Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5
May 8, 2013
Due date: May 20, 2013; Monday, by noon time (12:00 O'clock) (hardcopy is required)

1. PAT tree questions.

a. Create the PAT tree for the following bit string: 10101111101000111. What is the associated PAT array?

b. Explain how to use the PAT tree/array concept to answer a query such as the following: A <max n> B.Here A and B represent two different strings and <max n> indicates the condition that between A and B there can be at the most n number of bits.

2. Consider the paper " Automatic ranking of information retrieval systems using data fusion" by Nuray and Can article (*Information Processing and Management*, 2006). Consider four different information retrieval systems (A, B, C, D) ranking documents a,… f. Perform data fusion by using the reciprocal rank, Borda count, and Condorcet methods. Please show your steps.

A= (c, b, a, d)
B= (b, c, a, e)
C= (c, a, b, f)
D= (a, b = c) // b and c are assigned the same rank!

3. Consider the following signatures.

S1: 1100 1100
S2: 1100 0011
S3: 0011 1100
S4: 0000 1111
S5: 1011 0100
S6: 0100 1011

a. Use the fixed prefix method to partition the above signatures. Take k (key length) as 2.

b. Now consider the following queries.

Q1: 1110 0001
Q2: 0110 0011
Q3: 1100 1100
Q4: 0011 1100

Use the partitions of section-a to calculate the time needed (turnaround time) to process the queries in sequential and parallel environments. (Use the assumptions that we used in the class room, e.g., the processing of one page signature requires 1 time unit, etc.). What is the speed up ratio for the parallel environment?

4. Partition the signatures of question 3 using the following partitioning methods. (To answer this question consider the paper Lee and Leng 1989 paper in ACM TOIS, "Partitioned Signature Files: Design Issues and Performance Evaluation," or "Signature Files: An Integrated Access Method for Formatted and Unformatted Databases" by Aktug & Can. The second one is available on the Web site.

a. EPP (take z= 2).

b. FKP (take k= 2).

c. To process the following queries which pages need to be accessed and why?

Q1: 1110 0001
Q2: 0110 0011
Q3: 1100 1100
Q4: 0011 1100

5. Consider the signatures of question 3 show the expansion of a linear hashing file (LHSS file) as we insert these signatures one by one to the file. Assume that each page can store 3 signatures and also assume that the desired load factor we would like to maintain is 0.66. Begin with (hashing level) h= 1, (boundary value) bv=0.

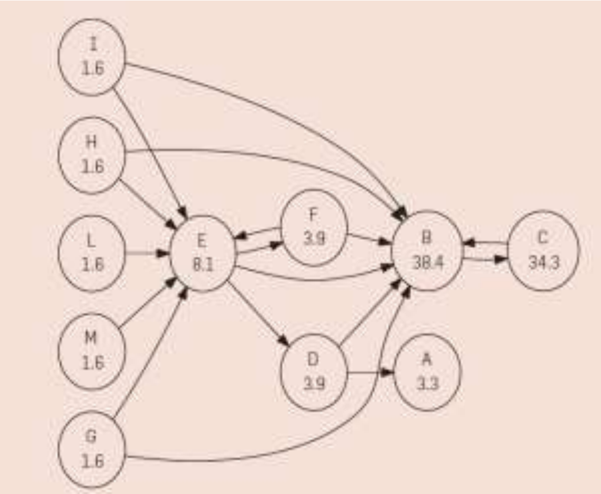
6. In an LHSS file assume that we have 12 pages (n= 12). For this file calculate h and bv values and also give the number of pages at hashing level h and h+1. Do the same calculation for n= 120.

7. Consider the following information filtering profiles used in a Boolean environment.
P1= a, b, c, d, e, f
P2= a, b
P3= b, c, e, f
P4= b, d
P5= a, c, e

Assume that when the terms are sorted in frequency order according to their number of occurrences in documents term a is the least frequently used term in the documents and is also the most frequently used term in the user profiles. The sorted term list continues as b, c … f.

Consider the ranked key method explained in the paper by Yan and Garcia-Molina (Index structures for selective dissemination of information under the Boolean model, *ACM TODS*) and draw the directory and the posting lists for the ranked key method and the tree method.

8. Consider the social network given below. In this network each node indicates the PageRank of that node. The PageRank of page *j* is the sum of the PageRank scores of pages *i* linking to *j*, weighted by the probability of going from *i* to *j*. Using this definition and also by following the additional explanation provided in the source paper of the figure (please see the figure subtitle) calculate the PageRank value of the nodes F and D.



Each node is labeled with its PageRank score. Scores have been normalized to sum to 100. We assumed α = 0.85. (Source: M. Franceschet, PageRank: Standing on the Shoulders of Giants. *Comm. of the ACM*, 54(6): 92-101, 2011.