

CS222: Principles of Database
Management
Fall 2007

Lecture 13: Estimating Result Sizes (cont)
Professor Chen Li

1

Size estimate for join $W = R1 \bowtie R2$

Let x = attributes of R1
 y = attributes of R2

Case 1 : $X \cap Y = \emptyset$. Same as $R1 \times R2$

CS222

Notes 13

2

Case 2: R1.A values uniformly distributed
over the R2.A values

$$W = R1 \bowtie R2 \quad X \cap Y = A$$

R1	A	B	C	R2	A	D

Assumption: one of the following is true

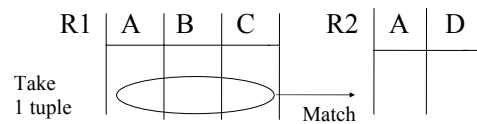
- Every A value in R1 is in R2 $\Rightarrow V(R1,A) \leq V(R2,A)$
– Example: foreign-key constraint from R1.A to R2.A
- **Or** every A value in R2 is in R1 $\Rightarrow V(R2,A) \leq V(R1,A)$

CS222

Notes 13

3

Computing $T(W)$ when $V(R1,A) \leq V(R2,A)$



1 tuple matches with $\frac{T(R2)}{V(R2,A)}$ tuples

$$\text{so } T(W) = \frac{T(R2)}{V(R2,A)} \times T(R1)$$

CS222

Notes 13

4

Generalization:

- $V(R1,A) \leq V(R2,A)$ $T(W) = \frac{T(R2) T(R1)}{V(R2,A)}$

- $V(R2,A) \leq V(R1,A)$ $T(W) = \frac{T(R2) T(R1)}{V(R1,A)}$

- A is the join attribute
- Which one to use depends on the assumption about the two attributes.

CS222

Notes 13

5

In general $W = R1 \bowtie R2$

$$T(W) = \frac{T(R2) T(R1)}{\max \{ V(R1,A), V(R2,A) \}}$$

CS222

Notes 13

6

Case 2 with alternate assumption

Values uniformly distributed over the same domain



This tuple matches $(R2)/\text{DOM}(R2,A)$ tuples

$$T(W) = \frac{T(R2) * T(R1)}{\text{DOM}(R2, A)} = \frac{T(R2) * T(R1)}{\text{DOM}(R1, A)}$$

Assume the same

CS222

Notes 13

7

In all cases: size of each resulting tuple

$$S(W) = S(R1) + S(R2) - S(A) \quad \leftarrow \text{size of attribute A}$$

CS222

Notes 13

8

Using similar ideas, we estimate sizes of

$\Pi_{AB}(R)$: Section. 7.4.2

$\sigma_{A=a \wedge B=b}(R)$: Section 7.4.3

$R \bowtie S$ with multiple join attributes: Section 7.4.5

Union, intersection, diff, Section 7.4.7

CS222

Notes 13

9

Note: for complex expressions, need intermediate T,S,V results.

$$\text{E.g. } W = [\underbrace{\sigma_{A=a}(R1)}] \bowtie R2$$

Treat as relation U

$$T(U) = T(R1)/V(R1,A) \quad S(U) = S(R1)$$

Also need $V(U, X)$, i.e., the number of distinct tuples in relation U on attribute X

CS222

Notes 13

10

To estimate $V(U, X)$

E.g., $U = \sigma_{A=a}(R1)$

Say R1 has attributes A,B,C,D

$V(U, A) = ?$

$V(U, B) = ?$

$V(U, C) = ?$

$V(U, D) = ?$

CS222

Notes 13

11

Example

R 1	<table border="1" style="display: inline-table; vertical-align: middle;"> <thead> <tr> <th>A</th> <th>B</th> <th>C</th> <th>D</th> </tr> </thead> <tbody> <tr> <td>cat</td> <td>1</td> <td>10</td> <td>10</td> </tr> <tr> <td>cat</td> <td>1</td> <td>20</td> <td>20</td> </tr> <tr> <td>dog</td> <td>1</td> <td>30</td> <td>10</td> </tr> <tr> <td>dog</td> <td>1</td> <td>40</td> <td>30</td> </tr> <tr> <td>bat</td> <td>1</td> <td>50</td> <td>10</td> </tr> </tbody> </table>	A	B	C	D	cat	1	10	10	cat	1	20	20	dog	1	30	10	dog	1	40	30	bat	1	50	10	$V(R1,A)=3$ $V(R1,B)=1$ $V(R1,C)=5$ $V(R1,D)=3$
A	B	C	D																							
cat	1	10	10																							
cat	1	20	20																							
dog	1	30	10																							
dog	1	40	30																							
bat	1	50	10																							

$$U = \sigma_{A=a}(R1)$$

Expected numbers:

- $V(U,A)=1$
- $V(U,B)=1$
- $V(U,C) = T(R1)V(R1,A)$.
– Reason: all C values. So # = expected number of selected tuples
- $V(U,D)$: somewhere in between 1 and $T(R1)V(R1,A)$.

CS222

Notes 13

12

Possible Guess $U = \sigma_{A=a}(R)$

$$V(U,A) = 1$$

$$V(U,B) = V(R,B)$$

CS222

Notes 13

13

For Joins $U = R1(A,B) \bowtie R2(A,C)$

- Assumption: each value of a nonjoin attribute is kept in the result.
- Called “preservation of value sets” in the textbook

$$V(U,A) = \min \{ V(R1, A), V(R2, A) \}$$

$$V(U,B) = V(R1, B)$$

$$V(U,C) = V(R2, C)$$

CS222

Notes 13

14

Example:

$$Z = R1(A,B) \bowtie R2(B,C) \bowtie R3(C,D)$$

R1 $T(R1) = 1000 \quad V(R1,A)=50 \quad V(R1,B)=100$

R2 $T(R2) = 2000 \quad V(R2,B)=200 \quad V(R2,C)=300$

R3 $T(R3) = 3000 \quad V(R3,C)=90 \quad V(R3,D)=500$

CS222

Notes 13

15

Partial Result: $U = R \bowtie S$

$$T(U) = \frac{1000 \times 2000}{200} \quad V(U,A) = 50$$

$$V(U,B) = 100$$

$$V(U,C) = 300$$

CS222

Notes 13

16

$$Z = U \bowtie R3$$

$$T(Z) = \frac{1000 \times 2000 \times 3000}{200 \times 300}$$

$$V(Z,A) = 50$$

$$V(Z,B) = 100$$

$$V(Z,C) = 90$$

$$V(Z,D) = 500$$

Summary

- Estimating size of results is an “art”
- Statistics must be kept up to date
 - We need to pay the cost