

Constructing an Optimal Clustering from Several Existing Clusterings

Jie Wang
Computer Science Department
UMass Lowell

Wednesday, 23 February 2005
Olsen 311

Refreshments at 2:30, Talk from 3:00-4:00

Let S be a data set. Suppose we are given several clusterings of S produced by different clustering algorithms. We are interested in, based on the information presented in these clusterings, constructing a new clustering of S that is optimal according to certain measures. Such a measure can be defined using a graph theoretic model or a set theoretic model. In this talk I will first explore the graph theoretic model. We define threshold graphs and formulate an integer programming problem to produce an optimal clustering. The problem is, unfortunately, NP-hard. Efficient ratio-8 approximation exists. We further obtain a ratio-4 approximation. The approximation algorithm is efficient provided that we can efficiently solve a minimization problem with a linear objective and non-linear constraints in the form of $x \leq \max\{y, z\}$. I will then explore the set theoretic model based on symmetric difference. If time permits, I'll show that a minimum clustering problem in this model is NP-hard.

This is joint work with M.-Y. Kao (Northwestern University) and N. Zhong (UMass Lowell).