Abstract— Data mining is an efficient technique that extracts significant, available information from massive, complicated data. The development of data mining techniques brings serious threat to the security of individual sensitive information. Privacy Preserving Utility Mining is one of the important approaches in Privacy Preserving Data Mining, where it is an integration of both Utility Pattern Mining and Privacy Preserving Data Mining. Utility Pattern Mining will extract useful knowledge from database with high utility issues whereas Privacy Preserving Data Mining is used to modify the data in such a way so as to perform the data mining algorithm effectively without compromising the security of sensitive information contained in the data. So, Privacy Preserving Utility Mining is used to improve efficiency of the algorithm and ensures the database privacy. In the previous approach, A Fast Perturbation Algorithm is used with Tree structure and Tables as a database which avoids costly candidate generation and it also achieves efficiency and better performance than apriori like algorithm. However, the database needs to be scanned twice to get the Fast Perturbation Algorithm Using Tree structure and Tables and repeatedly checking the frequency of occurrence against the support threshold. Therefore, this can be time consuming while new data are added to the existing database. To solve this problem, Pattern Tree algorithm is proposed based on a new data structure Pattern Tree and a new technique, which can get P-Tree through only one scan of database and obtain the corresponding Fast Perturbation Tree with a specified support threshold. Updating a P-Tree with new data needs one scan of the data only and the existing data need not be rescaned. In addition the proposed algorithm generates better scalability and presents a large database in a highly condensed format and avoids the secondary database scan.

Index Terms— Data mining, Privacy Preserving Data Mining (PPDM), Privacy Preserving Utility Mining (PPUM), Utility Pattern Mining, Fast Perturbation Using Tree Structure and Tables (FPDTT), Pattern Tree (P-Tree).

I. INTRODUCTION

Data mining is the process of analyzing the data from different perspectives and summarizing it into useful information. Data mining software is a analytical tool for analyzing data from different dimensions or angles, categorize it and summarize the relationship identified. Technically, data mining is the process of finding patterns among different fields in large relational database. Pattern mining uses data mining algorithm to discover unexpected, useful and interesting patterns from the database. Pattern mining algorithm is applied on different types of data such as sequential database, strings, streams, transactional database, graphs, spatial data, etc. This is been designed to discover various patterns such as associations, trends, rules, lattices, sub graphs, periodic patterns, sequential patterns and so on. Researchers define interesting patterns such that top patterns, the patterns which appears frequently in a database, rare patterns or the patterns with high confidence. Utility pattern mining is one type of frequent pattern mining which extracts useful knowledge from database with high utility issues. The basic meaning of utility is importance/ interestedness/ profitability of items to the user. The utility of items in a database consists of two different aspects such as external utility which defines the importance of distinct items and the internal utility defines the importance of items in a transaction. High utility itemset is defined as that its utility is no less than a user-specified threshold. Mining high utility itemsets from the databases is not an easy task since the downward closure property used in frequent itemset mining cannot be applied here. The challenge is to effectively prune the search space and efficiently capture all high utility itemsets with no miss. Privacy-preserving data mining (PPDM) is one of the trends in privacy, security and data mining research. In privacy-preserving data mining (PPDM), data mining algorithms are analyzed for the side-effects they incur in data privacy, and the main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and knowledge remain private even after the mining process. By doing this the unauthorized user cannot access the data and the privacy is maintained. A number of techniques such as Data perturbation technique, Trust Third Party, Secure Multiparty Computation and game theoretic approach have been used in order to perform privacy preserving data mining. Privacy Preserving Utility Mining is the integration of both Utility Pattern Mining (UPM) as well as Privacy Preserving Data Mining (PPDM). Privacy Preserving Utility Mining (PPUM) uses two algorithms to achieve the goal of hiding sensitive itemsets so that the unauthorized user cannot mine them from the modified database. In addition, Privacy Preserving Utility Mining reduce the impact on the sanitized database in the process of hiding sensitive itemsets. Privacy Preserving Utility Mining can deal with more important,
sensitive information because of the characteristics of utility pattern mining. A majority of the efficient PPUM methods have been applied to database perturbation for effectively hiding sensitive utility items or patterns by removing certain items from original databases.

II. PREVIOUS RELATED WORK

Reversible privacy preserving data mining: A combination of difference expansion and privacy preserving [2] Tung-Shou Chen, Wei-Bin Lee, Jeanne Chen, Yuan Hung Kao, Pei-Wen Hou uses Reversible privacy preserving data mining (RPPDM) in order to solve the problem of PPDM. Privacy Difference Expansion (PDE) method is used so that the original data is perturbed and embedded with a fragile watermark to accomplish privacy preserving and data integrity of mined data and also to recover the original data. PDE identifies whether there is any modification of data and also secure data integrity and improve the performance by reducing information loss and privacy disclosure risk.

Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases [4] Vincent S. Tseng, Bai-En Shieh, Cheng-Wei Wu, and Philip S. Yu. Here UP-growth and UP-growth+ algorithm is used for mining high utility itemset with a set of effective strategies for pruning candidate itemset. The information of high utility itemset is maintained in a tree structure called UP-tree such that candidate itemset can be generated efficiently with only two scans of database. UP-growth+ not only reduce the number of candidates efficiently but also outperform runtime especially when database contain lots of long transactions. UP-growth+ is the fastest performance among utility pattern mining based on tree structures.

Enabling Multi-level Trust in Privacy Preserving Data Mining [5] Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang uses the new dimension of Multi-Level Trust which poses new challenges for perturbation based Privacy Preserving Data Mining, in contrast to the single-level trust. In single-level trust only one perturbed copy is released, now multiple different perturbed copies of the same data is available to data miners at different trusted levels. The key challenge lies in preventing the data miners from combining copies at different trust levels to jointly reconstruct the original data more accurate than what is allowed by the data owner. This is done by properly correlating noise across copies at different trust level (noise covariance matrix with corner-wave property). Advantage is that the privacy of the data is preserved and disadvantage is that the resources and time taken to preserve the data in multilevel is high when compared with other methods.

Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining [9] Oliveira, S. R. M., & Zaiane which has two algorithms namely (round robin and random) were introduced for balancing privacy and knowledge discovery in association rule mining. Sanitization algorithm requires only two database scans (i.e., first scan for building index to speed up the process and second scan is to sanitize the original database). Advantage is that the algorithm alters the data which enables flexibility to tune them and also guarantee privacy and do not introduce false data. Disadvantage of the algorithm is that it is very costly to implement.

HHUIF and MSICF: Novel algorithms for privacy preserving utility mining [7] Jieh-Shan Yeh, Po-Chiang Hsu. The algorithms Hiding high utility item first algorithm (HHUIF) and Maximum sensitive itemsets conflict first algorithm (MSICF) are the two algorithms used to reduce the impact on the source database of privacy preserving utility mining. The algorithm is used to modify the database transactions containing sensitive itemset which reduces the utility value while preventing reconstructing of original database from the sanitized one. The algorithm will hide the sensitive itemset so that the unauthorized users cannot mine them from the modified database. The algorithm balances both the privacy and knowledge discovery in sharing data. Disadvantage is that the database scanning takes place many times and have to search the whole database for every scan.

III. PROBLEM DEFINITION

Mining frequent patterns with an FPUTT avoids costly candidate generation and repeatedly occurrence frequency checking against the support threshold. Hence FPUTT achieves better efficiency and performance than Apriori algorithm. The main disadvantage is to construct an FPUTT tree the database needs to be scanned twice and also it is time-consuming when new data are added to an existing database. To overcome this problem a modified Pattern tree data structure and a modified technique is used, where the Pattern tree can be constructed in a single scan and the corresponding FPUTT tree can be generated by a specified support threshold. Updating a Pattern tree with new data needs one scan of the new data only, and the existing data need not be re-scanned. Thus, this approach constructs a more compact tree structure than FPUTT tree and avoids the second scan of the database.

IV. SYSTEM ARCHITECTURE

The system architecture is shown in Figure 1 that clearly outlines every module. The module broadly classifies various sub topics within each of the modules. The input and output of the software forms the boundaries in the given figure. This work consists of three main modules such as 1. Utility Pattern Mining 2. Pattern Tree 3. Perturbed database. The overall process of P-Tree approach is to overcome the drawback of existing technique and also to refine the frequent itemset to most frequently used itemset.

A. Utility Pattern Mining

The Database or dataset is collected from different organizations like marketing, weather forecasting, insurance agencies, hospitals, etc. (E.g. Medical Dataset). Input will be the original database and sensitive itemset discovered from the database with the help of UP-growth+ (Utility Algorithm). The UP-growth+ conducts utility pattern mining through a tree structure for pruning candidate itemset. The UP-growth+ approach provide the fastest performance among utility pattern mining algorithm based on tree structures.

B. Pattern Tree Mining

In P-Tree we are going to use prefixspan algorithm. This
algorithm is based on sequential pattern mining and the input will be given in a sequential format and the output will be the frequent patterns. These patterns are considered to be the sensitive itemset.

We can obtain a P-tree through one scan of the database and get the corresponding FPUTT from the P-tree later. The construction of a P-tree can be divided into two steps as well:

1. First step is to sort the transaction in some order such as alphabetic, numerical or any other specific order, and then we can generate a P-Tree by inserting the transaction one by one and meanwhile we can record the support count of every item into the list L.

2. After the only scan of the database, we can sort List L according to the min_sup value.

The reconstruction of Pattern Tree follows similar insertions in the first step. The only difference is that one needs to sort the path according to before inserting it into a new P-tree.

The following algorithm mines the sequential frequent pattern according to recursive mining process:

1. Find length-1 sequential patterns.
2. Divide search space.
3. Find subset of sequential patterns.
4. Recursively partition the length-2 sequential patterns into n-subset.
5. Construct the projected database and mine the subset.
6. Recursively mine the sequential patterns.

C. Perturbed Database

Here perturbation can be many types such as sorting, sketching and watermarking. This perturbed database contains sensitive itemset which have been perturbed. The goal of the perturbation is that the sensitive itemset is perturbed and is embedded with a fragile watermark to accomplish privacy preserving and data integrity of mined data and also to recover the original data.

V. METHODOLOGY

The following algorithm mines the sequential frequent pattern according to recursive mining process:

- **Input:**
  - A sequence database S, and min_sup.
- **Output:**
  - The complete set of sequential frequent patterns.

**Begin**

1. Find length-1 sequential patterns.
   - i. Scan the sequential database.
   - ii. Find all frequent items from the database in sequence.
   - iii. List each frequent items which is of length -1.
2. Divide search space.

VI. ANALYSIS

PrefixSpan algorithm is more efficient than the existing algorithm because this uses pseudoprojection technique where this technique may reduces the number and size of projected data.

Pattern Tree is more efficient and compact such that memory space is reduced and total time to find out the frequent itemset is reduced.

<table>
<thead>
<tr>
<th>Tree structure</th>
<th>Max Memory(Mb)</th>
<th>Frequent count</th>
<th>Total time(ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPUTT</td>
<td>2.659</td>
<td>72</td>
<td>46</td>
</tr>
<tr>
<td>P-TREE</td>
<td>1.359</td>
<td>53</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 1. Comparison between FPUTT and P-Tree

Fig. 1. Overall process of the Pattern Tree approach
to generate frequent patterns. This technique avoids costly candidate generation and repeated frequent itemset generation. It is more efficient and reduces total time taken to generate the frequent (sensitive) itemset.

REFERENCES


VII. CONCLUSION

In this paper, a P-Tree is used with a modified techniques and data structure to obtain an efficient tree by one database scan. It also presents a large database in a highly condensed format and avoids the second scan of the database.
She has published her research articles in two international journals and one international conference.

**Dr. Shiram. R** is currently working as Professor in the Department of Computer Science & Engineering, B.S. Abdur Rahman University, Chennai. His areas of research interest include Pervasive Computing, Mobile Computing, Language Computing, Grid Computing, Education and Information Retrieval. He has published research articles in more than 15 international journals and 13 international conferences.

**Sasirekka. S** obtained her B.Tech. in Information Technology in the year 2014 from Sree Sastha Institute of Engineering and Technology affiliated to Anna University. She is currently pursuing her M.Tech. in Information Technology in B.S. Abdur Rahman University, Chennai.