Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 6
May 18, 2012
Due date: May 25, 2012; Friday, by noon time (12:00 O'clock) (hardcopy is required)

Final Exam Date & Place: May 28, Monday, 3:40-5:30 pm; EE517. I plan to send you a study guide for the final exam in the middle of next week.

Notes: Handwritten answers are not acceptable.

1. Consider a database containing 40,000 objects. The signature of an object requires 256 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 6 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 BS we place bit slices one after the other.

3. Consider the following signatures.
S1: 1100 0110
S2: 1010 0011
S3: 1100 0011
S4: 0000 1111
S5: 1011 0100
S6: 1011 0100
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 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 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 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.

6. Obtain the Huffman coding for the following vocabulary A: 0.33, B: 0.20, C: 0.15, D: 0.12. E: 0.08, F: 0.08, G: 0.04. After each number the probability of occurrence of the letter is provided. Please show your work.

7. Show the delta (δ) and gamma (γ) code for the following integers: 7, 17, 23, 50.

We have Huffman coding which provides optimal compression: why do we need delta (δ)and gamma (γ) compression?

Compare the advantages/disadvantages of delta coding and gamma coding with respect to each other. (The following article may be helpful: Justin Zobel, Alistair Moffat, Ron Sacks-Davis: Searching Large Lexicons for Partially Specified Terms using Compressed Inverted Files. VLDB 1993: 290-301