Declarative Debugging of Wrong and Missing Answers for SQL Views [⋆]

Rafael Caballero[†], Yolanda García-Ruiz[†], and Fernando Sáenz-Pérez[‡]

[†]Departamento de Sistemas Informáticos y Computación
[‡]Dept. de Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid, Spain
{rafa,fernan}@sip.ucm.es and ygarciar@fdi.ucm.es

Abstract. This paper presents a debugging technique for diagnosing errors in SQL views. The debugger allows the user to specify the error type, indicating if there is either a missing answer (a tuple was expected but it is not in the result) or a wrong answer (the result contains an unexpected tuple). This information is employed for slicing the associated queries, keeping only those parts that might be the cause of the error. The validity of the results produced by sliced queries is easier to determine, thus facilitating the location of the error. Although based on the ideas of declarative debugging, the proposed technique does not use computation trees explicitly. Instead, the logical relations among the nodes of the trees are represented by logical clauses that also contain the information extracted from the specific questions provided by the user. The atoms in the body of the clauses correspond to questions that the user must answer in order to detect an incorrect relation. The resulting logic program is executed by selecting at each step the unsolved atom that yields the simplest question, repeating the process until an erroneous relation is detected. Soundness and completeness results are provided. The theoretical ideas have been implemented in a working prototype included in the Datalog system DES.

1 Introduction

SQL (Structured Query Language [18]) is a language employed by relational database management systems. In particular, the SQL select statement is used for querying data from databases. Realistic database applications often contain a large number of tables, and in many cases, queries become too complex to be coded by means of a single select statement. In these cases, SQL allows the user to define *views*. A SQL view can be considered as a virtual table, whose content is obtained executing its associated SQL select query. View queries can rely on previously defined views, as well as on database tables. Thus, complex queries can be decomposed into sets of correlated views. As in other programming paradigms, views can have bugs. However, we cannot infer that a view is

⋆ Work partially supported by the Spanish projects STAMP (TIN2008-06622-C03-01), Prometidos-CM (S2009TIC-1465) and GPD (UCM-BSCH-GR35/10-A-910502)

incorrectly defined when it computes an unexpected result, because it might be receiving erroneous input data from other database tables or views. Given the high-abstraction level of SQL, usual techniques like trace debugging are difficult to apply. Some tools as [2, 13] allow the user to trace and analyze the stored SQL procedures and user defined functions, but they are of little help when debugging systems of correlated views. *Declarative Debugging*, also known as *algorithmic debugging*, is a technique applied successfully in (constraint) logic programming [16], functional programming [12], functional-logic programming [5], and in deductive database languages [3]. The technique can be described as a general debugging schema [11] which starts when an *initial error symptom* is detected by the user, which in our case corresponds to an unexpected result produced by a view. The debugger automatically builds a tree representing the erroneous computation. In SQL, each node in the tree contains information about both a relation, which is a table or a view, and its associated computed result. The root of the tree corresponds to the initial view. The children of a node correspond to the relations (tables or views) occurring in the definition of its associated query. After building the tree, it is *navigated* by the debugger, asking to the user about the validity of some nodes. When a node contains the expected result, it is marked as *valid*, and otherwise it is marked as *nonvalid*. The goal of the debugger is to locate a *buggy node*, which is a nonvalid node with valid children. It can be proved that each buggy node in the tree corresponds to either an erroneously defined view, or to a database table containing erroneous data. A debugger based on these ideas was presented in [4]. The main criticism that can be leveled at this proposal is that it can be difficult for the user to check the validity of the results. Indeed, even very complex database queries usually are defined by a small number of views, but the results returned by these views can contain hundreds or thousands of tuples. The problem can be easily understood by considering the following example:

Example 1. The loyalty program of an academy awards an intensive course for students that satisfy the following constraints:
- The student has completed the basic level course (level = 0).
- The student has not completed an intensive course.
- To complete an intensive course, a student must either pass the *all in one* course, or the three initial level courses (levels 1, 2 and 3).

The database schema includes three tables: *courses(id,level)* contains information about the standard courses, including their identifier and the course level; *registration(student,course,pass)* indicates that the *student* is in the *course*, with *pass* taking the value *true* if the course has been successfully completed; and the table *allInOneCourse(student,pass)* contains information about students registered in a special intensive course, with *pass* playing the same role as in *registration*. Figure 1 contains the SQL views selecting the award candidates. The first view is *standard*, which completes the information included in the table *Registration* with the course level. The view *basic* selects those *standard* students that have passed a basic level course (level 0). View *intensive* defines as intensive students those in the *allInOneCourse* table, together with the students that have

```
create or replace view standard(student, level, pass) as
   select R.student, C.level, R.pass
   from courses C, registration R
   where C.id = R.course;

create or replace view basic(student) as
   select S.student
   from standard S
   where S.level = 0 and S.pass;

create or replace view intensive(student) as
   (select A.student from allInOneCourse A  where A.pass)
   union
   (select a1.student
    from standard A1, standard A2, standard A3
    where A1.student = A2.student and A2.student = A3.student
          and
          a1.level = 1  and a2.level = 2 and a3.level = 3);

create or replace view awards(student) as
  select student from  basic
  where student not in (select student from  intensive);
```

Fig. 1. Views for selecting award winner students

completed the three initial levels. However, this view definition is erroneous: we have forgotten to check that the courses have been completed (flag *pass*). Finally, the main view *awards* selects the students in the basic but not in the intensive courses. Suppose that we try the query `select * from awards;`, and that in the result we notice that the student *Anna* is missing. We know that *Anna* completed the basic course, and that although she registered in the three initial levels, she did not complete one of them, and hence she is not an intensive student. Thus, the result obtained by this query is nonvalid. A standard declarative debugger using for instance a top-down strategy [17], would ask first about the validity of the contents of *basic*, because it is the first child of *awards*. But suppose that *basic* contains hundreds of tuples, among them one tuple for *Anna*; in order to answer that *basic* is valid, the user must check that *all* the tuples in the result are the expected ones, and that there is no missing tuple. Obviously, the question about the validity of *basic* becomes practically impossible to answer.

The main goal of this paper is to overcome or at least to reduce this drawback. This is done by asking for more specific information from the user. The questions are now of the type "Is there a missing answer (that is, a tuple is expected but it is not there) or a wrong answer (an unexpected tuple is included in the result)?" With this information, the debugger can:

- Reduce the number of questions directed at the user. Our technique considers only those relations producing/losing the wrong/missing tuple. In the example, the debugger checks that *Anna* is in *intensive*. This means that either *awards* is erroneous or *Anna* is wrong in *intensive*. Consequently, the debugger disregards *basic* as a possible error source, reducing the number of questions.
- The questions directed at the user about the validity in the children nodes can be simplified. For instance, the debugger only considers those tuples that are needed to produce the wrong or missing answer in the parent. In the example, the tool would ask if *Anna* was expected in *intensive*, without asking for the validity of the rest of the tuples in this view.

Another novelty of our approach is that we represent the computation tree using Horn clauses, which allows us to include the information obtained from the user during the session. This leads to a more flexible and powerful framework for declarative debugging that can now be combined with other diagnosis techniques. We have implemented these ideas in the system DES [14, 15].

Next section presents some basic concepts used in the rest of the paper. Section 3 introduces the debugging algorithm that constitutes the main contribution of our paper, including the theoretical results supporting the proposal. The implementation is discussed in Section 4. Finally, Section 5 presents the conclusions and proposes future work.

2 Preliminaries

This section introduces some basic concepts about databases, interpretations and types of errors which are used in the rest of the paper. A *table schema* has the form $T(A_1, \ldots, A_n)$, with T being the table name and A_i the attribute names for $i = 1 \ldots n$. We refer to a particular attribute A by using the notation $T.A$. Each attribute A has an associated type. An *instance* of a table schema $T(A_1, \ldots, A_n)$ is determined by its particular *tuples*. Each tuple contains values of the correct type for each attribute in the table schema. The notation t_i represents the i-th element in the tuple. In our setting, *partial* tuples are tuples that might contain the special symbol $\perp$ in some of their components. The set of defined positions of a partial tuple s, *def(s)*, is defined by $p \in \mathit{def(s)} \Leftrightarrow s_{\mathrm{p}} \neq \perp$. Tuples s with $\mathit{def(s)} = \emptyset$ are *total* tuples. Membership with partial tuples is defined as follows: if s is a partial tuple, and S a set of total tuples with the same arity as s, we say that $s \in S$ if there is a tuple $u \in S$ such that $u_p = s_p$ for every $p \in (\mathit{def(s)} \cap \mathit{def(u)})$. Otherwise we say that $s \notin S$.

A *database schema* D is a tuple $(\mathcal{T}, \mathcal{V})$, where $\mathcal{T}$ is a finite set of tables and $\mathcal{V}$ a finite set of views. *Views* can be thought of as new tables created dynamically from existing ones by using a SQL query. The general syntax of a SQL view is: create view $\mathsf{V}(A_1, \ldots, A_n)$ as Q, with Q a query and $V.A_1, \ldots V.A_n$ the names of the view attributes. A *database instance* d of a database schema is a set of table instances, one for each table in $\mathcal{T}$. The notation $d(T)$ represents the instance of a table T in d. The *dependency tree* of any view V in the schema is a tree with V labeling the root, and its children the dependency trees of the relations occurring

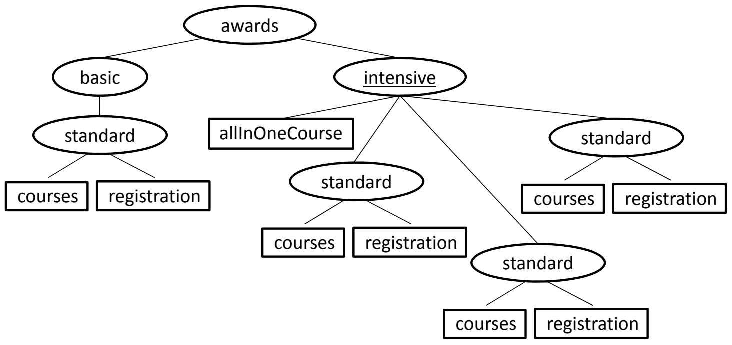

Fig. 2. Example of Computation Tree

in its query. Figure 2 shows the dependency tree for our running example. In general, the name *relation* refers to either a table or a view. The syntax of SQL queries can be found in [18]. We distinguish between *basic queries* and *compound queries*. A basic query Q contains both select and from sections in its definition with the optional where, group by and having sections. For instance, the query associated to the view *standard* in the example of Figure 1 is a basic query. A compound query Q combines the results of two queries Q_1 and Q_2 by means of set operators union [all], except [all] or intersect [all] (the keyword all indicates that the result is a multiset). For convenience, our debugger transforms basic queries into compound queries when necessary. We also assume that the queries defining views do not contain subqueries. Translating queries into equivalent definitions without subqueries is a well-known transformation (see for instance [6]). For instance, the query defining view *awards* in the Figure 1 is transformed into:

```
select  student  from  basic
except
select  student  from  intensive;
```

The semantics of SQL assumed in this paper is given by the Extended Relational Algebra (ERA) [10], an operational semantics allowing aggregates, views, and most of the common features of SQL queries. Each relation R is defined as a multiset of tuples. The notation $|R|_t$ refers to the number of occurrences of the tuple t in the relation R, and Φ_R represents the ERA expression associated to a SQL query or view R, as explained in [8]. A query/view usually depends on previously defined relations, and sometimes it will be useful to write $\Phi_R(R_1,\ldots,R_n)$ indicating that R depends on $R_1,\ldots,R_n$. Tables are denoted by their names, that is, $\Phi_T = T$ if T is a table. The *computed answer* of an ERA expression Φ_R with respect to some schema instance d is denoted by $\parallel \Phi_R \parallel_d$, where:

- If R is a database table, $\parallel \Phi_R \parallel_d = d(R)$.
- If R is a database view or a query and $R_1,\ldots,R_n$ the relations defined in R, then $\parallel \Phi_R \parallel_d = \Phi_R(\parallel \Phi_{R_1} \parallel_d,\ldots,\parallel \Phi_{R_n} \parallel_d)$.

The parameter d indicating the database instance is omitted in the rest of the presentation whenever is clear from the context.

Queries are executed by SQL systems. The answer for a query Q in an implementation is represented by $\mathcal{SQL}(Q)$. The notation $\mathcal{SQL}(R)$ abbreviates $\mathcal{SQL}(\textsf{select * from } R)$. In particular, we assume in this paper the existence of *correct* SQL implementations. A *correct* SQL implementation verifies that $\mathcal{SQL}(Q) = \parallel \Phi_Q \parallel$ for every query Q. In the rest of the paper, D represents the database schema, d the current instance of D, and R a relation defined in D. We assume that the user can check if the computed answer for a relation matches its intended answer. The *intended answer* for a relation R w.r.t. d, is a multiset denoted as $\mathcal{I}(R)$ containing the answer that the user expects for the query select * from R in the instance d. This concept corresponds to the idea of *intended interpretations* employed usually in algorithmic debugging. We say that $\mathcal{SQL}(R)$ is an *unexpected answer* for a query R if $\mathcal{I}(R) \neq \mathcal{SQL}(R)$. An unexpected answer can contain either a *wrong tuple*, when there is some tuple t in $\mathcal{SQL}(R)$ s.t. $|\mathcal{I}(R)|_t < |\mathcal{SQL}(R)|_t$, or a *missing tuple*, when there is some tuple t in $\mathcal{I}(R)$ s.t. $|\mathcal{I}(R)|_t > |\mathcal{SQL}(R)|_t$. For instance, the intended answer for *awards* contains *Anna* once, which is represented as $|\mathcal{I}(\textsf{awards})|_{\textsf{(Anna)}} = 1$. However, the computed answer does not include this tuple: $|\mathcal{SQL}(\textsf{awards})|_{\textsf{(Anna)}} = 0$. Thus, *('Anna')* is a missing tuple for *awards*. In order to define the key concept of erroneous relation we need the following auxiliary concept. Let R be either a query or a relation. The *expectable* answer for R w.r.t. d, $\mathcal{E}(R)$, is defined as:

1. If R is a table, $\mathcal{E}(R) = d(R)$, with d the database schema instance.
2. If R is a view, then $\mathcal{E}(R) = \mathcal{E}(Q)$, with Q the query defining R.
3. If R is a query $\mathcal{E}(R) = \Phi_R(\mathcal{I}(R_1), \ldots, \mathcal{I}(R_n))$ with $R_1, \ldots, R_n$ the relations occurring in R.

Thus, in the case of a table, the expectable answer is its instance. In the case of a view V, the expectable answer corresponds to the computed result that would be obtained assuming that all the relations R_i occurring in the definition of V contain the intended answers. Then, $\mathcal{I}(R) \neq \mathcal{E}(R)$ indicates that R does not compute its intended answer, even assuming that all the relations it depends on contain their intended answers. Such relation is called *erroneous*. In our running example, the real cause of the missing answer for the view *awards* is the erroneous definition of the view *intensive*.

3 Debugging Algorithm

In this section we present the algorithm that defines our debugging technique, describing the purpose of each function. Although the process is based on the ideas of declarative debugging, this proposal does not use computation trees explicitly. Instead, our debugger represents computation trees by means of Horn clauses, denoted as $\mathsf{H} \leftarrow \mathsf{C}_1, \ldots, \mathsf{C}_n$, where the comma represents the conjunction, and $\mathsf{H}, \mathsf{C}_1, \ldots, \mathsf{C}_n$ are positive atoms. As usual, a *fact* H stands for the clause $\mathsf{H} \leftarrow \mathsf{true}$. Next, we describe the functions that define the algorithm, although the

Code 1 debug(V)

```
Input: V: view name
Output: A list of buggy views
1: A := askOracle(all V)
2: P := initialSetOfClauses(V, A)
3: while getBuggy(P)=[ ] do
4:     LE := getUnsolvedEnquiries(P)
5:     E := chooseEnquire(LE)
6:     A := askOracle(E)
7:     P := P $\cup$ processAnswer(E,A)
8: end while
9: return (getBuggy(P))
```

code of some basic auxiliary functions is omitted for the sake of space. This is the case of *getSelect*, *getFrom*, *getWhere*, and *getGroupBy* which return the different sections of a SQL query. In *getFrom*, we assume that every relations has an alias. The result is a sequence of elements of the form *R* as *R'*. A Boolean expression like *getGroupBy(Q)*=[] is satisfied if the query Q has no group by section. Function *getRelations(R)* returns the set of relations involved in R. It can be applied to queries, tables and views: if R is a table, then *getRelations(R)* $= \{R\}$, if R is a query, then *getRelations(R)* is the set of relations occurring in the definition of the query, and if R is a view, then *getRelations(R)* = *getRelations(Q)*, with Q the query defining R. The function *generateUndefined(R)* generates a tuple whose arity is the number of attributes in R containing only undefined values $(\perp, \ldots, \perp)$.

The general schema of the algorithm is summarized in the code of function *debug* (Code 1). The debugger is started by the user when an unexpected answer is obtained as computed answer for some SQL view V. In our running example, the debugger is started with the call *debug(awards)*. Then, the algorithm asks the user about the type of error (line 1). The answer A can be simply *valid*, *nonvalid*, or a more detailed explanation of the error, like *wrong(t)* or *missing(t)*, indicating that t is a wrong or missing tuple respectively. In our example, A takes the initial value *missing(('Anna'))*. During the debugging process, variable P keeps a list of Horn clauses representing a logic program. The initial list of clauses P is generated by the function *initialSetofClauses* (line 2). The purpose of the main loop (lines 3-8) is to add information to the program P, until a buggy view can be inferred. The function *getBuggy* returns the list of all the relations R such that *buggy(R)* can be proven w.r.t. the logic program P. The clauses in P contain enquiries that might imply questions to the user. Each iteration of the loop represents the election of an enquiry in a body atom whose validity has not been established yet (lines 4-5). Then, an enquiry about the result of the query is asked to the user (line 6). Finally, the answer is processed (line 7). Next, we explain in detail each part of this main algorithm.

Code 2 corresponds to the initialization process of line 2 from Code 1. The function *initialSetofClauses* gets as first input parameter the initial view V. This

Code 2 initialSetofClauses(V, A)

Input: V: view name, A: answer
Output: A set of clauses

```
1: P := ∅
2: P := initialize(V)
3: P := P ∪ processAnswer((all V), A)
4: return P
```

createBuggyClause(V)
Input: V: view name
Output: A Horn clause

```
1: [$R_1$, ..., $R_n$] := getRelations(V)
2: return { buggy(V)← state((all V), nonvalid),
                state((all $R_1$), valid), ..., state((all $R_n$), valid)). }
```

initialize(R)
Input: R: relation
Output: A set of clauses

```
1: P := createBuggyClause(R)
2: for each $R_i$ in getRelations(R) do
3:     P := P ∪ initialize($R_i$)
4: end for
5: return P
```

view has returned an unexpected answer, and the input parameter A contains the explanation. The output of this function is a set of clauses representing the logic relations that define possible buggy relations with predicate *buggy*. Initially it creates the empty set of clauses and then it calls the function *initialize* (line 2), a function that traverses recursively all the relations involved in the definition of the initial view *V*, calling *createBuggyClause* with V as input parameter. *createBuggyClause* adds a new clause indicating the enquiries that must hold in order to consider V as incorrect: it must be nonvalid, and all the relations it depends on must be valid. Next is part of the initial set of clauses generated for the running example of this paper:

```
buggy(awards)    :- state(all(awards),nonvalid),
                    state(all(basic),valid), state(all(intensive),valid).
buggy(basic)     :- state(all(basic),nonvalid), state(all(standard),valid).
buggy(intensive) :- state(all(intensive),nonvalid),
                    state(all(allInOneCourse),valid), state(all(standard),valid).
    ...
```

The correlation between these clauses and the dependency tree is straightforward. Finally, in line 3, function *processAnswer* incorporates the information that can be extracted from A into the program *P*. The information about the validity/nonvalidity of the results associated to enquiries is represented in our setting with predicate *state*. The first parameter is an enquiry *E*, and the second one can be either *valid* or *nonvalid*. Enquiries can be of any of the following forms: *(all R)*, *(s ∈ R)*, or *(R' ⊆ R)* with *R*, *R'* relations, and *s* a tuple with the same schema as relation *R*. Each enquiry *E* corresponds to a specific question with a possible set of answers and an associated complexity $\mathcal{C}(E)$:
- If $E \equiv$ *(all R)*. Let $\mathsf{S} = \mathcal{SQL}(\mathsf{R})$. The associated question asked to the user is *"Is S the intended answer for R?"* The answer can be either *yes* or *no*. In the case of *no*, the user is asked about the type of the error, *missing* or *wrong*,

Code 3 processAnswer(E,A)

```
Input: E: enquiry, A: answer obtained for the enquiry
Output: A set of new clauses
if A ≡ yes then
   P := {state(E,valid).}
else if A ≡ no or A ≡ missing(t) or A ≡ wrong(t) then
   P := {state(E,nonvalid).}
end if
if E ≡ (s ∈ R) then
   if (s ∈ SQL(R) and A ≡ no) then
      P:= P ∪ processAnswer((all R),wrong(s))
   else if (s ∉ SQL(R) and A ≡ yes) then
      P:= P ∪ processAnswer((all R),missing(s))
   end if
else if E ≡ (V ⊆ R) and (A ≡ wrong(s)) then
   P:= P ∪ processAnswer((all R), A)
else if E ≡ (all V) with V a view and (A ≡ missing(t) or A ≡ wrong(t)) then
   Q := SQL query defining V
   P := P ∪ slice(V,Q,A)
end if
return P
```

giving the possibility of providing a witness tuple t. If the user provides this information, the answer is changed to *missing(t)* or *wrong(t)*, depending on the type of the error. We define $\mathcal{C}(E) = |S|$, with $|S|$ the number of tuples in S.

-If $E \equiv (R' \subseteq R)$. Let $S = \mathcal{SQL}(R')$. Then the associated question is *"Is S included in the intended answer for R?"* As in the previous case the answer allowed can be *yes* or *no*. In the case of *no*, the user can point out a wrong tuple $t \in S$ and the answer is changed to *wrong(t)*. $\mathcal{C}(E) = |S|$ as in the previous case.

- If $E \equiv (s \in R)$. The question is *"Does the intended answer for R include a tuple s?"* The possible answer can be *yes* or *no*. No further information is required from the user. In this case $\mathcal{C}(E) = 1$, because only one tuple must be considered.

In the case of *wrong*, the user typically points to a tuple in the result R. In the case of *missing*, the tuple must be provided by the user, and in this case *partial* tuples, i.e., tuples including some undefined attributes are allowed. The answer *yes* corresponds to the state *valid*, while the answer *no* corresponds to *nonvalid*. An atom *state(q,s)* occurring in a clause body, is a *solved enquiry* if the logic program P contains at least one fact of the form *state(q, valid)* or *state(q, nonvalid)*, that is, if the enquiry has been already solved. The atom is called an *unsolved enquiry* otherwise. The function *getUnsolvedEnquiries* (see line 4 of Code 1) returns in a list all the unsolved enquiries occurring in P. The function *chooseEnquiry* (line 5, Code 1) chooses one of these enquiries according to some criteria. In our case we choose the enquiry E that implies the smaller complexity value $\mathcal{C}(E)$, although other more elaborated criteria could be defined without affecting the theoretical results supporting the technique. Once the enquiry has been chosen, Code 1 uses the function *askOracle* (line 6) in order to ask for the associated question, returning the answer of the user. We omit the definitions of these simple functions for the sake of space.

The code of function *processAnswer* (called in line 7 of Code 1), can be found in Code 3. The first lines (1-5) introduce a new logic fact in the program with the state that corresponds to the answer A obtained for the enquiry E. In our running example, the fact *state(all(awards), nonvalid)* is added to the program. The rest of the code distinguishes several cases depending on the form of the enquiry and its associated answer. If the enquiry is of the form *($s \in R$)* with answer *no* (meaning $s \notin \mathcal{I}(R)$), and the debugger checks that the tuple s is in the computed answer of the view R (line 7), then s is wrong in the relation R. In this case, the function *processAnswer* is called recursively with the enquiry *(all R)* and *wrong(s)* (line 8). If the answer is *yes* and the debugger checks that s does not belong to the computed answer of R (line 10), then s is missing in the relation R. For enquiries of the form *($V \subseteq R$)* and answer *wrong(s)*, it can be ensured that s is wrong in R (line 13). If the enquiry is *(all V)* for some view V, and with an answer including either a wrong or a missing tuple, the function *slice* (line 16) is called. This function exploits the information contained in the parameter A (*missing(t)* or *wrong(t)*) for slicing the query Q in order to produce, if possible, new clauses which will allow the debugger to detect incorrect relations by asking simpler questions to the user. The implementation of *slice* can be found in Code 4. The function receives the view V, a subquery Q, and an answer A as

Code 4 slice(V,Q,A)

```
Input: V: view name, Q: query, A: answer
Output: A set of new clauses
 1: P := $\emptyset$;   $S$= $\mathcal{SQL}(Q)$;   $S_1$= $\mathcal{SQL}(Q_1)$;   $S_2$= $\mathcal{SQL}(Q_2)$
 2: if (A $\equiv$ wrong(t) and Q $\equiv$ $Q_1$ union [all] $Q_2$) or
       (A $\equiv$ missing(t) and Q $\equiv$ $Q_1$ intersect [all] $Q_2$) then
 3:     if $|S_1|_t = |S|_t$ then P:= P $\cup$ slice(V, $Q_1$, A)
 4:     if $|S_2|_t = |S|_t$ then P:= P $\cup$ slice(V, $Q_2$, A)
 5: else if A $\equiv$ missing(t) and Q $\equiv$ $Q_1$ except [all] $Q_2$ then
 6:     if $|S_1|_t = |S|_t$ then P:= P $\cup$ slice(V, $Q_1$, A)
 7:     if Q $\equiv$ $Q_1$ except $Q_2$ and $t \in S_2$ then P :=P$\cup$ slice(V,$Q_2$,wrong(t))
 8: else if basic(Q) and groupBy(Q)=[] then
 9:     if A $\equiv$ missing(t) then P := P $\cup$ missingBasic(V, Q, t)
10:     else if A $\equiv$ wrong(t) then P := P $\cup$ wrongBasic(V, Q, t)
11: end if
12: return P
```

parameters. Initially, Q is the query defining V, and A the user answer, but this situation can change in the recursive calls. The function distinguishes several particular cases:
- The query Q combines the results of Q_1 and Q_2 by means of either the operator union or union all, and A is *wrong(t)* (first part of line 2). Then query Q produces

too many copies of t. Then, if any Q_i produces as many copies of t as Q, we can blame Q_i as the source of the excessive number of t's in the answer for V (lines 3 and 4). The case of subqueries combined by the operator intersect [all], with $A \equiv$ *missing(t)* is analogous, but now detecting that a subquery is the cause of the scanty number of copies of t in $\mathcal{SQL}(V)$.

- The query Q is of the form Q_1 except [all] Q_2, with $A \equiv$ *missing(t)* (line 5). If the number of occurrences of t in both Q and Q_1 is the same, then t is also missing in the query Q_1 (line 6). Additionally, if query Q is of the particular form Q_1 except Q_2, which means that we are using the difference operator on sets (line 7), then if t is in the result of Q_2 it is possible to claim that the tuple t is wrong in Q_2. Observe that in this case the recursive call changes the answer from *missing(t)* to *wrong(t)*.

- If Q is defined as a basic query without group by section (line 8), then either function *missingBasic* or *wrongBasic* is called depending on the form of A.

Both *missingBasic* and *wrongBasic* can add new clauses that allow the system to infer buggy relations by posing questions which are easier to answer. Function *missingBasic*, defined in Code 5, is called (line 9 of Code 4) when A is *missing(t)*. The input parameters are the view V, a query Q, and the missing

Code 5 missingBasic(V,Q,t)

```
Input: V: view name, Q: query, t: tuple
Output: A new list of Horn clauses
 1: P := $\emptyset$;    S := $\mathcal{SQL}$(SELECT getSelect(Q) FROM getFrom(Q) )
 2: if t $\notin$ S then
 3:    for (R AS S) in (getFrom(Q)) do
 4:       s = generateUndefined(R)
 5:       for i=1 to length(getSelect(Q)) do
 6:          if $t_i \neq \perp$ and member(getSelect(Q),i) = S.A, A attrib., then s.A = $t_i$
 7:       end for
 8:       if s $\notin \mathcal{SQL}$(R) then
 9:          P := P $\cup$ { (buggy(V) $\leftarrow$ state((s $\in$ R), nonvalid).) }
10:       end if
11:    end for
12: end if
13: return P
```

tuple t. Notice that Q is in general a component of the query defining V. For each relation R with alias S occurring in the from section, the function checks if R contains some tuple that might produce the attributes of the form $S.A$ occurring in the tuple t. This is done by constructing a tuple s undefined in all its components (line 4) except in those corresponding to the select attributes of the form $S.A$, which are defined in t (lines 5 - 7). If R does not contain a tuple matching s in all its defined attributes (line 8), then it is not possible to obtain the tuple t in V from R. In this case, a buggy clause is added to the program

Code 6 wrongBasic(V,Q,t)

```
Input: V: view name, Q: query, t: tuple
Output: A set of clauses
1: P := $\emptyset$
2: F := getFrom(Q)
3: N := length(F)
4: for i=1 to N do
5:    $R_i$ as $S_i$ := member(F,i)
6:    relevantTuples($R_i$,$S_i$,$V_i$, Q, t)
7: end for
8: P := P $\cup$ { (buggy(V) $\leftarrow$ state(($V_1 \subseteq R_1$), valid), ..., state(($V_n \subseteq R_n$), valid).) }
9: return P
```

Code 7 relevantTuples(R_i,R',V,Q,t)

```
Input: $R_i$: relation, R': alias,
    V: new view name, Q: Query, t: tuple
Output: A new view in the database schema
1: Let $A_1$, ..., $A_n$ be the attributes defining $R_i$
2: $\mathcal{SQL}$(create view V as
      (select $R_i.A_1$, ..., $R_i.A_n$ from $R_i$)
          intersect all
      (select R'.$A_1$, ..., R'.$A_n$ from getFrom(Q)
      where getWhere(Q) and eqTups(t,getSelect(Q))))
```

eqTups(t,s)

```
Input: t,s : tuples
Output: SQL condition
1: C := true
2: for i=1 to length(t) do
3:    if $t_i \neq \perp$ then
4:       C:= C AND $t_i = s_i$
5: end for
6: return C
```

P (line 9) meaning that if the answer to the question *"Does the intended answer for R include a tuple s?"* is *no*, then V is an incorrect relation.

The implementation of *wrongBasic* can be found in Code 6. The input parameters are again the view V, a query Q, and a tuple t. In line 1, this function creates an empty set of clauses. In line 2, variable F stands for the set containing all the relations in the from section of the query Q. Next, for each relation $R_i \in F$ (lines 4 - 7), a new view V_i is created in the database schema after calling the function *relevantTuples* (line 6), which is defined in Code 7. This auxiliary view contains only those tuples in relation R_i that contribute to produce the wrong tuple t in V. Finally, a new buggy clause for the view V is added to the program P (line 8) explaining that the relation V is buggy if the answer to the question associated to each enquiry of the form $V_i \subseteq \mathrm{R}_i$ is *yes* for $i \in \{1 \ldots n\}$.

The following theoretical results guarantee that the technique is reliable.

Theorem 1. *Let* R *be a relation. Then:*
Correctness: If the call debug(R) *returns a list* L*, then all relation names contained in* L *are erroneous relations.*
Completeness: Let A be the answer obtained after the call to askOracle(all R) *in line 1 of Code 1. If A is of the form* nonvalid, wrong(t) *or* missing(t)*, then the call* debug(R) *(defined in Code 1) returns a list L containing at least one relation.*

Thus, the algorithm always stops pointing to some user view (completeness) which is incorrectly defined (correctness).

4 Implementation

The algorithm presented in Section 3 has been implemented in the Datalog Educational System (DES [14, 15]). The debugger is started when the user detects that *Anna* is not among the (large) list of student names produced by view *awards*. The command `/debug_sql` starts the session:

```
1: DES-SQL> /debug_sql awards
2: Info: Debugging view 'awards': {   1 - awards('Carla'), ... }
3: Is this the expected answer for view 'awards'?  m'Anna'
4: Does the intended answer for 'intensive' include  ('Anna')       ? n
5: Does the intended answer for 'standard'  include  ('Anna',1,true) ? y
6: Does the intended answer for 'standard'  include  ('Anna',2,true) ? y
7: Does the intended answer for 'standard'  include  ('Anna',3,false)? y
8: Info: Buggy relation found: intensive
```

The user answer *m'Anna'* in line 3 indicates that *('Anna')* is missing in the view *awards*. In line 4 the user indicates that view *intensive* should not include *('Anna')*. In lines 5, 6, and 7, the debugger asks three simple questions involving the view *standard*. After checking the information for *Anna*, the user indicates that the listed tuples are correct. Then, the tool points out *intensive* as the buggy view, after only five simple questions. Observe that intermediate views can contain hundreds of thousands of tuples, but the slicing mechanism helps to focus only on the source of the error. Next, we describe briefly how these questions have been produced by the debugger.

After the user indicates that *('Anna')* is missing, the debugger executes a call *processAnswer(all(awards),missing((Anna)))*. This implies a call to *slice(awards,* Q_1 except Q_2*, missing(('Anna')))* (line 16 of Code 3). The debugger checks that Q_2 produces *('Anna')* (line 7 of Code 4), and proceeds with the recursive call *slice(awards,* Q_2*, wrong(('Anna')))* with $Q_2 \equiv$ select student from intensive. Query Q_2 is basic, and then the debugger calls *wrongBasic(awards,* Q_2*, ('Anna'))* (line 10 of Code 4)). Function *wrongBasic* creates a view that selects only those tuples from *intensive* producing the wrong tuple *('Anna')* (function *relevantTuples* in Code 7):

```
create view intensive_slice(student) as
(select * from intensive)
intersect all
(select * from intensive I where  I.student = 'Anna');
```

Finally the following buggy clause is added to the program P (line 8, Code 6):

```
buggy(awards) :- state(subset(intensive_slice,intensive),valid).
```

By enabling development listings with the command `/development on`, the logic program is also listed during debugging. The debugger chooses the only body atom in this clause as next unsolved enquiry, because it only contains one tuple.

The call to *askOracle* returns *wrong(('Anna'))* (the user answers *'no'* in line 4). Then *processAnswer(subset(intensive_slice,intensive), wrong(('Anna')))* is called, which in turn calls to *processAnswer(all(intensive),wrong(('Anna')))* recursively. Next call is *slice(intensive, Q, wrong(('Anna')))*, with $Q \equiv Q_3$ union Q_4 the query definition of *intensive* (see Figure 1). The debugger checks that only Q_4 produces *('Anna')* and calls to *slice(intensive,* Q_4*, wrong(('Anna')))*. Query Q_4 is basic, which implies a call to *wrongBasic(intensive,* Q_4*, ('Anna'))*. Then *relevantTuples* is called three times, one for each occurrence of the view *standard* in the from section of Q_4, creating new views:

```
create view standard_slicei(student,level,pass) as
   ( select R.student, R.level, R.pass from standard as R)
       intersect all
   (select  A1.student, A1.level, A1.pass
    from standard as A1,  standard as A2, standard as A3
    where (A1.student = A2.student and A2.student = A3.student
       and A1.level = 1 and A2.level = 2 and A3.level = 3)
       and A1.student = 'Anna');
```

for $i = 1 \ldots 3$. Finally, the clause:

```
buggy(intensive) :- state(subset(standard_slice1,standard),valid),
                    state(subset(standard_slice2,standard),valid),
                    state(subset(standard_slice3,standard),valid).
```

is added to P (line 8, Code 6). Next, the tool selects the unsolved question with less complexity that correspond to the questions of lines 5, 6, and 7, for which the user answer *yes*. Therefore, the clause for buggy(intensive) succeeds and the algorithm finishes.

5 Conclusions

We have presented a new technique for debugging systems of SQL views. Our proposal refines the initial idea presented in [4] by taking into account information about *wrong* and *missing* answers provided by the user. Using a technique similar to dynamic slicing [1], we concentrate only in those tuples produced by the intermediate relations that are relevant for the error. This minimizes the main problem of the technique presented in [4], which was the huge number of tuples that the user must consider in order to determine the validity of the result produced by a relation. Previous works deal with the problem of tracking provenance information for query results [9, 7], but to the best of our knowledge, none of them treat the case of missing tuples, which is important in our setting.

The proposed algorithm looks for particular but common error sources, like tuples missed in the from section or in *and* conditions (that is, *intersect* components in our representation). If such shortcuts are not available, or if the user only answers *yes* and *no*, then the tools works as a pure declarative debugger.

A more general contribution of the paper is the idea of representing a declarative debugging computation tree by means of a set of logic clauses. In fact, the algorithm in Code 1 can be considered a general debugging schema, because it is independent of the underlying programming paradigm. The main advantage of this representation is that it allows combining declarative debugging with other diagnosis techniques that can be also represented as logic programs. In our case, declarative debugging and slicing cooperate for locating an erroneous relation. It would be interesting to research the combination with other techniques such as the use of assertions.

References

1. H. Agrawal and J. R. Horgan. Dynamic program slicing. *SIGPLAN Not.*, 25:246–256, June 1990.
2. ApexSQL Debug , 2011. `http://www.apexsql.com/sql_tools_debug.aspx/`.
3. R. Caballero, Y. García-Ruiz, and F. Sáenz-Pérez. A theoretical framework for the declarative debugging of datalog programs. In *SDKB 2008*, volume 4925 of *LNCS*, pages 143–159. Springer, 2008.
4. R. Caballero, Y. García-Ruiz, and F. Sáenz-Pérez. Algorithmic Debugging of SQL Views. In *Proceedings of the 8th International Andrei Ershov Memorial Conference, PSI 2011*, volume 7162. Springer LNCS, 2012.
5. R. Caballero, F. López-Fraguas, and M. Rodríguez-Artalejo. Theoretical Foundations for the Declarative Debugging of Lazy Functional Logic Programs. In *Proc. FLOPS'01*, number 2024 in LNCS, pages 170–184. Springer, 2001.
6. S. Ceri and G. Gottlob. Translating SQL Into Relational Algebra: Optimization, Semantics, and Equivalence of SQL Queries. *IEEE Trans. Softw. Eng.*, 11:324–345, April 1985.
7. Y. Cui, J. Widom, and J. L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25:179–227, June 2000.
8. H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2008.
9. B. Glavic and G. Alonso. Provenance for nested subqueries. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 982–993, New York, NY, USA, 2009. ACM.
10. P. W. Grefen and R. A. de By. A multi-set extended relational algebra: a formal approach to a practical issue. In *ICDE'94*, pages 80–88. IEEE, 1994.
11. L. Naish. A Declarative Debugging Scheme. *Journal of Functional and Logic Programming*, 3, 1997.
12. H. Nilsson. How to look busy while being lazy as ever: The implementation of a lazy functional debugger. *Journal of Functional Programming*, 11(6):629–671, 2001.
13. Rapid SQL Developer Debugger, 2011. `http://docs.embarcadero.com/products/rapid_sql/`.
14. F. Sáenz-Pérez. Datalog Educational System v3.0, March 2012. `http://des.sourceforge.net/`.
15. Fernando Sáenz-Pérez, Rafael Caballero, and Yolanda García-Ruiz. A deductive database with Datalog and SQL query languages. In 9th Asian Symposium, APLAS 2011, LNCS 7078. Springer, 2011.
16. E. Shapiro. *Algorithmic Program Debugging.* ACM Distiguished Dissertation. MIT Press, 1982.
17. J. Silva. A survey on algorithmic debugging strategies. *Advances in Engineering Software*, 42(11):976–991, 2011.
18. SQL, ISO/IEC 9075:1992, third edition, 1992.