Journal of Theoretical and Applied Information Technology
20th March 2013. Vol. 49 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRIVACY PRESERVING SENSITIVE UTILITY PATTERN MINING

[1]C.SARAVANABHAVAN, [2]R.M.S.PARVATHI
[1]Research Scholar & Asst Professor, Department of CSE,
Annai Mathammal Sheela Engineering College, Tamil Nadu, India.
[2]Principal & Professor, Department of CSE,
Sengunthar College of Engineering for Women, Tamil Nadu, India

E-MIAL: saravanabhavanphd@gmail.com

ABSTRACT

Data mining services require accurate input data for their results to be meaningful, but privacy concerns may influence users to provide spurious information. The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. The main feature of the most PPDM algorithms is that they usually modify the database through insertion of false information or through the blocking of data values in order to hide sensitive information. In this paper, we have incorporated the privacy preserving concept into the previously developed weighted utility mining approach. In this, we have presented an efficient algorithm for mining of privacy preserving high utility item sets by considering the sensitive item sets. The algorithm comprise of three major steps to attain the aim of our research includes, 1) Data sanitization, 2) Construction of sensitive utility FP-tree and, 3) Mining of sensitive utility item sets. The experimentation has carried out using real as well as the synthetic dataset and the performance of the proposed algorithm is evaluated with the aid of the evaluation metrics such as Miss cost and Database difference ratio.

Keywords: *Data Mining, Privacy Preserving Data, Utility Mining, Sanitized Data, Sensitive Item, Data Sanitization, Miss Cost, Database Difference Ratio.*

1. INTRODUCTION

Data mining mainly focuses on the problem of discovering unknown or hidden patterns from data. It comprises building models on data, providing statistical summary of data in a human comprehensible form, deciding upon strategies based on the mined information [1]. Recently, integrating utility constraints into data mining tasks has drawn much attention among the researchers. Several researchers have proposed many algorithms and techniques for mining high utility item sets. Moreover, researchers from the data mining area have highly utilized the qualitative aspects of attributes such as significance, utility than considering only the quantitative ones (e.g. number of appearances in a database, etc.,) for the reason that qualitative properties are required in order to completely use the attributes present in the dataset. Mining high utility item sets improves the standard frequent item set mining framework because it utilizes the intuitively defined utility rather than statistics-based support measure [2]. Utility mining is widely used in many practical applications. Naturally, utility is a measure of how “useful” (i.e. “gainful”) an item set is. The local transaction utility and external utility are employed to evaluate the utility of an item or item set. The local transaction utility of an item is defined based on the information stored in a transaction, such as the number of the item sold in the transaction, whereas the external utility of an item is based on the information from resources besides transactions, like a profit table [3].

In some business environments, the data mining may need to be processed among databases. Although the data may be distributed among several sites, the sites are not allowed to reveal its database to another site. For instance, some insurance companies have their own databases which contain their insured one’s information. For mutual benefit, these companies decided to work with insurance fraud detection by distributed data mining. The data mining model must be high accurate to identify fraud, because a fault leads to huge loss of income or great amounts of pay. Also, insurance companies cannot share the data about their clients with other companies, because of the restriction laws (and having a high competitive edge). They can share information about the fraudulent insurance records, but not their data. Each company have made an effort to share their "black-box" models to find out more interesting rules on the entire shared information than that on their own database, and can defend the private records that other companies may find [18, 19]. Privacy considerations may prevent this approach.

Privacy preservation is becoming more and more a serious problem for future progress of data mining techniques with great potential access to datasets having private, sensitive, or confidential information [8]. The major challenge for existing data mining algorithms is extracting accurate data mining results while still maintaining privacy of datasets. Due to the increasing concern on privacy, a new category of data mining called privacy preserving data mining (PPDM) has been introduced. But, the privacy-preserving data mining has turned into a major problem in recent years because of the huge amount of private data which is tracked by several business applications. In many situations, the users are reluctant to provide personal information unless the privacy of sensitive information is assured [4, 5]. PPDM was first introduced by Agrawal and Srikant in 2000 [6]. PPDM algorithms are developed by integrating privacy protection mechanism to conceal sensitive data before executing data mining algorithms. Then several different branches with different goals have been developed. Privacy preserving classification techniques prohibit a miner from building a classifier which is capable of forecasting the personal data [7, 9].

The main consideration in privacy preserving data mining is the sensitive nature of raw data. The data miner, while mining for comprehensive statistical information about the data, should not be able to access data in its original form with all the sensitive information. This necessitates more robust techniques in privacy preserving data mining that intentionally alter the data to conceal sensitive information as well as protect the inherent statistics of the data which is vital for mining purpose [1]. The latest trend in business collaboration is they are keen to share data or mined results to gain mutual benefit. But, it has also increased a potential threat of disclosing sensitive information when releasing the data. Data sanitization is the process, which hides the sensitive item sets present in the source database with proper modifications and discloses the modified database [11]. In this paper, we have presented an efficient privacy preserving algorithm for mining of high utility item sets which extends our previously proposed weighted utility item sets mining approach. Most of the privacy-preserving data mining techniques apply a transformation which reduces the efficiency of the original data when it is applied to data mining techniques or algorithms. Also, there is a natural tradeoff between privacy and accuracy, but this tradeoff is suffered by some specific algorithm which is employed for privacy-preservation. Therefore, the key issue is to sustain maximum utility of the data by satisfying the basic privacy constraints.

The main contributions of this research includes,

- We have transformed the original transaction database into the sanitized database, in order to hide the sensitive information.
- We have built a sensitive utility FP-tree using the sanitized database with sensitivity factor.
- We have mined the sensitive utility patterns by hiding the sensitive itemsets.
- We have evaluated our proposed algorithm in terms of standard evaluation metrics such as hiding failure, miss cost and the Database difference ratio.

The rest of the paper is organized as follows: a brief review of some of the literature works in privacy preserving data mining is presented in Section 2. The problem formulation of proposed technique is given in Section 3. The proposed efficient privacy preserving utility mining technique is detailed in Section 4. The experimental results and the performance evaluation discussion are provided in Section 5. Finally, the conclusions are summed up in Section 6.

2. RELATED WORKS

The new generation techniques of PPDM will definitely plays a major role in the future process of data mining. Some efficient techniques of PPDM available in the literature are reviewed briefly in this section.

Privacy is a key element in data mining such that the private information is not disclosed after mining. However data quality is important such that no forged information is provided and privacy is not jeopardized. R. R. Rajalaxmi and A. M.

Natarajan [11] have proposed a Conflict based Utility Item set Sanitization (CUIS) approach for preserving privacy in sensitive high utility item set mining. The proposed approach has intentionally altered the database to reduce the utility of the sensitive item sets. The approach has been iteratively applied in a greedy fashion until the utility of each sensitive item set becomes lower than the threshold MUT, whereas the amount of non-sensitive item sets with utility less than the MUT has been reduced. Several experimental results have demonstrated that the data sanitization approach has done minimum number of modifications to the database and removed a minimal amount of non-sensitive item sets from the database.

An important problem in data mining is to find a balance between privacy protection and knowledge discovery in the sharing process. Jieh-Shan Yeh, Po-Chiang Hsu [12] have paid attention on privacy preserving utility mining (PPUM) and proposed two algorithms called HHUIF (Hiding High Utility Item First Algorithm) and MSICF (Maximum Sensitive Item sets Conflict First Algorithm), in order to achieve the objective of hiding sensitive item sets so that the opponents cannot extract them from the modified database. They have also reduced the effect on the sanitized database of hiding sensitive item sets. The experimental results have proved that the HHUIF has achieved lower miss costs when compared to MSICF on two synthetic datasets. Alternatively, MSICF algorithm normally has a lower difference ratio than the HHUIF between the original and sanitized databases.

Publishing the data about the individuals, without disclosing their sensitive private information is an important issue. Also, due to proximity and divergence attack, privacy is not 100% safe. Thus, E. Poovammal and M. Ponnavaikko [13] have proposed an approach to design micro data sanitization technique for protecting privacy from malicious attack as well as to protect the utility of the data for any type of mining task. A graded grouping transformation and a mapping table based transformation have been applied on numerical sensitive attribute and categorical sensitive attribute respectively, by the proposed approach. They have performed experiments on adult dataset and compared the results of original and transformed table in order to prove that their proposed task independent technique has the ability to protect the privacy, information and utility.

Generally, two approaches called statistics-based and crypto-based approaches are used to deal with PPDM. One advantage of statistics-based approach is it efficiently handles a huge amount of datasets. Patrick Sharkey *et al.* [14] have proposed a technique for statistics-based PPDM. Their technique was entirely different from the existing techniques because it allows the data owners to share with each other the knowledge models that are mined from their own private datasets, instead of allowing the data owners to publish any of their own private datasets (not even in any sanitized form). Here, the knowledge models obtained from the individual datasets have been utilized to produce some pseudo-data and such data has been then used for extracting the superior “global" knowledge models. There are some technical delicacies while instrumental, so it needs to be carefully addressed. Particularly, they have proposed an algorithm for producing pseudo-data based on the paths of a decision tree, a technique for adapting anonymity measures of datasets to evaluate the privacy of decision trees, and an algorithm to reduce a decision tree in order to assure a given privacy requirement. Through an experimental study performed on different environments with several types of datasets, predictive models, and utility measures have proved that the predictive models learned using the proposed technique are much more precise than those learned using the existing l-diversity technique.

Since the organizations are gathering and sharing data increasingly about their customers, the infringement of customer privacy is increasing very rapidly. Although some of the sharing is for the use of general public such as to identify the disease behavior in medical research, individuals are worried about the intrusion of their privacy. To avoid such violation, the sensitive attributes of data are mapped to another domain such that the original values are not disclosed and yet the original associations are preserved. Mukkamala, R *et al.* [15] have compared a set of fuzzy-based mapping methods in terms of their privacy-preserving property and their potential to preserve the same relationship with other fields. Particularly, their contribution is on four fronts: 1) Alteration in the fuzzy function definition, 2) Introducing seven ways to merge the diverse functional values for a data item into a single value, 3) Comparing the original data with the mapped data using various similarity metrics, 4) Evaluating the effect of mapping on derived association rule.

The k-anonymity based method is the most commonly used method in PPDM, for achieving data mining objectives while preserving privacy. The most common approach for achieving compliance with k-anonymity is to restore certain values with less specific but semantically reliable values. Nissim Matatov *et al.* [16] have introduced an approach for achieving k-anonymity by dividing the original dataset into many projections so that each one of them follows k-anonymity. Furthermore, any effort to rejoin the projections, results in a table that still adheres to k-anonymity. A classifier has been trained on each projection and then, an unlabelled instance has been classified by combining the classifications of all classifiers. Based on classification accuracy and k-anonymity constraints, a genetic algorithm has been used by the proposed data mining privacy by decomposition (DMPD) algorithm to seek the best feature set partitioning. Ten different datasets have been used with DMPD to evaluate its classification performance with other k-anonymity-based methods. The results have shown that performance of DMPD was better when compared to other existing k-anonymity-based algorithms and there is no need for using domain dependent knowledge. They have also evaluated the tradeoff between the two inconsistent objectives in PPDM: privacy and predictive performance, by using multi-objective optimization techniques.

Since the total number of traffic data in networks has been increasing at a shocking rate, a substantial body of research has been made that tries to mine the traffic data in order to obtain the valuable information. For example, there are some studies based on the identification of Internet worms and trespasses by determining the abnormal traffic patterns. However, as the network traffic data have the information about the Internet usage patterns of users, network users' privacy may be weakened during the mining process. Seung-Woo Kim *et al.* [17] have proposed a robust technique, which preserves the privacy during the sequential pattern mining on network traffic data. Their proposed technique has used an N-repository server model, which operates as a single mining server and a retention replacement method, which changes the answer to a query probabilistically so as to find the frequent sequential patterns without breaching privacy. Also, the technique has expedited the overall mining process by maintaining the meta tables in each site in order to find out quickly whether the candidate patterns have ever occurred in the site or not. The accuracy and effectiveness of their proposed technique have been shown by performing experiments using real-world network traffic data.

In recent days, different methods based on random perturbation of data records have been introduced for protecting the privacy of the user in data mining process. K. Srinivasa Rao and V. Chiranjeevi [21] have concentrated on an improved distortion process, which attempts to improve the accuracy by selectively altering the list of items. In typical distortion process, tuning the probability parameters for balancing the privacy and accuracy parameters was very difficult, and the presence or absence of each item was altered with an equal probability. But, in this improved distortion method, the frequent one item-sets and non-frequent one item-sets have been modified with a diverse probabilities controlled by two probability parameters such as *fp, nfp* respectively. These two probability parameters (*fp* and *nfp)* have been tuned flexibly by the proprietor of the data based on his/her requirement for privacy and accuracy. The experiments performed on real time datasets have proved that there is a considerable increase in the accuracy at a very marginal cost in privacy.

3. PROBLEM FORMULATION

The problem of mining of privacy preserving utility itemsets is discussed and some basic definitions are described in this subsection. Let $I=\{i_1,i_2,...,i_m\}$ be a set of items and $DB=\{t_1,t_2,...,t_n\}$ be a transaction database where the items of each transaction t_i is a subset of I. Also, let $I*$ denote a set of sensitive items that need to be hidden according to some security policies, i.e., $I*$ is a subset of I. The problem is to transform *DB* into a sanitized database *DB** such that only the patterns belong to $I*$ can be mined from *DB**. The sanitized database is the database, which has been modified for hiding sensitive patterns with privacy concern. Moreover, the non-sensitive patterns should be preserved as many as possible. The utility of item i_p in transaction t_q, denoted as $U(i_p,t_q)$ is defined as $Iu(i_p,t_q)\times Eu(i_p)$. Let an itemset X be a subset of I. The utility of X in transaction t_q, denoted by $U(X,t_q)$ is defined as $U(X,t_q)=\sum_{i_p\in X}U(i_p t_q)$.

The task of high utility mining is to find all items that have utility above a user-specified min_utility. Since utility is not anti-monotone, the concept of

Frequency Weighted Utility (FWU) is used to prune the search space of high utility itemsets.

Definition: The internal utility or local transaction utility value $Iu(i_p, t_q)$ represents the quantity of item i_p in transaction t_q. The external utility $Eu(i_p)$ represents the unit profit value of item i_p.

Definition: Utility $U(i_p, t_q)$ is the quantitative measure of utility for item i_p in transaction t_q defined by $U(i_p, t_q) = Iu(i_p, t_q) \times Eu(i_p)$.

Definition: The utility of an itemset X in transaction t_q, $U(X, t_q)$, is defined by $U(X, t_q) = \sum_{i_p \in X} U(i_p, t_q)$; where $X = \{i_1, i_2, ...i_k\}$ is a k-itemset, $X \subseteq t_q$ and $1 \le k \le m$.

4. PROPOSED METHODOLOGY FOR MINING OF PRIVACY PRESERVING HIGH UTILITY ITEMSETS

Privacy preserving against mining algorithms is a new research area that examines the side effects of data mining techniques that obtained from the privacy diffusion of persons and organizations [10]. The objective of PPDM algorithms is to mine the significant knowledge from huge amounts of data while preserving sensitive personal information at the same time. Recent research made in this area has given much effort to establish a trade-off between the right to privacy and the need of knowledge discovery. It is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Hence, it is important to give a complete view on a set of metrics related to existing privacy preserving algorithms so that we can gain insights on how to design more robust measurement and PPDM algorithms [20]. The utility of the data, at the end of the privacy preserving process, is an important problem, because in order for sensitive information to be hidden, the database is essentially modified by the insertion of forged information or by the blocking of data values. It is important to note here that some of privacy preserving techniques, like the use of sampling, do not change the information stored in the database, but still, the utility of the data falls, because the information is not complete in this case. Thus, an evaluation parameter for the data utility should be the amount of information that is lost after the application of privacy preserving process. Also, the measure employed to calculate the information loss depends on a particular data mining technique with respect to which a privacy algorithm is performed.

Issues regarding privacy-preserving data mining have emerged globally. The recent development in PPDM techniques is evident. Assume that, we have a server and multiple clients, where each client has a set of data. The clients require the server to collect statistical information about the relationship among items in order to provide recommendations to the customers. But, the clients do not want the server to know about some sensitive patterns. Sensitive pattern is the frequent itemset that have a sensitive knowledge. So, when a client sends its database to the server, some sensitive patterns are concealed from its database based on some particular privacy policies. Hence, the server only can collect the statistical information from the modified database [23].

By considering these facts, in this paper, we have presented an efficient approach for mining of privacy preserving high utility item sets from the sanitized database. The proposed approach is executed with the aid of three major steps:3

1) Transformation of Original database into sanitized database.
2) Construction of sensitive utility FP-Tree using sanitized database.
3) Mining of sensitive utility item sets from the sensitive utility FP-tree.

1) Transformation of Original database into sanitized database

In general, the construction of the FP-tree and the mining of frequent patterns from the FP-tree are the major important steps in the frequent pattern tree algorithm. Here, the proposed approach includes the sensitive items and the utility factors while mining the high utility patterns. Before the construction of the FP-tree, the ordering of the transaction is important, since each path of the FP-tree follows it. Here, the ordering mainly depends on the frequency weighted utility FWU of an item along with the sensitive factor. At first, the sensitive item is hidden by decreasing the FWU by the computed sensitive factor. Then, the items are sorted out in descending order for each transaction based on the frequency weighted utility value and the items which have utility values are being in ordered transaction.

Definition: Frequency-weighted utility FWU of an item i_p, denoted by $FWU(i_p)$, is computed using the transaction frequency (TF), transaction weightage and the external utility.

$$FWU(i_p) = \frac{TF(i_p) * TW(i_p) * EU(i_p)}{U_F}$$

Definition: The transaction frequency of an item $TF(i_p)$ denotes the actual number of occurrences of i_p in all the transactions.

Definition: Transaction weightage $TW(i_p)$ is defined as the overall quantity of the item i_p in all transactions.

Definition: Utility factor U_F is the overall sum of the profit of each items present in the database.

After the computation of the frequency weighted utility of the items, we have pushed to hide the sensitive information for that we need to find the sensitive items from the original database since the original database contains sensitive items. Thus, in order to ensure the privacy, we have incorporated the computation of the sensitive factor for all sensitive items for hiding the information by sanitization process. The collected original data *DB* is transformed into the sanitized data *DB** with the help of the sensitive factor subsequently that *DB**can be shared and published. The transformation, called sanitization, ensures that privacy of individuals is not compromised and that the data is useful for analytical purposes. Privacy is usually measured using some form of disclosure risk, while the data utility is traditionally measured as information loss between the original data and the transformed sanitized data. The sensitivity factor α is computed with the aid of the following formula,

$$sensitive\ \ factor, \alpha = \frac{\sum_{1=1}^{m'} TW(i'_p) + Su(i'_p)}{\sum_{1=1}^{m'} Eu(i_p)}$$

Where, n is the number of transactions, m' is the number of sensitive items and c is the constant and m is the number of items in the database.

Definition: If i_p is said to be a sensitive item, it should satisfy the condition, $FWU(i_p) \geq \min\_util$.

Definition: (sensitive itemsets) Let $DB = \{t_1, t_2, ..., t_n\}$ be a set of transactions, ε be the minimum utility threshold, and *L* be a set of all high utility itemsets for each item. Let $I^* = \{i'_1, i'_2, ..., i'_m\}$ be a subset of *L*, where , called sensitive itemset, is an itemset that should be hidden according to some security policies.

Definition: The sensitive utility weightage $Su(i'_p)$ is denoted as the average value of the external utility of the sensitive items.

The items in the transactions have been ordered based on the sensitive factor. Initially, we calculate the Frequency-weighted utility FWU of all the items in the transactions. Subsequently, we identity the sensitive item with the help of min-utility value after that the sensitive factor is computed only by considering the sensitive items to hide the sensitive information. The sensitive item is hidden by decreasing the FWU value by the sensitive factor. The example of a transaction database is given in table 1, Table 2 shows the external utility value of all the items.

Table. 1. Example Of A Transaction Database

Item TID	A	B	C	D	E
01	2	2	0	1	0
02	3	0	2	2	0
03	0	2	3	1	0
04	5	5	1	0	1
05	0	0	2	0	2

Table 2. External Utility Values

Item	Profit ($)
A	25
B	20
C	15
D	10
E	5

For example: Let us consider database consist the items present in table 1. The FWU of item 'A', is computed as follows: the $TF(i_p)$ of the item 'A' is 3, the $TW(i_p)$ is 10 and the profit records for that item is 25. Also, the total profit value, called utility factor is found out to be 75 in this case. Now, $FWU(A) =$ 10, $FWU(B) =$ 7.2, $FWU(C) =$ 6.4 and $FWU(D) =$ 1.6 and $FWU(E) =$ 0.4. Based on the above FWU values we make the ordered database, the following table 3 shows that the ordered database.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3. Ordered Database

TID	Frequent items
01	ABD
02	ACD
03	BCD
04	ABCE
05	CE

Now consider the value of the min-utility as 6.5 then the sensitive items are A and B. we need to reduce the sensitivity of the sensitive items with the help of the sensitive factor α . The value of the sensitive factor is 2.78 now the value. With the help of the sensitive factor, we make the privacy on the sensitive items by reducing the *FWU* value by the sensitive factor until the value of the *FWU* is satisfy the min-utility value, this process is known as sanitization process. The result of the sanitization process becomes sensitive items becomes the normal item sets. Now the $FWU(A) =$ 4.4, $FWU(B) =$ 4.2, the following table 4 shows that the ordered item set of the sanitized database.

Table 4. Ordered Sanitized Database

TID	Frequent items
01	ABD
02	CAD
03	CBD
04	CABE
05	CE

The second major step in the proposed approach notifies the construction of the sensitive utility FP-Tree using sanitized database. Here, the sensitive utility FP-tree is constructed using the frequency weighted utility rather than the frequency value along with the sensitive information. In addition to this, the mining process utilizes pattern growth methodology, where the support is computed based on the frequency weighted utility rather than the frequency. In this step, the utility FP-tree is constructed by inserting the ordered transactions so that it only necessitates two scans on the transaction database and works in a divide and conquer manner. In the first scan, the proposed algorithm generates the 1-length frequent weighted utility items based on the frequent weighted utility measure. In the second scan, the transaction database is compressed into a utility FP- tree. In this section, we describe the construction process of our proposed sensitive utility FP- tree structure based on the transaction weighted utility and the sensitive factor as the frequent items described in table 4.

Example: The insertion of each transaction is processed as follows: In the first transaction, the frequent weighted utility items (A, B, D) are processed. The root of the tree is initially fixed as null. Then, this transaction is attached as a first branch of the root node. Each node of the branch is attached with their frequency weighted utility values. Similarly, the items (C, A, D), (C, B, D), (C, A, B, E), (C, E) are being processed. In the next item, (C, A, D) does not share any prefix path in the previous tree after executing the first transaction, so the new nodes (C: 6.4) is attached with the root node as its child. Also, the other new node (A: 4.44) and (D: 1.6) is created and linked with their children. Subsequently, the next items (C, B, D) shows that the (C: 12.8) shares the same prefix path (C: 6.4), in which B: 4.42 and D: 1.6 are added as their own sub-nodes. Likewise, (C, A, B, E) and (C, E) are added in the sensitive FP-tree with their weighted utility values. The results after the insertion of all the nodes in the sensitive utility FP-tree are shown in fig. 1

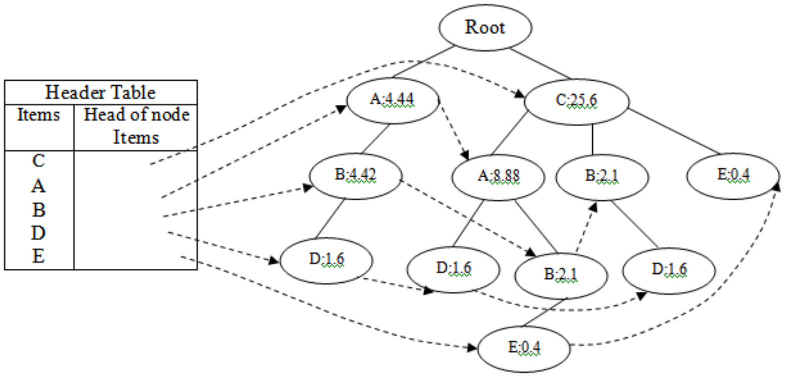

Figure 1. Sensitive Utility FP-Tree

3) Mining of sensitive utility item sets from the sensitive utility FP-tree

Another important step in our proposed methodology is carried out the mining process based on the constructed sensitive utility FP-tree as shown in fig. 2. The mining process of utility itemsets from the sensitive utility FP-tree based on the pattern growth methodology [24] is explained in the following. Sensitive utility FP-tree derives the utility itemsets directly from the sensitive utility FP-tree and do not necessitate generation of candidate itemsets for mining. It recursively processes the utility items one by one and bottom-up with regard to the Header Table. By constructing a conditional utility FP-tree for each utility item, high utility itemsets are mined recursively from it. This process is executed until all the items in the sensitive utility FP-tree get processed. The mining of sensitive patterns involved in two cases,

1) Generation of conditional sensitive utility FP-tree
2) Mining of sensitive patterns

Case 1: Generation of conditional sensitive utility FP-tree

After the sensitive utility FP-tree is constructed from an ordered sanitized database, the mining procedure starts with the generation of the conditional utility pattern base and the conditional utility FP-tree. Here, we start with the mining process from the bottom nodes of the sensitive FP-tree and their corresponding prefix paths are extracted from it. Then, their relevant utility pattern base and conditional utility FP-tree are generated in order to mine n-length sensitive utility patterns. . The conditional pattern-bases and the conditional FP-trees generated are summarized in table 5.

Table 5. *Generated Conditional FP-Trees*

Item	Conditional FP-tree
C	C
A	A, CA
B	B, AB, CB, CAB
D	D, AD, BD, CD, ABD, CAD, CBD
E	E, CE, AE, BE, CAE, CBE, ABE, CABE

Case 2: Mining of sensitive utility patterns

After the generation of the conditional sensitive utility fp-tree, sensitive utility patterns are mined from it based on the minimum support threshold. Here, utility patterns are mined recursively from the conditional utility FP tree so that all length patterns having frequency weighted utility greater than the minimum threshold are obtained. The patterns are said to be frequent weighted utility patterns if their support is greater than min_util. The obtained sensitive utility patterns are shown in table 6.

Table 6. *Shows That The Sensitive Utility Patterns*

Sensitive utility patterns			
A: 13.32	CA: 8.88	CAB: 4.42	CABE: 0.4
B: 13.26	AB: 8.84	ABD: 1.6	
C: 25.6	CB: 4.42	CAD: 1.6	
D: 4.8	AD: 3.2	CBD: 1.6	
E: 0.8	BD: 3.2	CAE: 0.4	
	CD: 1.6	CBE: 0.4	
	CE: 0.8	ABE: 0.4	
	AE: 0.4		
	BE: 0.4		

5. EXPERIMENTAL RESULTS AND DISCUSSUION

The results obtained from the experimentation of the proposed privacy preserving utility itemsets mining with different datasets are presented in this section. Privacy is usually measured using some form of disclosure risk, while the data utility is traditionally measured as information loss between the original data and the transformed sanitized data [22]. We have implemented our proposed utility mining algorithm using Java (jdk 1.6). The dataset utilized in our experimental results are real-world data obtained from various fields and widely-accepted synthetic data.

5.1 Experimental Environment and Dataset Description

This experimental environment of proposed utility mining algorithm is Windows XP Operating system at a 2 GHz dual core PC machine with 2 GB main memory running a 64-bit version of Windows 7. We have tested our algorithm in two different datasets, namely T10I4D100K and Retail [25, 26]. For real life data, we have used Retail dataset, a real market basket data and synthetic data T10I4D100K is obtained from the IBM dataset generator. ***Dataset Description: T10I4D100K:*** This dataset contains 100,000 transactions and 999 distinct items. T10I4D100K denotes the Average size of the transactions (T), Average size of the maximal potentially large itemsets (I) and the number of transactions (D). ***Retail Dataset:*** This dataset contains 88,162 transactions and 16,470 distinct items. This dataset was donated by Tom Brijs and contains the (anonymized) retail market basket data from an anonymous Belgian retail store. Table 7 gives the descriptions of the two test datasets.

Table 7. *Test Dataset Description*

Dataset	Size	No. of transactions	No. of Items

T10I4D100K	3.93MB	100,000	999
Retail	4.07MB	88,162	16,470

5.2 Evaluation Metrics

The hiding failure parameter is evaluated by the percentage of sensitive information that is revealed before and after the sanitizing process. Majority of the developed privacy preserving algorithms are designed with the aim of achieving zero hiding failure. Hence, they conceal all the patterns considered as sensitive. However, it is eminent that if we conceal more sensitive information, then more non-sensitive information we miss. To evaluate the efficiency of the proposed algorithms, our investigation use the performance measures introduced in [12]. The performance measures are described as follows:

1) ***Miss cost (MC):*** The difference ratio of valid item sets found in the original and the sanitized databases. The miss cost is calculated as follows:

$$MC = \frac{|\sim U(ODB') - \sim U(ODB)|}{|U(ODB)|}$$

Where ~ *U(ODB)* and *~(ODB')* indicate the non-sensitive item sets revealed from the ordered database *ODB* and sanitized ordered database *ODB'*, respectively.

2) ***Database difference ratio (DBDR):*** the difference ratio between the original database DB and the sanitized database DBDR is given by:

$$DBDR = \frac{|U(ODB') - U(ODB)|}{No.of\ items}$$

a) Performance evaluation

The experimentations on both datasets T10I4D100K and Retail datasets are undertaken by the privacy preserving sensitive utility pattern mining. With different levels of min-utility value, Miss Cost and Database Different Ratio are evaluated for above large database are analyzed.

b) Retail

The performance of the retail dataset is obtained by varying the min-utility value. The following figures shows that the performance of the evaluation metrics. From the following figure 2, the miss cost has decreased when the min-utility get increase also from the figure 3 we conclude that the variation of min-utility value didn't affect the database different ratio.

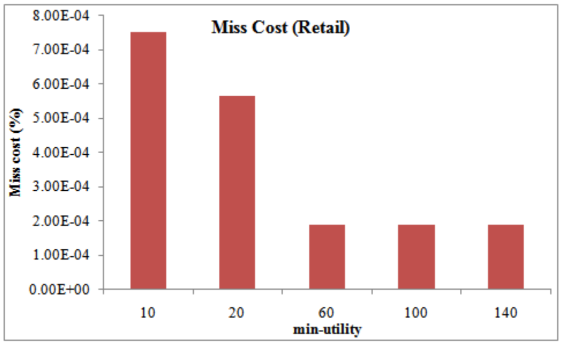

Figure: 2 Shows That The Miss Cost Of Retail Dataset

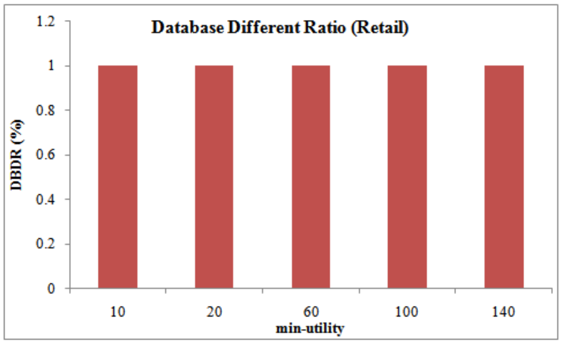

Figure: 3 Shows That The Database Different Ratio Of Retail Dataset

c) T10I4D100K

The performance of the T10I4D100K dataset is obtained by varying the min-utility value. The following figures shows that the performance of the evaluation metrics of the T10I4D100K dataset. From the following figure 4, the miss cost has decreased when the min-utility get increase also from the figure 5, the database different ratio is decreased when the min-utility value get increase up to the min-utility value reaches 100, for the min-utility value 140 the database different ratio get increase.

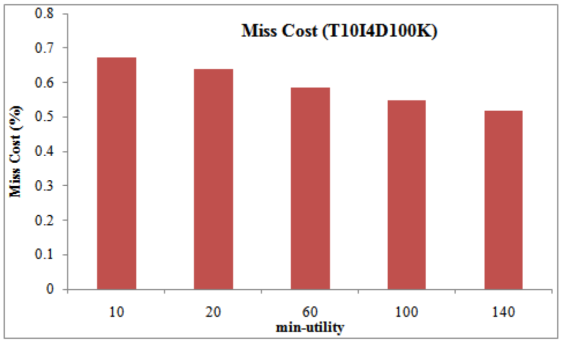

Figure: 4 Shows That The Miss Cost Of T10I4D100K Dataset

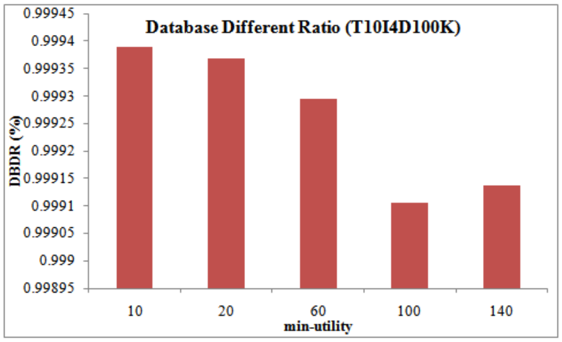

Figure: 5 Shows That The Database Different Ratio Of T10I4D100K Dataset

6. CONCLUSION

In this paper, we have incorporated the privacy preserving concept into the previously developed weighted utility mining approach. In this, we have presented an efficient algorithm for mining of privacy preserving high utility item sets by considering the sensitive item sets. The sensitive items are found based on the FWU value subsequently we make the privacy on the sensitive item by reducing the value of the FWU value with the help of the sensitive factor. The algorithm comprise of three major steps to attain the aim of our research includes, 1) Data sanitization, 2) Construction of sensitive utility FP-tree and, 3) Mining of sensitive utility item sets. The experimentation has carried out using real dataset and the performance of the proposed algorithm is evaluated with the aid of the evaluation metrics such as Miss Cost and Database difference ratio.

REFERENCES

[1] Rajesh Kumar Boora, Ruchi Shukla, A. K. Misra, "An Improved Approach to High Level Privacy Preserving Itemset Mining", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[2] Yao H., Hamilton H. J., Geng L., "A Unified framework for Utility based Measures for Mining Itemsets", Second International Workshop on Utility-Based Data Mining, Philadelphia, Pennsylvania, 2006.

[3] Yao H., Hamilton, H. J. and Butz, C. J., “A Foundational Approach to Mining Itemset Utilities from Databases”, in proceedings of SIAM International Conference on Data Mining, pp. 482-486, 2004.

[4] Shibnath Mukherjee, Zhiyuan Chen, Aryya Gangopadhyay, "A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms", The VLDB Journal, Vol: 15, No. 4, pp: 293–315, 2006.

[5] Murat Kantarcioglu, Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1026 – 1037, 2004.

[6] R. Agrawal and R. Srikant. “Privacy-preserving data mining”, In Proc. of ACM SIGMOD’00, pp: 439–450, Dallas, Texas, USA, 2000.

[7] Jie Wang, Jun Zhang, Shuting Xu, Weijun Zhong, "A Novel Data Distortion Approach via Selective SSVD for Privacy Protection,” International Journal of Information and Computer Security, Vol. 2, No. 1, pp. 48-70, 2007.

[8] Verykios,V. S., Bertino, E., Fovino,I. N., Provenza, L. P., Saygin, Y., Theodoridis, Y. “State-of-the-art in privacy preserving data mining,” SIGMOD Record, Vol. 33, No. 1, pp:50-57, 2004.

[9] Bertino, E., Fovino, I. N., Provenza, L. P., ‘A framework for evaluating privacy preserving data mining algorithms’, Data Mining and Knowledge Discovery, Vol. 11, No. 2, pp.121-154, 2005.

[10] Mohammad Naderi Dehkordi, Kambiz Badie, Ahmad Khadem Zadeh, "A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms", Journal on Software, vol. 4, no. 6, pp: 555- 562, 2009.

[11] R.R.Rajalaxmi, A.M.Natarajan, "A Novel Sanitization Approach for Privacy Preserving Utility Itemset Mining",

[12] Jieh-Shan Yeh, Po-Chiang Hsu, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining", Expert Systems with Applications, Vol: 37, pp: 4779–4786, 2010.

[13] E. Poovammal, M. Ponnavaikko, "Privacy and Utility Preserving Task Independent Data Mining", International Journal of Computer Applications, Vol:1, No. 15, pp: 104-111,

[14] Patrick Sharkey, Hongwei Tian, Weining Zhang, and Shouhuai Xu, "Privacy-Preserving Data Mining through Knowledge Model Sharing",

[15] Mukkamala, R.; Ashok, V.G.; , "Fuzzy-based Methods for Privacy-Preserving Data Mining", Information Technology: New Generations

(ITNG), 2011 Eighth International Conference on, Las Vegas, NV, pp: 348 - 353, 2011.

[16] Nissim Matatov, Lior Rokach, Oded Maimon, "Privacy-preserving data mining: A feature set partitioning approach", Information Sciences, Volume 180, Issue 14, 15 July 2010, Pages 2696-2720

[17] Seung-Woo Kim, Sanghyun Park, Jung-Im Won, Sang-Wook Kim, "Privacy preserving data mining of sequential patterns for network traffic data", Information Sciences Volume 178, Issue 3, 1 February 2008, Pages 694-713.

[18] Chin-Chen Chang, Jieh-Shan Yeh, and Yu-Chiang Li, "Privacy-Preserving Mining of Association Rules on Distributed Databases", IJCSNS International Journal of Computer Science and Network Security, Vol.6 No.11, November 2006.

[19] Veloso, A.A., Meira Jr., W., Parthasarathy, S. and de Carvalho, M.B., “Efficient, accurate and privacy preserving data mining for frequent itemsets in distributed databases,” Proceedings of the Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, October, pp.281-292, 2003.

[20] Elisa Bertino, Dan Lin, and Wei Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms",

[21] K.Srinivasa Rao, V.Chiranjeevi, "Distortion Based Algorithms For Privacy Preserving Frequent Item Set Mining", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.4, July 2011.

[22] Michal Sramka, "Data Mining As a Tool in Privacy-Preserving Data Publishing",

[23] Guanling Lee, Chien-Yu Chang and Arbee L.P Chen, "Hiding Sensitive Patterns in Association Rules Mining",

[24] Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, Vol: 8, pp: 53–87, 2004.

[25] M. Sulaiman Khan, Maybin Muyeba, Frans Coenen, "A Weighted Utility Framework for Mining Association Rules", Second UKSIM European Symposium on Computer Modeling and Simulation, Liverpool, pp: 87 - 92, 2008.

[26] Bay Vo, Huy Nguyen, Bac Le, "Mining High Utility Itemsets from Vertical Distributed Databases", in proceedings of the International Conference on Computing and Communication Technologies, Da Nang, pp: 1-4, 2009.

[27] Oliveira, S. R. M., & Zaïne, O. R.,"A framework for enforcing privacy in mining frequent patterns", Technical Report, TR02-13, Computer Science Department, University of Alberta, Canada, June, 2000.