

4ο Σύνολο Ασκήσεων
Ημερομηνία Παράδοσης: 13/1/2003

Εξηγήστε τις απαντήσεις σας.

Θεματική Ενότητα: Αποθήκευση. Ευρετήρια. Επεξεργασία και Βελτιστοποίηση Ερωτήσεων.

1. Έστω ένα B+-δέντρο τάξης 400 (και για τους εσωτερικούς κόμβους και για τα φύλλα).

(i) Ποιος είναι ο μικρότερος αριθμός κλειδιών που μπορεί ένα τέτοιο B+-δέντρο να κρατήσει (ευρετηριοποιήσει); Ποιος είναι ο μεγαλύτερο αριθμός κλειδιών που μπορεί ένα τέτοιο B+-δέντρο να κρατήσει (ευρετηριοποιήσει);

(ii) Θεωρείστε ότι κάθε κόμβος του B+-δέντρου είναι γεμάτος κατά 70%. Πόσα επίπεδα χρειάζονται για $6 * 10^9$ κλειδιά; Ποιο ποσοστό του συνολικού χώρου (blocks) είναι φύλλα σε αυτήν την περίπτωση;

2.

(i) Διορθώστε το Επεκτατό Ευρετήριο Κατακερματισμού (extendible hashing) του Σχήματος 1. Στη συνέχεια να εισάγετε το 18, 30, 62 και 63.

(ii) Για το Γραμμικό Ευρετήριο Κατακερματισμού (linear hashing) του Σχήματος 2:

(α) Ποιός είναι ο μεγαλύτερος αριθμός εγγραφών που μπορούμε να εισάγουμε χωρίς να γίνει διάσπαση.

(β) Ποιός είναι ο μικρότερος αριθμός εισαγωγών εγγραφών που μπορεί να οδηγήσει σε διάσπαση και των 4 κάδων.

3. Θεωρήστε μια σχέση $R(a, b, c, d)$ που περιέχει 100.000 εγγραφές και κάθε block της σχέσης περιέχει 10 εγγραφές. Το R είναι αποθηκευμένο σε αρχείο σωρού. Υποθέστε ότι το a είναι πρωτεύον κλειδί για την R και παίρνει τιμές από το 0 έως το 999.999 (ομοιόμορφα). Για τις παρακάτω ερωτήσεις (α) έως (δ), ποια από τις τέσσερις μεθόδους προσπέλασης έχει το μικρότερο κόστος (εξηγείστε):

1. Προσπέλαση (scan) του αρχείου απευθείας
2. Χρήση ενός ευρετηρίου B+ δέντρου στο γνώρισμα $R.a$
3. Χρήση ενός ευρετηρίου ενός επιπέδου στο γνώρισμα $R.a$
4. Χρήση γραμμικού κατακερματισμού στο γνώρισμα $R.a$

Οι ερωτήσεις είναι

(α) $\sigma_{a>1.000}(R)$

$$(\beta) \sigma_{a < 1.000}(R)$$

$$(\gamma) \sigma_{a = 1.000}(R)$$

$$(\delta) \sigma_{a = 1.000 \text{ OR } b > 100}(R)$$

Επαναλάβετε την ερώτηση αν η σχέση R είναι αποθηκευμένη σε ταξινομημένο στο a αρχεία με φθίνουσα διάταξη.

4. Θεωρείστε τη πράξη $R \bowtie_{R.a=S.b} S$. Το κόστος που μας ενδιαφέρει είναι ο αριθμός των μπλοκ που μεταφέρονται μεταξύ δίσκου και μνήμης. Αγνοείτε το κόστος της εγγραφής του αποτελέσματος στο δίσκο. Υποθέστε ότι: το μέγεθος του block είναι 512 bytes, η σχέση R έχει 7.000 πλειάδες και κάθε πλειάδα έχει μέγεθος 90 bytes, η σχέση S έχει 350.000 πλειάδες και κάθε πλειάδα έχει μέγεθος 100 bytes. Υποθέστε ότι κάθε πλειάδα του R συνενώνεται με ακριβώς 50 πλειάδες της σχέσης S . Θεωρήστε μη εκτεινόμενη αποθήκευση. Τα γνώρισμα a και b έχουν μέγεθος 8 bytes. Το γνώρισμα a είναι το πρωτεύον κλειδί για τη σχέση R . Ένας δείκτης block όπως και ένας δείκτης εγγραφής έχει μήκος 7 bytes. Υποθέστε ότι η σχέση S είναι αποθηκευμένη σε αρχείο διατεταγμένο ως προς το b .

(i) Πόσοι καταχωρητές μνήμης μεγέθους ενός block χρειάζονται για να χωρέσει όλο το ευρετήριο στο b στη μνήμη;

(ii) Έστω N η απάντησή σας στο προηγούμενο ερώτημα. Υποθέστε ότι υπάρχουν $N + 2$ καταχωρητές (δηλαδή αν η απάντησή σας στο ερώτημα (α) είναι 49, θεωρήστε 51 καταχωρητές). Υποθέστε ότι αρχικά το ευρετήριο στο b μεταφέρεται (φορτώνεται) όλο στη μνήμη. Δώστε μια σύντομη περιγραφή του αλγορίθμου συνένωσης. Υπολογίστε το κόστος της συνένωσης. Είναι αυτό το κόστος βέλτιστο;

(iii) Πόσο γίνεται το κόστος της συνένωσης αν υπάρχουν μόνο τρεις καταχωρητές; Δώστε τον αλγόριθμο.

5. (σύντομες απαντήσεις)

(i) Δώστε τουλάχιστον 2 λόγους γιατί τα B+ δέντρα πλεονεκτούν του κατακερματισμού.

(ii) Θεωρήστε τους 3 τύπους οργάνωσης αρχείων που συζητήσαμε στο μάθημα (αρχεία σωρού, διατεταγμένα αρχεία και αρχεία κατακερματισμού). Δώστε ένα παράδειγμα ενός συνόλου ερωτήσεων για τα οποία το καθένα είναι κατάλληλο.

(iii) Έστω μια σχέση R αποθηκευμένη σε m blocks όπου σε κάθε block αποθηκεύονται n πλειάδες. Ξέρουμε ότι μία επιλογή επιστρέφει τουλάχιστον m πλειάδες ομοιόμορφα κατανομημένες στα block του αρχείου. Με άλλα λόγια, περιμένουμε ότι κάθε block του αρχείου περιέχει τουλάχιστον μια πλειάδα που ανήκει στο αποτέλεσμα της επιλογής. Ο βασικός τρόπος εκτέλεσης της επιλογής είναι σειριακή αναζήτηση (scan) του αρχείου, δηλαδή ανάγνωση του αρχείου ένα block τη φορά, αυτό έχει κόστος m I/Os. Πόσο μπορεί αυτό να βελτιωθεί με τη χρήση κάποιου ευρετηρίου;