Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2024-2025

Πρώτη Σειρά Ασκήσεων

Ανάθεση: 13-03-2025
Παράδοση: 27-03-2025 Ώρα (23:55)

Οδηγίες

- *Η πρώτη σειρά ασκήσεων είναι **ατομική** και **υποχρεωτική**.*
- *Η υποβολή της εργασίας πρέπει να γίνει στο eclass.*
- *Το παραδοτέο σας θα πρέπει να είναι ένα αρχείο PDF με όνομα AM.pdf (όπου AM είναι ο αριθμός μητρώου σας. π.χ. "3220001.pdf").*
- *Τα διαγράμματα πρέπει να είναι κατασκευασμένα σε κάποιο πρόγραμμα (της επιλογής σας) και όχι σκαναρισμένα χειρόγραφα.*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*
- ***Για την επίλυση των ασκήσεων να μελετήσετε τις διαφάνειες των διαλέξεων του μαθήματος.***

  Η συνολική βαθμολογία των ασκήσεων ανέρχεται σε 105 μονάδες (100+5 μονάδες bonus).

Άσκηση 1 [25 μονάδες]

Ένας σκληρός δίσκος έχει τα παρακάτω χαρακτηριστικά:

- 4 πλακέτες (platters) διπλής όψης
- 4096 ίχνη (tracks) ανά επιφάνεια
- 1024 τομείς ανά ίχνος (sectors/track)
- Μέγεθος τομέα 512 bytes
- Μέσος Χρόνος Μετακίνησης Κεφαλής (Average Seek Time) = 10 ms
- Ταχύτητα περιστροφής 7200 rpm
- Μέγεθος μπλοκ (block) 4096 bytes.

a. Ποιος είναι ο μέγιστος αριθμός εγγραφών που είναι δυνατόν να αποθηκευτούν στον δίσκο, αν το μέγεθος κάθε εγγραφής είναι 256 bytes;

b. Στον δίσκο είναι αποθηκευμένο ένα αρχείο με 1.000.000 εγγραφές μεγέθους 256 bytes έκαστη. Το αρχείο καταλαμβάνει συνεχόμενα μπλοκ στον δίσκο. Θέλουμε να διαβάσουμε Χ τυχαίες εγγραφές από το αρχείο. Μπορούμε να διαβάσουμε τυχαία Χ μπλοκ ή να διαβάσουμε ολόκληρο το αρχείο και να αναζητήσουμε τις Χ εγγραφές. Για ποιες τιμές του Χ η ανάγνωση ολόκληρου του αρχείου είναι πιο αποδοτική από Χ τυχαία διαβάσματα;

Σημείωση: Για λόγους απλούστευσης θεωρείστε ότι στην περίπτωση διαβάσματος όλου του αρχείου **ο χρόνος μετακίνησης στο επόμενο ίχνος** είναι μηδαμινός (δηλαδή δεν λαμβάνεται υπόψη).

Άσκηση 2 [25 μονάδες]

Έστω μια σχέση R (**A**,B,C,D) με πρωτεύον κλειδί το γνώρισμα **A**. Η σχέση περιέχει 100.000 πλειάδες και είναι αποθηκευμένη σε ένα αρχείο το οποίο είναι διατεταγμένο στον δίσκο με βάση το πρωτεύον κλειδί. Το γνώρισμα A έχει μέγεθος 6 bytes, τα γνωρίσματα B και C έχουν μέγεθος 25 bytes το καθένα, και το γνώρισμα D έχει μέγεθος 8 bytes. Θεωρείστε ότι όλες οι τιμές του γνωρίσματος D είναι μοναδικές και ότι το μέγεθος του μπλοκ (σελίδας) είναι 512 bytes.

a. Πόσα μπλοκ καταλαμβάνει το αρχείο με τις πλειάδες της σχέσης R;

b. Θέλουμε να κατασκευάσουμε ένα δευτερεύον (non-clustered) B+ ευρετήριο με πεδίο ευρετηρίασης το γνώρισμα D. Υπολογίστε το **μικρότερο** και το **μεγαλύτερο** αριθμό των μπλοκ (σελίδων) που θα καταλαμβάνει αυτό το ευρετήριο, υποθέτοντας ότι το μέγεθος δείκτη δέντρου (δείκτης προς κόμβο του δέντρου) είναι 4 bytes και το μέγεθος δείκτη δεδομένων είναι 5 bytes.

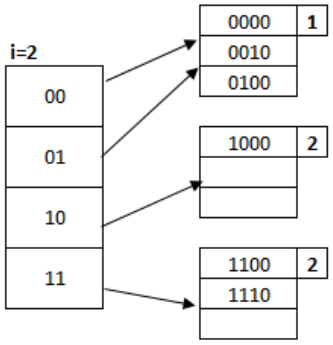
c. Έστω ότι 1000 εγγραφές της σχέσης R ικανοποιούν την συνθήκη d1<D<d2, όπου d1 και d2 είναι τιμές του πεδίου τιμών του γνωρίσματος D. Πόσα μπλοκ πρέπει να διαβαστούν για να ανακτηθούν οι παραπάνω εγγραφές χρησιμοποιώντας το μικρότερο B+ δέντρο που υπολογίσατε στο ερώτημα b;

Να θεωρήσετε ότι:

- Το ευρετήριο B+ δέντρο είναι αποθηκευμένο στον δίσκο.
- Μπορεί να χρησιμοποιηθεί η συνολική χωρητικότητα του μπλοκ τόσο για την αποθήκευση των εγγραφών του ευρετηρίου όσο και για την αποθήκευση των εγγραφών της σχέσης R.
- Κάθε εγγραφή (ευρετηρίου και δεδομένων) αποθηκεύεται ολόκληρη σε ένα μπλοκ.

Άσκηση 3 [20 Μονάδες]

Θεωρείστε το παρακάτω ευρετήριο επεκτατού κατακερματισμού, με χώρο τριών κλειδιών ανά κάδο και μία συνάρτηση κατακερματισμού η οποία επιστρέφει 4 bits για κάθε κλειδί. Οι κάδοι περιέχουν τις κατακερματισμένες τιμές των κλειδιών.



Ζητείται:

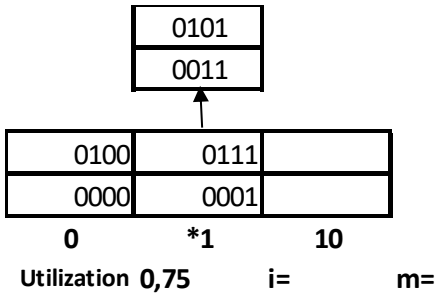
a. Ποιος είναι ο μέγιστος αριθμός τιμών που μπορούν να καταχωρηθούν στο ευρετήριο δίχως να αυξηθεί το ολικό του βάθος; Να αιτιολογήσετε την απάντησή σας.

b. Ποιος είναι ο ελάχιστος αριθμός τιμών που θα οδηγούσε σε διπλασιασμό του ευρετηρίου; Να αιτιολογήσετε την απάντησή σας.

c. Να εισαγάγετε τις ακόλουθες τιμές με την σειρά που δίνονται: **[1111, 1001, 1101, 0101]** Να δείξετε την μορφή του ευρετηρίου μετά την εισαγωγή κάθε τιμής.

Άσκηση 4 [25 μονάδες]

Έστω ένα αρχείο ευρετηρίου που χρησιμοποιεί την μέθοδο του γραμμικού κατακερματισμού με κάδους χωρητικότητας δύο εγγραφών. Για την κατανομή των τιμών χρησιμοποιούνται τα **i** λιγότερο σημαντικά bits. Ο αριθμός των κάδων πρέπει να αυξάνεται όταν το utilization του ευρετηρίου γίνει μεγαλύτερο ή ίσο του **80%**. Το **i** αυξάνεται μόνο όταν κρίνεται απαραίτητο. Επίσης, δεν υπάρχει όριο στον αριθμό σελίδων υπερχείλισης. Κάθε σελίδα υπερχείλισης χωράει και αυτή δύο εγγραφές. Στο ευρετήριο έχουν εισαχθεί έξι κλειδιά (αναγράφεται η τιμή h(x) αντί για το κλειδί x):



Ζητείται:

a. Να προσδιορίσετε την τιμή των **i** και **m**

b. Να εισαγάγετε τις παρακάτω τιμές με την σειρά που δίνονται ξεκινώντας από αριστερά προς τα δεξιά. Να εμφανίσετε την μορφή του ευρετηρίου μετά από κάθε εισαγωγή κλειδιού δείχνοντας και όσα ενδιάμεσα βήματα απαιτούνται. Κάθε πράξη εισαγωγής πρέπει να εκτελείται στο αποτέλεσμα της προηγούμενης και όχι στο αρχικό ευρετήριο.

[0010, 0011, 0100, 0100, 0111, 0110, 0111]

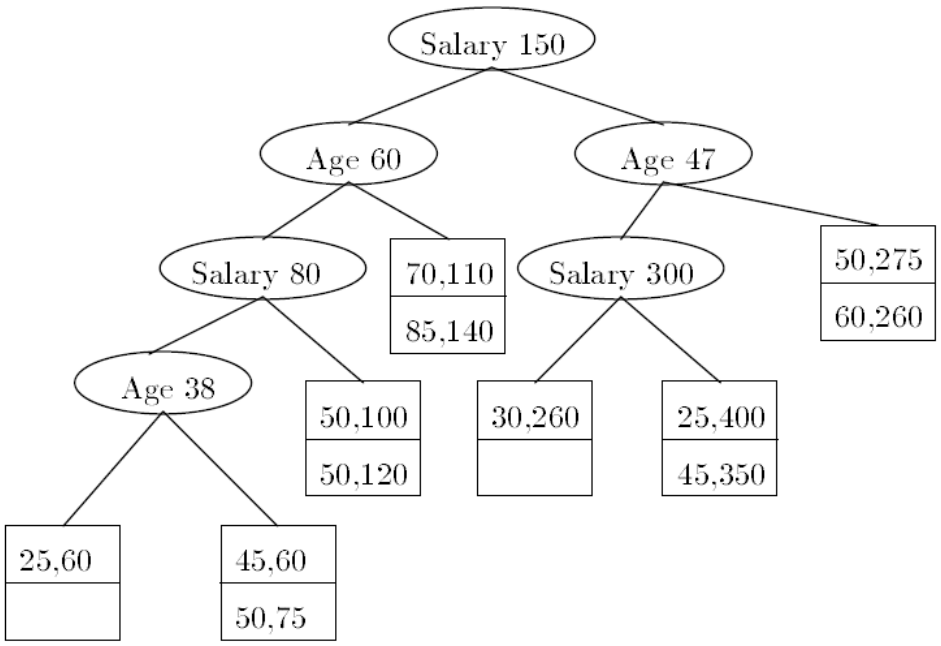
c. Να υπολογίσετε τον μέσο αριθμό προσπελάσεων για την ανάκτηση μιας εγγραφής όταν δίνεται η τιμή του κλειδιού. Να θεωρήσετε ότι:

- Οι τιμές του κλειδιού είναι μοναδικές και το κλειδί προς αναζήτηση υπάρχει στο ευρετήριο.
- Η αναζήτηση γίνεται στην τελική μορφή του ευρετηρίου όπως αυτή έχει προκύψει μετά την εισαγωγή των τιμών του ζητήματος b.
- Στους κάδους του ευρετηρίου αποθηκεύονται οι τιμές των κλειδιών και όχι τα hash codes.

Να δείξετε τον τρόπο υπολογισμού και όχι μόνο το τελικό αποτέλεσμα.

Άσκηση 5 [10 μονάδες]

Έστω το ακόλουθο ευρετήριο kd-tree



Γράψτε τις λογικές εκφράσεις που προσδιορίζουν **καλύτερα** το σύνολο των σημείων που δύναται να εισαχθούν στο φύλλο με τιμές α) με τιμές (30,260) και β) (50,100) και (50,120) του παραπάνω kd-tree; Να αιτιολογήσετε την απάντησή σας.