Candidate Name: Payel Banerjee

(Index no: 147/16/Phys./25)

THESIS TITLE: Design and Evaluation of fast and Efficient Clustering algorithms for high data volume using hardware/software Methodologies

Abstract: In recent years, the major challenge to data mining experts is to deal with the increasing avalanche of data piling up at such a high speed. Apart from collecting and storing data with limited space and memory, another major challenge is to extract useful information from it especially when the dataset is unlabeled. Various data mining methodologies are there to discover hidden patterns in the data with the most famous one being the 'Clustering' technique which is specially used for unsupervised data mining applications. But the traditional clustering algorithms fail to handle such gigantic data volume resulting in making data analysis difficult. We aim to design and modify clustering algorithms using hardware and software approaches to make those suitable for large data volume by reducing time Complexity without increasing the existing space complexity. Apart from dealing with large data, another major problem is analyzing databases that are "frequently changing" by nature. Our aim also includes developing Clustering algorithms that can not only analyze large data but also dynamic databases. Out of various Clustering techniques, one of the most commonly used technique is Hierarchical Clustering which can create a hierarchy of clusters in a tree-like structure without any requirement of the apriori specification of the cluster number. Hierarchical Clustering algorithm can be subdivided into various linkage methods based on the definition of similarity between clusters. Our research work mainly focuses on Single and Complete Linkage algorithms because of their wide applicability and high convergence time . We have also developed a hardware-implemented accelerated version of the Leader algorithm which is widely used in the pre-clustering phase of various other well known clustering algorithms resulting in speeding up of those algorithms too. We have used real-world data (collected from various online repositories) for the purpose of testing and validation.

Signature of the Supervisors

TKBallabh

Department of Physics
Jadavpur University
Kolkata – 700 032

Director
A. K. Choudhury
School of IT
University of Calcutta

Signature of the Candidate

Payel Banerjee

01.03.2023