption)

there is no "wrong" data in the available database.
Partial Database Example

- \( D = (D^a, D^i) \)
  - is partial database with two students (Oliver & Wu) in two different classes (2b & 2a).

- **Ideal** Database \( D^i = \{
  \text{Student}(\text{Oliver}, "\text{EMCL}"), \text{Class}(\text{Oliver}, 2, b),
  \text{Student}(\text{Wu}, "\text{ICCL}") , \text{Class}(\text{Wu}, 2, a)\} \)

- **Available** Database \( D^a = D^i \setminus \text{Class}(\text{Oliver}, 2, a) \)

**Note**
Available database is missing the fact that Oliver is a second year student.
What does it mean for a query $Q$ to be complete?

**Definition**

$Q$ is said to be complete written as $Compl(Q)$:

$$(D^i, D^a) \models Compl(Q) \iff Q(D^i) = Q(D^a)$$

Intuition: a query $Q$ is complete if query evaluation over available database is the same as over ideal one.
Peter confirmed:

"Workshop database contains all 2 year students" \(^2\)

We formalize this as a **table completeness statement**:

\[
\text{Student}^i(N,M), \text{Class}^i(N,2,C) \rightarrow \text{Student}^a(N,M)
\]

or shortly **Compl**(student(N,M) ; class(N,2,C))

General notation:

\[
\text{Compl}(R(\bar{s}); G)
\]

where query \( Q(\bar{s}) = R(\bar{s}) \), \( G \) is safe

\(^2\)It is actually not true, right Martin?
Main question in the project how to implement the problem:

When completeness of small parts of the database entail completeness of the query?

Formally:
TC-QC: table completeness entails query completeness

\[
\text{Compl}(R_1, G_1), \ldots, \text{Compl}(R_n, G_n) \models \text{Compl}(Q)
\]

Example
All students in Dresden, Vienna, Bolzano and Lisbon are good, does it mean that all ECML students are good?
Query Containment

• Definition (Query Containment: $Q_1$ is contained in $Q_2$ written as: $Q_1 \subseteq Q_2$)

$$Q_1(D) \subseteq Q_2(D) \quad \forall D - \text{db instances}$$

• Studied for conjunctive queries (CQ).
  - Correspond to single-block select-from-where SQL query
  - Query that ask for good EMCL students:

$$Q(\text{Name}) \leftarrow \text{Student}(\text{Name}, "EMCL"), \text{Good}(\text{Name}).$$

• Extensions: CQs with comparisons ($\geq$, $>$), finite domains, unions of CQs.

• Complexity: from $NP$ to $\Pi^P_2$. \(^3\)

\(^3\) Free Complexity Class tonight in the pub
Containment example

Given two queries $Q_1$ and $Q_2$

\[
Q_1(Name) \leftarrow \text{Student}(Name,"EMCL"), \text{Good}(Name).
\]
\[
Q_2(Name) \leftarrow \text{Student}(Name,"EMCL").
\]

$Q_1 \subseteq Q_2$ ?

The question whether all good EMCL student are among EMCL student?
And the answer is, of course, yes.
Opposite does not hold:
It is hard to believe but there might exist not good EMCL students.
Algorithm for the TC-QC

- TC-QC problem can be reduced to the variants of query containment.

**Intuition:**
- Query needs parts \( \{P_i\} \) of the relation \( R_i \) to be complete
- Is \( P_i \) contained in the parts \( S_1, \ldots, S_n \) stated to be complete?

so containment:

\[
P_i \subseteq S_1 \cup S_2 \cup \cdots \cup S_n
\]

- Query containment can be in reduced to evaluation task of different reasoning engines.
Implementation

Query containment can be in principle reduced to the

- ASP: done in DLV for Relational Case
- SMT: partially studied for comparisons in Z3.
- QBF: alternative approach in the future.
Future Work

- Investigate different faces of the problem e.g. finite domain constraint (now in progress)

- Develop different implementations: SMT, DLV, ASP+Difference logic, QBF.

- Create a uniform benchmark for different classes of languages (RQ, LQ, CQ, UCQ)
Evaluation of the project

A detailed report with complete results is going to be submitted to ESSLLI 2012 as an article and a poster.
<joke>
- **Sir Humphrey**: If local authorities don’t send us statistics, Government figures will be a nonsense.
- **Hacker**: Why?
- **Sir Humphrey**: They’ll be incomplete.
- **Hacker**: Government figures are a nonsense, anyway.
- **Bernard**: I think Sir Humphrey wants to ensure they’re a complete nonsense.
</joke>

Thank you for your attention.