Tecnologie delle Basi di Dati M

Appello del 21/10/2011

Esercizio 1 (2 punti)

Data la relazione con schema:

`Personale(codice, cognome, nome, annodinascita, stipendio)`

si effettui una stima del numero di livelli (e di nodi) di un R-tree costruito sulla coppia di attributi (`annodinascita, stipendio`). Si supponga di avere pagine di dimensione 4 KB, di cui 96 B riservati per il page header, e si considerino i seguenti valori:

- Numero di tuple = 11K
- Numero di pagine = 100
- Dimensione `annodinascita` = dimensione `stipendio` = 8 byte
- Numero di chiavi (`annodinascita`) = 50
- Numero di chiavi (`stipendio`) = 20
- Dimensione RID = 4 byte
- Dimensione PID = 3 byte
- Percentuale di riempimento foglie = 82.5%
- Percentuale di riempimento nodi = 87.5%

Si ricordi che in un R-tree le foglie contengono coppie (chiave, lista di RID), mentre i nodi interni contengono coppie (MBB, PID) e che per la memorizzazione di un MBB richiede una coppia di valori per ciascun attributo.

Esercizio 2 (5 punti)

Data la relazione con schema:

`Personale(matricola, nome, data, luogo, stipendio, responsabile)`

si ottimizzi l'esecuzione della seguente interrogazione SQL:

```
SELECT P.matricola, R.matricola
FROM Personale P, Personale R
WHERE P.responsabile=R.matricola
  AND R.stipendio < 40000
  AND P.luogo IN ('Milano', 'Bologna', 'Torino')
```

tenendo conto che dai cataloghi della base di dati risulta:

- Numero di tuple `Personale` = 20K
- Numero di pagine `Personale` = 1K
- Numero di responsabili = 100
- Indice clustered su `luogo`: numero foglie = 100, numero chiavi = 20
- Indice unclustered (TID ordinate) su `stipendio`: numero foglie = 50, valore minimo = 20000, valore massimo = 220000
- Indice unclustered su `matricola`: numero foglie = 3000

Si disegni infine l'albero corrispondente al piano di accesso di costo minimo e stimi il numero di risultati dell'interrogazione.

Suggerimento: per la formula di Cardenas si utilizzino i seguenti valori, validi per P = 1000:

R	Φ(R, P)
100	95.20785289
200	181.3511705
300	259.2929678
400	329.814094
500	393.6210551
600	451.3530925
700	503.5885866
800	550.8508514
900	593.6133775
1000	632.3045752

R	Φ(R, P)
1100	667.3120671
1200	698.9865709
1300	727.6454132
1400	753.5757086
1500	777.0372363
1600	798.2650423
1700	817.4717945
1800	834.849913
1900	850.5734982
2000	864.8000746

R	Φ(R, P)
2100	877.6721692
2200	889.3187393
2300	899.8564645
2400	909.3909155
2500	918.0176119
2600	925.822979
2700	932.8852139
2800	939.2750686
2900	945.0565589
3000	950.287606

Esercizio 3 (5 punti)

Si illustri il funzionamento dell'ottimizzatore delle interrogazioni di System R, evidenziandone le peculiarità e le possibili cause di sub-ottimalità della soluzione.

Esercizio 4 (3 punti)

Il protocollo di recovery del sistema ARIES si basa sull'assunzione che l'operazione di checkpoint sia un'azione atomica, assunzione ovviamente non realistica. Vi saranno, perciò, due record di inizio e fine checkpoint ed è quindi possibile che alcune transazioni scrivano record nel log durante l'effettuazione del checkpoint. Si illustri come debbano essere trattati tali record in un'eventuale fase di analisi seguita ad un crash del sistema.

Soluzione Esercizio 1

Numero di coppie (`annodinascita, stipendio`) = 20 x 50 = 1000, mediamente ci sono 11K/1K = 11 tuple per ogni valore di chiave.
Dimensione di ogni record (foglia) = 8 + 8 + 11 x 4 = **60B**
Dimensione "reale" foglia = (4096 – 96) x 0.825 = **3300B**
Numero di record per foglia = 3300/60 = **55**
Numero di foglie = 1K/55 = **19**
Dimensione di ogni record (nodo interno) = 3 + 2 x (8 + 8) = **35B**
Dimensione "reale" nodo = (4096 – 96) x 0.875 = **3500B**
Numero di record per nodo interno = 3500/35 = **100**
Numero nodi livello 1 = 19/100 = **1**
L'R-tree corrispondente si compone quindi di 2 livelli per un totale di 1 nodi interni e 19 foglie.

Soluzione Esercizio 2

Selettività dei predicati:
Predicato su `stipendio` = (40000 – 20000)/(220000 – 20000) = **0.1**
Predicato su `luogo` = 1/20 = **0.05** per ogni valore di `luogo`
Predicato di join = 1/50K (chiave esterna)

Accesso a P:
Costo scan sequenziale = **1000**
Costo indice su `luogo`: 3 x (NL x 0.05 + NP x 0.05) = 3 x (5 + 50) = **165**
Numero tuple residue = 3 x NT x 0.05 = **3000**

Accesso a R:
Costo scan sequenziale = **1000**
Costo indice su `stipendio`: NL x 0.1 + Φ(NT x 0.1, NP) = 50 x 0.1 + Φ(20K x 0.1, 1K) = 5 + Φ(2K, 1K) = 5 + 865 = **870**
Costo indice su `matricola`: 1 + 1 = **2**
Numero tuple residue = NT x 0.1 = **2000**

Costi di join:
`P` esterna: costo = costo indice `luogo` + 3000 x costo indice `matricola` = 165 + 3000 x 2 = **6165**
`R` esterna: costo = costo indice `stipendio` + 2000 x costo indice `luogo` = 870 + 2000 x 165 = **330870**
Il numero di risultati dell'interrogazione è 20K x 0.15 x 0.1 = **300**