Homework Assignment 2  
(due Thursday, October 27, 2011 - in class)

Suggested Reading:

Chapters 4, 5, and 7 of the “Foundations of Databases” books. Material about SQL can also be found in the deck of slides titled “The Influence of Relational Algebra and Relational Calculus on SQL (available at the course webpages). The last part of this deck contains material about the syntax and semantics of aggregate functions in SQL; an understanding of this material is needed for Problem 6 below. You should also start reading the classical papers by A.K. Chandra and P.M. Merlin on conjunctive queries and by M.Y. Vardi on the complexity of relational query languages (both available at the course webpages).

Problems:

1. For each of the following operations on relations, give a relational calculus expression that defines it.

   (a) The symmetric part $R^*$ of a binary relation $R$, where $R^*$ consists of all pairs $(a, b)$ such that both $(a, b)$ and $(b, a)$ are in $R$.

   (b) The symmetric difference $R \Delta S$ of two ternary relations, where $R \Delta S$ consists of all tuples that belong to exactly one of $R$ and $S$.

   (c) The composition $R_1 \circ R_2$ of two binary relations $R_1$, $R_2$, where $R_1 \circ R_2$ consists of all pairs $(a, c)$ for which there is an element $b$ such that $(a, b) \in R_1$ and $(b, c) \in R_2$.

   (d) Recall from the previous homework assignment that the semijoin $R \bowtie S$ of two relations $R$ and $S$ is the relation consisting of all tuples $t$ in $R$ such that there is at least one tuple in $S$ that agrees with $t$ in all attributes that $R$ and $S$ have in common.

   Assume now that the attributes of $R$ are $A, B, C$ and the attributes of $S$ are $B, C, D$. Give a relational calculus expression for $R \bowtie S$.

2. An airline maintains a FLIGHTS database that includes a table called DIRECT with two attributes FROM and TO containing information about direct flights between two cities. Give a relational calculus expression for the relation AT-MOST-TWO consisting of all pairs $(c, d)$ of cities such that one can travel from city $c$ to city $d$ with at most two intermediate stops.

3. Consider the beer drinkers database consisting of relations over the following relational schema:

   FREQUENTS(drinker, bar), LIKES(drinker, beer), SERVES(bar, beer)

   Give relational calculus expressions for the following queries:

   (a) “List all drinkers who frequent at least one bar that serves at least one beer they like”.

   (b) “List all drinkers who frequent only bars that serve at least one beer they like.”

   (c) “List all drinkers who frequent every bar that serves ANCHOR STEAM”.

Optional Problem: Give SQL expressions for the above queries.
4. For each of the following relational calculus expressions determine whether or not it is domain independent. If it is, give a translation to an equivalent relational algebra expression; if it is not, give an example of two different domains yielding different answers on the same database instance.

(a) \{ (x, y) : \exists z (R(y, z) \land x \neq z) \}
(b) \{ x : \forall y (R(x, y) \rightarrow T(x, y)) \}
(c) \{ x : P(x) \land (\forall y (R(x, y) \rightarrow T(x, y))) \}
(d) \{ x : \exists y \forall z Q(y, z, x) \}.

5. Suppose that you work with a full-fledged implementation of SQL; in particular, the system directly supports both the difference operation \( R - S \) and the intersection operation \( R \cap S \) using the EXCEPT and INTERSECT constructs, respectively. Suddenly, both these features are disabled in your system. Explain how you can use the remaining constructs of SQL to write expressions that express \( R - S \) and \( R \cap S \).

6. The Computing Research Association (CRA) collects data on graduate students and faculty in Ph.D.-granting computer science departments in North America. Based on these data, CRA publishes the Taulbee Survey each year, which can be found at http://www.cra.org/resources/taulbee/; you may want to take a look at Table 26-1, p. 35 of the most recent Taulbee Survey, as this homework problem is based on that table.

CRA maintains a database that includes the following two tables:

GRAD-STIPEND\( (\text{university}, \text{dpt-type}, \text{id-no}, \text{title}, \text{amount}) \), RANKING\( (\text{university}, \text{rank}) \)

The values of dpt-type are “US-CS”, “US-CE”, “US Information”, and “Canadian”; the values of id-no are strings encoding names of graduate students receiving support; the values of title are “TA” (for Teaching Assistant) and “RA” (for Research Assistant); the values of rank are positive integers denoting the ranking of each university according to the latest survey by the National Research Council.

Express the following queries in SQL:

(a) Find the average, maximum and minimum stipend paid to each title in each department type.
(b) Find the average, maximum and minimum salary paid to each title in universities ranked 1-12 (included).
(c) For each university ranked 1-12 (included), find the average, minimum and maximum salary paid to each title.

7. Consider the following decision problem, called NATURAL JOIN NON-EMPTINESS: given relations \( R_1, \ldots, R_m \), is their natural join \( R_1 \bowtie \cdots \bowtie R_m \) non-empty? In this version, both the number of relations given and the arity of each relation are arbitrary positive integers.

(a) Show that NATURAL JOIN NON-EMPTINESS is in NP.
(b) Show that NATURAL JOIN NON-EMPTINESS is an NP-hard problem (Hint: Use a reduction from one of the NP-complete problems discussed in class).
(c) Show that NATURAL JOIN NON-EMPTINESS is NP-hard even if the given relations are of arity at most 3 (Hint: Use a reduction from one of the NP-complete problems discussed in class). Note that a solution to this part of the problem gives also a solution to the preceding part, so you need only produce a solution for this part.

(d) Let $M$ and $K$ be fixed positive integers. Consider the restriction of NATURAL JOIN NON-EMPTINESS in which we are given at most $M$ relations and each relation has arity at most $K$. What can you say about the computational complexity of this problem? Is it NP-hard? Is it solvable in polynomial time?

8. Recall that a 2CNF-formula is a Boolean formula in conjunctive normal form such that each of its conjuncts is the disjunction of two literals, i.e., each conjunct has one of the following forms: $(x \lor y)$, $(\neg x \lor y)$, $(\neg x \lor \neg y)$. Recall also that 2Sat is the following decision problem: given a 2CNF formula $\varphi$, does $\varphi$ have a satisfying assignment? (i.e., an assignment of Boolean values to the variables of $\varphi$ so that each conjunct of $\varphi$ evaluates to 1 ("true").

(a) Show that 2Sat can be viewed as a special case of the HOMOMORPHISM PROBLEM. This entails coming up with a suitable relational database schema $S$ and showing that the question of whether a given 2CNF-formula $\varphi$ is satisfiable is the same question as to whether there is a homomorphism between two particular database instances over $S$. In fact, the homomorphisms between the two database instances should be precisely the satisfying assignments of $\varphi$.

(b) What changes will you have to make to your solution so that 3Sat can be viewed as a special case of the HOMOMORPHISM PROBLEM?

9. Let $R$ be a relational schema with attributes $A, B, C$. Consider the relational calculus expression

$$\pi_A(\pi_{A,C}(R) \bowtie \pi_{B,C}(R)).$$

(a) Give a conjunctive query written as a relational calculus expression that is equivalent to the above expression.

(b) Give a conjunctive query written as rule that is equivalent to the above expression.

Note 1: All written assignments, term project report or term paper, and final examination are to be submitted in a typeset format (preferably in \LaTeX).

Note 2: The purpose of the homework assignments is to help you develop a better understanding of the material covered in class and also to expose you to some material that lack of time prevents us from covering in class. You should try to work out these problems on your own using your notes from the class. You may also want to use the “Foundations of Databases” book and the papers posted at the course webpages as references. If you discuss any of the problems with other students in the class, you are still expected to produce your own write-up of the assignment; moreover, you are expected to state the names of all other students with whom you have shared ideas about the problems. Finally, please state what published or online sources you used, if any, in the solutions of the problems.