Solutions for Exercise Sheet 2

Richter, Yannick
MTK 03741982
ge78tup@mytum.de

Rodrigues, Diogo
MTK 03770446
diogo.rodrigues@tum.de

TUM – Query Optimization 2022/23
4th November 2022

Our solutions for Exercise Sheet 2.

Exercise 1

For relations R_1, R_2, predicates p_1, p_2, $\mathcal{F}(p_1) \subseteq \mathcal{A}(R_1)$ and assuming set semantics.

- Prove the following equivalence. If you use other equivalences, prove them as well:

$$\sigma_{p_1}(R_1 \bowtie_{p_2} R_2) \equiv \sigma_{p_1}(R_1) \bowtie_{p_2} R_2$$

- Does the equivalence also hold for outer joins? Justify your answer.

$$\sigma_{p_1}(R_1 ⟗_{p_2} R_2) \equiv \sigma_{p_1}(R_1) ⟗_{p_2} R_2$$

Definition 1.1. $\sigma_p(R) = \{t \mid t \in R \wedge p(t)\}$

Definition 1.2. $R_1 \bowtie_p R_2 = \{t_1 \circ t_2 \mid t_1 \in R_1 \wedge t_2 \in R_2 \wedge p(t_1 \circ t_2)\}$

Lemma 1.3. *If* $\mathcal{F}(p) \subseteq \mathcal{A}(R_1)$*, then* $\forall t_1 \in R_1, \forall t_2 \in R_2, p(t_1 \circ t_2) = p(t_1)$

Proof. If none of the free variables of p is an attribute of R_2, then we do not need the t_2 component of $t_1 \circ t_2$ when evaluating $p(t_1 \circ t_2)$. This lemma is thus trivial. ∎

Theorem 1.4. $\sigma_{p_1}(R_1 \bowtie_{p_2} R_2) \equiv \sigma_{p_1}(R_1) \bowtie_{p_2} R_2$

Proof.

$$\begin{aligned}
\sigma_{p_1}(R_1 \bowtie_{p_2} R_2) &\equiv \{t \mid t \in R_1 \bowtie_{p_2} R_2 \wedge p_1(t)\} && \text{(Apply 1.1)}\\
&\equiv \{t_1 \circ t_2 \mid t_1 \in R_1 \wedge t_2 \in R_2 \wedge p_1(t_1 \circ t_2) \wedge p_2(t_1 \circ t_2)\} && \text{(Apply 1.2)}\\
&\equiv \{t_1 \circ t_2 \mid t_1 \in R_1 \wedge t_2 \in R_2 \wedge p_1(t_1) \wedge p_2(t_1 \circ t_2)\} && \text{(Apply 1.3)}\\
&\equiv \{t_1 \circ t_2 \mid t_1 \in \sigma_{p_1}(R_1) \wedge t_2 \in R_2 \wedge p_2(t_1 \circ t_2)\} && \text{(Apply 1.1)}\\
&\equiv \sigma_{p_1}(R_1) \bowtie_{p_2} R_2 && \text{(Apply 1.2)}
\end{aligned}$$

∎

We will present a counterexample. Consider the following relations:

- $R_1 = \{(a : 1, b : 1)\}$
- $R_2 = \{(b : 1, c : 1), (b : 2, c : 2)\}$

We will now evaluate $\sigma_{a \text{ is not null}}(R_1 ⟗ R_2)$ and $\sigma_{a \text{ is not null}}(R_1) ⟗ R_2$ for these two relations

$$
\begin{aligned}
&\sigma_{a \text{ is not null}}(R_1 ⟗ R_2) \\
&= \sigma_{a \text{ is not null}}((R_1 ⟕ R_2) \cup (R_2 ⟕ R_1)) \\
&= \sigma_{a \text{ is not null}}(\{(a:1, b:1, c:1)\} \cup (\{(a:1, b:1, c:1), (a:\text{null}, b:2, c:2)\}) \\
&= \sigma_{a \text{ is not null}}\{(a:1, b:1, c:1), (a:\text{null}, b:2, c:2)\} \\
&= \{(a:1, b:1, c:1)\}
\end{aligned}
$$

$$
\begin{aligned}
&\sigma_{a \text{ is not null}}(R_1) ⟗ R_2 \\
&= \sigma_{a \text{ is not null}}\{(a:1, b:1)\} ⟗ \{(b:1, c:1), (b:2, c:2)\} \\
&= \{(a:1, b:1)\} ⟗ \{(b:1, c:1), (b:2, c:2)\} \\
&= (\{(a:1, b:1)\} ⟕ \{(b:1, c:1), (b:2, c:2)\}) \cup (\{(b:1, c:1), (b:2, c:2)\} ⟕ \{(a:1, b:1)\}) \\
&= \{(a:1, b:1, c:1)\} \cup \{(a:1, b:1, c:1), (a:\text{null}, b:2, c:2)\} \\
&= \{(a:1, b:1, c:1), (a:\text{null}, b:2, c:2)\}
\end{aligned}
$$

As it can be seen, the two queries have different results. This is sufficient evidence to show they are not equivalent. ■

Exercise 2

A multi-set R can be denoted as a collection of tuple and frequency pairs $(t, R(t))$ where $R(t)$ denotes the number of occurrences of t in R. The domain of R is the set of all t such that $R(t) > 0$:

$$dom(R) = \{t | R(t) > 0\}$$

We can redefine relational algebra operators to support multi-sets as inputs. Given multi-sets L, R:

$$\begin{aligned} L \cap R &= \{(t,\ \min(L(t),\ R(t)) \mid t \in dom(L) \cap dom(R)\} \\ L \cup R &= \{(t,\ L(t) + R(t)) \mid t \in dom(L) \cup dom(R)\} \\ L - R &= \{(t,\ \max(L(t) - R(t), 0)) \mid t \in dom(L)\} \end{aligned}$$

- Prove the following equivalence or find a counterexample. If you use other equivalences, prove them as well:

$$L \cap R = L - (L - R)$$

- Define the full outer join using multi-set semantics.

$$L ⟗_p R = ?$$

Proof.

$$\begin{aligned} L - (L - R) &= \{(t, \max(L(t) - \max(L(t) - R(t), 0), 0) \mid t \in dom(L) \quad\quad \} \\ L \cap R \quad\quad &= \{(t, \min(L(t), R(t)) \quad\quad\quad\quad \mid t \in dom(L) \cap dom(R)\} \end{aligned}$$

Now we want to prove that $\max(L(t) - \max(L(t) - R(t), 0), 0) = \min(L(t), R(t))$.

$$\begin{aligned} &\max(L(t) - \max(L(t) - R(t), 0), 0) && \text{Apply: } \max(a, b) = -\min(-a, -b) \\ &= \max(L(t) + \min(-L(t) + R(t), 0), 0) && \text{Apply: } \max(a, b) = -\min(-a, -b) \\ &= -\min(-(L(t) + \min(-L(t) + R(t)), 0), 0) && \text{Apply: } c + \min(a, b) = \min(a + c, b + c) \\ &= -\min(-\min(R(t), L(t)), 0) && \text{Apply: } \max(a, b) = -\min(-a, -b) \\ &= \max(\min(L(t), R(t)), 0) && \text{Because } L(t), R(t) \geq 0 \\ &= \min(L(t), R(t)) \end{aligned}$$

■

$$\begin{aligned} L \bowtie_p R &= \{(x \circ y, L(x) * R(y)) \mid x \in L \wedge y \in R \wedge p(x \circ y)\} \\ L \ltimes_p R &= \{(x, L(x) * R(y)) \mid x \in L \wedge y \in R \wedge p(x \circ y)\} \\ L \triangleright R &= \{(x, L(x)) \mid x \in L \wedge \nexists y \in dom(R) : p(x \circ y)\} \\ L ⟕_p R &= (L \bowtie_p R) \cup \{(x \circ \circ_{a \in A(R)}(a : \text{null}), L(x)) \mid x \in (L \triangleright_p R)\} \\ L ⟗_p R &= (L ⟕_p R) \cup (R ⟕_p L) - (L \bowtie_p R) \end{aligned}$$

Exercise 3

Given are two relations R and S, with sizes 1,000 and 100,000 pages respectively. Each page has 50 tuples. The relations are stored on a disk, the average access time for the disk is 10 ms and the transfer speed is 10,000 pages/sec. How long does it take to perform the Nested Loops Join of R and S? How long does it take to perform the Block Nested Loops Join with a block size of 100 pages? Assume that CPU costs are negligible and ignore I/O costs for the join output.

With:

p_R as amount of pages in relation R

n_R as amount of tuples in relation R

t_A as amount of time for action A

- $R \bowtie^{\mathsf{NL}} S$:

$$\begin{aligned} T &= (p_R + n_R * p_S) * (t_{access} + t_{transfer}) \\ &= (1000 + 50\,000 * 100\,000) * (10\,\mathsf{ms} + 0.1\,\mathsf{ms}) \\ &= 50\,500\,010\,100\,\mathsf{ms} \approx 584\,\mathsf{d} \end{aligned}$$

- $R \bowtie^{\mathsf{BNL}} S$:

$$\begin{aligned} T &= \frac{p_R}{b} * (t_{access} + b * t_{transfer} + p_S * (t_{access} + t_{transfer})) \\ &= \frac{1000}{100} * (10\,\mathsf{ms} + 100 * 0.1\,\mathsf{ms} + 100\,000 * (10\,\mathsf{ms} + 0.1\,\mathsf{ms})) \\ &= 10 * (20\,\mathsf{ms} + 100\,000 * 10.1\,\mathsf{ms}) \\ &= 10\,100\,200\,\mathsf{ms} \approx 2.8\,\mathsf{h} \end{aligned}$$