Πανεπιστημιο Ιωαννινων
Τμήμα Πληροφορικής

Βάσεις Δεδομένων
Ακαδημαϊκό Έτος 2000-2001

4ο Σύνολο Ασκήσεων

Ημερομηνία Παράδοσης: 11/1/2001 πριν το μάθημα

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

Το άριστα είναι το 100.

1. [25] Θεωρήστε ότι το μέγεθος του δείκτη εγγραφής είναι 8 bytes, το μέγεθος του δείκτη block είναι 7 bytes, το μέγεθος του block 512 bytes, το μέγεθος του γνωρίσματος $R.a$ 10 bytes και το μέγεθος του γνωρίσματος $S.b$ 9 bytes. Έστω μια σχέση $R(a, b, c, d, e)$ με 5.000.000 εγγραφές (πλειάδες) όπου κάθε block της σχέσης έχει 10 εγγραφές (παράγοντας ομαδοποίησης του αρχείου δεδομένων). Υποθέστε ότι το $R.a$ είναι υποψήφιο κλειδί και ότι η σχέση είναι αποθηκευμένη σε ένα αρχείο ταξινομημένο με βάση το $R.a$. Το $R.a$ παίρνει τιμές από 0 έως 4.999.999. Για τις παρακάτω εκφράσεις τις σχεσιακής άλγεβρας, ποιά από τις τρεις μεθόδους έχει το μικρότερο κόστος:

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

(α) $\sigma_{a<50.000}(R)$

(β) $\sigma_{a=50.000}(R)$

(γ) $\sigma_{a>50.000 \ AND \ a<50.010}(R)$

(δ) $\sigma_{a\neq 50.000}(R)$

Θέμα 2 [Μονάδες 20] 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. Αγνοείστε το κόστος της εγγραφής του αποτελέσματος στο δίσκο (εκτός αν αναφέρεται ρητά). Υποθέστε ότι: το μέγεθος του μπλοκ είναι 512 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 έχει μέγεθος 8 bytes και το γνώρισμα b 12 bytes. Ένας δείκτης block έχει μήκος $P = 6$ bytes και ένας δείκτης εγγραφής έχει μήκος $P_R = 7$ bytes. Υπάρχουν 48 καταχωρητές μεγέθους ενός block.

(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. Υπολογίστε το μέγεθος αυτών των ευρετηρίων.

(γ) Υποθέστε ότι κάθε πλειάδα του R συνενώνεται με ακριβώς 5 πλειάδες της σχέσης S. Υπολογίστε το κόστος της συνένωσης αν χρησιμοποιηθεί το ευρετήριο στο a. Δώστε τον αλγόριθμο της συνένωσης τους σε ψευτογλώσσα. Μη ξεχάστε να υπολογίστε το κόστος ανάγνωσης των απαραίτητων μπλοκ του ευρετηρίου.

(δ) Επαναλάβετε το ερώτημα (γ) χρησιμοποιώντας το ευρετήριο στο b.

3. [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.

4. [35] Σας ζητούν να σχεδιάστε μια μηχανή αναζήτησης για το web (κάτι σαν το hotbot ή το yahoo). Η απάντηση σας θα πρέπει να περιλαμβάνει τουλάχιστον τα παρακάτω και να μην ξεπερνά τις 3 σελίδες.

Το σύστημα σας δε θα πρέπει να χρησιμοποιεί κάποιο Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ).
(i) Δώστε μια γενική περιγραφή της αρχιτεκτονικής του συστήματος που θα υλοποιούσατε (κατά προτίμηση σχηματική).
(ii) Τι είδους ευρετήρια θα χρησιμοποιούσατε; Ποια θα ήταν η μορφή των εγγραφών του ευρετηρίου;
(iii) Βασισμένοι στο είδος των ερωτήσεων που υποστηρίζει η μηχανή αναζήτησής σας δείξτε τη βελτίωση που παρέχει το ευρετήριο.
(iv) Πότε θα ενημερώνεται το ευρετήριο και ποιο είναι το κόστος ενημέρωσής του;

Τώρα υποθέστε ότι χρησιμοποιείται ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων (π.χ., Oracle).
(i) Δώστε μια γενική περιγραφή της αρχιτεκτονικής του συστήματος που θα υλοποιούσατε (κατά προτίμηση σχηματική).
(ii) Δώστε 3 μειονεκτήματα και 3 πλεονεκτήματα της προσέγγισης με χρήση ΣΔΒΔ.

5. [15] (α) Θεωρείστε to B+ δέντρο της εικόνας.

(i) Δώστε το δέντρο που προκύπτει μετά την εισαγωγή των εγγραφών με κλειδί 10, 11, και 12.
(iii) Δώστε το δέντρο μετά τη διαγραφή της εγγραφής με κλειδί 80.

(β) Για τα B+ δέντρα προσπαθούμε να βάλουμε όσο περισσότερα κλειδιά/δείκτες σε κάθε κόμβο ώστε ο κόμβος να χωρά σε μια σελίδα δίσκου (block). Στόχος είναι να διατηρηθεί το ύψος του δέντρου όσο το δυνατόν μικρότερο. Γιατί στην περίπτωση των δομών δεδομένων για "μνήμη" (π.χ., ισοζυγισμένα δέντρα) δε χρησιμοποιούνται επίσης "μεγάλοι" κόμβοι (συνήθως τα δέντρα αυτά έχουν 2 ή 3 δείκτες ανά κόμβο); Αυτό συμβαίνει γιατί οι άνθρωποι που ασχολούνται με τις βάσεις δεδομένων είναι πιο έξυπνοι ώστε να καταλάβουν ότι μεγαλύτεροι κόμβοι έχουν σαν αποτέλεσμα δέντρα με μικρότερο ύψος ή/και για κάποιον άλλο λόγο;