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 (15 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.

Λύση

α) Η έννοια της πληροφοριακής εντροπίας περιγράφει την ποσότητα της πληροφορίας που βρίσκεται μέσα σε ένα σήμα.

Παράδειγμα: Έστω ένα κουτί από μπάλες. Αν όλες οι μπάλες έχουν διαφορετικό χρώμα τότε είμαστε αβέβαιοι στο μέγιστο βαθμό όσο αφορά την απάντηση στο ερώτημα τι χρώμα μπάλα θα πάρει κάποιος από το κουτί. Αν όμως ένα χρώμα υπάρχει σε μεγαλύτερη ποσότητα από κάποια άλλα τότε η αβεβαιότητα μας μικραίνει.

Τα συστατικά στοιχεία του λόγου του Σενέκα είναι:

Λέξη	Πιθανότητα εμφάνισης
Ο	2/22
γνωστικός	2/22
γνωρίζει	2/22
τον	2/22
αμαθή	1/22
γιατί	2/22
υπήρξε	2/22
και	1/22
ο	1/22
ίδιος	1/22
αμαθής	2/22
δεν	2/22
γνωστικό	1/22
ποτέ	1/22

Υπολογισμός εντροπίας:

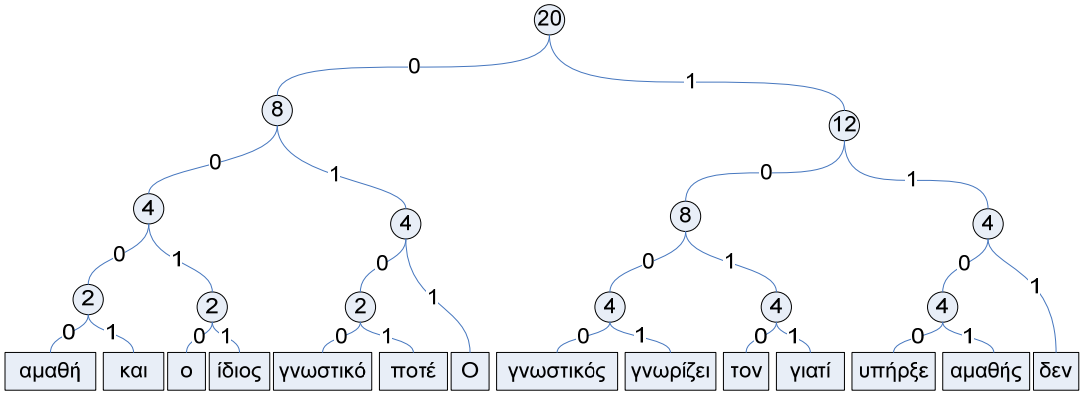
$$H(x) = \sum_{i=1}^{n} p(i) \log_2 \left(\frac{1}{p(i)}\right) = -\sum_{i=1}^{n} p(i) \log_2 p(i).$$

=3,731 bits

β) Κωδικοποίηση Huffman

Μεθοδολογία:

1. Επιλέγουμε δύο λέξεις με το χαμηλότερο αριθμό εμφανίσεων. Αν υπάρχουν περισσότερες από δύο, διαλέγουμε τυχαία.
2. Ενώνουμε τις δύο λέξεις με ένα κοινό πατρικό κόμβο, κατασκευάζοντας ένα υποδένδρο. Αναθέτουμε στον πατρικό κόμβο έναν αριθμό, ο οποίος είναι το άθροισμα των αριθμών των θυγατρικών του κόμβων.
3. Επαναλαμβάνουμε από το πρώτο βήμα έως ότου δεν υπάρχουν ασύνδετες μεταξύ τους λέξεις.

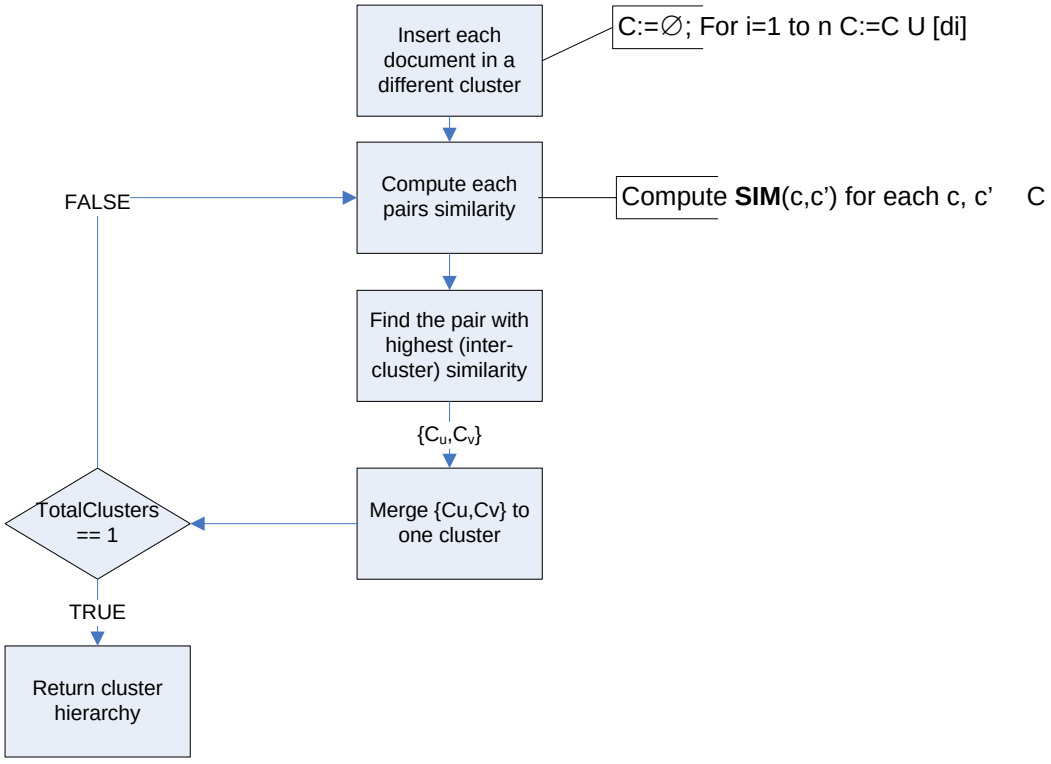


Λέξη	**Huffman code**
Ο	110
γνωστικός	0001
γνωρίζει	1001
τον	0101
αμαθή	0000
γιατί	1101
υπήρξε	0011
και	1000
ο	0100
ίδιος	1100
αμαθής	1011
δεν	111
γνωστικό	0010
ποτέ	1010

Άσκηση 2 (15 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**

Λύση



α) **SingleLink**

Βήμα 1: Βάζουμε κάθε έγγραφο σε διαφορετικό Cluster
C_A = {A}, C_B = {B}, C_C = {C}, C_D = {D}, C_E = {E}
Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

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

Βήμα 3: Βρίσκουμε το ζευγάρι με την υψηλότερη (inter-cluster) ομοιότητα.
C_A, C_E
Βήμα 4: Συγχωνεύουμε τα ζευγάρια
C_{AE} = {C_A, C_E}

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των *ζευγαριών*

AE				
B	min{6, 8}			
C	min{4.2, 5}	5		
D	min{4, 6}	2	5	
	AE	**B**	**C**	**D**

Βήμα 3: Βρίσκουμε το ζευγάρι με την υψηλότερη (inter-cluster) ομοιότητα.
C_B, C_D
Βήμα 4: Συγχωνεύουμε τα ζευγάρια
C_{BD} = {C_B, C_D}

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AE			
BD	min{4, 6}		
C	4.2	min{5, 5}	
	AE	**BD**	**C**

Βήμα 3: Βρίσκουμε το ζευγάρι με την υψηλότερη (inter-cluster) ομοιότητα.
C_{AE}, C_{BD}

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
C_{AEBD} = {C_{AE}, C_{BD}}

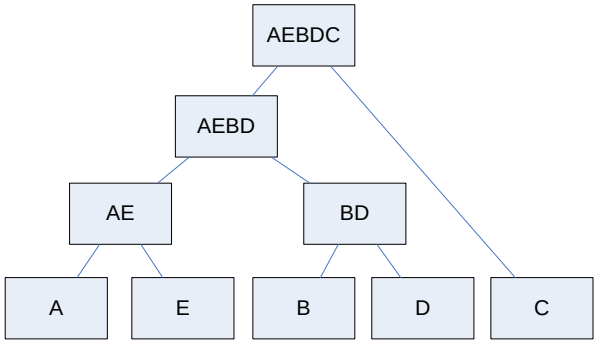
Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AEBD		
C	min{4.2, 5}	
	AEBD	**C**

Βήμα 3: Βρίσκουμε το ζευγάρι με την υψηλότερη (inter-cluster) ομοιότητα.
C_{AEBD}, C_C

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{AEBDC} = \{C_{AEBD}, C_C\}$

Βήμα 5: Έχουμε μόνο ένα Cluster



b) **<u>CompleteLink</u>**

Βήμα 1: Βάζουμε κάθε έγγραφο σε διαφορετικό Cluster
$C_A = \{A\}$, $C_B = \{B\}$, $C_C = \{C\}$, $C_D = \{D\}$, $C_E = \{E\}$

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

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

Βήμα 3: Βρίσκουμε το ζευγάρι με την μικρότερη (inter-cluster) ομοιότητα.
C_A, C_E

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{AE} = \{C_A, C_E\}$

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AE				
B	max{6, 8}			
C	max{4.2, 5}	5		
D	max{4, 6}	2	5	
	AE	**B**	**C**	**D**

Βήμα 3: Βρίσκουμε το ζευγάρι με την μικρότερη (inter-cluster) ομοιότητα.
C_B, C_D

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{BD} = \{C_B, C_D\}$

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AE			
BD	max{8, 6}		
C	5	max{5, 5}	
	AE	**BD**	**C**

Βήμα 3: Βρίσκουμε το ζευγάρι με την μικρότερη (inter-cluster) ομοιότητα.
C_{AE}, C_C

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{AEC} = \{C_{AE}, C_C\}$

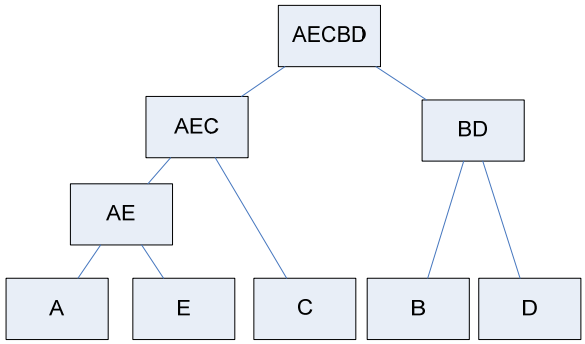
Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AEC		
BD	max{8, 5}	
	AEC	**BD**

Βήμα 3: Βρίσκουμε το ζευγάρι με την μικρότερη (inter-cluster) ομοιότητα.
C_{AEC}, C_{BD}

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{AECBD} = \{C_{AEC}, C_{BD}\}$

Βήμα 5: Έχουμε μόνο ένα Cluster



c) AverageLink

Βήμα 1: Βάζουμε κάθε έγγραφο σε διαφορετικό Cluster
C_A = {A}, C_B = {B}, C_C = {C}, C_D = {D}, C_E = {E}

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

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

Βήμα 3: Βρίσκουμε το ζευγάρι με την μέση (inter-cluster) ομοιότητα.
C_A, C_E

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
C_{AE} = {C_A, C_E}

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AE				
B	avg{6, 8}			
C	avg{4.2, 5}	5		
D	avg{4, 6}	2	5	
	AE	**B**	**C**	**D**

Βήμα 3: Βρίσκουμε το ζευγάρι με την μέση (inter-cluster) ομοιότητα.
C_B, C_D

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
C_{BD} = {C_B, C_D}

Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AE			
BD	avg{7, 5}		
C	4.6	avg{5, 5}	
	AE	**BD**	**C**

Βήμα 3: Βρίσκουμε το ζευγάρι με την μέση (inter-cluster) ομοιότητα.
C_{AE}, C_{C}

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{AEC} = \{C_{AE}, C_{C}\}$

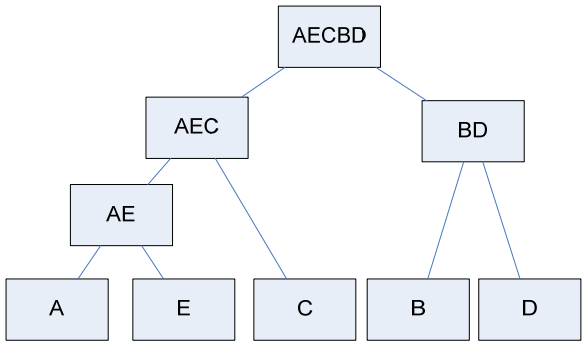
Βήμα 2: Υπολογίζουμε την ομοιότητα μεταξύ όλων των ζευγαριών

AEC		
BD	avg{6, 5}	
	AEC	**BD**

Βήμα 3: Βρίσκουμε το ζευγάρι με την μέση (inter-cluster) ομοιότητα.
C_{AEC}, C_{BD}

Βήμα 4: Συγχωνεύουμε τα ζευγάρια
$C_{AECBD} = \{C_{AEC}, C_{BD}\}$

Βήμα 5: Έχουμε μόνο ένα Cluster



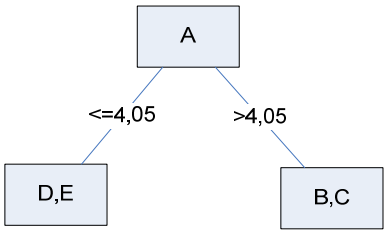
<u>Άσκηση 3</u> (10 βαθμ.)
Έστω ότι έχουμε 5 εικόνες A, B, C, D, E των οποίων οι αποστάσεις είναι αυτές που δίνονται στον πίνακα της Άσκησης 2. Προκειμένου να μπορούμε να απαντήσουμε επερωτήσεις γρήγορα θέλουμε να φτιάξουμε ένα μετρικό ευρετήριο, συγκεκριμένα ένα Vantage-Point-Tree (VTP).
Σχεδιάστε το VTP που προκύπτει:
α) αν επιλέξουμε την εικόνα A ως κεντρική (pivot),
β) αν επιλέξουμε την εικόνα C ως κεντρική (pivot).

Λύση

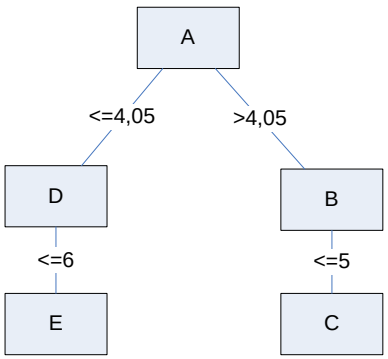
α) Επιλέγοντας την εικόνα A σαν pivot έχουμε:

$$M = \frac{d(A,B)+d(A,C)+d(A,D)+d(A,E)}{4} = \frac{6+4.2+4+2}{4} = \frac{16,2}{4} = 4.05$$

Άρα τα στοιχεία με απόσταση μικρότερη ή ίση του M εισάγονται στο αριστερό υποδένδρο, ενώ τα υπόλοιπα στο δεξί.



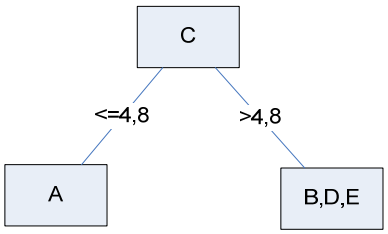
Αναδρομικά, υπολογίζουμε το μέσο όρο των αποστάσεων του αριστερού και του δεξιού υποδένδρου αντίστοιχα. Τα στοιχεία αναδιατάσσονται και το τελικό δένδρο που προκύπτει είναι:



β) Επιλέγοντας την εικόνα C ως κεντρική (pivot), υπολογίζουμε τον μέσο όρο των αποστάσεων από αυτή την εικόνα:

$$M = \frac{d(C,A)+d(C,B)+d(C,D)+d(C,E)}{4} = \frac{4.2+5+5+5}{4} = 4{,}8$$

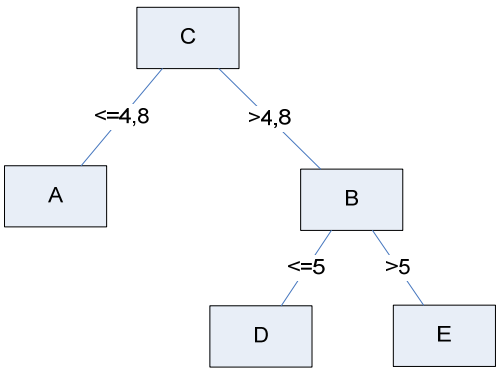
Επομένως, τα στοιχεία με απόσταση μικρότερη ή ίση του 4.8 εισάγονται στο αριστερό υποδένδρο, ενώ τα υπόλοιπα στο δεξί.



Αναδρομικά, υπολογίζουμε τον μέσο όρο των αποστάσεων μόνο του δεξιού υποδένδρου, αφού το αριστερό περιέχει μόνο ένα στοιχείο.

$$M' = \frac{d(B,D)+d(B,E)}{2} = \frac{2+8}{2} = 5$$

Τα στοιχεία του δεξιού υποδένδρου αναδιατάσσονται και το δένδρο που προκύπτει είναι,



<u>Άσκηση 4</u> (60 βαθμοί)
Θεωρείστε τα ακόλουθα έγγραφα όπου τα γράμματα A-E συμβολίζουν λέξεις.
d1 = «ΑΒ Γ », d2 = «ΒΕ Β»
d3 = «ΔΒ », d4 = «ΓΕΓ»
d5 = «ΔΓΕΓ», d6 = «ΓΕ»
d7 = «ΒΔΒ», d8 = «ΕΒ»
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 M πάνω από αυτά τα συστήματα.
(α) Για την επιλογή πηγής ο M θέλει να περιγράψει τα περιεχόμενα της κάθε πηγής με ένα διάνυσμα. Δώστε τα διανύσματα πηγών των S1, S2 και S3.

(β) Έστω ότι ο M 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 νοποιημένη 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(Α)=2, h(Β)=3, h(Γ)=6 h(Δ)=6, h(Ε)=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)

Λύση

α)
S1: d1= «ΑΒΓ », d5= «ΔΓ ΕΓ», d6= «Γ Ε»
S2: d2= «ΒΕΒ», d4= «Γ ΕΓ»
S3: d3= «ΔΒ», d7= «ΒΔΒ», d8= «ΕΒ»

term i	Fi1	Fi2	Fi3	Tfi1	Tfi2	Tfi3	idfi	Tfi1*idfi	Tfi2*idfi	Tfi3*idfi
A	1	0	0	¼	0/2	0/4	3/1	¾	0	0
B	1	2	4	¼	2/2	4/4	3/3	¼	1	1
Γ	4	2	0	4/4	2/2	0/4	3/2	3/2	3/2	0
Δ	1	0	2	¼	0/2	2/4	3/2	3/8	0	¾
E	2	2	1	2/4	2/2	1/4	3/3	1/2	1	1/4

S1=< 3/4, 1/4, 3/2, 3/8, 1/2>
S2=<0,1, 3/2,0,1>
S3=<0,1, 0, 3/4,1/4>

β)

q= <1*3,0*3,1*3/2,0*3 /2,0*3/ 2> = <3,0,1.5,0,0>
|S1| = sqrt(0.752^2+0.25^2+1.5^2+0.375^2+0.5^2)= 1.8
|S2| = sqrt(1^2+1.5^2+1^2) = 2.06
|S3| = sqrt(1^2+0.75^2+0.25^2) = 1.27
|q| = sqrt(3^2+1.5^2) = 3.35

$SimCos_1$=(0,75*3)+1,5*1,5/(1.8*3.35) = 0.74
$SimCos_2$=(1.5*1.5)/(2.06*3.35) = 0.33
SimCos3= 0

Άρα η επερώτηση θα προωθηθεί στην S1.

γ) Round robin interleaving: < {d1, d4, d7}, {d6, d2, d8}, {d5, d3} >

δ) Παραλείποντας τους λογάριθμους τα idf κάθε όρου στην καθολική συλλογή εγγράφων:

A: 8/1, B: 8/5, Γ: 8/4, Δ: 8/3, E: 8/5

ε) Παίρνουμε τις εξής απαντήσεις για την επερώτησή μας για τα συστήματα S1 S4

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

Borda:
d1 = 1 + 2 = 3
d5 = 2 + 3 = 5
d6 = 3 + 1 = 4

Άρα κορυφαίο έγγραφο είναι το d1.

Condorset:
d1:d5 = 2:0
d5 d6 = 1:1
d1:d6 = 1:1

KemenyDistance (S1,S2) = 2
(d1<S1 d6, d6<S2 d1) (d5<S1 d6, d6<S2 d5)

στ)

i)
Ο πίνακας δρομολόγησης του κόμβου S1 θα είναι:
finger[1] = succ(h(IPaddress(S1)) + 20) = succ(2) = 2 (S2)
finger[2] = succ(h(IPaddress(S1)) + 21) = succ(3) = 4 (S3)
finger[3] = succ(h(IPaddress(S1)) + 22) = succ(5) = 5 (S6)

Ο πίνακας δρομολόγησης του κόμβου S3 θα είναι:
finger[1] = succ(h(IPaddress(S3)) + 20) = succ(5) = 5 (S6)
finger[2] = succ(h(IPaddress(S3)) + 21) = succ(6) = 7 (S5)
finger[3] = succ(h(IPaddress(S3)) + 22) = succ(8) = 1 (S1)

ii) Το ανεστραμμένο αρχείο είναι:
A: <d1,1>
B: <d1,1>, <d2,2>, <d3,1>, <d7,2>, <d8,1>
Γ: <d1,1>, <d4,2>, <d5,2>, <d6,1>
Δ: <d3,1>, <d5,1>, <d7,1>
E: <d2,1>, <d4,1>, <d5,1>, <d6,1>, <d8,1>

Για κάθε όρο έχουμε:
Ο Α στον κόμβο S2, γιατί h(A) = 2 και h(IPaddress(S2)) = 2
Ο Β στον κόμβο S3, γιατί h(B) = 3 και h(IPaddress(S3)) = 4
Ο Γ στον κόμβο S5, γιατί h(Γ) = 6 και h(IPaddress(S5)) = 7
Ο Δ στον κόμβο S5, γιατί h(Δ) = 6 και h(IPaddress(S5)) = 7
Ο Ε στον κόμβο S6, γιατί h(E) = 5 και h(IPaddress(S6)) = 5

Άρα η κατανομή τους ανεστραμμένου ευρετηρίου στους κόμβους θα είναι:

S2: A: <d1,1>
S3: B: <d1,1>, < d2,2>, <d3,1>, <d7,2>, <d8,1>
S5: Γ: <d1,1>, <d4,2>, <d5,2>, <d6,1>
S5: Δ: <d3,1>, <d5,1>, <d7,1>
S6: E: <d2,1>, <d4,1>, <d5,1>, <d6,1>, <d8,1>