Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 3
March 14, 2013
Due date: March 28, 2013; Thursday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable. The next assignment may overlap with this on

1. What is cluster hypothesis? Does it make sense? Please briefly explain why.

2. What is the distinction between clustering algorithm and clustering method? Please refer to *Classification* chapter in vanRijsbergen book *Information Retrieval* (available on the web).

3. Consider the similarity measures such as cosine, Dice, Euclidian distance. Do you see any difficulty with the use of these similarity measures for "information retrieval"? Do some research on this.

4. Please consider the following document by term D matrix.

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

a. Obtain the corresponding single-link clustering structure i.e., the dendrogram (commonly used misspelling dendogram). 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.

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

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

d. How can we use the simulation experiment (i.e., Monte Carlo experiment, Monte Carlo analysis) to show that the correlation coefficient value we obtain is significant or not. (For Monte Carlo experiment: see class notes or the Jain & Dubes book *Algorithms for Clustering Data*).

5. Prove that complete link is order dependent, i.e., it is possible to obtain different dendrograms depending on our decisions as we obtain links among existing clusters.

What can we say about single-link, is it order dependent or not? Why.

6. Consider the D matrix given above.

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. **Bonus (optional)**: Repeat the steps a-h but this time using a weighted D matrix of your own (again 5 x 6). Note that for a weighted D matrix the seed power has a slightly different definition, please see the paper.

7. 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.

8. Indexing-clustering relationships.

Apply the clustering-indexing relationships formulas to the D given above to estimate the 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? Try to be creative and realistic.

9. 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.

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

b. What is data mining? What is the relationship between data mining and clustering.

c. This paper is one of the most frequently cited papers in computer science. How can we explain this?

d. Identify another computer science paper which is also cited very frequently and explain the reason why it is frequently cited.

11. K-means

a. Define k-means briefly.

b. Consider a very large collection. How can we use Apache Hadoop for the implemenatation of k-means in such an environment? Try to design a step by step scenario. A brief explanation is enough you do not need to mention the details. I am not an expert. If you create an implementation please send it in a zipped file to me with the necessary readme file.

12. 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)	= 5

a. Assume that each cluster has the same size. Assuming that we have random distribution of documents among clusters what is the expected number of clusters to be accessed to retrieve all relevant documents of a query with 5 (k) number of relevant documents? (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.

b. Is there any other paper available in the literature that we can use for the same purpose?

.

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.

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