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

4η Σειρά ασκήσεων
(Συμπίεση, Ομαδοποίηση, Ευρετηρίαση Πολυμέσων, Κατανεμημένη Ανάκτηση)
Ημερομηνία Ανάθεσης: 5/5/2006
Ημερομηνία Παράδοσης: 5/6/2006

Άσκηση 1 (1.5 βαθμ.)

Θεωρείστε τα εξής λόγια του Σενέκα:
«Ο γνωστικός γνωρίζει τον αμαθή γιατί υπήρξε και ο ίδιος αμαθής. Ο αμαθής δεν γνωρίζει τον γνωστικό γιατί δεν υπήρξε ποτέ γνωστικός».
α) Θεωρώντας την κάθε λέξη ως σύμβολο του αλφαβήτου, ποια είναι η εντροπία του αλφαβήτου;
β) Δώστε τη συμπιεσμένη μορφή του κειμένου χρησιμοποιώντας κανονικοποιημένους κώδικες Huffman.

Άσκηση 2 (1.5 βαθμ.)
Θεωρείστε 5 έγγραφα A, B, C, D, E και έστω ότι οι αποστάσεις μεταξύ τους είναι αυτές του παρακάτω πίνακα. Δώστε το δενδρικό διάγραμμα που προκύπτει εφαρμόζοντας ιεραρχική ομαδοποίηση εγγράφων τύπου: (α) SingleLink, (β) CompleteLink, και (γ) Average Link.

A					
B	6				
C	4.2	5			
D	4	2	5		
E	2	8	5	6	
	A	B	C	D	E

Άσκηση 3 (1 βαθμ.)
Έστω ότι έχουμε 5 εικόνες A, B, C, D, E των οποίων οι αποστάσεις είναι αυτές που δίνονται στον πίνακα της Άσκησης 2. Προκειμένου να μπορούμε να απαντήσουμε επερωτήσεις γρήγορα θέλουμε να φτιάξουμε ένα μετρικό ευρετήριο, συγκεκριμένα ένα Vantage-Point-Tree (VTP). Σχεδιάστε το VTP που προκύπτει:
α) αν επιλέξουμε την εικόνα A ως κεντρική (pivot),
β) αν επιλέξουμε την εικόνα C ως κεντρική (pivot).

Άσκηση 4 (60 βαθμοί)
Θεωρείστε τα ακόλουθα έγγραφα όπου τα γράμματα A-E συμβολίζουν λέξεις.
d1 = «Α Β Γ », d2 = «Β Ε Β»
d3 = «Δ Β », d4 = «Γ Ε Γ»
d5 = «Δ Γ Ε Γ», d6 = «Γ Ε»
d7 = «Β Δ Β», d8 = «Ε Β»
Έστω ότι τα d1,d5, d6 ανήκουν σε ένα σύστημα S1, τα d2,d4 σε ένα σύστημα S2, και τα υπόλοιπα (d3,d7,d8) σε ένα σύστημα S3. Θέλουμε να φτιάξουμε έναν μεσίτη Μ πάνω από αυτά τα συστήματα.

(α) Για την επιλογή πηγής ο Μ θέλει να περιγράψει τα περιεχόμενα της κάθε πηγής με ένα διάνυσμα. Δώστε τα διανύσματα πηγών των S1, S2 και S3.

(β) Έστω ότι ο Μ έχει ήδη τα διανύσματα πηγών των S1, S2, S3 και λαμβάνει την επερώτηση q= "Α Γ". Αν θέλει να προωθήσει την επερώτηση q σε μία μόνο πηγή, ποια θα επιλέξει;

(γ) Ο Μ λαμβάνει μια επερώτηση, την προωθεί σε όλες τις πηγές, και λαμβάνει τα εξής αποτελέσματα από την κάθε μια:

S1: <d1, d6, d5>
S2: <d4, d2>
S3: <d7, d8, d3>

Δώστε την ενοποιημένη διάταξη κατά round robin interleaving

(δ) Προκειμένου ο μεσίτης να λαμβάνει από τις πηγές απαντήσεις με συγκρίσιμα σκορ, αποφασίζει να κάνει αποτίμηση επερωτήσεων σε δυο φάσεις ώστε οι πηγές να λαμβάνουν τα καθολικά στατιστικά που χρειάζονται για τον σωστό υπολογισμό των σκορ. Δώστε το idf του κάθε όρου στην καθολική συλλογή εγγράφων.

(ε) Ο μεσίτης βρίσκει άλλο ένα σύστημα S4 το οποίο έχει την ίδια συλλογή με αυτήν του S1, δηλαδή και αυτό παρέχει πρόσβαση στα έγγραφα d1, d5, d6. Έστω ότι ο Μ προωθεί μια επερώτηση q στα S1 και S4 και λαμβάνει τις εξής απαντήσεις:

S1: <d1, d5, d6>
S4: <d6, d1, d5>

Ποιο είναι το κορυφαίο έγγραφο αν ενοποιήσουμε τις διατάξεις: (i) κατά Borda, (ii) κατά Condorcet;
Ο Μ αποφασίζει να δίνει στο χρήστη όχι μόνο την ενοποιημένη διάταξη, αλλά και την Kemeny distance μεταξύ των διατάξεων που έλαβε από τα υποσυστήματα (προκειμένου ο χρήστης να παίρνει μια γεύση για το βαθμό συμφωνίας των πηγών). Ποια είναι αυτή η απόσταση στην προκειμένη;

(στ) Τα συστήματα S1, S2, S3 δεν θέλουν πλέον να έχουν ανάγκη τον Μ και αποφασίζουν να «ανεξαρτητοποιηθούν» φτιάχνοντας ένα σύστημα ομοτίμων (P2P), συγκεκριμένα ένα δομημένο σύστημα τύπου Chord. Προσελκύουν μάλιστα άλλα δυο συστήματα S5 και S6 (τα οποία δεν έχουν καμία συλλογή εγγράφων).
Αποφασίζουν να χρησιμοποιήσουν μια συνάρτηση κατακερματισμού h των 3 bits, και έστω ότι

h(IPaddress(S1))=1, h(IPaddress(S2))=2, h(IPaddress(S3))=4,
h(IPaddress(S5))=7, h(IPaddress(S6))=5

Αποφασίζουν να διανείμουν το ανεστραμμένο ευρετήριο θεωρώντας κάθε όρο σαν κλειδί και έστω ότι

h(A)=2, h(B)=3, h(Γ)=6
h(Δ)=6, h(E)=5

Δώστε (i) τους πίνακες δρομολόγησης των κόμβων S1 και S3 και (ii) πως θα κατανεμηθεί το ανεστραμμένο ευρετήριο στους κόμβους του δικτύου (δείξτε τι ακριβώς θα έχει κάθε κόμβος)