



UA

Unidad 2: Procesamiento y Optimización de Consultas

*Bases de Datos Avanzadas, Sesión 7 :
Uso de Estadísticas y Estimación de
tuplas de una consulta*

*Iván González Diego
Dept. Ciencias de la Computación
Universidad de Alcalá*



INDICE

- *Introducción.*
- *Información del Catálogo para la estimación del coste*
- *Estimación del tamaño de operaciones*
- *Estimación del número de valores diferentes*

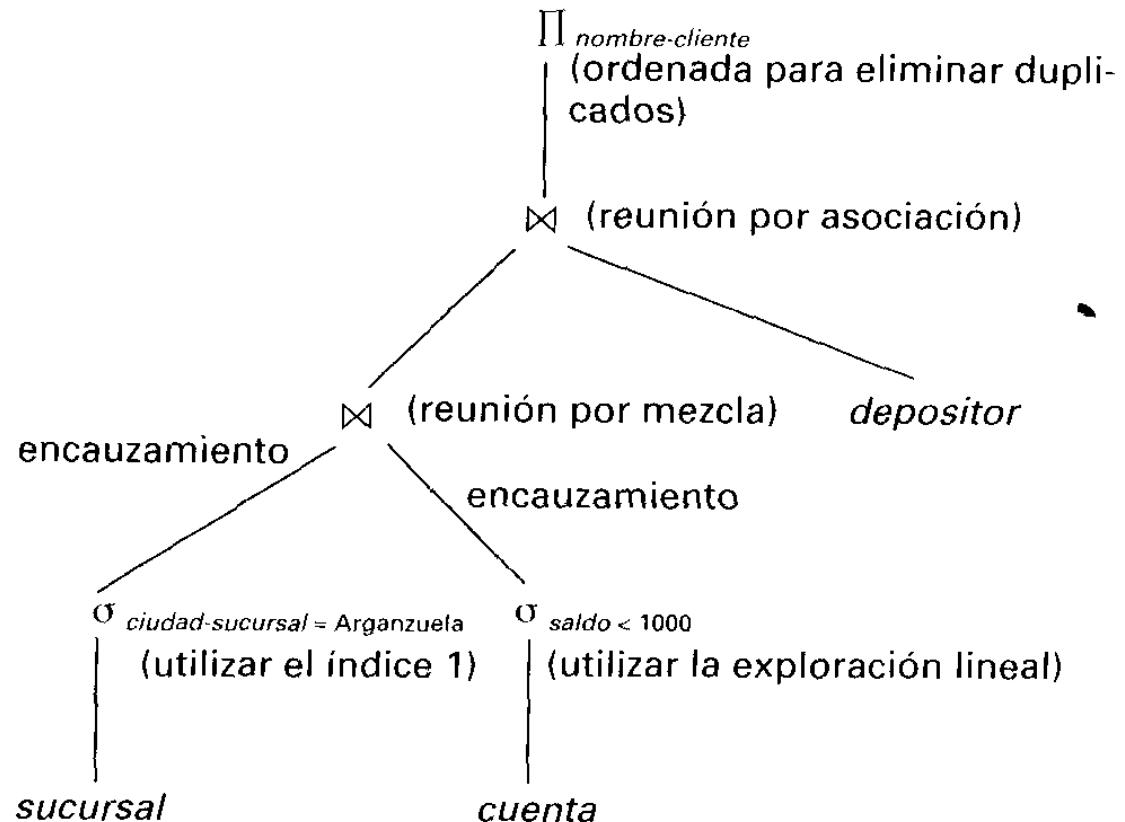
Referencias: *Silberschatz 4^a Ed.*
Elmasri, 3^a Ed.



Introducción

UA

- Necesario estimar las tuplas de entrada y salida de cada operación individual de un plan de evaluación





Información estadística para la Estimación del Coste

- n_r : número de tuplas en la relación r .
- b_r : número de bloques de r .
- s_r : tamaño en bytes de una tupla de r .
- f_r : factor de bloques de r — número de tuplas en un bloque.
- $V(A, r)$: número de valores distintos que aparecen en r para un atributo A ; mismo tamaño que $\Pi_A(r)$.
- Si las tuplas de r se almacenan juntas físicamente en un fichero, entonces:

$$b_r = \left\lceil \frac{n_r}{f_r} \right\rceil$$

- Estadísticas sobre los índices \Rightarrow alturas árboles, número de bloques de los índices, etc
- Histogramas



UA

Estimación del tamaño de la Selección

- **Selección de igualdad** $\sigma_{A=v}(r)$
 - n_{rc} : número de registros que satisfarán la condición.
 - $\lceil n_{rc}/f_r \rceil$ — número de bloques que ocuparán.
 - *Ejemplo: Estimación coste para la búsqueda binaria:*

$$\text{Coste} = \lceil \log_2(b_r) \rceil + \left\lceil \frac{n_{rc}}{f_r} \right\rceil - 1$$

- *Igualdad para un atributo clave:* $n_{rc} = 1$
- *Distribución uniforme de valores*
 - $n_{rc} = n_r / V(A,r)$ tuplas



UA

Selecciones con Comparación

- *Selecciones de la forma $\sigma_{A \leq v}(r)$ (caso de $\sigma_{A \geq v}(r)$ es simétrico)*
- *n_{rc} es el número de tuplas estimado que satisfacen la condición*
 - *Si $\min(A, r)$ y $\max(A, r)$ están disponibles en el catálogo*
 - $n_{rc} = 0$ si $v < \min(A, r)$
 - $n_{rc} = n_r$ si $v \geq \max(A, r)$
 - $$n_{rc} = n_r \cdot \frac{v - \min(A, r)}{\max(A, r) - \min(A, r)}$$
 - *En ausencia de información estadística $\Rightarrow n_{rc} = n_r / 2$.*
 - *Contar el número de valores que satisfacen la condición n_v y multiplicar por el número de tuplas que devuelve un valor.*
 - $n_{rc} = n_v * n_r / V(A, R)$
 - $n_r \rightarrow$ Número de tuplas de la tabla r



UA

Selecciones Complejas

- La **selectividad** de una condición θ_i es la probabilidad de que una tupla de r satisfaga θ_i . Si s_i es el número de tuplas que la satisfacen $\Rightarrow s_i / n_r$
- **Conjunción:** $\sigma_{\theta_1 \wedge \theta_2 \wedge \dots \wedge \theta_n}(r)$. El número estimado de tuplas es:

$$n_r * \frac{s_1 * s_2 * \dots * s_n}{n_r^n}$$

- **Disyunción:** $\sigma_{\theta_1 \vee \theta_2 \vee \dots \vee \theta_n}(r)$. Número estimado de tuplas:

$$n_r * \left(1 - \left(1 - \frac{s_1}{n_r} \right) * \left(1 - \frac{s_2}{n_r} \right) * \dots * \left(1 - \frac{s_n}{n_r} \right) \right)$$

- **Negación:** $\sigma_{\neg \theta}(r)$. Número estimado de tuplas:

$$n_r - \text{size}(\sigma_\theta(r))$$



UA

Estimación del tamaño de las Reuniones

- *El producto cartesiano $r \times s$ contiene $n_r * n_s$ tuplas; ocupando cada tupla $s_r + s_s$ bytes.*
- *Si $R \cap S = \emptyset \Rightarrow r \bowtie s$ es el mismo que $r \times s$.*
- *Si $R \cap S$ es clave de $R \Rightarrow$ una tupla de s se combinará con una tupla de r*
 - *El número de tuplas de $r \bowtie s$ no es mayor que el número de tuplas de s .*
- *Si $R \cap S$ es una clave ajena de S referenciando a R, \Rightarrow número de tuplas de $r \bowtie s$ es el mismo que de s .*
 - *El caso de $R \cap S$ siendo una clave ajena referenciando a S es simétrico.*



UA

Estimación del tamaño de las Reuniones

- Si $R \cap S = \{A\}$ no es clave para R ni S .
si se asume que cada tupla t de r produce tuplas en $R \bowtie S \Rightarrow$ N° de tuplas estimado:

$$\frac{n_r * n_s}{V(A,s)}$$

Para el caso contrario:

$$\frac{n_r * n_s}{V(A,r)}$$

En general:

$$\frac{n_r * n_s}{\max \{V(A,r), V(A,s)\}}$$



Estimación del tamaño de otras operaciones

- Proyección: tamaño estimado $\Pi_A(r) = V(A, r)$
- Agregación : tamaño estimado ${}_A\mathbf{g}_F(r) = V(A, r)$
- Operaciones de conjuntos
 - Para uniones/intersecciones de selecciones en la misma relación r : reescribir y usar tamaño estimado para las selecciones
 - Ejemplo $\sigma_{\theta_1}(r) \cup \sigma_{\theta_2}(r) \Rightarrow \sigma_{\theta_1 \vee \theta_2}(r)$
 - Para operaciones sobre diferentes relaciones:
 - $r \cup s = \text{tamaño de } r + \text{tamaño de } s.$
 - $r \cap s = \text{mínimo } \{ \text{tamaño } r, \text{tamaño } s \}.$
 - $r - s = r.$
- Reunión externa:
 - $r \bowtie s = \text{tamaño } r \bowtie s + \text{tamaño } r$
 - $r \bowtie s = \text{tamaño } r \bowtie s + \text{tamaño } r + \text{tamaño } s$



Estimación del número de valores distintos

- *Cuando se eliminan tuplas → Puede cambiar $V(A, r)$*

Selecciones: $\sigma_\theta(r)$

- *Si θ obliga A tomar un valor: $V(A, \sigma_\theta(r)) = 1$.*
 - *Ejemplo: A = 3*
- *Si θ obliga A tomar uno del conjunto de valores:
 $V(A, \sigma_\theta(r)) = n^o$ de valores especificados.*
 - *Ejemplo, (A = 1 V A = 3 V A = 4),*
- *Si θ es de la forma A op v: $V(A, \sigma_\theta(r)) = V(A, r) * s$*
 - *donde s es la selectividad.*
- *Para otros casos: $\min(V(A, r), n_{\sigma\theta(r)})$*
 - *Más exactitud ⇒ teoría de probabilidad*



Estimación del número de valores distintos

Reuniones: $r \bowtie s$

- Si todos los atributos de A proceden de r :
 $V(A, r \bowtie s) = \min(V(A, r), n_{r \bowtie s})$
- Si A contiene atributos de A1 de r y A2 de s :
 $V(A, r \bowtie s) = \min(V(A1, r) * V(A2 - A1, s), V(A1 - A2, r) * V(A2, s), n_{r \bowtie s})$

Proyecciones: $V(A, \Pi_{A(r)}) = V(A, r)$.

Para valores agregados: ${}_G g_{F(A)}(r)$

- asumir todos los valores agregados son distintos y usar $V(G, r)$



UA

Ejemplo:

- Tablas: *Impositor (nombre_cliente, numero_cuenta)*
Cliente (nombre_cliente, apellidos, dirección)
- *Impositor* \bowtie *cliente*

$N_{cliente} = 10000$ tuplas

$N_{impositor} = 5000$ tuplas

$V(nombre_cliente, impositor) = 2500$

1. Determinar el tamaño de la reunión utilizando las claves.
2. Determinar el tamaño utilizando la expresión general.