

Tecnologie delle Basi di Dati M

Appello del 22/10/2010

Esercizio 1 (2 punti)

Data la relazione con schema:

Dipartimenti(nome, sede, budget, direttore)

si effettui una stima del numero di foglie di indici B⁺-tree costruiti sugli attributi (nome) e (budget). Si supponga di avere pagine di dimensione 4KB, di cui 96 B riservati per il page header, e si considerino i seguenti valori:

- Numero di tuple = 10K
- Numero di chiavi (budget) = 500
- Dimensione budget = 8 byte
- Dimensione nome = 28 byte
- Dimensione RID = 4 byte
- Dimensione PID = 4 byte
- Percentuale di riempimento foglie (nome) = 80%
- Percentuale di riempimento foglie (budget) = 100%

Esercizio 2 (5 punti)

Data la relazione con schema:

Dipartimenti(nome, sede, budget, direttore)

si ottimizzi l'esecuzione della seguente interrogazione SQL:

```
SELECT *
FROM Dipartimenti
WHERE sede IN ('Milano', 'Bologna', 'Roma', 'Firenze')
AND NOT (nome = 'Amministrazione')
AND budget BETWEEN 500 AND 1000
AND direttore LIKE 'P%'
```

tenendo conto che dai cataloghi della base di dati risulta:

- Numero di tuple Dipartimenti = 10K
- Numero di pagine Dipartimenti = 100
- Numero di chiavi per l'attributo sede = 100
- Indice unclustered (TID disordinate) su nome: numero foglie = 80
- Indice unclustered (TID ordinate) su sede: numero foglie = 20, numero chiavi = 100
- Indice unclustered (TID ordinate) su budget: numero foglie = 50, valore minimo = 0, valore massimo = 5000
- Indice clustered su direttore: numero foglie = 40, i valori iniziano con una consonante dell'alfabeto italiano.

Si stimi infine il numero di risultati dell'interrogazione e si indichi sotto quali condizioni risulterebbe vantaggioso l'uso di un indice (eventualmente B⁺-tree) sull'attributo nome.

Suggerimento: per la formula di Cardenas si utilizzino i seguenti valori, validi per P = 100:

R	$\Phi(R, P)$
10	9.561792
20	18.20931
30	26.02996
40	33.10282
50	39.49939
60	45.28434
70	50.51613
80	55.24768
90	59.5268
100	63.39677

R	$\Phi(R, P)$
110	66.89669
120	70.06196
130	72.92457
140	75.51347
150	77.85482
160	79.9723
170	81.8873
180	83.6192
190	85.1855
200	86.60203

R	$\Phi(R, P)$
210	87.88312
220	89.04171
230	90.08952
240	91.03714
250	91.89415
260	92.66921
270	93.37017
280	94.0041
290	94.57741
300	95.09591

Esercizio 3 (5 punti)

Si descriva l'algoritmo di inserimento di oggetti in un R-tree, stimando i costi attesi, compreso il caso di split di nodi dell'albero.

Esercizio 4 (3 punti)

Gli algoritmi presentati per la valutazione efficiente di join sono adatti ad essere utilizzati nel caso in cui la condizione di join sia una condizione di uguaglianza. Si discuta sulla possibile estensione degli algoritmi *sort-merge join* e *hash join* per la valutazione efficiente di outer join (destro, sinistro e full).

Soluzione Esercizio 1

Dimensionamento indice (nome):

Numero di chiavi = 10K.

Dimensione di ogni record (foglia) = $28 + 4 = 32B$

Dimensione "reale" foglia = $(4096 - 96) \times 0.80 = 3200B$

Numero di foglie = $10K \times 32/3200 = 100$

Dimensionamento indice (budget):

Numero di chiavi = 200, quindi mediamente ci sono $10K/500 = 20$ tuple per ogni valore di chiave.

Dimensione di ogni record (foglia) = $8 + 4 \times 20 = 88B$

Dimensione "reale" foglia = $(4096 - 96) = 4000B$

Numero di foglie = $500 \times 88/4000 = 11$

Soluzione Esercizio 2

Selettività dei predicati:

Predicato su sede = $100/10K = 0.01$ per ogni valore di sede

Predicato su nome = $1 - 1/10K = 0.9999$

Predicato su budget = $(1000 - 500)/(5000 - 0) = 0.1$

Predicato su direttore = $1/16 = 0.0625$

Accesso a Dipartimenti:

Costo scan sequenziale = **100**

Costo indice su sede: $4 \times (NL \times 0.01 + \Phi(NT \times 0.01, NP)) = 4 \times (20 \times 0.01 + \Phi(10K \times 0.01, 100))$

$= 4 \times (1 + \Phi(100, 100)) = 4 \times (1 + 64) = 260$

Costo indice su nome: $NL \times 0.9999 + NT \times 0.9999 = 80 \times 0.9999 + 10K \times 0.9999 = 80 + 9999 = 10079$

Costo indice su budget: $NL \times 0.1 + \Phi(NT \times 0.1, NP) = 50 \times 0.1 + \Phi(10K \times 0.1, 100) = 5 + \Phi(1000, 100) = 5 + 100 = 105$

Costo indice su direttore: $NL \times 0.0625 + NP \times 0.0625 = 40/16 + 100/16 = 3 + 7 = 10$

Il numero di risultati dell'interrogazione è $10K \times 0.04 \times 0.1 \times 0.0625 = 2.5$