Semantic Errors in SQL Queries: Exam Evaluation 2003-03

Christian Goldberg
Martin-Luther-Universität Halle-Wittenberg
goldberg@informatik.uni-halle.de

Abstract

We investigate classes of SQL queries which are syntactically correct, but certainly not intended, no matter for which task the query was written. For instance, queries that are contradictory, i.e. always return the empty set, are obviously not intended. Current database management systems, e.g. Oracle, execute such queries without any warning.

In this evaluation, we give a statistic of such errors for one special exam and list the concerning SQL exercises and their possible solutions. Section 1 contains important data of the analyzed exam. In section 2 we explain the database scheme(s) that is/are used in the listed exercises together with their possible solutions in section 3. Section 4 conducts a survey on the number and sorts of occurred semantic errors.

1 Exam Data

Lecture Title	:	Database Systems I
Term	:	Winter 2003/2004 - midterm exam
Lecturer	:	Prof. Dr. Stefan Brass
University	:	Martin-Luther-University Halle, Germany
Analysis	:	Christian Goldberg
Date of Analysis	:	September 2006
Error Code Reference	:	[1]

2 Underlying Database Scheme

In the following exercises, we use a database scheme for a online learning system which stores information about sections, terms/topics and sources:

```
SECTION(SID, TITLE, URL, PAGES)
TERM(TID, HANDLE, TOPIC→TERM)
SOURCE(SID→SECTION, TID→TERM, TYPE)
```

The type of source can be 'DEF' for a definition of that term, 'EX' for an example or 'KNOWN' if the term is used and it is assumed to be known. The column TOPIC can be null.

3 Analyzed Exercises and Possible Solutions

The midterm exam "Database Systems I" in winter 2003/2004 contained 13 exercises about integrity constraints, relational calculus and SQL. The 3 analyzed SQL queries resulted in 9 out of 23 points. The 153 participating students had 60 minutes to solve the exercises and were allowed to use the lecture script or other notes but no electronic resources.

It was pointed out that unnecessary complications, unnecessary `DISTINCT` and many duplicates may result in a deduction of points.

3.1 Exercise 3a)

Which section title contains both "SQL" and "queries" as substrings. Request the ID, title and URL of these sections.

```
SELECT SID, TITLE, URL
FROM   SECTION
WHERE  TITLE LIKE '%SQL%'
AND    TITLE LIKE '%queries%'
```

3.2 Exercise 3b)

Request `SID` and title of all sections that define at least one term that is a top level topic. This means that `TOPIC` contains a null value for that term.

```
SELECT S.SID, S.TITLE
FROM   SECTION S
WHERE  EXISTS (SELECT *
               FROM   TERM T, SOURCE O
               WHERE  O.SID=S.SID
               AND    T.TID=O.TID
               AND    O.TYPE='DEF')
```

3.3 Exercise 3c)

Write a query to list all terms that are used in a section A with a lower `SID` than the section B in which they were defined. List the `HANDLE` of the term and the `SID` for both sections A and B. The columns shall be named "`TERM`", "`ASSUMED_IN`" and "`DEFINED_IN`".

```
SELECT T.HANDLE, A.SID, B.SID
FROM   TERM T, SOURCE A, SOURCE B
WHERE  A.SID<B.SID
AND    T.TID=A.TID AND T.TID=B.TID
AND    A.TYPE='KNOWN' AND B.TYPE='DEF'
```

4 Statistics

The list of error types mentioned in [1] is based on our experience from grading a large number of exams and homeworks. After this error taxonomy was finished, we analyzed the solutions of the SQL exercises in several exams of the course "Databases I" at the University of Halle The results for the midterm exam in winter term 2003/2004 are shown in Figure 1. The exercises are numbered with the numbers and letters from section 3, Further course material and exam exercises are available from the project web page ([4]).

Error	3a	3b	3c	$\sum$
1	3	4	14	21
2	2	-	5	7
4	-	-	2	2
5	-	-	2	2
6	-	-	25	25
7	5	-	8	13
8	2	2	6	10
9	-	6	-	6
12	2	3	2	7
23	1	-	-	1
27	1	3	29	33
31	-	5	1	6
34	11	-	-	11
37	-	46	-	46
Correct	104	37	30	37.3%
Only Semantic	20	50	50	26.1%
Wrong Task	12	30	9	11.1%
Not Counted	4	8	32	9.6%
Syntax and Semantic	6	11	23	8.7%
Only Syntax	7	17	9	7.2%

Figure 1: Error statistics for midterm winter exam 2003/2004

The number of exams that contained at least one semantic error is the sum of the entries "Only semantics" and "Syntax and Semantic". Of course we counted only semantic errors from our list in [1], i.e. that are detectable without knowing the task of the query. "Wrong task" lists the number of exams that can only be detected as incorrect if the goal of the query is known. "Not counted" lists exams that did not try the particular exercise, or that contained so severe syntax errors that looking at semantic errors in detail was not possible. In this exam that we analyzed with this error taxonomy, the occurred semantic errors are (percentages are relative to all detected semantic errors):

1.	24.2 %	Error 37: Many duplicates
2.	17.4 %	Error 27: Missing join condition
3.	13.2 %	Error 6: Unnecessary join
4.	11.1 %	Error 1: Inconsistent condition
5.	6.8 %	Error 7: Tuple variables are always identical
6.	5.8 %	Error 34: Wildcards without `LIKE`
7.	5.3 %	Error 8: Implied, tautological or inconsistent subcondition

References

[1] Stefan Brass and Christian Goldberg. Semantic Errors in SQL Queries: A Quite Complete List. In: *Elsevier's Journal of Systems and Software 79(5)*, 2006.

[2] Stefan Brass and Christian Goldberg. Proving the Safety of SQL Queries. In: *Fifth International Conference on Quality Software (QSIC'05)*, IEEE Computer Society Press, 2005.

[3] Stefan Brass and Christian Goldberg. Detecting Logical Errors in SQL Queries. In: *16th Workshop on Foundations of Databases (GvD'04)*, 2004.

[4] Stefan Brass and Christian Goldberg. SQLLint: Detecting Logical Errors in SQL Queries. Project website: http://dbs.informatik.uni-halle.de/sqllint/