Πανεπιστημιο Ιωαννινων
Τμήμα Πληροφορικής

Βάσεις Δεδομένων
Ακαδημαϊκό Έτος 2001-2002

4ο Σύνολο Ασκήσεων

Ημερομηνία Παράδοσης: 14/1/2002 πριν το μάθημα

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

Το άριστα είναι το 100.

Θέμα 1 [15] Υποθέστε ότι έχετε 10.000.000 εγγραφές μεγέθους 100 bytes που είναι αποθηκευμένες σε ένα ταξινομημένο αρχείο. Θεωρείστε μη εκτεινόμενη οργάνωση. Το μέγεθος του block είναι 4096 bytes. Κατασκευάζουμε ένα μη πυκνό B+ ευρετήριο όπου τα φύλλα έχουν δείκτες στην πρώτη εγγραφή κάθε block. Θεωρείστε ότι το κλειδί ευρετηρίου είναι 20 bytes, οι δείκτες δέντρου και αρχείου (δεδομένων) είναι 20 bytes.

(i) Δώστε τον αριθμό των κόμβων του δέντρου για τη χειρότερη και την καλύτερη περίπτωση (δηλαδή την περίπτωση που χρησιμοποιούνται τα περισσότερα και τα λιγότερα blocks αντίστοιχα).
(ii) Θεωρείστε ότι χρησιμοποιούμε το δέντρο για να υπολογίσουμε μια συνθήκη διαστήματος, και ότι αυτή επιστρέφει 860 εγγραφές. Υπολογείστε τον αριθμό των blocks που χρειάζεται να μεταφερθούν για τη χειρότερη και την καλύτερη περίπτωση.

Θέμα 2 [Μονάδες 15] Θεωρείστε επεκτατό δυναμικό κατακερματισμό.

(i) Υποθέστε ότι κάθε κάδος (bucket) χωρά το πολύ 3 κλειδιά. Δώστε τη δομή που προκύπτει με την εισαγωγή των 1, 2, ..., 14 με αυτή τη σειρά. Δώστε το τοπικό και το ολικό βάθος για τον κατάλογο. Υπόδειξη: χρησιμοποιείστε τα τελευταία bits της δυαδικής αναπαράστασης.
(ii) Ο διπλασιασμός του καταλόγου απαιτεί την εξέταση όλων των κάδων με τοπικό βάθος ίσο με το ολικό βάθος; Εξηγείστε.
(iii) Αν η συνάρτηση κατακερματισμού δεν είναι καλή (οδηγεί σε μη ομοιόμορφη κατανομή των δεδομένων σε κάδους) τι συμπέρασμα μπορείτε να βγάλετε για το μέγεθος του καταλόγου; Τι μπορείτε να πείτε για την πυκνότητα των εγγραφών στα blocks του αρχείου δεδομένων (δηλαδή για το πόσο γεμάτα είναι τα blocks);

Θέμα 3 [Μονάδες 15] Θεωρείστε το B+ δέντρο του Σχήματος 1. Απαντείστε στα παρακάτω και αποδείξτε τις απαντήσεις σας.

(i) Δώστε μια ακολουθία από 5 τιμές που:

- η εισαγωγή τους με αυτήν τη σειρά ακολουθούμενη από την διαγραφή τους με την ανάποδη σειρά (π.χ., εισαγωγή a, εισαγωγή b, διαγραφή b, εισαγωγή α) έχει ως αποτέλεσμα το αρχικό δέντρο
- η εισαγωγή τους με αυτήν τη σειρά ακολουθούμενη από την διαγραφή τους με την ανάποδη σειρά (π.χ., εισαγωγή a, εισαγωγή b, διαγραφή b, εισαγωγή α) έχει ως αποτέλεσμα ένα διαφορετικό δέντρο

(ii) Ποιος είναι ο μικρότερος αριθμός εισαγωγών τιμών που θα οδηγούσε το ύψος του δέντρου του Σχήματος 1 από την τωρινή τιμή του (δηλαδή από 1) σε 3;

Θέμα 4 [Μονάδες 15] Θεωρήστε μια σχέση $R(a, b, c, d)$ που περιέχει 1.000.000 εγγραφές και κάθε block της σχέσης περιέχει 10 εγγραφές. Το R είναι αποθηκευμένο σε αρχείο σωρού με πυκνά δευτερεύοντα ευρετήρια και οι εγγραφές του R είναι αποθηκευμένες στο αρχείο τυχαία. Υποθέστε ότι το a είναι ένα υποψήφιο κλειδί για την R με τιμές από το 0 έως το 999.999. Για τις παρακάτω τέσσερεις ερωτήσεις ποιά από τις τρεις μεθόδους προσπέλασης είναι το πιο πιθανό να έχει το μικρότερο κόστος (δηλαδή να προκαλέσει τις λιγότερες προσπελάσεις block):

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

Οι ερωτήσεις είναι

(α) Βρείτε όλες τις πλειάδες της R

(β) $\sigma_{a<50}(R)$

(γ) $\sigma_{a=50}(R)$

(γ) $\sigma_{a>50 \ AND \ a<100}(R)$

Επαναλάβατε την ερώτηση αν (i) το a δεν είναι υποψήφιο κλειδί και (ii) η σχέση R είναι αποθηκευμένη σε ταξινομήμενο αρχείο ως προς το a.

Θέμα 5 [Μονάδες 35] Θεωρείστε τη πράξη $R \bowtie_{R.a=S.b} S$. Το κόστος που μας ενδιαφέρει είναι ο αριθμός των blocks που μεταφέρονται μεταξύ δίσκου και μνήμης. Αγνοείστε το κόστος της εγγραφής του αποτελέσματος στο δίσκο (εκτός αν ζητείτε ρητά). Υποθέστε ότι: το μέγεθος του block είναι 1024 bytes, η σχέση R έχει 10.000 πλειάδες και κάθε πλειάδα έχει μέγεθος 50 bytes, η σχέση S έχει 2.000 πλειάδες και κάθε πλειάδα έχει μέγεθος 48 bytes. Το γνώρισμα a είναι το πρωτεύον κλειδί για τη σχέση R. Το γνώρισμα a και το γνώρισμα b έχουν μέγεθος 8 bytes. Ένας δείκτης block έχει μήκος $P = 6$ bytes και ένας δείκτης εγγραφής έχει μήκος $P_R = 7$ bytes.

(α) Ποιός είναι ο μέγιστος αριθμός πλειάδων που παράγεται από τη συνένωση και ποιό το κόστος εγγραφής αυτού του αποτελέσματος στο δίσκο;

(β) Υποθέστε ότι η σχέση S είναι αποθηκευμένη σε μη ταξινομημένο αρχείο, ενώ η σχέση R είναι αποθηκευμένη σε ταξινομημένο αρχείο ως προς το γνώρισμα a. Υπάρχει ένα ευρετήριο ως προς a για τη σχέση R και ένα ευρετήριο ως προς b για τη σχέση S. Υποθέστε ότι κάθε πλειάδα του R συνενώνεται με ακριβώς 5 πλειάδες της σχέσης S. Θεωρείστε ότι υπάρχουν 25 καταχωρητές μεγέθους ενός block.

(β1) Υπολογίστε το μέγεθος αυτών των ευρετηρίων.

(β2) Δώστε τον αλγόριθμο της συνένωσης σε ψευτογλώσσα χρησιμοποιώντας το ευρετήριο στο a.

(β3) Υπολογίστε το κόστος της συνένωσης αν χρησιμοποιηθεί ο αλγόριθμος (β2). Μη ξεχάστε να υπολογίστε το κόστος ανάγνωσης των απαραίτητων blocks του ευρετηρίου.

(β4) Επαναλάβετε τα ερωτήματα (β2) και (β3) χρησιμοποιώντας το ευρετήριο στο b.

(γ) Υποθέστε ότι η σχέση S είναι αποθηκευμένη σε μη ταξινομημένο αρχείο, η σχέση R είναι αποθηκευμένη σε ταξινομημένο αρχείο ως προς το γνώρισμα a και ότι υπάρχει κατακερματισμός στο γνώρισμα a. Θεωρείστε ότι υπάρχουν 42 καταχωρητές μεγέθους ενός block.

(γ1) Δώστε σε ψευτογλώσσα έναν αλγόριθμο που να υπολογίζει τη συνένωση χρησιμοποιώντας τη συνάρτηση κατακερματισμού στο a.

(γ2) Υπολογίστε το κόστος της συνένωσης αν χρησιμοποιηθεί ο αλγόριθμος (γ1).

(δ) Εξηγείστε αν το αποτέλεσμα των αλγορίθμων (β2), (β4) και (γ1) είναι ταξινομημένο ή όχι.

Θέμα 6 [10] Άσκηση 16.11, ερώτημα (β) σελίδα 185 του 2ου τόμου του βιβλίου για την ερώτηση E1B. Δώστε ένα δέντρο που θα είχε μικρότερο κόστος κάτω από συγκεκριμένες συνθήκες.