Selettività di predicati locali

Dario Maio

http://bias.csr.unibo.it/maio/

Selettività di predicati locali

- Dato un predicato p, il suo **fattore di selettività** f_p è dato dal rapporto tra il numero di tuple che soddisfano p e il numero di tuple della relazione a cui p si riferisce.
- Il numero di tuple residue che risultano dall'applicazione di p alla relazione R è pertanto valutato come

$$ET_R = f_p \times NT_R$$

- Denotiamo con p(A) un predicato su un attributo A.
- La selettività di p(A), ipotizzando l'uniformità delle distribuzioni di valori, è data da:

$$f_{p(A)} = EK_A / NK_A$$

- dove EK_A rappresenta il numero di chiavi residue e NK_A è il numero di valori distinti di A

Selettività di predicati locali: casi notevoli

- Predicato "=" $f_{(A=v)} = 1/NK_A$
- Predicato "IN" $f_{(A \in set)} = card(set)/NK_A$
- Predicato "<" $f(A<v) = \frac{v - \min(A)}{\max(A) - \min(A)} \times \frac{NK_A - 1}{NK_A}$

  → per attributi con molti valori si può approssimare a:

  $$f(A<v) = \frac{v - \min(A)}{\max(A) - \min(A)}$$

- Predicato "between" $f(A \in [v_1, v_2]) = \frac{v_2 - v_1}{\max(A) - \min(A)} \times \frac{NK_A - 1}{NK_A} + \frac{1}{NK_A}$

  → per attributi con molti valori si può approssimare a:

  $$f(A \in [v_1, v_2]) = \frac{v_2 - v_1}{\max(A) - \min(A)}$$

Calcolo di selettività: esempio

$NT_{Employee}$ = 20000, NK_{DeptNo} = 100, NK_{Job} = 10
NK_{Sex} = 2, NK_{Salary} = 10
min(Salary) = 5000, max(Salary) = 50000

Predicato	Selettività	Tuple residue
DeptNo = 51	1/100	200
Salary > 10000	(50 - 10)/(50 - 5) × 9/10 = 8/10	16000
Job = 'Clerk'	1/10	2000
Sex = 'Female'	1/2	10000

Calcolo di selettività: esempio

- Nel caso di condizioni composte da più predicati si ha, grazie all'ipotesi di indipendenza tra attributi:

  f(p1 AND p2) = fp1 $\times$ fp2

  f(p1 OR p2) = fp1 + fp2 - fp1 $\times$ fp2 (con p1 e p2 su attributi distinti)

Condizione	Selettività
DeptNo = 51 and Salary > 10000	1/100 $\times$ 8/10 = 8/1000
Job = 'Clerk' or Sex = 'Female'	1/10 + 1/2 - 1/20 = 11/20

Piani di accesso

- Scansione sequenziale

$$C(\text{seq } R) = NP_R + \alpha \times NT_R$$

- Accesso con indice IX(R.Ai)

$$Ca(IX(R.Ai)\ \{clus|uncl\}) = NI_{R.Ai} + EL_{p(R.Ai)} + EP_{p(R.Ai)}$$

dove

- $EL_{p(R.Ai)}$: numero atteso di pagine foglia a cui bisogna accedere

$$EL_{p(R.Ai)} = \lceil f_{p(R.Ai)} \times NL_{(R.Ai)} \rceil$$

- $EP_{p(R.Ai)}$: numero atteso di pagine dati cui bisogna accedere
  - » clustered: $EP_{p(R.Ai)} = \lceil f_{p(R.Ai)} \times NP_R \rceil$
  - » unclustered: $EP_{p(R.Ai)} = EK_{p(R.Ai)} \times \Phi(NT_R/NK_{R.Ai}, NP_R)$

Piani di accesso: esempio

$NP_{Employee}$	= 2000	(alias E per Employee)
D	= 4096 byte	(capacità delle pagine dati e indice)
u	= 0.69	(fattore di riempimento delle foglie)
L(DeptNo)	= 2 byte	(alias D per DeptNo)
L(Salary)	= 4 byte	(alias S per Salary)
L(Job)	= 10 byte	(alias J per Job)
L(TID)	= 4 byte	(lunghezza di un TID)
NI	= 2	(per ogni indice)

Ricordando che il numero di foglie si calcola come:

$$NL_{R.Ai} = \left\lceil \frac{NK_{R.Ai} \times L(R.Ai) + NT_R \times L(TID)}{D \times u} \right\rceil$$

si ottiene $NL_{E.D} = NL_{E.S} = NL_{E.J} = 29$.

Piani di accesso: esempio

- **Scansione sequenziale**
  - $C(\text{Seq } E) = 2000 + \alpha \times 20000$
- **IX(E.D) unclustered e IX(E.S) clustered**
  - C(IX(E.D) uncl) $= 2 + \lceil 0.01 \times 29 \rceil + \Phi(0.01 \times 20000, 2000) + \alpha \times 200$
    $= 2 + 1 + 191 + \alpha \times 200 = 194 + \alpha \times 200$
  - C(IX(E.S) clus) $= 2 + \lceil 0.8 \times 29 \rceil + \lceil 0.8 \times 2000 \rceil + \alpha \times 16000$
    $= 2 + 24 + 1600 + \alpha \times 16000 = 1626 + \alpha \times 16000$
- **IX(E.D) clustered e IX(E.S) unclustered**
  - C(IX(E.D) clus) $= 2 + 1 + \lceil 0.01 \times 2000 \rceil + \alpha \times 200$
    $= 3 + 20 + \alpha \times 200 = 23 + \alpha \times 200$
  - C(IX(E.S) uncl) $= 2 + 24 + 8 \times \Phi(0.1 \times 20000, 2000) + \alpha \times 16000$
    $= 26 + 8 \times 1265 + \alpha \times 16000 = 10146 + \alpha \times 16000$