Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2005-2006 Εαρινό Εξάμηνο

3η Σειρά Ασκήσεων
Ανάθεση: 1 Απριλίου
Παράδοση: 14 Απριλίου

Άσκηση 1 *(3 βαθμοί) (Ενότητα: Ευρετηρίαση Κειμένου)*
Θεωρείστε ένα έγγραφο το οποίο περιέχει τα εξής λόγια του Βενιαμίν Φραγκλίνου:
«***Οποιος θυσιάζει την ελευθερία για την ασφάλεια δεν αξίζει ούτε την ελευθερία ούτε την ασφάλεια.***»
(α) Φτιάξτε το ανεστραμμένο ευρετήριο αυτού του εγγράφου.
(β) Φτιάξτε το δένδρο καταλήξεων του εγγράφου θεωρώντας ως σημεία ευρετηρίου (index points) τις αρχές των λέξεων (μπορείτε να δώσετε κατευθείαν το PATRICIA tree).
(γ) Σχολιάστε το μέγεθος των (α) και (β) ως προς το μέγεθος του εγγράφου.

Άσκηση 2 *(5 βαθμοί) (Ενότητα: Ευρετηρίαση Κειμένου)*
Θέλετε να σχεδιάσετε ένα ΣΑΠ που να βασίζεται στο διανυσματικό μοντέλο για μια συλλογή κειμένων συνολικού μεγέθους στο δίσκο 1 Gigabyte. Έστω ότι το μέσο μέγεθος των λέξεων που εμφανίζονται στα κείμενα είναι 10 χαραχτήρες και ότι το πλήθος των διαφορετικών λέξεων της συλλογής είναι 10.000.
(α) Ποιο το αναμενόμενο (μέγιστο) μέγεθος του ανεστραμμένου ευρετηρίου για τη συλλογή αυτή;
(β) Ποιο το αναμενόμενο (μέγιστο) μέγεθος του ανεστραμμένου ευρετηρίου αν χρησιμοποιήσετε block addressing με μέγεθος block ίσο με 200 λέξεις;
Τι ποσοστό μείωσης έχουμε, σε σχέση με το (α);
(γ) Αν έπρεπε το ευρετήριο να καταλαμβάνει το πολύ 1 MB (π.χ. για να χωράει στην κύρια μνήμη ενός κινητού τηλεφώνου) πως θα σχεδιάζατε το ανεστραμμένο ευρετήριο;
(δ) Αν έπρεπε να καταλαμβάνει το πολύ 100 K τι θα κάνατε;
(ε) Αν έπρεπε να καταλαμβάνει το πολύ 10 K τι θα κάνατε;
(στ) Αν έπρεπε να υποστηρίξετε και phrase queries (μέγιστου μήκους 4 διαδοχικών λέξεων) και είχατε στη διάθεση σας μόνο 1 Mbyte μνήμης και επιλέγατε αντί για ανεστραμμένο ευρετήριο να είχατε αρχείο υπογραφών πως θα το σχεδιάζατε;

Άσκηση 3 *(2 βαθμοί)*
Το Google θέλει να υποστηρίξει εξατομικευμένη διαβάθμιση των ιστοσελίδων στους χρήστες του. Κάθε χρήστης να μπορεί να ορίζει ένα ή περισσότερα προφίλ. Κάθε προφίλ (ενός χρήστη) θα έχει ένα όνομα (π.χ. ψυχαγωγία, my Master, MyMusic, computer science, κλπ) το οποίο ο χρήστης θα δίδει αρχικά. Η διαδικασία της αναζήτησης θα τροποποιηθεί ως εξής: Δίπλα στο πλαίσιο διατύπωσης επερωτήσεων του Google, θα υπάρχει ένα μενού το οποίο θα εμφανίζει όλα τα ονόματα προφίλ που έχει δηλώσει ο χρήστης καθώς και την προεπιλεγμένη επιλογή «ΧωρίςΠροφίλ». Η αναζήτηση με επιλογή «ΧωρίςΠροφίλ» θα διενεργείται όπως γίνεται σήμερα. Αν ο χρήστης έχει επιλέξει ένα προφίλ, τότε δίπλα σε κάθε στοιχείο της απάντησης (σύνδεσμο προς ιστοσελίδα) της κάθε απάντησης θα εμφανίζονται δυο κομβία: ένα good και ένα bad. Ανάλογα με το περιεχόμενο της κάθε σελίδα και τις προτιμήσεις του χρήστη (και για το συγκεκριμένο προφίλ), ο χρήστης θα μπορεί να πατήσει το good ή το bad για όποιες σελίδες το επιθυμεί. Βάσει αυτής της ανατροφοδότησης το Google πρέπει να παρέχει εξατομικευμένη διαβάθμιση σελίδων..
(α) Πως θα αξιοποιούσατε τα good και bad εισόδους του χρήστη ώστε να υποστηρίξετε εξατομικευμένη ανάκτηση; Μπορείτε να περιγράψετε παραπάνω από ένα τρόπους.
(β) Για κάθε προφίλ ενός χρήστη τι θα αποθηκεύατε; Λάβετε υπόψη ότι εκατομμύρια χρήστες χρησιμοποιούν το Google καθώς και το γεγονός ότι το Google ευρετηριάζει περίπου 9 δις σελίδες.
(γ) Ποιο θα ήταν το μεγαλύτερο τεχνικό πρόβλημα για την υποστήριξη αυτής της λειτουργικότητας;
Αν προτείνατε παραπάνω από μια μέθοδο απαντήστε αυτό το ερώτημα για κάθε μία μέθοδο και αναφέρετε τη μέθοδο που θα επιλέγατε (και δικαιολογείστε).