Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 2
March 16, 2008
Due date: March 24, Monday, 23:59 (Electronic submission); March 25, Tuesday, 11:59 (Hardcopy)

Notes:

- In this assignment, there are two sections.
- Section A contains regular questions based on our class discussions. The following paper is useful in this section: Can, F., Ozkarahan, E. A. "Concepts and effectiveness of the cover coefficient based clustering methodology for text databases." *ACM Transactions on Database Systems*. Vol. 15, No. 4 1990, pp. 483-517. See the appendices for some useful formulas (on p. 3). To verify your results you may use the "cluster.exe" program available on our course Web site.
- Section B is based on the paper by A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review" *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.
- Please send your work by email by the specified deadline given above. Your email subject must be "CS533, HW2 Submitted by" followed by your name. You must also submit a hardcopy on March 25, by noon time.

QUESTIONS

Section A

1. Consider the following D matrix.

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

Obtain the corresponding single-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at the similarity level 0.45 (note that you will obtain a partitioning structure). For similarity calculation use the Dice coefficient.

2. Consider the D matrix of question no. 1. Obtain the corresponding complete-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at the similarity level 0.45. For similarity calculation use the Dice coefficient.

3. Consider the the D matrix given in question 1.

a. Construct the corresponding C matrix (can be obtained either by matrix multiplication or the related formula), you may just give the C matrix.
b. Calculate the number of clusters.
c. Find the seed power of all documents.
d. Determine the cluster seeds. Explain your reasoning.
e. Construct IISD (Inverted Index for Seed Documents).
f. Use the IISD data structure to cluster d_5. Show your computations explicitly.
g. Construct the clusters.
h. In an efficient implementation of the C^3M how many entries of the C matrix do we have to calculate? Answer this question (1) in general using the symbols such as m, n, n_c, etc.; and (2) for the D matrix of this question.

4. Questions based on the clustering-indexing relationships implied by the cover-coefficient concept (for possible questions please refer to the related paper).

a. For the D matrix of question number 4, calculate the number of clusters by using the clustering-indexing relationships implied by the cover-coefficient concept.
Answer the following in a general context.

b. According to these relationships what happens to the number of clusters as depth of indexing (average number of terms per document) changes from its possible minimum value to its possible maximum value. Show it in a graph (n_c versus x_d).

c. According to these relationships what happens to the number of clusters as term specificity –also called "term generality"- (average number of documents per term) changes from its possible minimum value to its possible maximum value. Show it in a graph (n_c versus $t_{g)}$.

5. How can we use the concepts of C^3M for cluster maintenance?
Hint: Refer to Can, F. Incremental Clustering for Dynamic Information Processing, *ACM Trans. on Information Systems*. Vol. 11, No. 2 (April 1993), pp. 143-164. A short paragraph is enough.

6. Consider the following specifications for a document database:

m (No. of documents)	= 200
n_C (No. of clusters)	= 10
k (No. of relevant documents for a given query)	= 5

Assume that (a) documents are randomly distributed among the clusters; (b) each cluster has the same size. What is the expected number of clusters to be accessed to retrieve all relevant documents of the query? (Use Yao's formula, see the related paper: Yao, S. B., "Approximating block accesses in database organizations." *Communication of the ACM*, Vol. 20, No. 4, 1977, pp. 260-261.

7. Obtain the similarity matrix implied by the dendrogram of question number 1. Calculate the "product moment correlation coefficient" (see Appendix B below) between the corresponding elements of the implied similarity matrix and the original similarity matrix obtained by using the given D matrix.

Section B In this part consider the paper by A. K. Jain, M. N. Murty, P. J. Flynn.(Has already been covered, this part is cancelled) March 20, 2008.

APPENDIX

A. The definitions of c_{ij} and c_{ij}' are as follows.

$$c_{ij} = \alpha_i . \sum_{k=1}^{n} \left(d_{ik} . \beta_k . d_{jk} \right)$$

$$c_{ij}^{'} = \beta_i . \sum_{k=1}^{m} \left(d_{ki} . \alpha_k . d_{kj} \right)$$

B. The product moment correlation between X and Y is defined as follows.

$$r = r(X,Y) = \frac{\mathrm{cov}(X,Y)}{\left[\mathrm{var}(X).\mathrm{var}(Y)\right]^{\frac{1}{2}}} = \frac{\sum (x_i - x_{avg})(y_i - y_{avg})}{\left[\left[\sum (x_i - x_{avg})^2\right]\left[\sum (y_i - y_{avg})^2\right]\right]^{\frac{1}{2}}}$$