Novel Approach to Discover High Utility Itemsets from Transactional Database

Supriya P. Bhosale

Abstract- **Mining high utility itemsets from a transactional database refers to the identify the itemsets with high utility like profits. Although a number of Algorithm's have been proposed but there is problem like it generate large number of candidate Itemsets for High Utility Itemsets. Large number of Itemsets degrades the performance of mining in terms of execution time and space requirement. This situation may worse when database contains large number of Transaction.**

In proposed system for UP-Growth algorithm information of high utility itemsets is maintained in tree based data structure named Utility Pattern Tree. With the help of UP-Tree candidate itemsets can be generated with only two scans of database. Proposed algorithm, not only reduce number of candidate itemsets but also work efficiently when database contains lots of long transactions.

Keywords: **Candidate itemsets, high utility itemsets, utility mining, data mining**

I INTRODUCTION

A. Overview

The purpose of proposed systems is towards finding high utility itemset. Here, the meaning of itemset utility is interestingness, importance, or printability of an item to users. Utility of items in a transaction database consists of two aspects: 1) The importance of distinct items, which is called external utility, and 2) The importance of items in transactions, which is called internal utility. Utility of an itemset is defined as the product of its external utility and its internal utility. An itemset is called a high utility itemset if its utility is no less than a user specified minimum utility threshold; otherwise, it is called a low-utility itemset. The goal of frequent itemset mining is to finds items that co-occur in a transaction database above a user given frequency threshold, without considering the quantity or weight such as profits of the items. However, quantity and weight are significant for addressing real world decision problems that require maximizing the utility in an organization. The high utility itemset mining problem is to finds all itemsets that have utility larger than a user specified value of minimum utility.

B .Problem Definition

Data mining is the process of retrieving itemsets from database. Proposed system use transactional database and mine high utility itemsets. High utility itemsets is nothing but the itemsets which have highest profit.In existing System, HUP Algorithm is used to mining High Utility Itemsets from database but there are some disadvantages like, it generates huge set of PHUIs. This system use UP-Growth Algorithm. Main advantages of this Algorithm are, it scan database only two times and it generates less set of PHUIs.

II LITERATURE SURVEY

Study on Mining Frequent Pattern

1. Fast Algorithms for Mining Association Rules

Author: R. Agrawal and R. Srikant

Year: Proc. 20th Intl Conf. Very Large Data Bases (VLDB), pp. 487- 499, 1994

Description: Apriori is a great improvement in the history of association rule mining, Apriori al-gorithm was first proposed by Agrawal in 1994. Apriori is more efficient during the candidate generation process for two reasons, Apriori employs a different candidate's generation method and a new pruning technique. There are two processes to finds out all the large itemsets from the database in Apriori algorithm. First the candidate itemsets are generated, and then the database is scanned to check the actual support count of the corresponding itemsets. During the first scanning of the database the support count of each item is calculated and the large 1 -itemsets are generated by pruning those itemsets whose supports are below the predefined threshold. In each pass only those candidate itemsets that include the same specified number of items are generated and checked.

Advantages:

1] Uses large itemset property.

2] Easily parallelized.

3] Easy to implement.

4] It doesn't need to generate conditional pattern bases.

Disadvantages:

1] It requires multiple database scans.

2] Assumes transaction database is memory resident.

3] Generating candidate itemsets.

ISSN: 2278 – 1323

2. Mining Frequent Pattern without Candidate generation.

Author: J. Han, J. Pei, and Y. Yin item

Year: Proc. ACM-SIGMOD Intl Conf. Management of Data, pp. 1-12, 2000.

Description: Mining frequent patterns in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. The previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist long pat-terns. This method used a novel frequent pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree-based mining method, FP-growth , for mining the complete set of frequent patterns by pattern fragment growth. Efficiency of mining is achieved with three techniques: (1) a large database is compressed into a highly condensed, much smaller data structure, which avoids costly, repeated database scans, (2) FP-tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of candidate sets, and (3) a partition-ing based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining conned patterns in conditional databases, which dramatically reduces the search space. FP-growth method is efficient and scalable for mining both long and short frequent patterns

Advantages:

1] It finds frequent itemsets without generating any candidate itemset

2] Scans database just twice.

3] Does not generate candidate itemsets.

Disadvantages:

1] It treats all items with the same importance/weight/price.

2] Consumes more memory and performs badly with long pattern data sets.

From this algorithm we get one Advantage in our proposed system i.e. scans database only two times.

Study on Weighted Association Rule Mining

1 .Mining Association Rules with Weighted Items

Author: C.H. Cai, A.W.C. Fu, C.H. Cheng, and W.W. Kwong

Year: Proc. Intl Database Eng. and Applications Symp. (IDEAS 98), pp. 68-77, 1998

Description: This method extends the tradition association rule problem by allowing a weight to be associated with each item in a transaction, to reflect interest/intensity of the item within the transaction. This provides us in turn with an opportunity to associate a weight parameter with each item in the resulting association rule. We call it weighted association rule (WAR). WAR not only improves the confidence of the rules, but also provides a mechanism to do more effective target marketing by identifying or segmenting customers based on their potential degree of loyalty or volume of purchases.

Advantages:

1] Cai et al first propose concept of weighted items and weighted association rules

Disadvantages:

1] This framework does not have downward closure property so mining performance cannot be improved.

2. Weighted Association Rule Mining Using Weighted Sup-port and Significance Framework

Author: F. Tao, F. Murtagh, and M. Farid

Year: Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD 03), pp. 661-666, 2003

Description: This addresses the issues of discovering significant binary relationships in transaction datasets in a weighted setting. Traditional model of association rule mining is adapted to handle weighted association rule mining problems where each item is allowed to have a weight. The goal is to steer the mining focus to those significant relationships involving items with significant weights rather than being flooded in the combinational explosion of insignificant relationships. This system identifies the challenge of using weights in the iterative process of generating large itemsets. The problem of invalidation of the"downward closure property" in the weighted setting is solved by using an improved model of weighted support measurements and exploiting a"weighted downward closure property". A new algorithm called WARM (Weighted Association Rule Mining) is developed based on the improved model.

Advantages:

1] The algorithm is both scalable and efficient in discovering significant relationships in weighted settings.

2] Tao et al proposed the concept of weighted downward closure property

Disadvantages:

1] Weighted association rule mining considers the importance of item, in some applications such as transaction databases items quantities in transactions are not taken into considerations.

From this algorithm we get one Advantage in our proposed system i.e. Consider weight of item.

ISSN: 2278 – 1323

Study on High utility itemset Mining

1. A Fast High Utility Itemsets Mining Algorithm

Author: Y. Liu, W. Liao, and A. Choudhary

Year: Proc. Utility-Based Data Mining Workshop,2005.

Description: Association rule mining (ARM) identifies frequent itemsets from databases and generates association rules by considering each item in equal value. However, items are actually different in many aspects in a number of real applications, such as retail marketing, network log, etc. The difference between items makes a strong impact on the decision making in these applications. Therefore, traditional ARM cannot meet the demands arising from these applications. By considering the different values of individual items as utilities, utility mining focuses on identifying the itemsets with high utilities. As downward closure property doesn't apply to utility mining, the generation of candidate itemsets is the most costly in terms of time and memory space. this paper present a Two-Phase algorithm to efficiently prune down the number of candidates and can precisely obtain the complete set of high utility itemsets. In the first phase, a model that applies the transaction-weighted downward closure property on the search space to expedite the identification of candidates. In the second phase, one extra database scan is performed to identify the high utility itemsets. It performs very efficiently in terms of speed and memory cost

Advantages:

1] It performs very efficiently in terms of speed and memory cost

Disadvantages:

1] Generate too many candidates to obtain HTWUI require multiple database scan.

.2 Isolated Items Discarding Strategy for Discovering High Utility Itemsets

Author: Y.-C. Li, J.-S. Yeh, and C.-C. Chang

Year: Data and Knowledge Eng., vol. 64, no. 1, pp. 198-217, Jan. 2008

Description: traditional methods of association rule mining consider the appearance of an item in a transaction, whether or not it is purchased, as a binary variable. However, customers may purchase more than one of the same item, and the unit cost may vary among items. Utility mining, a generalized form of the share mining model, attempts to overcome this problem. Since the Apriori pruning strategy cannot identify high utility itemsets, developing an efficient algorithm is crucial for utility mining. This paper proposes the Isolated Items Discarding Strategy (IIDS), which can be applied to any existing level-wise utility mining method to reduce candidates and to improve performance.

Advantages:

1] R3weduce candidates and to improve performance.

Disadvantages:

1] This algorithm still scan database for several times.

3. Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases

Author: C.F. Ahmed, S.K. Tanbeer, B.-S. Jeong, and Y.-K. Lee

Description: It provide an efficient research method for high utility pattern mining for handling incremental databases, while considering many insertions, deletions, and medications with the currently available memory size. Three variations of our tree structure have been proposed.

1] Incremental HUP Lexicographic Tree (IHUPL-Tree), is arranged according to an items lexicographic order. It can capture the incremental data without any restructuring operation.

2] IHUP Transaction Frequency Tree (IHUPTF-Tree), which obtains a compact size by arranging items according to their transaction frequency.

3] IHUP-Transaction-Weighted Utilization Tree (IHUPTWU-Tree) is designed based on the TWU value of items in descending order.

All of the tree structures have the build once mine many properties and are highly suit-able for interactive mining. All three tree structures require maximum two database scans.

Advantages:

1] Ability to consider the no binary frequency values of items in transactions and different profits values for every item.

2] Incremental and interactive data mining provide the ability to use previous data structures and mining results in order to reduce unnecessary calculations when a database is updated, or when the minimum threshold is changed.

3] Build once mine many.

Disadvantages:

1] It generates huge set of PHUIs.

2] Their mining performance is degraded consequently.

This situation may become worse when database contain many long transactions or low thresholds are set.

From this algorithm we get following Advantages in our proposed system i.e.

1] Ability to consider the no binary frequency values of items in transactions and different profits values for every item.

2] Reduce unnecessary calculations when a database is updated, or when the minimum threshold is changed.

ISSN: 2278 – 1323

III Proposed System

If you consider Existing methods, in that Algorithm multiple number of HTWUIs are generated. For removing this drawback of existing system, new Algorithm is proposed i.e. UP-Growth Algorithm. Main aim of this system is reducing item sets overestimated utilities.

Algorithm Used:

1. UP-Growth Algorithm

System Architecture:

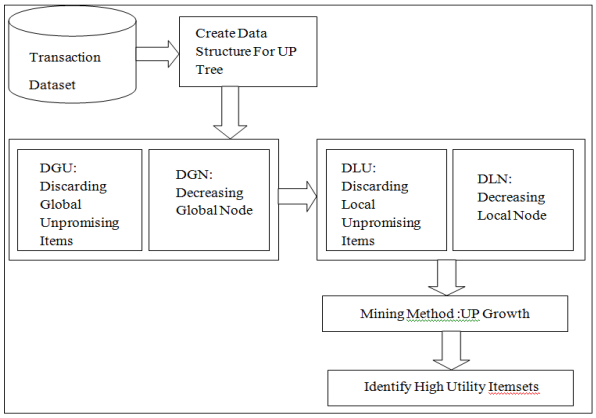

Figure System Architecture

Figure contains the following blocks:

1. Create Data Structure For UP Tree: In this block one UP Tree by using the data structure which consist of node name, node count, node utility, node parent, node link and set of child node.
1. DGU: Discarding Global Unpromising Items: After constructing UP tree the items which transaction weighted utility is less than the minimum utility threshold called unpromising items are discarded from item set.
2. DGN: Decreasing Global Node: After discarding the unpromising items the node utilities are decreased.
3. DLU: Discarding Local Unpromising Items: The items which transaction weighted utility is less than the minimum utility threshold called unpromising items are discarded from item set and construct conditional pattern base.
4. DLN: Decreasing Local Node: In DLN construct the local UP tree.
5. Mining Method: UP Growth: This block generates the fewer candidates from given transactional dataset.
6. Identify High Utility Itemsets: identify high utility itemsets and their utilities form the set of given candidates.

Advantages:

1] Number of generated candidates can be highly reduced.

2] High utility item sets can be identifies more efficiently.

IV Conclusion

Proposed system has a tree-based algorithm, called UP-Growth, for efficiently mining High utility itemsets from databases. We take Data Structure UP-Tree for maintaining the information of high utility itemsets and four effective strategies, DGU, DGN, DLU and DLN, to reduce search space and the number of candidates for utility mining. PHUIs can be efficiently generated from UP-Tree with only two database scans. UP-Growth Algorithm is faster than existing algorithms when database contains lots of long transactions.

REFERENCES

[1] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow,"Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases",IEEE Transaction on knowledge and data engineering, vol. 25, no. 8, Aug 2013.

[2] R. Agrawal and R. Srikant,"Mining Sequential Patterns ", Proc.11th Intl Conf. Data Eng. , pp. 3-14, Mar. 1995.

[3] J. Han, J. Pei, and Y. Yin,"Mining Frequent Patterns without Candidate Generation ", ACM-SIGMOD Intl Conf. Manage-ment of Data ,pp. 1-12, 2000.

[4] C.H. Cai, A.W.C. Fu, C.H. Cheng, and W.W. Kwong,"Mining Association Rules with Weighted Items ", Proc. Intl Database Eng.and Applications Symp. (IDEAS 98) , pp. 68-77, 1998.

[5] F. Tao, F. Murtagh, and M. Farid,"Weighted Association Rule Mining Using Weighted Support and Significance Framework ",ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD 03) , pp. 661-666, 2003.

[6] Y. Liu, W. Liao, and A. Choudhary, "A Fast High Utility Itemsets Mining Algorithm ", Proc. Utility-Based Data Mining Workshop 2005.

[7] Y.-C. Li, J.-S. Yeh, and C.-C. Chang,"Isolated Items Discarding Strategy for Discovering High Utility Itemsets ", Data and Knowl-edge Eng vol. 64, no. 1, pp. 198-217, Jan. 2008.

[8] C.F. Ahmed, S.K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases", IEEE Trans. Knowledge and Data Eng., vol. 21, no. 12,pp. 1708-1721, Dec. 2009

ISSN: 2278 – 1323