Procesamiento y optimización de consultas (II)

Outline - Query Processing

- Relational algebra level
  - transformations
  - good transformations
- Detailed query plan level
  - estimate costs
  - generate and compare plans

• Estimating cost of query plan

(1) Estimating size of results

(2) Estimating # of IOs

Estimating result size

- Keep statistics for relation R
  - T(R) : # tuples in R
  - S(R) : # of bytes in each R tuple
  - B(R): # of blocks to hold all R tuples
  - V(R, A) : # distinct values in R
    for attribute A

Estimación del tamaño del resultado

- Debemos mantener estadísticas para cada relación R:
  - T(R) : # tuplas in R
  - S(R) : # de bytes en cada tupla de R
  - B(R): # de bloques para guardar todas las tuplas de R
  - V(R, A) : # valores diferentes para el atributo A en R

Example

R

A	B	C	D
cat	1	10	a
cat	1	20	b
dog	1	30	a
dog	1	40	c
bat	1	50	d

A: 20 byte string

B: 4 byte integer

C: 8 byte date

D: 5 byte string

T(R) = 5   S(R) = 37

V(R,A) = 3   V(R,C) = 5

V(R,B) = 1   V(R,D) = 4

Size estimates for $W = R1 \times R2$

$T(W) = \quad T(R1) * T(R2)$

$S(W) = \quad S(R1) + S(R2)$

Size estimates for $W = \sigma_{A=cat}(R)$

$S(W) = S(R)$

$T(W) = ?$

Example

R

A	B	C	D
cat	1	10	a
cat	1	20	b
dog	1	30	a
dog	1	40	c
bat	1	50	d

$V(R,A)=3$

$V(R,B)=1$

$V(R,C)=5$

$V(R,D)=4$

$$W = \sigma_{A=cat}(R) \quad T(W) = \frac{T(R)}{V(R,A)}$$

Assumption:

Values in select expression Z = val
are uniformly distributed
over possible V(R,Z) values.

What about with $W = \sigma_{z \geq val}(R)$?

$T(W) = ?$

- Solution 1:

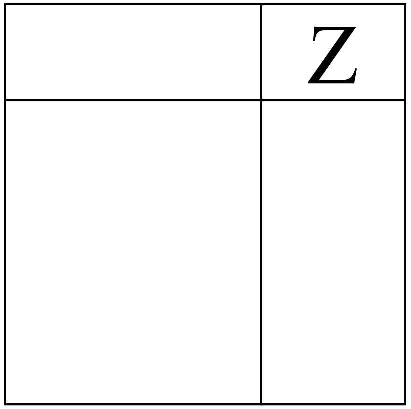
  $T(W) = T(R)/2$

- Solution 2:

  $T(W) = T(R)/3$

Solution 3: Estimate values in range

R



$T(R)=16$

Min=1

Max=20

$V(R,Z)=10$

$W = \sigma_{Z \geq 15}(R)$

$$f = \frac{20-15+1}{20} = \frac{6}{20} = 0.3 \quad \text{(fraction of range)}$$

$$f \times V(R,Z) = 0.3 \times V(R,Z) = 0.3 \times 10 = 3$$

$$T(R) / V(R,Z) = 16/10 = 1.6$$

Finally:

$$T(W) = [f \times V(R,Z)] \times \frac{T(R)}{V(R,Z)} = f \times T(R)$$

T(W)=0.3 x 1.6= 4.8

Size estimate for $W = R1 \bowtie R2$

Let x = attributes of R1

y = attributes of R2

$\boxed{\text{Case1}}$

$$X \cap Y = \emptyset$$

Same as R1 x R2

Case 2 $W = R1 \bowtie R2$ $X \cap Y = A$

R1

A	B	C

R2

A	D

Assumption:

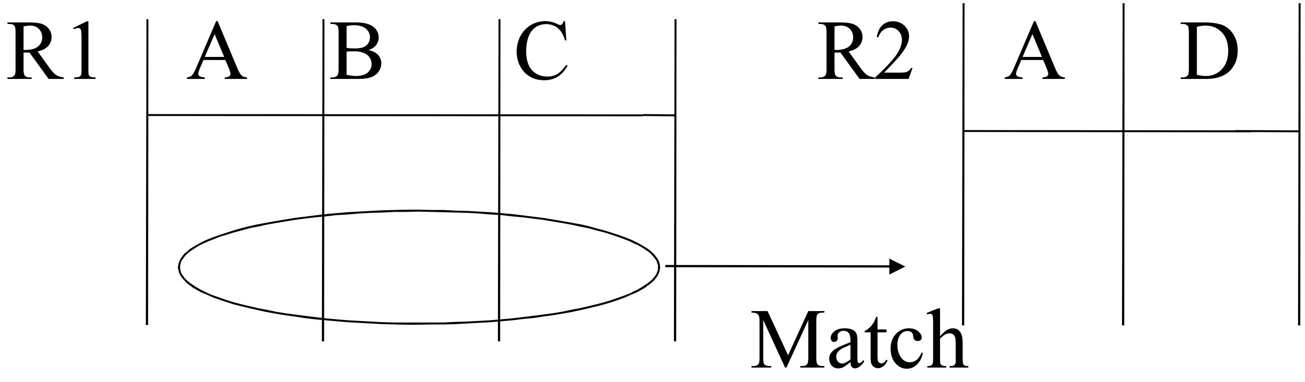
$V(R1,A) \leq V(R2,A) \Rightarrow$ Every A value in R1 is in R2

$V(R2,A) \leq V(R1,A) \Rightarrow$ Every A value in R2 is in R1

If R1 and R2 are related by integrity constraints

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

Take 1 tuple



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

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

Example: A is PK in R2

R2

A	B	C
7		
11		
9		
10		
5		
8		

R1

A	D
7	
11	
11	
8	
5	

An R1 tuple matches with T(R2)/V(R2,A)=6/6=1 tuples...

Example: A is not PK in R2

R2

A	B	C
7		
11		
9		
10		
5		
8		
8		
11		

R1

A	D
7	
11	
11	
8	
5	

An R1 tuple matches with $T(R2)/V(R2,A)=8/6=1.33$ tuples...

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

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

[A is common attribute]

In General $\quad W = R1 \bowtie R2$

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

In all the cases:

$$S(W) = S(R1) + S(R2) - S(A)$$

size of attribute A

Another example:

$$W = R(a,b,c,d) \bowtie S(c,d,f,g)$$

R(a,b,c,d)	**S(c,d,f,g)**
T(R) = 1000	T(S) = 2000
V(R,c) = 20	V(S,c) = 50
V(R,d) = 100	V(S,d) = 50

What would be the estimate size for T(W)?

SOLUTION:

$$T(W) = \frac{\left\{ \dfrac{T(R)\,T(S)}{\max\{ V(R,c), V(S,c) \}} \right\}}{\max\{ V(R,d), V(S,d) \}}$$

SOLUTION:

$$T(W) = \frac{1000 \times 2000}{50 \times 100}$$

Using similar ideas, we can estimate sizes for:

$W = \Pi_{AB}(R)$ ….. Sec. 16.4.2

In this case, the $T(W) = T(R)$ but $S(W)$ and $B(W)$ would change.

Union, intersection, diff, …. Sec. 16.4.7

For complex expressions, need intermediate T,S,V values:

$$W = \underbrace{[\sigma_{A=a}(R1)]}_{\text{Relation } U} \bowtie R2$$

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

WE ALSO NEED $V(U, *)$!!

To estimate Vs

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

says that R1 has attributes A,B,C,D

$V(U, A) =$

$V(U, B) =$

$V(U, C) =$

$V(U, D) =$

Example

R1

A	B	C	D
cat	1	10	10
cat	1	20	20
dog	1	30	10
dog	1	40	30
bat	1	50	10

$V(R1,A)=3$

$V(R1,B)=1$

$V(R1,C)=5$

$V(R1,D)=3$

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

For sure : $V(U,A)=1$ $V(U,B)=1$

$V(U,C)$ and $V(U,D)$ will be between 1 and $T(R1)/V(R1,A)$

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

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

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

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

Example:

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

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

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

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

Partial Result: $U = R1 \bowtie R2$

$$T(U) = \frac{1000 \times 2000}{200}$$

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

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

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

$$Z = U \bowtie R3$$

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

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

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

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

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

Summary

- Estimating size of results is an “art”

- Don’t forget:

  Statistics must be kept up to date…

  (cost?)

Outline - Query Processing

- Estimating cost of query plan
  - – Estimating size of results		done!
  - – Estimating # of IOs		next…

- Generate and compare plans