17-18 SCOPUS 8

Mining Top-K High Utility Itemsets without Generating Candidates

Lekha I. Surana[1] Prof. V. B. More[2]
[1]P.G. Student
[1,2]Department of Computer Engineering
[1,2]MET's Institute of Engineering Nashik, Maharashtra, India

Abstract— In the domain of data mining, utility mining is the new development area. Due to anti-monotonicity property utility mining is the complicated with itemset share framework. There exist several techniques on mining utility with two phase candidate generation approach. However, such approach is inefficient as well as not scalable for huge dataset. In two phase candidate generation approach, scalability issue is detected in case of huge number of candidates. The proposed system is efficiently identifying high utility itemsets without candidate generation. Reverse enumeration tree is introduced to reduce search space by utility upper bound. d2 HUP enables to compute tight bound efficient pruning & directly discovers the HUI in scalable as well as efficient way. Along with the proposed work, system can also discover the top-k itemsets from extracted results. With the experimental results, we have to show that the time required for pattern enumeration is very less using d2HUP algorithm than the existing techniques of candidate set generation.

Key words: Data Mining, Utility Mining, High Utility Patterns, Frequent Patterns, Pattern Mining

I. INTRODUCTION

Due to rapid development of database techniques which facilitates the storage and usage of massive data from business, corporations, governments and scientific organizations, there is certain of attentions has received that how to collect valuable information from different databases. Among all issues HUI i.e. high utility itemset mining problem is derived from the problem of frequent itemset mining. Frequent itemset mining is an approach of discovering frequent itemsets from transaction databse. Itemset frequency is measured with the support of itemset such as, number of transactions involved in itemset. There are many applications of mining high itemsets from database to identify interestingness of valuable data. In frequent itemset mining [4] [7] pattern is remarked as interesting if its frequency pass user specified threshold value. Frequent pattern mining is to discover frequently purchased products by customers. While purchasing products user interest may relate to many factors which required express the frequency of product. For instance, manager of supermarket may interest in extracting combinations of products with high profits and interests. It relates to unit profits and purchased product quantities which are not considered in mining frequent pattern mining. Itemset utility mining is proposed in shared framework for itemset mining. It addressed the limitations of frequent pattern mining i.e. FPM. There are several techniques existed which analyses the different domain of interestingness. Frequent pattern mining algorithm has anti-monotonicity property which is not applied on utility mining approach and also it gets complicated in shared framework. Previously, for utility itemset mining two-phase candidate generation algorithm is proposed which causes inefficiency and scalability issues for huge dataset [4] [15]. In two phase candidate generation approach, in first phase pattern having high utility get identified where is in second phase, database is scanning one more time to discover any remaining candidate of high utility pattern. The first phase contains many long transactions or sometimes minimum utility small threshold. It is inefficient task and required strong pruning of vertical data structure. To mine frequent patterns it explores regular set enumeration in a reverse lexicographic order. HUIMiner algorithm is less efficient than two phase algorithm when mining large databases due to inefficient join operations, lack of strong pruning, and scalability issue with its vertical data structure. There D^{2}HUP algorithm is proposed to address key challenges in existing approaches. It is used for utility mining with the itemset share framework. In proposed system following are the contributions such as:

To directly discovers high utility patterns in a single phase without generating high TWU patterns (candidates). For this, powerful pruning technique is utilized. It is based on tight upper bounds on utilities. It is based on closure and singleton property which deals with dense data and it enhances the efficiency.

A linear data structure known as, "CAUL" is proposed to represent the original utility information in raw data.

System reads the transaction from input dataset file generates XU table and reverse set enumeration tree which will be further used in DFS approach. XUT is the eXternal Utility Table which contains different items and their respective prices. Using this XUT table Transaction weighted utility is calculated for each item. And CAUL is generated for each transaction having items value reater than min util.

Rest part of this paper contains existing work of utility mining, proposed system, it's algorithm and mathematical representation etc.

II. RELATED WORK

D^{2}HUP, algorithm seems to be novel solution for mining utility itemsets in distributed framework. Basically, it addresses the issues like, scalability and efficiency. It extracts HUP from huge transactional databases i.e. TWU. Pruning approach is used to maintain strenghth of proposed algorithm.

The proposed algorithm utilises the singleton property to improve the efficiency of dense data. CAUL is the linear data structure which is used to represent the information of unrefined data utility. Frequent pattern mining outputs the Constraints based mining data structure [1].

To solve the problem in existing methods Apriori and AprioriTid are used. By merging both algorithm