Unidad 2: Procesamiento y Optimización de Consultas

Bases de Datos Avanzadas, Sesión 7 :

Uso de Estadísticas y Cálculo del Coste 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 de estadísticas*
- *Transformación de expresiones relacionales*

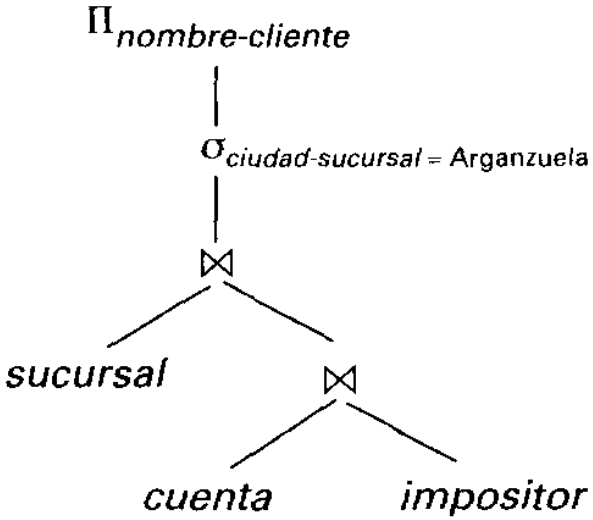
Referencias: Silberschatz 4ª Ed. Pp 319 - 341
Elmasri, 3ª Ed. Pp 553 - 595

Introducción

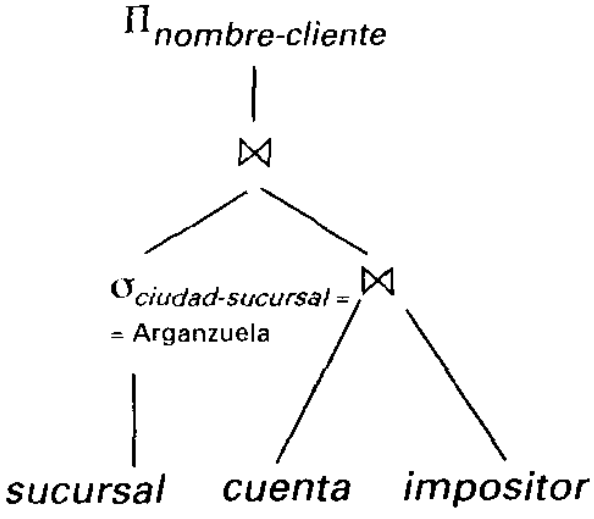
- *Muchas maneras de evaluar una consulta*
  - *Expresiones equivalentes*
  - *Diferentes algoritmos para cada operación*
- *La diferencia del coste entre una buena o mala manera de evaluar una consulta puede ser grande:*
  - *Ejemplo: realizar r X s seguido de una selección r.A = s.B puede ser más lento que realizar una reunión en la misma condición.*
- *Necesidad de estimar el coste de las operaciones*
  - *Depende de la información estadística sobre las relaciones*
    - *Número de tuplas, número de valores distintos para los atributos de reunión, etc*
  - *Necesidad de estimar operaciones intermedias para analizar el coste de una expresión compleja*

Introducción

- *Las relaciones generadas por dos expresiones equivalentes tienen el mismo conjuntos de atributos y contienen el mismo conjunto de tuplas, aunque sus atributos pueden estar ordenados diferentemente*



(a) Árbol inicial de la expresión



(b) Árbol transformado de la expresión

Introducción

- *Pasos para la generación de planes de evaluación:*

  1. *Generar expresiones lógicas equivalentes*
     - *Usar reglas de equivalencia para transformar una expresión en una equivalente*
  2. *Anotación de las expresiones resultantes para obtener planes de evaluación alternativos*
  3. *Elegir el plan con menor coste estimado*

- *Todo el proceso se llama Optimización basada en coste*

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*

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:*

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

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

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 → Número de tuplas de la tabla r*

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:*** $\sigma_{\theta 1 \wedge \theta 2 \wedge \ldots \wedge \theta n}(r)$. *El número estimado de tuplas es:*

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

- ***Disyunción:*** $\sigma_{\theta 1 \vee \theta 2 \vee \ldots \vee \theta n}(r)$. *Número estimado de tuplas:*

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

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

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

Estimación del tamaño de las Reuniones

- *El producto cartesiano* r *x* s *contiene* $n_r * n_s$ *tuplas; ocupando cada tupla* $s_r + s_s$ *bytes.*
- *Si* $R \cap S = \varnothing \Rightarrow r \bowtie s$ *es el mismo que* r *x* 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.*

Estimación del tamaño de las Reuniones

- *Si* R ∩ S = *{*A*} no es clave para* R *ni* S.
  si se asume que cada tupla t *de* r *produce tuplas es* R⋈ S ⇒ *Nº 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 = *tamaño de* r + *tamaño de* s.
    - r $\cap$ s = *mínimo* { *tamaño* r ,*tamaño* s}.
    - r – s = r.
- *Reunión externa:*
  - r $\exists\bowtie$ s = tamaño r $\bowtie$ s + tamaño r
  - r $\exists\bowtie\subset$ s = tamaño r $\bowtie$ s + *tamaño* r + *tamaño* s

Estimación del número de valores distintos

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

Selecciones: σ_θ(r)

- *Si* θ *obliga* A *tomar un valor:* V(A,σ_θ(r)) = 1.
  - *Ejemplo:* A = 3
- *Si* θ *obliga A tomar uno del conjunto de valores:* V(A,σ_θ(r)) = *nº de valores especificados.*
  - *Ejemplo,* (A = 1 V A = 3 V A = 4),
- *Si* θ *es de la forma* A op v: V(A,σ_θ(r)) = V(A,r) * s
  - *donde* s *es la selectividad.*
- *Para otros casos:* min(V(A,r), $n_{\sigma\theta(r)}$)
  - *Más exactitud* $\Rightarrow$ *teoría de probabilidad*

Estimación del número de valores distintos

Reuniones: r ⋈ 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)

Ejemplo:

- *Impositor ⋈ cliente*

  Ncliente= 10000 tuplas
  Fcliente= 25
  Nimpositor=5000 tuplas
  Fimpositor=50
  V(nombre,impositor)=2500

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

Transformación de Expresiones Relacionales

- *Dos expresiones del álgebra relacional son **equivalentes** si para cada instancia de la base de datos legal, las dos expresiones generan el mismo conjunto de tuplas*
- *En SQL $\Rightarrow$ entradas y salidas son un multiconjunto de tuplas*
- *Una **regla de equivalencia** afirma que las expresiones de las dos formas son equivalentes*
  - *Se puede reemplazar la expresión de la primera forma por la segunda ó viceversa.*

Reglas de Equivalencia

1. *Operaciones de selección conjuntiva se pueden dividir en una secuencia de selecciones individuales.*

$$\sigma_{\theta_1 \wedge \theta_2}(E) = \sigma_{\theta_1}(\sigma_{\theta_2}(E))$$

2. *Operaciones de selección son conmutativas.*

$$\sigma_{\theta_1}(\sigma_{\theta_2}(E)) = \sigma_{\theta_2}(\sigma_{\theta_1}(E))$$

3. *Sólo la última secuencia de operaciones de proyección se necesitan, las otras se pueden omitir:*

$$\Pi_{t_1}(\Pi_{t_2}(\ldots(\Pi_{tn}(E))\ldots)) = \Pi_{t_1}(E)$$

4. *Las selecciones se pueden combinar con productos cartesianos y reuniones zeta*

   a. $\sigma_{\theta}(E_1 \times E_2) = E_1 \bowtie_{\theta} E_2$
   b. $\sigma_{\theta 1}(E_1 \bowtie_{\theta 2} E_2) = E_1 \bowtie_{\theta 1 \wedge \theta 2} E_2$

Reglas de Equivalencia

5. *Las operaciones de reunión zeta son conmutativas*

$$E_1 \bowtie_\theta E_2 = E_2 \bowtie_\theta E_1$$

6. *(a) Operaciones de reunión natural son asociativas:*

$$(E_1 \bowtie E_2) \bowtie E_3 = E_1 \bowtie (E_2 \bowtie E_3)$$

(b) Reuniones zeta son asociativas de la siguiente manera:

$$(E_1 \bowtie_{\theta 1} E_2) \bowtie_{\theta 2 \wedge \theta 3} E_3 = E_1 \bowtie_{\theta 1 \wedge \theta 3} (E_2 \bowtie_{\theta 2} E_3)$$

donde θ_2 *involucra sólo atributos de* E_2 *y* E_3.

Reglas de Equivalencia

7. *La operación de selección se distribuye por la operación de reunión zeta bajo las dos condiciones siguientes:*

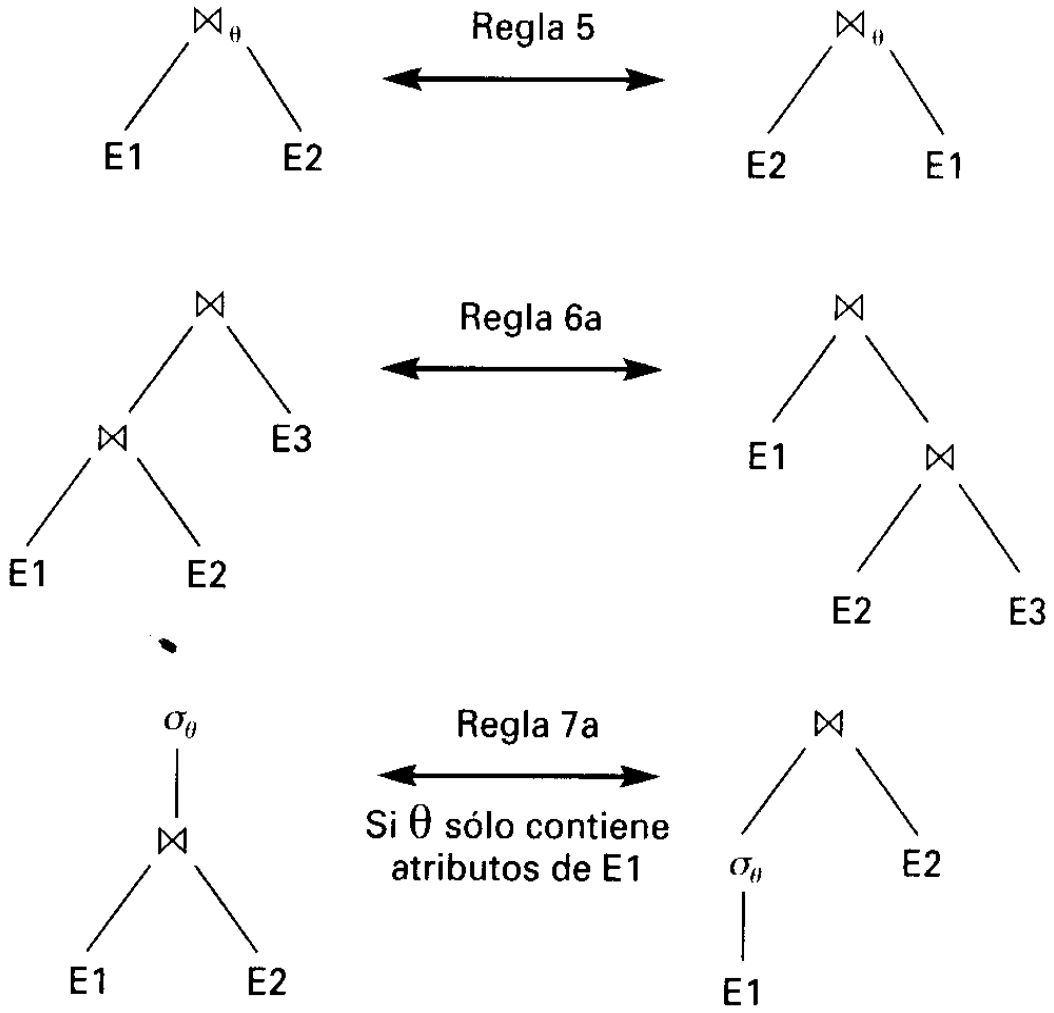
   (a) cuando todos los atributos θ_0 implican sólo los atributos de una de las expresiones (E_1) que están reuniendo.

$$\sigma_{\theta 0}(E_1 \bowtie_{\theta} E_2) = (\sigma_{\theta 0}(E_1)) \bowtie_{\theta} E_2$$

   (b) Cuando θ_1 implica sólo los atributos de E_1 y θ_2 implica sólo los atributos de E_2.

$$\sigma_{\theta 1 \wedge \theta 2}(E_1 \bowtie_{\theta} E_2) = (\sigma_{\theta 1}(E_1)) \bowtie_{\theta} (\sigma_{\theta 2}(E_2))$$

Reglas de Equivalencia



Reglas de Equivalencia

8. *Las operaciones de proyección se distribuyen por la operación de reunión zeta bajo las condiciones:*

   (a) Si θ *implica sólo los atributos de* $L_1 \cup L_2$*:*

$$\prod_{L_1 \cup L_2} (E_1 \bowtie_\theta E_2) = (\prod_{L_1} (E_1)) \bowtie_\theta (\prod_{L_2} (E_2))$$

   (b) Considerar una reunión $E_1 \bowtie_\theta E_2$*.*

   - L_1 *y* L_2 *conjunto de atributos de* E_1 *y* E_2*, respectivamente.*
   - L_3 *atributos de* E_1 *que están implicados en la condición de reunión* θ*, pero no están incluidos en* $L_1 \cup L_2$
   - L_4 *atributos de* E_2 *que están implicados en la condición de reunión* θ*, peor no están incluidos en* $L_1 \cup L_2$*.*

$$\prod_{L_1 \cup L_2} (E_1 \bowtie_\theta E_2) = \prod_{L_1 \cup L_2} ((\prod_{L_1 \cup L_3} (E_1)) \bowtie_\theta (\prod_{L_2 \cup L_4} (E_2)))$$

Reglas de Equivalencia

9. *Operaciones de conjuntos de unión e intersección son conmutativas*

$$E_1 \cup E_2 = E_2 \cup E_1$$
$$E_1 \cap E_2 = E_2 \cap E_1$$

10. *Unión e intersección de conjuntos son asociativas.*

$$(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3)$$
$$(E_1 \cap E_2) \cap E_3 = E_1 \cap (E_2 \cap E_3)$$

11. *La operación de selección se distribuye sobre* $\cup$, $\cap$ *y* $-$.

$$\sigma_\theta (E_1 - E_2) = \sigma_\theta (E_1) - \sigma_\theta(E_2)$$

Similarmente para $\cup$ *y* $\cap$ *en lugar de* $-$

También: $\sigma_\theta (E_1 - E_2) = \sigma_\theta(E_1) - E_2$

y similarmente para $\cap$ *, pero no para* $\cup$

12. *Operación de proyección se distribuye sobre la operación de unión*

$$\Pi_L(E_1 \cup E_2) = (\Pi_L(E_1)) \cup (\Pi_L(E_2))$$

Ejemplo

Sucursal(nombre_sucursal,ciudad_sucursal,activo)
Cuenta(numero_cuenta,nombre_sucursal,saldo)
Impositor(nombre_cliente,numero_cuenta)

Optimizar:

$$\Pi_{nombre\_cliente}(\sigma_{ciudad='Arganzuela'}(sucursal \bowtie (cuenta \bowtie impositor)))$$

$$\Pi_{nombre\_cliente}(\sigma_{ciudad='Arganzuela' \wedge saldo>1000}(sucursal \bowtie (cuenta \bowtie impositor)))$$