Computer Engineering Department
Bilkent University

CS 533: Information Retrieval Systems
Assignment No. 6
May 7, 2010
Due dates: May 21, 2010, Friday, 5:00 pm
Final Exam Date & Place: May 26, Wednesday, 10:40-11:30 pm; EA502

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. 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.

Minimum requirement: if your Bilkent ID is an even number solve at least five even numbered questions, otherwise solve at least five odd numbered questions.

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 (for calculating no. of clusters use the approximate formula implied by the indexing-clustering relationships: (m x n) / t).

b. similarity concept where similarity is obtained using the Dice coefficient (use the approximate method based on the collection centroid approach).

2. Consider the following Turkish word corpus: “bilmek, bilgiç, bilgin, bilgisayar, bilim, bilinç, bilişim, bilişsel”. What is the stem of the word “bilgin” using the successor variety method?

3. PAT tree questions.

a. Create the PAT tree for the following bit string: 011011100010111000. What is the associated PAT array? (Create the tree for the first 12 sistrings.)

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.

4. Consider a database containing 10,240 objects (e.g., documents). The signature of an object requires 192 bits. What are signature file sizes using the following signature file organization methods?

a. Sequential Signatures (SS),

b. Bit-sliced Signatures (BS).

5. In the database environment of question 4 consider a query with 5 bit positions equal to one. These bit positions are 1, 2, 50, 51, 140. (The leftmost position of a signature is bit position 1.) For filtering (i.e., for query signature - document signatures matching) how many pages (disk blocks) need to be accessed in the case of SSF (sequential signature file) and BSF (bit sliced signature file)? Use the following assumption: in SSF the document signatures are placed in the pages one after the other (at the beginning we have the signature of the first document and immediately after that we have the document signature of the second document); in BSF a similar signature placement technique is used for the bit slices (first bit slice is immediately followed by the second bit slice, etc.).

Note that signature of a document (or a bit slice) may cross the page boundaries. (Page size is 0.25 K bytes.)

6. Consider a bit sliced signature environment for 2^{30} documents. In this environment F (signature size) is 16000 bits and the document signature bit density is $(1/2)^4$, i.e., only 6.25% of document signature bits are equal to 1.

a. How many documents we need to access for false drop resolution after processing 1 bit of the query signature? Answer the same question after processing 2, 3, and 4 query signature bits.

b. In this environment does it make sense not to process all query bit positions equal to 1 and switch to false drop resolution? Please explain your answer. (Hint: The paper Kocberber, S., Can, F. "Partial evaluation of queries for bit-sliced signature files." Information Processing Letters. Vol. 60 (1996), pp. 305-311 provides further information about this. Remembering our class discussion is enough to answer this question.)

7. 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
For each query indicate the partitions selected and also give the PAR (partition activation ratio: no. of partitions activated / total no. of partitions) and SAR (signature activation ratio: no. of signatures activated /total no. of signatures).

8. Partition the signatures of question 7 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

9. 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?

10. 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.

11. 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. (This question is related to our term discrimination value discussion in the classroom.)