Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 3
March 9, 2012
Due date: ~~March 19 ,2012; Monday~~, March 26, 2012; Monday by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable.

Section A

1. Consider the following D matrix.

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

Obtain the corresponding single-link clustering structure i.e., the dendrogram for this case (another common spelling for the same structure is dendogram). For similarity calculation use the overlap coefficient (for the definition of the overlap coefficient you may refer to van Rijsbergen's book). Use similarity values for cluster construction.

2. Consider the D matrix of question no. 1. Obtain the corresponding complete-link clustering structure (dendrogram). For similarity calculation use the overlap coefficient. Use similarity values for cluster construction.

For which cut similarity threshold value the dendrogram can be used to obtain two clusters? If not possible explain why?

3. A clustering algorithm is referred to as order dependent if it can produce more than one clustering structure. Is the complete link algorithm order dependent when we use the similarity matrix that you obtain for the above D matrix? If not so modify the S matrix by hand (or use another handmade S matrix) to show that the complete link algorithm may generate different clustering structure based on your decision to connect the nodes.

4. Show how do you obtain the document by document similarity matrix by using the method that uses the term inverted indexes. It is the most efficient method we discussed in the classroom.

5. Consider the D matrix given in question 1. In this question consider the cover coefficient-based clustering methodology C^3M. " (For easy reference Appendix A below provides the cover coefficient formulas.)

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.

6. How can we use the concepts of C^3M for cluster maintenance?
To answer this question read the paper "Incremental clustering for dynamic information processing" from *ACM Trans. on Information Systems* (1993).

7. Consider the following specifications for a document database:

m (No. of documents)	= 150
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 (using Yao's formula)?

8. 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.

9. van Rijsbergen has three comments about cluster file structures in his *Information Retrieval* book. Do they still make sense in present day computer technologies? Please state your own arguments.

Section B

In this part consider the classic paper of Jain, Murty, Flynn, "Data clustering: A review," *ACM Computing Surveys*, 1999.

1. What is the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification)? What kind of classification would you have in Information Filtering and why?

2. What are the components of a clustering task? Explain each step within the framework of an information retrieval environment.

3. What is the purpose of "cluster tendency" analysis?

4. What is k-Means clustering algorithm? How does it work? Did we study this algorithm in the classroom by using a different name?

5. What are the applications areas of clustering? List all of them. Explain two areas other than IR with one or two paragraphs. (That is explain how clustering is being used in these areas.)

Section C

In this part consider the paper by Carpineto, Osinski, Romano, and Weiss, "A survey of web clustering engines," *ACM Computing Surveys*, 2009.

1. What can be gained by search result clustering?

2. Try to find two search engines with search result clustering functionalities (Table III may help for this purpose). Experiment with these search engines by using at least seven queries and explain their advantages with respect to each other. Please specify your queries and provide some screen shots to make your points. A short report to make your points is sufficient.

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.

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