ABSTRACT

Data mining is interesting patterns and trend finding process in large database. Clustering is one of data mining functionality used for grouping objects into clusters, in which objects in the same cluster have high of similarity and high dissimilarity in different clusters.

Clustering problem is well known in the database literature for their use in numerous applications such as customer segmentation, classification and trends analysis. In high dimensional spaces not all dimensions may be relevant to a given cluster. One way to handling this is to pick the closely correlated dimensions and find clusters in the corresponding subspace. Traditional feature selection algorithm attempts to achieve this.

But this approach can lead to a loss of variables and not effectively identify clusters on different subspaces. Different sets of points may cluster better for different subsets of dimensions. The number of dimensions in each such specific subspace may also vary. Hence, the subspace clustering or projected clustering concept is needed, in which the subsets of dimensions selected are specific to the cluster themselves. This final project studies and analyzes how PROCLUS algorithm clusters projected cluster in high dimensional data.

Keywords *: data mining, PROCLUS algorithm, subspace clustering, projected clustering, high dimensional data*