CSE 414 Section 8

Cardinality Estimation Practice Solutions

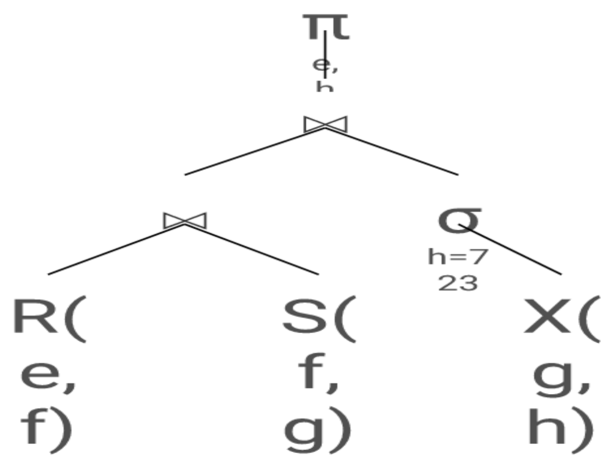
1. You're given the following relations and grocery store stats:

Safeway(id, name, category, price), T=1000, V(name)=900, V(category)=10, V(price)=200, Range(price) = [1,50)
QFC(id, name, category, price), T=2000, V(name)=1900, V(category)=12, V(price)=500

Estimate the cardinality for the following queries:

- Select * from Safeway where id = 45 -- **1 tuple**
- Select * from Safeway where name = 'Milk' -- **10/9 tuples**
- Select * from Safeway where price < 20 -- **(20-1)/(50-1) * 1000 tuples**
- Select * from Safeway S, Qfc Q where S.id = Q.id -- **1000 tuples**
- Select * from Safeway S, Qfc Q where S.name = Q.name -- **(1000*2000)/max{900,1900} tuples**

2. (Adapted from 414 SP 17 Final)



Consider the relations R(e, f), S(f, g), and X(g, h) in the query plan depicted above.

- Joins are natural joins that perform on matching attributes (e.g. R join S on R.f = S.f)
- Every attribute is integer-valued
- Assume uniform distributions on the attributes

Table	#tuples
R	1,000
S	5,000
X	100,000

Attribute	# distinct values	Minimum	Maximum
R.f	100	1	1,000
S.f	1,000	1	2,000
S.g	5,000	1	2,000
X.g	1,000	1	10,000
X.h	1,000	1	500,000

A. Estimate the number of tuples in the selection $\sigma_{h=723}(X)$.

We assume a uniform distribution of values for X.h. (Don't be confused—the first X is the selectivity factor, while T(X) is the number of tuples in the "X" table)

$$X = \frac{1}{1000}$$
$$T(X) \cdot X = 100000 \cdot X = 100$$

B. Estimate the number of tuples in the join R ⋈ S.

This natural join is the same as the equijoin R ⋈ (R.f = S.f) S:

$$X = \frac{1}{max(V(R.f), V(S.f))} = \frac{1}{max(100, 1000)} = \frac{1}{1000}$$
$$T(R) \cdot T(S) \cdot X = 1000 \cdot 5000 \cdot X = 5000$$

C. Estimate the cardinality of the final result.

(5000 * 100) / max(V(S, g), V(X, g)) = (5000 * 100) / max(5000, 100) = 100

Note: From Part A, the selection $\sigma_{h=723}$(X) results in 100 tuples. We assume the preservation of distinct values as much as possible and, thus, we assume that the 100 tuples from Part A are all distinct, including the values in attribute X.g. Thus, the final equation of max(V(S, g), V(X, g)) would be (5000, 100). This assumption enables us to simplify our calculations and estimate the cardinality easily.