Datastructuren

November 16, 2009
Due Date: December 21

Programming Assignment No 5

1. In this programming assignment you are asked to implement the single-link k-clustering algorithm discussed in class and for which you also have a handout. You may assume that the distance function on the objects to be classified is already given/designed. Furthermore choose for your single-link maximal spacing algorithm an implementation which is derived from Kruskal's MST algorithm. For efficiency's sake the implementation will be done with the union-find data structure. Your implementation of the union-find data structure should have the following characteristics: the Union operation takes O(1) time, the MakeUnionFind takes O(n) time and the Find operation O(log(n)) time. Your program will input a symmetric n-by-n array of non-negative numbers (the distance (or dissimilarity) matrix) and the number of desired clusters k and output the k lists of clusters. Furthermore it will output the minimum distance between the clusters (a quantity the algorithm maxiximizes). (The n objects to be classified are identified with the first n natural numbers $\{1, ..., n.\}$.)

   This problem breaks up very naturally into two subproblems: 1) construct an efficient implementation of the Union-Find data structure (you can also say in this context: Union-Find abstract data type), and 2) the implementation of the single-link k-clustering algorithm via a derivative of Kruskal's MST algorithm which uses the Union-Find data structure similar to the use of Union-Find in the algorithm of Kruskal.

2. Find an interesting clustering problem. Give a description of your problem of choice and describe how you designed the distance function on the objects to be clustered. In case your clustering problem comes equipped with a distance function, it suffices to provide its definition. Run your algorithm on some interesting instances of your problem of choice. Report what criteria you used for the choice of the parameter k. Interpret the results and discuss how well these fit your own way (that is, non-machine way) of clustering the data.

Turn your C++ code and the paragraph of text with the interpretation of your results as a zipped file to Minh Tran Ngoc (email: minhtn@liacs.nl) on or before , December 21. Make sure both the C++ files and the interpretation

contain the names of your team. Furthermore turn in a hardcopy of your files (C++ and text). You can work on this assignment in pairs. Refer questions to minhtn@liacs.nl or deutz@liacs.nl.