EFFICIENT ALGORITHM FOR HIGH UTILITY PATTERN MINING

[1]Ms. S. S. Naghate, [2]Dr. Mrs. S. S. Sherekar, [3]Dr. V. M. Thakare

[1]Student ME, [2]Professor, [3]Professor

[1] PG Department of Computer Science and Engineering,
[1]SGBAU, Amravati, India

Abstract: Utility Mining is new development of Data Mining Technology. This paper is focused on analysis of different high utility pattern mining algorithms, Such as mining for Transactional database, Concise and Lossless Representation, one phase without Generating Candidates, improving efficiency of Sequential Pattern Extraction, establishing manufacturing plans with Sliding Window Control. However there are some issues that need this paper and efficient methods to be resolved are discussed in "Mining method for High Utility" is proposed using the analysis of the various utility mining algorithms.

Index Terms - **Data Mining, Utility mining, High Utility Patterns.**

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a new powerful technology with great potential to help companies to focus on the most important information in data warehouses. The goal of data mining also known as KDD (Knowledge Discovery in Database) is to find interesting information or patterns from massive databases in automatic ways [1]. Discovering useful patterns hidden in a database plays an essential role in several data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining. The utility of an itemset represents its importance, which can be measured in terms of weight, profit, cost, quantity or other information depending on the user preference [2]. Among them, Utility mining emerges as an important topic in data mining field and has a variety of applications. For example, genome analysis, condition monitoring, cross marketing, and inventory prediction [3]. The problem of high-utility itemset mining is an extension of frequent pattern mining [4]. Frequent pattern mining is a popular problem in data mining, which consists in finding frequent patterns in transaction databases. For example, discovering combinations of products with high profits or revenues is much harder than other categories of utility mining problems [5].

This paper, addresses all the approaches in pattern mining, high utility pattern mining that has been employed to find a set of products creating high profits and minimum threshold by considering the purchase quantity and price of each product.

II. BACKGROUND

Many studies on utility mining have been done to develop the data mining algorithms in past years. Such algorithms are:
High utility pattern mining has been employed to find a set of products creating high profits by considering the purchase quantity and price of each product [1]. Mining high utility itemsets (HUIs) from databases is an important data mining task, which refers to the discovery of itemsets with high utilities (e.g. high profits). However, it may present too many HUIs to users, which also degrades the efficiency of the mining process [2].The novelties lie in a high utility pattern growth approach, a lookahead strategy, and a linear data structure. A novel algorithm that finds high utility patterns in a single phase without generating candidates [3]. In practice, most of the patterns identified by frequent pattern mining algorithms may not be informative to end-users. High Utility Sequential Pattern Extraction (HuspExt), which calculates the utilities of the child patterns based on that of the parents [4]. Frequent itemset mining is identifying set if items whose count in the transaction database is greater than a predefined minimum value. Frequent itemset mining is identifying set of items whose count in the transaction database is greater than a predefined minimum value [5].

The paper is organized as follows:

Section I Introduction. **Section II** discusses Background. **Section III** discusses previous work. **Section IV** discusses existing methodologies. **Section V** discusses attributes and parameters and how these are affected on mining techniques. **Section VI** proposed method and outcome of result. Finally **Section VII** Conclude this analytical paper.

III. PREVIOUS WORK DONE

In research literature, many data mining technology have been studied to provide various utility mining and improve the performance in terms of profitable product and high profits.

Unil Yun *et. al* (2017) [1] proposed an efficient algorithm named SHUPM for mining recent HUPs over data stream based on sliding window. This method can find useful pattern information that can allow users to understand the recent purchase preferences of customers.
Vincent S. Tseng *et. al* (2015) [2] proposed the performance of the AprioriHC and CHUD algorithms and compare them with two state-of-the-art algorithms UP-Growth and Two-Phase.
Junqiang Liu *et. al* (2016) [3] proposed many approaches for high utility pattern mining like, High utility pattern growth, Growing reverse set enumeration tree, Pruning by utility upper bounding, Avoiding enumeration by lookhead.
OznurKirmemisAlkan *et. al* (2015) [4] proposed efficient data structures and pruning technique which is based on Cumulated Rest of Match (CRoM) based upper bound and High Utility Sequential Pattern Extraction (HuspExt), which calculates the utilities of the child patterns based on that of the parents.

Vincent S. Tseng and Bai-En Shie *et. al* (2013) [5] proposed algorithms for mining high utility itemsets from large static datasets is given which uses a vertical approach for mining process and an incremental strategy for mining new set of high utility itemsets without scanning the original database again.

IV. EXISTING METHODOLOGIES

Many techniques and algorithms have been implemented over the last several decades. There are different methodologies that are implemented i.e. mining for establishing manufacturing plans with sliding window control, mining the concise and lossless representation, one phase without generating candidates, CRoM and HuspExt: Improving efficiency of high utility sequential pattern extraction, High utility itemset from Transactional Databases.

A. Mining for establishing manufacturing plans with sliding window control:

Two algorithms, named utility pattern growth (UPGrowth) and UP-Growth+, and a compact tree structure, called utility pattern tree (UP-Tree), for discovering high utility itemsets and maintaining important information related to utility patterns within databases are proposed [1].
In HUPM algorithm given by,
For each transaction t and for each item i in t
There is TWU (i) = TWU (i) + u (t)

B. Mining the concise and lossless representation:

To achieve high efficiency for the mining task and provide a concise mining result to users. Mining the concise and lossless representation proposed, three efficient algorithms named AprioriCH (Apriori-based algorithm for mining High utility closed itemsets), AprioriHC-D (AprioriHC algorithm with Discarding unpromising and isolated items) and CHUD (Closed High Utility Itemset Discovery) to find this representation [2].

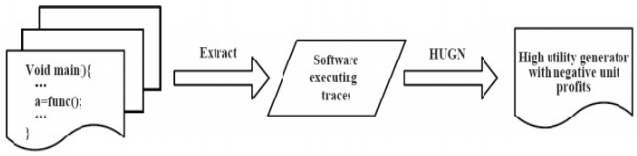

FIGURE 1: The main framework of concise and lossless representation

C. One phase without generating candidates:

One phase without generating candidates proposed a novel algorithm that finds high utility patterns in a single phase without generating candidates. The novelties lie in a high utility pattern growth approach, a lookahead strategy, and a linear data structure[3].
The novel algorithm given by,

1. if u(pat(N)) ≥ minU then output pat(N)
2. W ← {i|i≺ pat(N) ^ uB item(I, pat(N)) ≥ minU}
3. u (pat(N); t) ← u (pat(P), t) + u (I, t)
4. ∑←∑ + u (j, t)

D. CRoM and HuspExt : Improving efficiency of high utility sequential pattern extraction:

CRoM and HuspExtproposed efficient data structures and pruning technique which is based on Cumulated Rest of Match (CRoM) based upper bound. CRoM, by defining a tighter upper bound on the utility of the candidates, allows more conservative pruning before candidate pattern generation. In addition, an efficient algorithm developed High Utility Sequential Pattern Extraction (HuspExt), which calculates the utilities of the child patterns based on that of the parents [4].
In these, an itemset X is a subset of items sorted alphabetically, X U I. If | X|=r, the itemset X is called an r-itemset. I = {i1,i2,..,im} is a set of items, which may appear in sequences. For example, the itemset {ab} contains 2 items and is called a 2- itemset.

E. High utility itemset from Transactional Databases:

Frequent itemset mining is identifying set of items whose count in the transaction database is greater than a predefined minimum value. Frequent itemset mining is identifying set of items whose count in the transaction database is greater than a predefined minimum value. Frequent itemset mining follows downward closure property. According to this property if an itemset is infrequent then all the supersets of that itemset are also infrequent so it is not required to check the supersets of the infrequent itemsets thus preventing checking all the itemsets [5].
The calculation of transactional database given by,
Sum (UL (BX).i u's) +sum (UL (BX).ru's)≥ min_util

V. ANALYSIS AND DISCUSSION

Establishing manufacturing plans with sliding window control can find useful pattern information that can allow users to understand the recent purchase preferences of customers and developed several strategies to decrease overestimated utility and enhance the performance of utility mining [1].
The concise and lossless representation of mining addressed the problem of redundancy in high utility itemset mining by proposing a lossless and compact representation named closed high utility itemsets, which has been explored so far [2].
One phase without generating candidate's basic approach is to depth-first search the reverse set enumeration tree with pruning by basic upper bounding and enhanced significantly by the lookahead strategy that identifies high utility patterns without enumeration [3].
Improving efficiency of high utility sequential pattern extraction using CRoM and HuspExtreduces computational time as well as space requirements and utilizes the efficient data structures for keeping sequences and patterns in order to calculate CRoM values efficiently [4].

High utility itemset from Transactional Databases developed several strategies to decrease overestimated utility and enhance the performance of utility mining and it also spends time and memory to check and store minimal node utilities, which is more effective especially when there are many longer transactions in databases [5].

TABLE 1:Comparison between different utility mining techniques.

Proposed techniques	**Advantages**	**Disadvantages**
Mining for establishing manufacturing plans with sliding window control	With reference of algorithm it does not generate candidate patterns, requires less memory space and better scalability performance.	As these does not generates candidate patterns, the maintenance of information stored compactly.
Mining the concise and lossless representation	These algorithms are useful to the societies and it provides employment opportunities to communicate.	With these mining task decreases greatly for low minimum utility threshold or dealing with dense databases.
One phase without generating candidates	It facilities fast matching between candidates and transactions, and improve the efficiency of second phase as well as the pruning of lookhead strategy.	As the number of candidates and database is large the algorithm did not report the running time of second phase.
CRoM and HuspExt : Improving efficiency of high utility sequencial pattern extraction	The time and memory consumption together with the number of generated candidates and the pruned nodes increase as the minimum utility threshold decreases as the HuspExt is faster than the other state of art techniques.	The memory consumption and execution time is more than other datasets.
High utility itemset from Transactional Databases	Good scalability on runtime and the databases also contain lots of long transactions. A low minimum utility threshold is achieved.	Because of size of database increases, Memory usage increases and for identifying high utility itemset the performance is worse than the other.

VI. PROPOSED METHODOLOGY

High Utility Itemset Mining is a challenging task as in the recent times many algorithms have been proposed for mining high utility itemsets. High utility pattern mining is a technique that finds valuable patterns from large-sized databases with each item's importance and quantity of information associated with it. The representative utility pattern mining technique, high utility pattern mining, calculates the utilities of patterns by summating all the item utilities in the patterns.

Steps Of Algorithm:

Algorithm1: High_utility(D, t, B)

Input: A sequence of database D, a minimum utility threshold t, a set of data called batches B

Output: A set of High Utility Pattern P

1. Begin
2. Scan and Construct D into multiple batches B according to ascending entries of database D
3. For each batch B in D
4. If parent P.utility>= t
5. **Mining(B, P, t)**
6. Output parent P as high utility pattern
7. Elseif the parent P does not exist in a set of batch lists
8. Then
9. Construct a new list of parent P
10. End if
11. End

Algorithm2: Mining (B, P, t)

Input: A set of batches B, a set of high utility pattern P, a minimum utility threshold t

1. Begin
2. List := Set of Batches(B)
3. For each Batches(B)
4. If B(i)>=t
5. Add the patterns to parent P.utility
6. Else
7. Generate new batch from **High_Utility(D, t, B)**
8. End if
9. End

Diagrammatic representation of proposed method is shown as follows:

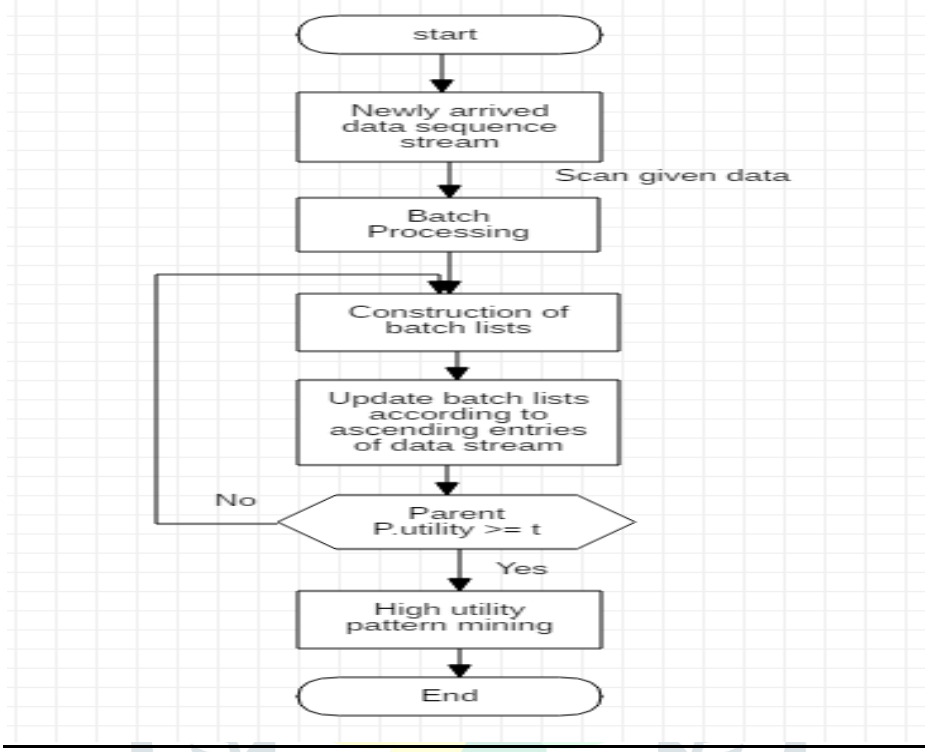

FIGURE 2: Flowchart of Proposed Methodology

VII. OUTCOMES AND POSSIBLE RESULT

In this way the proposed method performs for the high utility pattern mining when the data is in an information form. With the help of these algorithms, the proposed method calculates a minimum utility threshold from the batches made by scanning and updating batch list.

VIII. CONCLUSION

This paper focused on the comparative study of various Utility Mining Techniques i.e. for Transactional databases, Concise and Lossless Representations, one phase without Generating Candidates, improving efficiency of Sequential Pattern Extraction and establishing manufacturing plans with Sliding Window Control. But there are some issues for generating high utility patterns which reduces the number of redundant high utility patterns. With the help of proposed algorithm, significantly improves the memory requirements even in large-scale data with the utilization of low utility threshold values. It is also increases the performance of the system when dealing with the dense database and improves the performance of the system.

IX. FUTURE SCOPE

It is being observed that the mining of an arrangement of information will be improved in near future using High Utility Mining Techniques, which will be reducing the memory consumption and execution time significantly. With the advent of new mining techniques for low minimum utility threshold, dealing with dense databases will also be increased greatly.

REFERENCES

[1] Unil Yun, Gangin Lee, and Eunchul Yoon, Senior Member, IEEE, "Efficient High Utility Pattern Mining for Establishing Manufacturing Plans With Sliding Window Control", IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS,VOL. 64, NO.9, 7239-7249,September 2017

[2] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu, Fellow, IEEE, "Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 27, NO. 3,726-739, MARCH 2015

[3] Junqiang Liu, Member, IEEE, Ke Wang, Senior Member, IEEE, andBenjamin C.M. Fung, Senior Member, IEEE, "Mining High Utility Patterns in One Phase without Generating Candidates", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 28, No. 5, 1245-1257, May 2016

[4] OznurKirmemisAlkan and Pinar Karagoz, "CRoM and HuspExt: Improving Efficiency of High Utility Sequential Pattern Extraction", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 10, 2645-2657,October 2015

[5] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow, IEEE, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 8, 1772-1786, August 2013