Title: How to cluster different types of data towards big data analytics?

Abstract:
Every day, over 2.5 quintillion bytes of data are created from everywhere of our social life. Such huge amount of data makes difficult to handle them with traditional data management and processing tools. Therefore, to extract the most valuable pieces of information from such big data, novel and efficient analyzing technologies should be explored. As we know, clustering is a very useful technology for data analysis. However, it has suffered big challenges in big data environment. One of the problems is due to the variety of big data. That is, the data can be any type and the attributes of data can be numerical, categorical, or both. Unfortunately, most of the existing clustering approaches are applicable to purely numerical or categorical data only, but not the both. In general, it is a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes because there exists an awkward gap between the similarity metrics for categorical and numerical data. In this talk, we therefore introduce a general clustering framework and give a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. Moreover, the mechanism of determining the number of clusters for such mixed data is discussed as well. Consequently, a new clustering algorithm is presented with its performance demonstrated empirically.