CC42A – BASES DE DATOS

Profesores: Claudio Gutiérrez, Gonzalo Navarro

Auxiliar: Mauricio Monsalve

Guía de estimación de costos

Resumen rápido: *Las reglas de estimación de costos se reducen a las siguientes propiedades:*

- Los datos tienen distribución uniforme: Esto quiere decir que cada valor aparece repetido más o menos las mismas veces que otro valor del mismo tipo (o atributo). Ejemplo: R(A,B,C), si A="platelminto" aparece 15 veces, entonces A="gasterópodo" aparecerá 15 veces, o un número parecido a 15. Luego se asume (estimando) que aparece 15 veces.

- Las estimaciones son probabilísticas: Si se tiene una "entrada" de tamaño N (sean registros, bloques, bytes, etc.) y se efectúa una operación σ, entonces se asume que el resultado es N x P(Se cumpla la condición de σ).

- Probabilidad de la igualdad: Si se tiene una entrada de tamaño N y una selección del tipo A=b, entonces la probabilidad de que se cumpla la condición es el recíproco del número de valores distintos que asume A en la entrada, o sea, $V(Entrada,A)^{-1}$. Luego la salida será de tamaño N/V(Entrada,A).

- Probabilidad de los "AND": Si se hace selección de la forma "A y B y C y…" entonces la probabilidad de que un registro cumpla la condición es P(A)xP(B)xP(C)x…. La salida será N multiplicado por la probabilidad antes mencionada. Excepción: Si dos condiciones son excluyentes, la probabilidad de que se cumpla una es ***inmediatamente cero*** *(ejemplo: A=3 y A=5).*

- Probabilidad de los "OR": Por regla de probabilidades, y asumiendo independencia de las condiciones, una condición "A o B o C o…" se cumple si es que no se cumple su negación "NoA y NoB y NoC y…". Y NoA se cumple si es que no se cumple A. Por complemento de probabilidades (P(A)=1-P(NoA) y P(NoA)=1-P(A)), se tiene: P(A o B o C)=1-P(NoA y NoB y NoC)=1-P(NoA)P(NoB)P(NoC)=1-(1-P(A))(1-P(B))(1-P(C)). Por motivos de cálculo, es más fácil calcular P(NoA)=1-P(A) de antemano y evitarse hacer un cálculo muy grande con una expresión con muchos paréntesis.

- Número de tuplas de salida de un producto cruz: La salida de la operación AxB tiene tamaño $\boldsymbol{N_A \cdot N_B}$ ***registros*** *(1 byte por 1 byte no da un byte…). El tamaño del registro será la suma de los tamaños de los registros, sea* $\boldsymbol{b_A + b_B}$ ***bloques*** *o bien* $\boldsymbol{t_A + t_B}$ ***bytes*** *o kilobytes.*

Recomendación final: *Use algo de experiencia de probabilidades y suposiciones creíbles para hacer las estimaciones. No es la idea que memorice fórmulas complicadas; lo ideal es que las sepan por tacto o práctica.*

Ejercicios (1-4 *comprensión*):

1. Sea la consulta $\sigma_{A=7 \wedge B=4}(S)$, con *S(A,B,C), V(S,A)*=25, *V(S,B)*=80 y N_S=10000. Indique el número de registros que cumplen la condición pedida.
2. Sea la consulta $\sigma_{A=7 \wedge B=4}(S)$, con *S(A,B,C), V(S,A)*=25, *B* atributo llave y N_S=5000. ¿Qué puede decir de *V(S,B)*? Indique el número de registros que cumplen la condición pedida.
3. Sea la consulta $\sigma_{B=8}(R \times \sigma_{A=3}(S))$, con *S(A,C), R(B,D)*. Para este ejercicio no se dan número pues el objetivo es entregar una expresión para el número de registros de salida de esta consulta, escribiendo los resultados en función de *V(Tabla,Atributo)* y N_{TABLA}.
4. Escriba la probabilidad de que se cumpla la condición "*A=7 y (B=5 o C=4)*" dada una tabla *R(A,B,C)* donde *A* es llave, *V(R,B)=2V(R,C)*=500 y la tabla R posee 2500 entradas.

Soluciones ejercicios 1-4:

1. *La probabilidad de que se cumpla A es 1/25 y de que se cumpla B es 1/80. La operación lógica "Y" nos lleva a la multiplicación: P(AyB)=1/(25·80)=1/2000. Como la entrada es 10000, la salida es 10000/2000=10/2=5.*
2. *Como B es llave, todos los valores de B son distintos, luego hay tantos valores de B como registros en la tabla. Por eso, V(S,B)=N_S=5000. Ahora, la probabilidad de que se cumpla A=7 y B=4 es, a priori, 1/(5000·25). Entonces la salida es 5000/(5000·25)=1/25=0.04<1. Que la probabilidad haya dado menos que uno sale del hecho de que si B=4 se cumple, entonces sale sólo un registro, pues B es llave. Después no se sabe si A=7 será parte de ese registro, eso depende de la dependencia funcional.*
3. *Vamos desde adentro hacia fuera, en este caso bottom-up: La selección sobre S entrega N_S/V(S,A) registros. Considerando N_R registros para el producto cruz, salen $N_R N_S$/V(S,A) registros. Aplicando la última selección terminan por salir $N_R N_S$/(V(S,A)V(RxS,B)) registros. Ahora bien, como esa condición cae sobre B, entonces no cae sobre S, sólo sobre R, luego se puede estimar que V(RxS,B)=V(R,B) pues S no agrega nuevos valores para B. Entonces la salida es $N_R N_S$/(V(S,A)V(R,B)).*
4. *Empecemos por el paréntesis (bottom-up): B=5 o C=4. La probabilidad de que se cumpla es P(B=5 o C=4)=1-P(B!=5 y C!=4)=1-P(B!=5)P(C!=4). P(B!=5)=1-P(B=5)=1-1/500=0.998. P(C!=4)=1-P(C=4)=1-1/250=0.996. Entonces P(B=5 o C=4)=1-0.998·0.996=1-0.994008=0.005992=0.006-ε. Ahora, la probabilidad de que se cumpla A=7 es 1/2500=0.0004. Luego la probabilidad de que se cumpla la condición sobre algún registro de la tabla R es P=0.0004·0.006=0.0000024, lo que es razonable dado que se restringe sobre una llave.*

Ejercicios (5-10 *teoría y práctica*):

5. Se sabe que en *R(A,B)* el 25% de los valores de *A* son mayores a 50 y uno de cada tres valores de *B* es igual a *A*, y se sabe que *A* es llave. Estime la probabilidad de que se cumpla la condición "*A=B y A>50*". *Resp.: 1/12=0.8333.*
6. Sabiendo lo mismo de la parte anterior calcule la probabilidad de que se cumpla la condición "A=B y B<50". Resp.: ¼=0.24.
7. Si R(A,B,C,D) tiene 1000 filas, V(R,A)=2V(R,B)=3V(R,C), D es llave y la probabilidad de que un registro cumpla con B="dato" es 4%, estime los valores de V(R,AtributoSimple). Resp.: V(R,A)=50, V(R,B)=25, V(R,C)=17 y V(R,D)=1000.
8. ¿Qué información necesita para estimar el costo de la operación de la división? Hint: Repase guías antiguas, sobre todo los ejercicios de *cardinalidades* de las operaciones relacionales.
9. Suponga que se tienen dos consultas relacionales, la optimizada y la original. ¿Difiere el número de registros que debieran arrojar? ¿Qué (otras) cosas difieren entre ambas consultas, en términos de costos? *Hint: Arrojan los mismos registros, dé el motivo.*
10. ¿Qué relación tiene la carga de la consulta (costo en cada parte) con el tiempo de ejecución de la consulta? En general, ¿a qué otras situaciones se aplica este principio? *Hint: Se aplica, por ejemplo, al tratamiento de grandes cantidades de información.*