Diversity Maximization over Large Data Sets

Sepideh Mahabadi, Toyota Technological Institute at Chicago

In this paper we consider efficient construction of ``composable core-sets"" for the task of diversity maximization. A core-set is a subset of the data set that is sufficient for approximating the solution to the whole data set. A composable core-set is a core-set with the composability property: given a collection of data sets, the union of the core-sets for all data sets in the collection, should be a core-set for the union of the data sets. Using composable core-sets one can obtain efficient solutions to a wide variety of massive data processing applications, such as streaming algorithms and map-reduce model of computation.

We review several measures that capture the notion of diversity, including ""minimum pairwise distance"" and ""sum of pairwise distances"". However, the main focus of the talk will be on the ""determinant maximization"" problem which has recently gained a lot of interest for modeling diversity. The presented algorithms are simple to implement and we further show their effectiveness on standard data sets.