Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5
December 21, 2018
Due date/time: January 2, 2019; Wednesday, Final exam time

Notes: A word processor generated submission is required, handwritten submissions are unacceptable. Please bring your hard copy submission to the exam room. If you see something missing in a question please make reasonable assumptions and explain.

1. Consider the following signatures.
S1: 1000 1001
S2: 1010 0010
S3: 1100 1100
S4: 0000 1111
S5: 0111 0100
S6: 0101 1010
S7: 1100 0101
S8: 1000 1110
S9: 0011 0111

a. Use fixed prefix method to partition the above signatures. Let key length k equal to 3.

b. Consider the following queries.
Q1: 1101 0001
Q2: 0110 0011
Q3: 1100 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 (defined as ratio (parallel processing time for all queries / sequential processing time for all queries)?

2. Consider the signatures of question 1.

a. Use Extendible Hashing method to partition the signatures. Take block size as 2. Show intermediate steps as you insert the signatures.

b. For the following query which pages do we need to access?
Q: 1001 1001

3. Consider the signatures of question 1.

a. Use LHSS (Linear Hashing with Superimposed Signatures) method to partition the signatures. Take block size as 3 and LF to be maintained -desired load factor level- as 2/3. Show intermediate steps as you insert the signatures.

b. For the following query which pages do we need to access?
Q: 1101 0001

4. Prove that in LHSS when you come to the desired load factor level and then after adding (Bkfr x LF) number of records, and after that when you update the file structure again LF goes back to the desired level.

5. Partition the signatures of question 1 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.

a. EPP (take z= 2).
b. FKP (take k= 2).
c. To process the following queries which pages need to be accessed? Answer separately for EPP and FKP.
Q1: 1110 0001
Q2: 0110 0011
Q3: 1100 1100
Q4: 0011 1100
d. What is the intuition behind EPP and FKP methods? Explain briefly.

6. Consider the following information filtering profiles used in a Boolean environment.

P1= a, b, c, f
P2= a, d
P3= b, c, f
P4= c, d
P5= a, c, d

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.

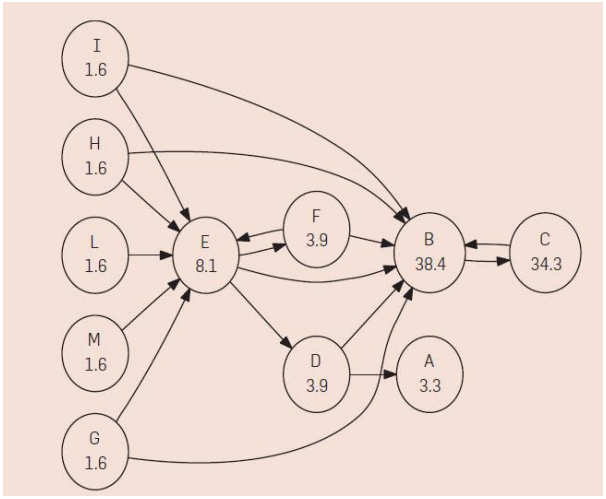
a. Consider the ranked key method explained in the paper by Yan and Garcia-Molina (Index structures for selective dissemination of information under the Boolean model) and draw the directory and the posting lists for the ranked key method.

b. What is the intuition behind the ranked key method: how does it improve the filtering efficiency?

c. Suggest an algorithm to automatically update user profiles in information filtering for better better performance? Why would you expect that you algorithm would improve performance.

7. Please read the paper: M. Franceschet, PageRank: Standing on the Shoulders of Giants. *Comm. of the ACM*, 54(6): 92-101, 2011.

a. Consider the social network, from the paper, 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 paper calculate the PageRank value of the nodes E and F. Each node is labeled with its PageRank score. Scores have been normalized to sum to 100. We assumed $\alpha = 0.85$.



8. In the Franceschet pagerank paper the author gives pagerank-like examples in other fields. Identify these fields and give an example of yours, other than the ones provided in the paper, for at least three fields.