1. INTRODUCTION

Clustering is an unsupervised method where the objective is to construct decision boundaries based on unlabeled training data to find the natural groups or clusters that exist in the data set (Hastie, Tibshirani, & Friedman, 2001; Kaufmann & Rousseeuw, 2005). Unsupervised learning or clustering is a typical problem because the number and shapes of clusters vary for datasets and different application areas however the basic aim is to maximize the similarity of the data points within each cluster and minimize it across clusters. The clustering technique has been extensively studied in many fields such as pattern recognition, image segmentation, data visualization and similarity search, signal processing and trend analysis (Jain, Murty, & Flynn, 1999; Dy & Brodley, 2004). As a consequence, several clustering algorithms have been proposed in the literature and new clustering algorithms continue to appear. The few reasons which have attracted the interest of researchers in cluster analysis are firstly; it can reduce the cost associated with collecting and labeling a large set of sample patterns. Secondly, it can be used as a preprocessing step to find out the natural groupings within the data and then it can be subjected to supervised learning for finding labels. And lastly it may be valuable to gain some insight into the inherent properties of the data
with clustering and later the authors can change our approach to handle different groups. Many algorithms exist on clustering, however most of them focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points (He, Xu, & Deng, 2002; Ester, Kriegel, Sander, & Xu, 1996). However in real world much of the data is categorical in nature, where attribute values can’t be naturally ordered as numerical values. An example of categorical attribute may be faculty whose values include humanities, science engineering and commerce etc. Hence due to the special properties of categorical attributes, the clustering of categorical data is very different from that of numerical data.

2. RELATED WORK

Clustering in data mining is a discovery process that groups a set of data such that the intracluster similarity is maximized and the intercluster similarity is minimized (Chen, Han, & Yu, 1996). These discovered clusters can be used to explain the characteristics of the underlying data. Clustering has found many business applications, it can be used to identify different customer segments and allow businesses to offer them customized solutions, or to predict customer buying patterns based on the properties of the cluster to which they belong.

Many clustering algorithms exist for various type of target datasets, most of the previous clustering algorithms exist for numerical data whose inherent geometric properties can be naturally analyzed to find out the distance function between data points such as k-means, DBSCAN, CURE, Wave Cluster (Queen, 1967; Nanopoulos & Theodoridis, 2001; Ester et al., 1996; Zhang et al., 1996; Sheikholeslami et al., 1998). Most traditional clustering algorithms are limited in handling datasets that contain categorical attributes. Clustering algorithms for numerical attributes don’t work well for the categorical attributes due to their different properties. A few algorithms have been proposed in recent years for clustering categorical data (Guha et al., 1998; Karypis et al., 1999; Huang, 1997; Zhang et al., 2000; Gibson et al., 1998). He et al., (2003) have proposed a k-histogram algorithm for categorical data which extends the k-means algorithm by replacing the means of clusters with histograms and dynamically updates histograms in the clustering process.

BIRCH is Balanced Iterative Reducing and Clustering using Hierarchies algorithm, suitable for very large databases. It incrementally clusters incoming multi-dimensional metric data points to try to produce the best quality clustering. It employs a tree or dendrogram to get all needed information to perform clustering. In this algorithm a CF tree is created (Zhang, Ramakrishnan, & Livny, 1996). A CF tree is a height balanced tree with two parameters: branching factor B and threshold T. Categorical Clustering Using Summaries (CACTUS) is a fast summarization based algorithm for categorical data. It consists of three phases: summarization, clustering and validation. It discover clusters in subspaces, which are especially useful when the data consists of a large number of attributes (Ganti, Gehrke, & Ramakrishnan, 1999).

Guha et al. have given RObust Clustering using links (ROCK) algorithm, which is an adaptation of agglomerative hierarchical clustering algorithm. It starts by assigning each tuple to a separated cluster and then clusters are merged repeatedly according to the proximity between clusters, where the proximity between clusters is defined as the sum of the number of links between all pairs of tuples, where the number of links is computed as the number of common neighbors between two tuples (Guha, Rastogi, & Shim, 1999).

K-modes is an algorithm which extends the k-means paradigm to categorical domain is introduced. New dissimilarity measures to deal with categorical data is conducted to replace means with modes, and a frequency based method is used to update modes in the clustering process to minimize the clustering cost function (Huang, 1999).

Zhang et al. (1999) have introduced a novel statistical procedure for clustering categorical data based on Hamming distance (HD) vectors.
Potential Cases, Database Types, and Selection Methodologies for Searching Distributed Text Databases
www.igi-global.com/chapter/potential-cases-database-types-selection/24152?camid=4v1a

A Proposal to Study of Cross Language Information Retrieval (CLIR) System Users' Information Seeking Behavior
www.igi-global.com/chapter/a-proposal-to-study-of-cross-language-information-retrieval-clir-system-users-information-seeking-behavior/198645?camid=4v1a