Computer Engineering Department
Bilkent University

CS 533: Information Retrieval Systems
Assignment No. 5
May 13, 2008
Due dates: May 27, 2008, Tuesday, 5:00 pm
Final Exam Date & Place: May 29, Thursday, 5:40-7:30 pm; EA502

Notes: In Section A 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. 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. In Section B: For your answers please use a word processor.

Section A

1. Consider the following D matrix. (Use File/Print Preview to see the matrix.)

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

What are the term discrimination values of the terms of the D matrix. Find the values using the

a. cover coefficient concept,

b. similarity concept where similarity is obtained using the Dice coefficient (use no approximation, i.e., do not use collection centroid).

2. What is the symbol tree for the following set of Turkish words: "bilmek, bilgiç, bilgin, bilgisayar, bilim, bilinç, bilişim, bilişsel"?

3. Consider the Turkish word corpus given in the above question. What is the stem of the word "bilgin" using the successor variety method?

4. For a word of m number of characters how many 2-grams and 3-grams can we have? In each case you should give a formula in terms of m. Assume that m > 3.

5. PAT tree questions.

a. Create the PAT tree for the following bit string: 011010100010111000. What is the associated PAT array?

b. Explain how to use the PAT tree concept to answer a query such as the following: A <max 20> B. Here A and B represent two different words and <max 20> indicates the condition that between A and B there can be at the most 20 characters.

6. Consider a database containing 10,240 objects. The signature of an object requires 128 bits. What are signature file sizes using the following signature file organization methods?

a. Sequential Signatures (SS),

b. Bit-sliced Signatures (BS).

7. 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.)

8. Consider the following signatures.
S1: 0110 1100
S2: 1010 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: 0001 1110
Q2: 1001 1100
Q3: 0011 0011
Q4: 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?

9. Partition the signatures of question 9 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.

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: 0000 1111
Q2: 1111 0000
Q3: 1001 1010
Q4: 1110 0001

10. Consider the following data structure for Ranked Key Method: (in the Word document version of this document please use "print preview" for proper display of the following figure). For this question please consider the paper Index structures for selective dissemination of information under the Boolean model, by T. W. Yan, T. W., H. Garcia-Molina (the course Web site has the link to acm.org/dl).

Directory	
a	
b	/
c	/
d	
e	
f	/
g	/

→ [P1, 2,e, g][P2, 1,c][P3, 2,b, d]

→ [P4, 2,e, f]
→ [P5, 1, f]

a. Show the elements for each profile,
b. A document satisfies P1 and P5 and contains 5 terms. Determine the document terms.
c. For the same profiles please show the tree organization for the ranked key method.
d. For a document that contains the terms a, b, c how many sub trees do we need to search in part c?

11. Consider the use of run length encoding for posting lists. Consider two extreme cases for the posting lists. The posting for the infrequent terms and posting lists for the frequent terms. For what type of posting lists we would have a better compression ratio (i.e., relative memory savings with respect to no compression)? Explain your answer.

12. Obtain the Huffman coding for the following vocabulary A: 0.35, B: 0.18, C: 0.15, D: 0.12. E: 0.10, F: 0.10. After each number the probability of occurrence of the letter is provided. Please show your work.

13. Show the delta (δ) and gamma (γ) code for the following integers: 5, 15, 23, 50.

We have Huffman coding which provides optimal compression ration. Then why do we need delta (δ)and gamma (γ) coding? There is a common sense answer to this question without searching the Web?

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

Section B

1. Read the paper "The anatomy of a large-scale hypertextual Web search engine" by Brin and Page (available on the web).

a. Describe their comments regarding number of Web pages and how the numbers changes(scaling up). Is their prediction correct? (Visit **http://searchenginewatch.com/** for new numbers.)
b. Define their method of page rank calculation and try to provide a simple example of your own.
c. Define their search engine in terms of its data structures. Try to make your presentation more intuitive by providing a figure.
d. How do they avoid disk seeks?

2. Consider the paper "Information retrieval on Turkish texts." by Can, Kocberber, Balcik, Kaynak, Ocalan., Vursavas, *Journal of the American Society for Information Science and Technology*. 59(3): 407-421 (**http://www.users.muohio.edu/canf/papers/JASIST2008offPrint.pdf**) Briefly explain the stemmers defined for Turkish in this paper, provide a brief description for them. How did the authors adapted the sucessor variety method to Turkish? Explain briefly.

3. Consider the paper "Another look at automatic text-retrieval systems" by Salton, *Comm. of the ACM*, 29(7): 648-656, 1986. Briefly explain the thesaurus and phrase transformations explained in the paper. Explain their purpose. State if you agree with them, please explain your answer briefly.