Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5
May 10, 2011
Due date: May 19, 2011; Thursday, by noon time (12:00 O'clock) (hardcopy is required)

Notes: Handwritten answers are acceptable (Word document will be appreciated). Provide your answers on standard sized paper and use only one side of each paper. Answers must be given in the same order as the questions. Please number answers properly for easy identification. Staple all papers on the left upper corner and write your name on the first page. No late assignment will be accepted. Answering all of the questions is optional; however, you have to answer at least four questions.

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 concept to answer a query such as the following: A < n> B.

Here A and B represent two different strings and for example < 10> indicates the condition that between A and B there must be 10 bits.

c. Explain how to use the PAT tree concept to answer a query such as the following: A < min 10> B.

Here A and B represent two different strings and < min 10> indicates that between A and B there must be at least 10 bits.

Possibly helpful link:
http://webcache.googleusercontent.com/search?q=cache:http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap05.htm

2. Consider a database containing 10,000 objects. The signature of an object requires 512 bits. Calculate the signature file sizes using the following signature file organization methods?
a. Sequential Signatures (SS),
b. Bit-sliced Signatures (BS).

3. In the database environment of question 2 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 are needed to be accessed in the case of SS and BS? Page size is 0.5 K bytes. Note that in SS files signatures are placed in the pages one after the other, in BS files the bit slices are placed in the pages one after the other in consecutive order: the second slice follows the first slice etc.

4. Consider the following signatures.
S1: 1000 1001
S2: 1100 0010
S3: 0011 1100
S4: 0000 1111
S5: 1011 0100
S6: 0100 1010
S7: 1100 0101

a. Use the fixed suffix method to partition the above signatures. Take k (key length) as 2. We didn't study this method in class but it is easy to imagine how it works (it is a version of the fixed prefix method but this time it is based on suffixes).

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?

5. Partition the signatures of question 4 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 our course Web site. If a password is needed use myirnotes

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

6. In a linear hashing environment assume that h is 12. For this case specify the minimum and maximum possible values for bv (boundary value variable), number of pages at hashing level h, number of pages at hashing level h+1? Are these values independent of the desired load factor that we want to maintain?

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

P1= a, b, c, d, f
P2= a, f
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) and draw the directory and the posting lists for the ranked key method.

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