

Implementation of DBMS

Exercise Sheet 13, Solutions

Klingemann, WS 2024 / 2025

1) Below are some statistics for four relations W, X, Y and Z. We assume in this task that a projection does not remove duplicates. We further assume that all values used in selection conditions actually appear in the corresponding instance of the relation.

W(a, b)	X(b, c)	Y(b, d)	Z(b, d)
T(W) = 100	T(X) = 200	T(Y) = 300	T(Z) = 400
V(W, a) = 20	V(X, b) = 40	V(Y, b) = 30	V(Z, b) = 40
V(W, b) = 60	V(X, c) = 100	V(Y, d) = 50	V(Z, d) = 100

Estimate the number of tuples of the relations that are the result of the following expressions:

- $\sigma_{a=10}(W)$
- $W \times Y$
- $\sigma_{d>10}(Z)$
- $W \bowtie X$
- $\sigma_{a=1 \text{ AND } b=2}(W)$
- $\sigma_{a=1 \text{ AND } b>2}(W)$
- $\sigma_{b=10}(W) \bowtie X$
- $\pi_b(\sigma_{a=20}(W))$
- $W \bowtie X \bowtie Y$
- $Y \bowtie Z$
- $W \bowtie_{a=d} Z$

Solutions:

- $T(W) / V(W, a) = 100 / 20 = 5$
- $T(W) T(Y) = 100 * 300 = 30000$
- $T(Z) / 3 = 400 / 3 = 133 + 1/3$
- We have exactly one common attribute namely b and therefore the estimate is:
 $T(W) T(X) / \max\{V(W, b), V(X, b)\} = 100 * 200 / \max\{60, 40\} = 333 + 1/3$
- We know that $\sigma_{a=1 \text{ AND } b=2}(W) = \sigma_{a=1}(\sigma_{b=2}(W))$. Let $U := \sigma_{b=2}(W)$ Then we can estimate
 $T(U) = T(W) / V(W, b) = 100 / 60 = 5/3$ and **as a is not involved in the selection condition:**
 $V(U, a) = V(W, a) = 20$.

Then our estimate for the final number of tuples is $T(U) / V(U, a) = (5/3) / 20 = 1/12$.

- We know that $\sigma_{a=1 \text{ AND } b>2}(W) = \sigma_{a=1}(\sigma_{b>2}(W))$. Let $U := \sigma_{b>2}(W)$ Then we can estimate
 $T(U) = T(W) / 3 = 100 / 3 = 33 + 1/3$ and as a is not involved in the selection condition:
 $V(U, a) = V(W, a) = 20$.

Then our estimate for the final number of tuples is $T(U) / V(U, a) = (100 / 3) / 20 = 5/3$.

- Let $U := \sigma_{b=10}(W)$. We can estimate $T(U) = T(W) / V(W, b) = 100 / 60 = 5/3$ and **$V(U, b) = 1$.**
U and X have exactly one common attribute namely b and therefore the estimate is:

$T(U) T(X) / \max\{V(U, b), V(X, b)\} = 5/3 * 200 / \max\{1, 40\} = 25/3 = 8 + 1/3$

- Let $U := \sigma_{a=20}(W)$. We can estimate $T(U) = T(W) / V(W, a) = 100 / 20 = 5$. As a projection is just changing the number of attributes, this is also the final estimate.

- Let $U := W \bowtie X$. We know from d) that $T(U) = 1000/3 = 333 + 1/3$.

For the number of distinct values we can estimate $V(U, b) = \min\{V(W, b), V(X, b)\} = 40$.

Then our estimate for the final number of tuples is

$T(U) T(Y) / \max\{V(U, b), V(Y, b)\} = 1000/3 * 300 / \max\{40, 30\} = 2500$

- Y and Z have two common attributes namely b and d. Therefore the estimate is:

$T(Y) T(Z) / (\max\{V(Y, b), V(Z, b)\} * \max\{V(Y, d), V(Z, d)\}) =$
 $300 * 400 / (\max\{30, 40\} * \max\{50, 100\}) = 30$

- We have an equijoin which compares a and d so that the estimate is:

$T(W) T(Z) / \max\{V(W, a), V(Z, d)\} = 100 * 400 / \max\{20, 100\} = 400$

2) Consider a query optimizer that uses statistical data. In particular, the following information is known about an attribute A of relation R. Attribute A is of type integer. Make the best use of the given the information.

- There are 100 tuples with A values between 1 and 10. In this range, there are 8 unique A values.
- There are 200 tuples with A values between 11 and 20. In this range, there are 5 unique A values.
- There are 300 tuples with A values between 21 and 30. In this range, there are 10 unique A values.
- There are 400 tuples with A values between 31 and 40. In this range, there are 10 unique A values.

a) Consider the query $\sigma_{A=7}(R)$. How many tuples are expected in the answer, assuming values are uniformly distributed over possible $V(R, A)$ values?

b) Consider the query $\sigma_{A=17}(R)$. How many tuples are expected in the answer, assuming values are uniformly distributed over possible domain values?

c) Consider the query $R \bowtie S$, where R has attributes R(A, B, C) and S has attributes S(A, D, E). Assume that S has the same number of tuples as R, and that the A attribute in S has the same distribution as A has in R. Assuming values are uniformly distributed over possible $V(R, A)$ values, how many tuples are expected in the answer?

Solutions:

a) Tuples with an A-value of 7 can only be found among those tuples which have A-values in the range between 1 and 10. Therefore, we can apply our estimate to this subset of the tuples. As we assume that 7 is one of the 8 values occur in this range for attribute A, the estimate is $100 / 8 = 12.5$.

b) Similar to a), we just consider tuples in the relevant value range. However, this time, we assume that 17 is some value from the domain, i.e., an integer value between 11 and 20. Therefore our estimate is $200 / 10 = 20$.

c) Like for the selections above, we can only find tuples with a particular A-value among those that have an A-value in the corresponding value range. This means that we can make separate calculations for the four value ranges. For example, the estimate for the number of tuples in the result relation that have A-values between 1 and 10 is $100 * 100 / 8$:

In total we therefore estimate:

$$100^2 / 8 + 200^2 / 5 + 300^2 / 10 + 400^2 / 10 = 1250 + 8000 + 9000 + 16000 = 34250$$