**Computer Engineering Department**
**Bilkent University**

CS533: **Information Retrieval Systems**
Assignment No. 3
March 7, 2011
Due date: March 14, 2011; Monday, by 5:00 pm

**Notes**: Normally handwritten answers are not acceptable, but, this time I will make an exception and will accept handwritten answers, a word processor output will be appreciated. Answer the questions in the order given here.

**1**. Consider the following document by term binary D matrix for m= 6 documents (rows), n= 6 terms (columns).

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

Obtain the similarity matrix S for the above D matrix (you don't need to show your intermediate steps). Use the Dice similarity coefficient. Use the S matrix to construct the dendrogram (cluster tree) structure corresponding to the single-link and complete link clustering methodologies.

**2**. Consider the above D matrix. Cluster the documents using the cover coefficient-based clustering methodology ($C^3M$). Please a) Show the double-stage probability experiment tree for the second document, and show the calculation of $c_{24}$ of the corresponding C matrix, b) obtain the C matrix (you do not need to show the intermediate steps), c) find the number of clusters implied by the C matrix – explain how-, d) find the cluster seeds, e) obtain the IISD (inverted index for seed documents), f) obtain the clusters and explain how you find them.

**3**. According to the cover coefficient concept if $c_{ij} > 0$ then $c_{ji} > 0$. Prove this property.

**4**. Consider the incremental version of $C^3M$: $C^2ICM$, Cover Coefficient-based Incremental Clustering Methodology, described in Can F, Incremental clustering for dynamic information processing, ACM TOIS, 1993).

**a.** Briefly explain the algorithm (one paragraph).

**b.** In the paper there is the concept of clustering similarity, explain its purpose within the context of $C^2ICM$.

**5.** The paper mentions a measure called Rand coefficient (and cites the classic book of Jain & Dubes: Algorithms For Clustering Data, http://www.cse.msu.edu/~jain/Clustering_Jain_Dubes.pdf, pp. 172-177). Obtain the (regular) Rand similarity of the clustering structures CS1= { {a, b}, {c, d, e}, {f, g}} and another clustering structure CS2= {{a, b}, {c, d}, {e, f, g}} -where the last cluster of CS2 contains the members e, f, and g-. Optional: you may also obtain the corrected Rand coefficient using these two clustering structure. Show the contingency table that needs to be corrected for the Rand coefficients. (For this question and for the following questions see http://www.cs.bilkent.edu.tr/~canf/CS533/clusterValidationNotes.pdf )

**6.** For the clusters of question 4 assume that CS1 is the ground truth, under this assumption calculate recall, precision and F measure values.

**7.** Consider a partitioning clustering structure that contains the following clusters. C1= {x, x, x, y} C2= {y, y, x), C3= {z, z, z, z, x, y}. This presentation means that in C1 there are three items of type x and one item of type y and we have similar interpretations for the contents of the other clusters. Calculate the cluster purity value for the above clustering structure.

**8.** In this part consider the paper A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review" *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.

**a.** Please explain the stages of clustering as defined in this paper.

**b.** Consider fuzzy clustering and introduce and idea that we can use fuzzy clustering approach in connection with $C^3M$.

**c.** In connection with simulated annealing the authors mention "tabu search." What does it mean? Explain its use within the context of simulated annealing-based clustering.

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

**e.** In connection with the above question (section d) please also explain what is meant by clustering tendency? Does it make sense to use clustering tendency in some stage(s) of clustering? What would you propose to use for identifying clustering tendency? Please try to be creative. For this purpose you may do a literature search and borrow some ideas and use them after some modification.

**9.** Assume that we have 100 pages and each page contains 20 records. We have a query with 5 relevant records. What is the expected number of pages to be accessed to retrieve all of these relevant records? Use Yao's formula (the paper is cited in our course web site).