

CSE 4125: Distributed Database Systems

Chapter – 6

(Part – A)

Optimization of Access Strategies.

Pre-requisites

- Knowledge of Chapter 2

Topics to be discussed -

- Database Profiles
- Estimating Database Profiles

Database Profiles

What are Database Profiles?

❑ *Statistical information of the database.*

❑ *Necessity:*

- To perform sequence of operations, relations must be transmitted over the network.
- It is important to estimate the size of the results to minimize the data transfers.
- We need the statistical information (**database profile**) to estimate.

Information in Database Profiles

For a relation $R(A_1, A_2, \dots, A_n)$ with fragments R_1, R_2, \dots, R_r the database profile contains following information.

- **card (R_i):** number of tuples of R_i .
- **size (A_i):** size or length (i.e. number of bytes) of attribute A_i .
- **size (R_i):** sum of the size of all attributes of R_i .

Information in Database Profiles

For a relation $R(A_1, A_2, \dots, A_n)$ with fragments R_1, R_2, \dots, R_r the database profile contains following information.

- **val ($A_i [R_i]$):** number of distinct values appearing for attribute A_i of R_i .
- **dom(A_i):** domain of an attribute.
- **site (R_i):** allocated site of the fragment R_i .

Database Profiles (example)

- $\text{card}(\text{DEPT}_1) = 10$ Number of rows
- $\text{site}(\text{DEPT}_1) = 2$ Dept 1 fragment is located at Site 2

	deptnum	name	area	mgrnum
size	2	15	1	7
val	10	10	2	10

- $\text{Q: size}(R_i) = ?$ $2+15+1+7 = 25$ byte

Database Profiles (example)

- $\text{card}(\text{DEPT}_1) = 10$
 - $\text{site}(\text{DEPT}_1) = 2$

	deptnum	name	area	mgrnum
size	2	15	1	7
val	10	10	2	10

- $Q: \text{size}(R_i) = 25$

Exercise 1

Given three relations R, S, T. We want to perform the following query Q.

Where, $Q = PJ_a(R \cup S)$

Card (R) = 3000

Card (S) = 1000

Card (T) = 4000

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	c	b	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R \cup S) has no duplicate values in attribute 'a'

Is the database profile correct? Mention and rewrite the errors if any.

1

Estimating profiles of results of algebraic operations

What to Estimate?

- Estimating the profiles of results of algebraic operations.
- This information is useful for optimization.
- Assume, R and S are input fragments and T is the result.
 - We will mostly estimate $\mathbf{card}(T)$ and $\mathbf{size}(T)$.
 - Example: If $\mathbf{card}(R)$ and $\mathbf{card}(S)$ is given, can we estimate $\mathbf{card}(T)$ for $T = R \text{ UN } S$

Selection

$$T = \sigma_{A = value} R$$

Cardinality

$$\text{card}(T) = \rho \times \text{card}(R)$$

Here ρ is selectivity.

$$\rho = 1 / \text{val}(A[R])$$

Selection

$$T = SL_{A = value} R$$

Cardinality

$$\text{card}(T) = \rho \times \text{card}(R)$$

Here ρ is selectivity.

$$\rho = 1 / \text{val}(A[R])$$

example

R

$$T = SL_{A1 = B} R$$

$$\rho = ?$$

$$\text{card}(T) = ?$$

A1	A2
A	E
B	F
C	G
D	H

Selection

$$T = SL_{A = value} R$$

Cardinality

$$\text{card}(T) = \rho \times \text{card}(R)$$

Here ρ is selectivity.

$$\rho = 1 / \text{val}(A[R])$$

example

R

$$T = SL_{A1 = B} R$$

$$\rho = ?$$

$$\text{card}(T) = ?$$

A1	A2
A	E
B	F
A	G
B	H

****Assuming, values are homogeneously distributed**

Selection

$$T = \text{SL}_{A = \textit{value}} R$$

Size

size (T) = ?

Selection

$$T = \sigma_{A = value} R$$

Size

$$\text{size}(T) = \text{size}(R)$$

➤ *Selection doesn't affect the size of relations.*

Projection

$$T = PJ_{A1} R$$

Cardinality

card (T) = ?

example

R

<i>A1</i>	<i>A2</i>
A	E
B	F
C	G
B	H

Projection

$$T = PJ_{A1} R$$

Cardinality

$$\text{card}(T) = \text{val}(A1[R])$$

example

R

<i>A1</i>	<i>A2</i>
A	E
B	F
C	G
B	H

Projection

$$T = PJ_{A1} R$$

Size

size (T) ? size (?)

example

R

<i>A1</i>	<i>A2</i>
A	E
B	F
C	G
B	H

Projection

$$T = PJ_{A1} R$$

Size

$$\text{size}(T) = \text{size}(A1)$$

example

R

<i>A1</i>	<i>A2</i>
A	E
B	F
C	G
B	H

Union

$$T = R \cup S$$

Cardinality

$$\text{card}(T) \leq \text{card}(R) + \text{card}(S)$$

***Equality holds when duplicate tuples are not eliminated.*

Union

$$T = R \cup S$$

Size

$$\text{size}(T) = \text{size}(R) = \text{size}(S)$$

$$\text{card (Result)} \leq \text{card (R)} + \text{card (S)}$$

$$\text{size (Result)} = \text{size (R)} = \text{size (S)}$$

Example: $R \cup S$

R		
A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S		
A	B	C
a	1	a
a	3	f

T		
B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result		
A	B	C
a	1	a
b	1	b
a	1	d
b	2	f
a	3	f

Difference

$$T = R \text{ } \Delta \text{ } S$$

Cardinality

$$\max (0, \text{card} (R) - \text{card} (S)) \leq \text{card} (T) \leq \text{card} (R)$$

Size

$$\text{size} (T) = \text{size} (R) = \text{size} (S)$$

$$\max (0, \text{card} (R) - \text{card} (S)) \leq \text{card} (\text{Result}) \leq \text{card} (R)$$

$$\text{size} (\text{Result}) = \text{size} (R) = \text{size} (S)$$

Example: *R* *DF* *S*

R		
A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S		
A	B	C
a	1	a
a	3	f

T		
B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result		
A	B	C
b	1	b
a	1	d
b	2	f

Cartesian Product

$$T = R \text{ CP } S$$

Cardinality

$$\text{card}(T) = \text{card}(R) \times \text{card}(S)$$

Size

$$\text{size}(T) = \text{size}(R) + \text{size}(S)$$

$$\text{card (Result)} = \text{card (R)} \times \text{card (S)}$$

$$\text{size (Result)} = \text{size (R)} + \text{size (S)}$$

Example: *R CPS*

R		
A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S		
A	B	C
a	1	a
a	3	f

T		
B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result					
R.A	R.B	R.C	S.A	S.B	S.C
a	1	a	a	1	a
b	1	b	a	1	a
a	1	d	a	1	a
b	2	f	a	1	a
a	1	a	a	3	f
b	1	b	a	3	f
a	1	d	a	3	f
b	2	f	a	3	f

Join

$$T = R \Join_{R.A = S.B} S$$

Cardinality

$$\begin{aligned} \text{card}(T) &= \text{selectivity} \times \text{card}(R \times S) \\ &= \rho \times (\text{card}(R) \times \text{card}(S)) \end{aligned}$$

Join

$$T = R \text{ JN}_{R.A = S.B} S$$

Cardinality

$$\begin{aligned} \text{card}(T) &= \text{selectivity} \times \text{card}(R \text{ CP } S) \\ &= \rho \times (\text{card}(R) \times \text{card}(S)) \\ &= 1 / \text{val}(A[R]) \times \text{card}(R) \times \text{card}(S) \\ &= (\text{card}(R) \times \text{card}(S)) / \text{val}(A[R]) \\ &= (\text{card}(R) \times \text{card}(S)) / \text{val}(B[S]) \end{aligned}$$

Join

$$T = R \bowtie_{R.A = S.B} S$$

Size

$$\text{size}(T) = \text{size}(R) + \text{size}(S)$$

$$\text{card (Result)} = (\text{card(R)} \times \text{card(T)}) / \text{val (C[R])}$$

$$\text{size (Result)} = \text{size (R)} + \text{size (S)}$$

Example: $R \Join_{R.C=T.C} T$

R		
A	B	C
a	1	a
b	1	b
a	1	d
b	2	f

S		
A	B	C
a	1	a
a	3	f

T		
B	C	D
1	a	1
3	b	1
3	c	2
1	d	4
2	a	3

Result					
A	R.B	R.C	T.B	T.C	D
a	1	a	1	a	1
a	1	a	2	a	3
b	1	b	3	b	1
a	1	d	1	d	4

Semi-Join

$$T = R \text{ SJ}_{R.A = S.B} S$$

Cardinality

$$\begin{aligned} \text{card}(T) &= \text{selectivity} \times \text{card}(R) \\ &= \rho \times \text{card}(R) \end{aligned}$$

Here,

$$\rho = \text{val}(A[S]) / \text{val}(\text{dom}(A))$$

Size

The size of the result of a semi-join is the same size of its first operand.

Estimating Example 1

Given three relations R, S, T. We want to perform the following query Q.

Where, $Q = (PJ_a(R \cup S)) \Join_{a=a} T$

Card (R) = 3000

Card (S) = 1000

Card (T) = 4000

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	c	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R \cup S) has no duplicate values in attribute 'a'

Estimate the cardinality of the result of query Q. Indicate necessary formulas applied to estimate. 4

Given Query: $(PJ_a(R \text{ UN } S)) \text{ CP } (T \text{ JN}_{a=a} S)$

Let,

$$X = R \text{ UN } S$$

$$Y = PJ_a(X)$$

$$Z = T \text{ JN}_{a=a} S$$

$$F = Y \text{ CP } Z$$

$$\begin{aligned} \text{Card}(X) &= \text{Card}(R) + \text{Card}(S) = 3000 + 1000 \\ &= 4000 \end{aligned}$$

$$\text{Card}(Y) = \text{val}(a[X]) = 4000$$

$$\begin{aligned} \text{Card}(Z) &= \rho \times (\text{card}(T) \times \text{card}(S)) \\ &= (1 / \text{val}(a[T])) * \text{Card}(T) * \text{card}(S) \\ &= (1 / 4000) * 4000 * 1000 \\ &= 1000 \end{aligned}$$

$$\begin{aligned} \text{Card}(F) &= \text{Card}(Y) * \text{card}(Z) \\ &= 4000 * 1000 = 4000000 \end{aligned}$$

Estimating Example 2

Given three relations R, S, T. We want to perform the following query Q.

Where, $Q = (PJ_a (R \text{ UN } S)) \text{ CP } (T \text{ JN}_{a=a} S)$

Card (R) = 3000

Card (S) = 1000

Card (T) = 4000

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	c	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Estimate the size of the result of query Q. Indicate necessary formulas applied to estimate.

4

Given Query: $(PJ_a(R \text{ UN } S)) \text{ CP } (T \text{ JN}_{a=a} S)$

Let,

$X = R \text{ UN } S$

$Y = PJ_a(X)$

$Z = T \text{ JN}_{a=a} S$

$F = Y \text{ CP } Z$

$$\text{size}(X) = \text{size}(R)$$

$$= \text{size}(a) + \text{size}(b) + \text{size}(c) + \text{size}(d)$$

$$= 6+7+2+10 = 25 \text{ bytes}$$

$$\text{size}(Y) = \text{size}(a) = 6 \text{ bytes}$$

$$\text{size}(Z) = \text{size}(T) + \text{size}(S)$$

$$= (6+5+4) + (6+7+2+10)$$

$$= 40 \text{ bytes}$$

$$\text{size}(F) = \text{size}(Y) + \text{size}(Z)$$

$$= 6 + 40 = 46 \text{ bytes} = 46 * 8 \text{ bits} = 368 \text{ bits}$$

Exercise 2

Given three relations R, S, T. We want to perform the following query Q.

Where, $Q = PJ_a(R \cup S)$

Card (R) = 3000

Card (S) = 1000

Card (T) = 4000

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	c	b	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R UN S) has no duplicate values in attribute 'a'

Is the database profile correct? Mention and rewrite the errors if any. 1

Estimate the cardinality of the result of query Q. Indicate necessary formulas applied to estimate. 4

What is the output of size(R SJ_{a=a} T) ? 1

Exercise 3

Given three relations R, S, T. We want to perform the following query Q.

Where, $Q = (PJ_a(R \cup S)) \Join_{a=a} T$

Card (R) = 3000

Card (S) = 1000

Card (T) = 4000

	a	b	c	d
Size	6	7	2	10
Val	3000	1000	30	500

	a	b	c	d
Size	6	7	2	10
Val	1000	20	500	100

	a	m	n
Size	6	5	4
Val	4000	10	5

Assume that , the result of (R \cup S) has no duplicate values in attribute 'a'

Estimate the cardinality & size of query Q. Indicate necessary formulas applied to estimate.

4