Self-Tuning Clustering: An Adaptive Clustering Method for Transaction Data

Ching-Huang Yun[1], Kun-Ta Chuang[2], and Ming-Syan Chen[1]

[1] Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan, ROC
mschen@cc.ee.ntu.edu.tw, chyun@arbor.ee.ntu.edu.tw
[2] Graduate Institute of Communication Engineering
National Taiwan University
Taipei, Taiwan, ROC
doug@arbor.ee.ntu.edu.tw

Abstract. In this paper, we devise an efficient algorithm for clustering market-basket data items. Market-basket data analysis has been well addressed in mining association rules for discovering the set of large items which are the frequently purchased items among all transactions. In essence, clustering is meant to divide a set of data items into some proper groups in such a way that items in the same group are as similar to one another as possible. In view of the nature of clustering market basket data, we present a measurement, called the small-large (SL) ratio, which is in essence the ratio of the number of small items to that of large items. Clearly, the smaller the SL ratio of a cluster, the more similar to one another the items in the cluster are. Then, by utilizing a self-tuning technique for adaptively tuning the input and output SL ratio thresholds, we develop an efficient clustering algorithm, *algorithm STC* (standing for *Self-Tuning Clustering*), for clustering market-basket data. The objective of algorithm STC is "*Given a database of transactions, determine a clustering such that the average SL ratio is minimized.*" We conduct several experiments on the real data and the synthetic workload for performance studies. It is shown by our experimental results that by utilizing the self-tuning technique to adaptively minimize the input and output SL ratio thresholds, algorithm STC performs very well. Specifically, algorithm STC not only incurs an execution time that is significantly smaller than that by prior works but also leads to the clustering results of very good quality.

Keywords: Data mining, clustering market-basket data, small-large ratios, adaptive self-tuning.

1 Introduction

Mining of databases has attracted a growing amount of attention in database communities due to its wide applicability to improving marketing strategies [3].

Y. Kambayashi, W. Winiwarter, M. Arikawa (Eds.): DaWaK 2002, LNCS 2454, pp. 42–51, 2002.

Among others, data clustering is an important technique for exploratory data analysis [6]. In essence, clustering is meant to divide a set of data items into some proper groups in such a way that items in the same group are as similar to one another as possible. Most clustering techniques utilize a pairwise similarity for measuring the distance of two data points. Recently, there has been a growing emphasis on clustering very large datasets to discover useful patterns and correlations among attributes [4] [10]. Note that clustering is a very service dependent issue and its potential applications call for their own specific requirements.

Itemset data, referred to as market-basket data, has been well studied in mining association rules for discovering the set of frequently purchased items [2]. Different from the traditional data, the features of market basket data are known to be high dimensionality, sparsity, and with massive outliers. It is noted that there are several clustering technologies which addressed the issue of clustering market-basket data [5][7][8][9]. ROCK [5] is an agglomerative hierarchical clustering algorithm by treating market-basket data as categorical data and using the links between the data points for clustering categorical data. The authors in [7] proposed an algorithm by using large items as the similarity measurement to divide the transactions into clusters such that similar transactions are in the same clusters. OAK [8] combines hierarchical and partitional clustering techniques. The work in [9] utilized a predetermined ratio for clustering market-basket data.

In view of the nature of clustering market-basket data, we present in [9] a measurement, called the *small-large (SL) ratio*, which is in essence the ratio of the number of small items to that of large items. With their formal definitions given in Section 2, a *large item* is basically an item with frequent occurrence in transactions whereas a *small item* is an item with infrequent occurrence in transactions. Clearly, the smaller the SL ratio of a cluster, the more similar to one another the items in the cluster are. Then, by utilizing a self-tuning technique for adaptively tuning the input and output SL ratio thresholds, we develop an efficient clustering algorithm, referred to as *algorithm STC* (standing for *Self-Tuning Clustering*), for clustering market-basket data. Algorithm STC consists of three phases, namely, *the pre-determination phase*, *the allocation phase*, and *the refinement phase*. In the pre-determination phase, the *minimum support* S and the *maximum ceiling* E are calculated according to a given parameter, called *SL distribution rate* β. In the allocation phase, algorithm STC uses the minimum support S to identify the large items and uses the maximum ceiling E to identify the small items. Explicitly, algorithm STC scans the database and allocates each transaction to a cluster for minimizing the SL ratio. In the refinement phase, each transaction will be evaluated for its status to minimize its SL ratio in the corresponding cluster. Algorithm STC utilizes two kinds of SL ratio thresholds, *output SL ratio threshold* α^o and *input SL ratio threshold* α^i, to evaluate the quality of the clustering. Explicitly, a transaction is moved from one cluster to the excess pool if its SL ratio is larger than α^o, and a transaction is moved from the excess pool to one cluster if the SL ratio is smaller than α^i. Detailed operations of STC will be described in Section 3 later. It is important to note that by utilizing $\alpha(U)$ to tune both α^i and α^o, STC is able to minimize the SL ratios of