CSE 414 Section 8

Cardinality Estimation Practice

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

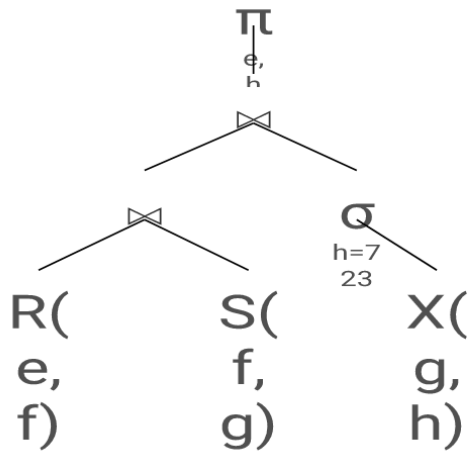
Safeway(id, name, category, price), T=1000, V(name)=900, V(category)=10, V(price)=200, Range(price) = [1,50)
QFC(id, name, category, price), T=2000, V(name)=1900, V(category)=12, V(price)=500

Estimate the cardinality for the following queries:

- Select * from Safeway where id = 45
- Select * from Safeway where name = 'Milk'
- Select * from Safeway where price < 20
- Select * from Safeway S, Qfc Q where S.id = Q.id
- Select * from Safeway S, Qfc Q where S.name = Q.name

Selectivity factors:
$a = c \rightarrow X \cong 1 / V(R, a)$
$a < c \rightarrow X \cong (c - \min(R, a)) / (\max(R, a) - \min(R, a))$
Join on cn $1=x2 \rightarrow X \cong (1 / (V(c1) \cdot V(c2))) \cdot \min(V(c1), V(c2))$
$\cong 1 / \max(V(c1), V(c2))$

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 $\sigma_{h=723}(X)$.

B. Estimate the number of tuples in the join $R \bowtie S$.

C. Estimate the cardinality of the final result.

Formula Guide for Cardinality Estimation

In cost estimation, we assume data is uniformly distributed such that each distinct value has the same number of tuples.

Selectivity factor (X), assuming table R(a, b) cartesian joined S(a,c) and constants x, x1, x2:

- R.a = x => $X \cong \frac{1}{V(R,a)}$
- R.a < x => $X \cong \frac{x - min(R.a)}{max(R.a) - min(R.a)}$
- R.a > x => $X \cong \frac{max(R.a) - x}{max(R.a) - min(R.a)}$
- x1 < R.a < x2 => $X \cong \frac{x2 - x1}{max(R.a) - min(R.a)}$
- R.a = S.a (equijoin) => $X \cong \frac{1}{max(V(R,a),V(S,a))}$
- cond1 AND cond2 =>$X = X_1 * X_2$

On deriving the selectivity of an equijoin:

Why R.a = S.a (equijoin) => $X \cong \frac{1}{max(V(R,a),V(S,a))}$?

Let say x0 a value such that R.a = S.a = x0, that means when we do selection R.a = x0 AND S.a = 0, the selectivity is:

$$X \cong \frac{1}{V(R,a)*V(S,a)}$$

But there can be as many as min(V(R,a), V(S,a)) distinct values of x0 (for example R has 100 value of a, S has 1000 value of a, the number of value of a after join is 100 because 100 < 1000, other S.a is filtered out. That means there can be 100 value of x0 such that R.a = S.a = x0)

Therefore, we multiply the above selectivity by min(V(R,a), V(S,a)) which means the min value is crossed out of the denominator, leaving the maximum value. Thus

$$X \cong \frac{1}{max(V(R,a),V(S,a))}$$

Note: this is the selectivity factor. To estimate the number of tuples in a join, multiply by the T_1T_2, the number of tuples in a Cartesian product.