Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 3 (Progressive: I may add a few more question, just may)
May 7, 2015
Due date: May 18, 2015; Monday, by noon time (12:00 O'clock) (hardcopy is required)

Final Exam Date & Place: May 21, Thursday, 12:30-14:30 pm; EB104. I plan to send you a study guide for the final exam in the middle of next week.

Notes: Handwritten answers are not acceptable. You have the option of solving any four of the questions.

1. Consider a document collection containing 60,000 objects. The signature of an object requires 512 bits. What are the signature file sizes using the following signature file organization methods?
a. Sequential Signatures (SS),
b. Bit-sliced Signatures (BS).

2. In the database environment of question 1 consider a query with 5 bit positions equal to one. These bit positions are 1, 2, 50, 51, 60. (The leftmost position of a signature is bit position 1.) For filtering (i.e., for query signature - document signatures matching) how many pages need to be accessed in the case of SS and BS? (Page size is 0.5 K bytes.) Note that in SS we place signatures one after the other and in the case of BS we place bit slices one after the other: Place the first bit slice and then right after that place the second bit slice and if there is room in the page allocated to slice 1 use the remaining space for the second bit slice and carry on like this.

3. Consider the following signatures.
S1: 1100 0110
S2: 1010 0011
S3: 1100 0011
S4: 0000 1111
S5: 1010 0110
S6: 1011 0100
S7: 1100 1010
a. Use the fixed prefix method to partition the above signatures. Take k (key length) as 2. Show the file structure (contents of the pages etc.).

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 average 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.
a. EPP (take z= 2).
b. FKP (take k= 2).
c. To process Q1 of question 3 which pages are need to be accessed with the EPP and FKP methods and why?

5. Partition the signatures of question 3 this time by using the extendible hashing algorithm (using prefixes). Assume that each data block can contain two signatures.

For Q1 and Q2 of question 3 please specify which pages need to be access and please explain briefly.

6. Partition the signatures of question 3 this time by using the linear hashing algorithm (using suffixes). Assume that each data block can contain two signatures. (Bkfr= 2) and LF that we want to maintain is 1/2.

For Q1 and Q2 of question 3 please specify which pages need to be access and please explain briefly.

For these queries again what is the recall and precision rates before false drop resolution? Does false drop resolution change the recall rate? Why?

7. Consider a Bloom filter with a bit array size (m) of 10 bits, number of hash functions (k) of 2, number of elements (n) of 3. Calculate the false drop probability. If the cost of false drop resolution is 1 ms what is the cost of processing 10,000 look ups. Assume that if there is a hit (true hit) query time is 0.5 ms.
Reference: Bloom, Burton H. (1970), "Space/Time Trade-offs in Hash Coding with Allowable Errors", *Com. of the ACM* 13 (7): 422–426,

8. Consider the paper " Automatic ranking of information retrieval systems using data fusion" by Nuray and Can (*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 with enough detail so that it can be followed.

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

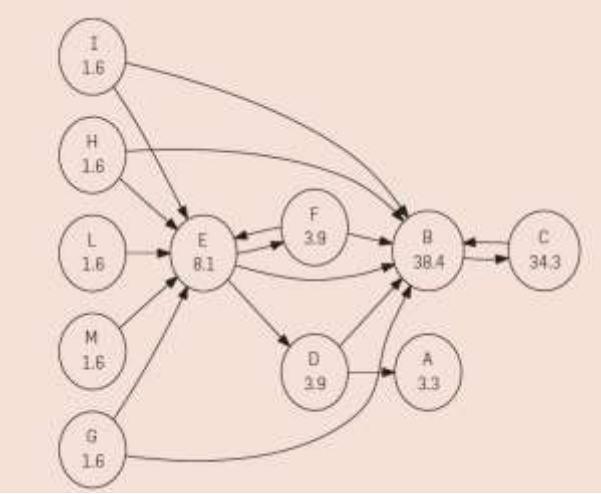
9. Consider the following information filtering profiles used in a Boolean environment.

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

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.

10. 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 A.



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.

11. Consider the following document collection containing four documents (rows) defined by four terms (columns).

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

A query is submitted and its vector is defined as follows: Q= [0 1 0 0]

Assume that we want to use the MMR algorithm for selecting the best matching first two documents. After each case what is the cohesiveness (similarity) and diversity among the selected documents and how can we measure it? Does the MMR algorithm provide what it promises. For each case please show your steps explicitly. <u>For similarity calculations use the Dice coefficient.</u>

a. Use λ = 1.00 and indicate the selected documents.

b. Use λ = 0.00 and indicate the selected documents. What is the diversity among the selected documents and how can we measure it? Do we have more diversity with respect to the first case above? Please explain.

c. Use λ = 0.50 and indicate the selected documents. What is the diversity among the selected documents and how can we measure it? Do we have more diversity with respect to the first case above? Please explain.