Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 3
March 18, 2009
Due date: March 24, 2009; Tuesday, 11:59 am

Please submit at least solutions to the three of the questions/items given below. A word processor output is expected.

1. Clustering concepts.

a. Give the definition of the cluster hypothesis,

b. Give the definition of he number of clusters hypothesis. (We briefly considered this in connection with the indexing-clustering relationships implied by the cover coefficient concept.)

c. In his *Information Retrieval* book van Rijsbergen mentions implications of clustering algorithms. Please briefly define three of them.

2. Define your own binary D matrix of size m= 7, n= 6. Please try to generate a D matrix that would have a value if used in an exam or an IR textbook.

a. Obtain the corresponding single-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at two different similarity levels (note that for each you will obtain partitioning clustering structures). For similarity calculation please use the Dice coefficient. (If you use the cluster.exe program please make sure that you use the Dice coefficient, as you would remember there is a confusion of similarity measures in that program.)

b. Obtain the corresponding complete-link clustering structure (dendrogram). Give the clustering structure if the dendrogram is cut at the the same similarity values. For similarity calculation please again use the Dice coefficient.

c. Obtain the similarity matrix implied by the dendrogram of Section a. For 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 your D matrix. Please show your steps, but do not exaggerate in terms of details.

d. Please repeat the Section c but this time by using the complete-link structure.

3. Consider the D matrix that you defined in question no. 2.

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 one of the documents. 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.

i. Repeat the steps a-h but this time using a weighted D matrix of your own (again with the same m and n values). Note that for a weighted D matrix the seed power has a slightly different definition, please see the paper.

4. Prove that according to the cover coefficient concept the number of cluster implied by documents (n_c) and terms (n_c') are equal to each other.

5. Indexing-clustering relationships.

Apply the clustering-indexing relationships formulas to the D matrix of question no. 2 to estimate number of clusters and average cluster sizes (in terms of number of members) for document and term clusters.

How can we use the clustering-indexing relationships implied by the cover coefficient concept for practical purposes? Describe its possible uses. Please state more than one purpose and be creative, I am not looking for specific answers to this question.

6. 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.

7. In this question please refer to the paper Jain, A. K., M., Murty, N., Flynn, P. J. Data Clustering: A Review. *ACM Computing Surveys*. (31(3): 264-323 (1999).

a. ~~According to the authors what are the steps of clustering?~~

b. What is the importance of parallel implementations of clustering algorithms?

c. ~~How can we use clustering for data mining?~~

8. Consider the following specifications for a document database:

m (No. of documents)	= 400
n_C (No. of clusters)	= 20
k (No. of relevant documents for a given query)	= 4

a. Assume that (1) documents are randomly distributed among the clusters; (2) 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.

9. For the database specifications given above this time assume that we have 10 clusters that contain 10 documents and the rest of the documents are uniformly distributed among the remaining 10 clusters. Perform the same computation but this time use the modified Yao's formula as presented in Can & Ozkarahan, ACM TODS, Dec. 1990.

10.. Briefly describe how to use the Yao's formula in cluster validation. You are free to suggest a method different from the one we discussed in the classroom.

Suggest a cluster validation method that can be used in a partitioning clustering environment designed to be used in an IR application.

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.

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