

Computer Science COMP2400, Lab 3

Solutions

September 11, 2007

1 Introduction

This exercise is an introduction to Relational Algebra. The notation used to express relational Algebra is:

π_A	Projection onto the set of attributes A
σ_p	Selection (restriction) by predicate p
$\rho_A(R)$	Rename a relation and/or attributes
\cup	Union of two relations
\cap	Intersection of two relations
\times	Cartesian product of two relations
\bowtie_C	Relational join of two relations with join condition C .
$\mathcal{R}\mathcal{F}_L$	Aggregation of unique values of R with group functions L

Please review the lecture notes and/or text to refresh your memory of the meaning of these operators.

We will be looking at relational queries against the company schema from the E&N and Lab 1. Lab3.zip on the course web site includes a full definition of this schema including the tables you were asked to create in the lab exercise, plus some more sample data.

2 Exercises

Work *in pairs* to complete the following exercises.

2.1 SQL to Relational Algebra

Translate the following SQL statements into relational algebra.

1. `SELECT * FROM employee`

Solution: `employee` (*trick question!*)

2. `SELECT fname, lname FROM employee`

Solution: `$\pi_{fname, lname}$ employee`

3. `SELECT fname, lname FROM employee WHERE enumber=1234`

Solution: `$\pi_{fname, lname}(\sigma_{enumber=1234}(\text{employee}))$`

4. `SELECT * FROM employee, department`

Solution: $\text{employee} \times \text{department}$

5. `SELECT * FROM employee, department
WHERE employee.dno = department.dnumber`

Solution:

$\text{employee} \bowtie_{\text{dno=dnumber}} \text{department}$

or

$\sigma_{\text{dno=dnumber}} (\text{employee} \times \text{department})$

6. `SELECT pname, sum(hours) FROM project p, works_on w
WHERE p.pnumber = w.pno
GROUP BY pname`

Solution:

$\text{pname} \mathcal{F}_{\text{sum(hours)}} (\text{project} \bowtie_{\text{pnumber=pno}} \text{works_on})$

7. `SELECT pname, dname, sum(hours)
FROM project p, works_on w, employee e, department d
WHERE p.pnumber = w.pno
AND e. enumber = w. eno
AND d.dnumber = e.dno
GROUP BY pname, dname`

Solution:

$t1 \leftarrow \text{project} \bowtie_{\text{pnumber=pno}} \text{works_on}$
 $t2 \leftarrow t1 \bowtie_{\text{eno=enumber}} \text{employee}$
 $t3 \leftarrow t2 \bowtie_{\text{dno=dnumber}} \text{department}$
 $\text{pname} \mathcal{F}_{\text{sum(hours)}} (t3)$

2.2 Relational Algebra to SQL

Translate the following relational algebra expressions into SQL statements.

1. $\pi_{\text{enumber, salary}} (\text{employee})$

Solution: `SELECT enumber, salary FROM employee`

2. $\pi_{\text{dname, fname, lname}} (\sigma_{\text{enumber=mgr_empid}} (\text{employee} \times \text{department}))$

Solution:

`SELECT dname, fname, lname
FROM employee, department
WHERE employee.enumber = department.mgr_eno`

- 3.

$\text{mgr_names} \leftarrow \pi_{\text{fname, lname, enumber}} (\text{employee} \bowtie_{\text{enumber=mgr_empid}} \text{department})$
 $\text{work_on_projects} \leftarrow \pi_{\text{eno, hours, pname}} (\text{works_on} \bowtie_{\text{pno=pnumber}} \text{project})$
 $\text{mgr_works_on} \leftarrow \text{department} \bowtie_{\text{mgr_empid=eno}} \text{works_on_projects}$
 $\text{mgr_hours_worked} \leftarrow \pi_{\text{mgr_empid, pname, hours}} (\text{mgr_works_on})$
 $\text{named_hours} \leftarrow \pi_{\text{fname, lname, hours}} \text{mgr_names} \bowtie_{\text{enumber=eno}} (\text{mgr_hours_worked})$

$\text{fname, lname} \mathcal{F}_{\text{sum(hours)}} \text{named_hours}$

Solution:

```
SELECT fname, lname, sum(hours)
FROM project p, works_on w, employee e, department d
WHERE p.pnumber = w.pno
AND d.mgr_empid = w.eno
AND e.enunder = w.eno
AND d.dnumber = e.dno
AND e.enunder = d.mgr_empid
GROUP BY fname, lname
```

In English, ‘list the first name, last name and total hours worked on all projects by the head of each department.’

2.3 Query Costs

Relational Algebra is an ‘operational’ language for defining queries, in that it specifies a precise sequence of operations for evaluating a query. For this reason, it can be used as an intermediate representation in real database systems, particularly when alternative execution strategies are evaluated. This exercise takes a brief look at how a database might do this.

In this exercise, we will look at alternative execution strategies for a moderately complex query, and decide which one will execute most quickly. To do this, we need to assign a cost to a relational algebra expression, by defining a real-valued function $C[\mathcal{E}]$ that gives the cost of the expression E . The definition we will use is:

$$\begin{aligned} C[\sigma_p E] &= \log_2 |E| \\ C[E \bowtie F] &= \log_2 |E| + \log_2 |F| \end{aligned}$$

All other operators are “free”.

1. Translate the following query into Relational Algebra, using project and join operations. Write the expression using intermediate values for the result of each ‘select’ and ‘join’ operation.

```
SELECT enumber, hours
FROM employee e, works_on w, project p, department d
WHERE e.enunder = w.eno
AND w.pno = p.pnumber
AND p.dnum = d.dnumber
AND d.dname = 'Information_Technology'
```

Begin by joining the employee relation with the works_on relation.

2. For each step in your RA expression, calculate the number of rows in the intermediate result.

Solution:

$$\begin{aligned} t_1 &\leftarrow \text{employee} \bowtie \text{works_on} & |t_1| &= 25 \\ t_2 &\leftarrow t_1 \bowtie \text{project} & |t_2| &= 25 \\ t_3 &\leftarrow t_2 \bowtie \text{department} & |t_3| &= 25 \\ t_4 &\leftarrow \sigma_{\text{dname}='I.T.'} t_3 & |t_4| &= 7 \\ & & \pi_{\text{enumber}, \text{hours}}(t_4) & \end{aligned}$$

3. Apply the cost function given above to your sequence of steps.

Solution:

$$\begin{aligned}
 \mathcal{C} [employee \bowtie works_on] &= \log_2 |employee| + \log_2 |works_on| \\
 &= \log_2(9) + \log_2(25) \\
 &= 7.8 \\
 \mathcal{C} [t_1 \bowtie project] &= \log_2 |t_1| + \log_2 |project| \\
 &= \log_2(25) + \log_2(10) = 8 \\
 \mathcal{C} [t_2 \bowtie department] &= \log_2 |t_2| + \log_2 |department| \\
 &= \log_2(25) + \log_2(3) = 6.2 \\
 \mathcal{C} [\sigma_{dname='I.T.'} t_3] &= \log_2 |t_3| \\
 &= \log_2(25) = 4.6 \\
 \mathcal{C} [total] &= 26.6
 \end{aligned}$$

4. Rearrange the query (ie write down a new sequence of steps) in a way that will produce a cheaper query. Calculate the cost of the new query.

Solution:

$$\begin{array}{llll}
 t'_1 \leftarrow \sigma_{dname='I.T.'} department & |t'_1| = 1 & \mathcal{C} = \log_2 3 = 1.6 \\
 t'_2 \leftarrow t'_1 \bowtie project & |t'_2| = 5 & \mathcal{C} = \log_2 10 + \log_2 1 = 3.3 \\
 t'_3 \leftarrow t'_2 \bowtie works_on & |t'_3| = 7 & \mathcal{C} = \log_2 5 + \log_2 25 = 6.9 \\
 t'_4 \leftarrow t'_3 \bowtie employee & |t'_4| = 7 & \mathcal{C} = \log_2 7 + \log_2 9 = 6 \\
 \pi_{enumber, hours}(t'_4) & & \mathcal{C} [total] = 17.8
 \end{array}$$

If you have time, try using the same analysis on query 3 in Section 2.2.

2.4 Query Equivalence using Relational Algebra—*Optional*

This exercise requires some inventiveness and mathematical creativity, and not everyone is expected to complete this. The attempt is worthwhile however, since expressing the first query in relational algebra is a good test of your understanding of SQL.

The SQL statements

```

SELECT fname, lname, dno FROM employee
WHERE enumber = (SELECT mgr_empid FROM department
                 WHERE dnumber = dno)

```

and

```

SELECT fname, lname, dno
FROM employee, department
WHERE enumber = mgr_empid

```

return the same result. Translate both queries into relational algebra, and you should be able to establish mathematically that these queries are identical.

Solution: I'm not 100% certain how to do this best, the problem is how to reflect the correlated subquery. I think the best way might be to do a union of a family of sets, indexed by the set $\pi_{dno} employee$, perhaps

$$\begin{aligned}
 depts &\leftarrow \pi_{dno} employee \\
 managers &\leftarrow \bigcup_{a \in depts} \pi_{mgr_empid}(\sigma_{dnumber=a} department)
 \end{aligned}$$

2.5 More SQL queries—*Optional*

Like most languages, SQL becomes easier with practice. Here are some more interesting SQL queries on the same schema.

1. Which employee has the highest salary ?

Solution:

```
SELECT enumber
FROM employee
WHERE employee.salary = (SELECT max(salary) FROM employee)
```

2. Which employees have not worked on the 'Red tape is Fun' project ?

Solution:

```
SELECT enumber FROM employee
WHERE NOT enumber IN (SELECT eno FROM works_on, project
                      WHERE pname = 'Red_tape_is_Fun'
                      AND pnumber = pno)
```

3. Assuming that employees are paid for a 35 hour week, 48 weeks/year, what is the cost of each project at current salary levels ?

Solution:

```
SELECT '$' || round(sum(hours * salary / 35 / 48), 2) AS cost, pname
FROM employee, works_on, project
WHERE employee.enumber = works_on.eno
AND works_on.pno = project.pnumber
GROUP BY pname
```

4. Which employees work on the most time consuming project ?

Solution:

```
SELECT eno FROM works_on
WHERE pno IN (SELECT pno FROM works_on
             GROUP BY pno
             HAVING sum(hours) =
                (SELECT max(sum_hours)
                 FROM (SELECT sum(hours) AS sum_hours
                      FROM works_on GROUP BY pno) hours))
```

5. Which employee has contributed the most hours to projects run by departments they do not belong to.

Solution:

```
select enumber
from employee, department, works_on w, project p
where dnumber != dno
and w.eno = enumber
and w.pno = p.pnumber
and p.dnum = dnumber
group by enumber
having sum(hours) = (select max(total_hours)
  from (select sum(hours) as total_hours
    from employee, department, works_on w, project p
      where dnumber != dno
      and w.eno = enumber
      and w.pno = p.pnumber
      and p.dnum = dnumber
      group by enumber) as hours)
```