Improving Security and Efficiency in Association Rule Mining using PFP-Growth Algorithm via Transaction Splitting

R. Syed Ali Fathima[1], M. John Basha[2], P.Saravanan[3]

[1]*PG Scholar, Deptartment of CSE,PTR Engineering college, Madurai, India*
[2]*Head of the department, Deptartment of CSE,PTR Engineering college, Madurai, India*
[3]*Assistant Professor, Deptartment of CSE,PTR Engineering college, Madurai, India*

Abstract- *Data Mining is a technique which is used to discover hidden information from a large database. Frequent item set mining is also an important fundamental problem in data mining. Nowadays, most of the researchers are used association rule mining to find correlation between items and items sets resourcefully. Security is also important problem in data mining. To this end, we propose a transaction splitting based on PFP-growth algorithm and frequent items should keep as secured with the help of cryptography algorithms. PFP-Growth algorithm is advanced to FP-growth algorithm. It consists of both preprocessing phase and mining phase. In the preprocessing phase, we used smart transaction splitting method to improve the utility and tradeoff. In the mining phase, the transformed database and user specified threshold value helps to estimate the number of support computations, so that we can gradually reduce the amount of noise required and the information loss caused by transaction splitting. Using frequent item, we find the global association rules based on association rule mining. In this paper, cryptography technique (AES - Advanced Encryption Standard algorithm) is used to secure the frequent item set. Trusted party should preserve the privacy of individual data while the data is distributed among different sites.*

Keywords: *Data mining, frequent itemset mining, transaction splitting, cryptography technique*

I. INTRODUCTION

Data mining is an inter-disciplinary field with roots in enterprise decision support. Data mining is not for analyzing small datasets. It is the task of discovering interesting and hidden patterns/data from large amounts of data where the data can be stored in databases, data warehouses, OLAP or other repository information [1]. Data mining also involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, neural networks, fuzzy and rough set theory, knowledge representation, inductive logic programming, information retrieval and etc. In data mining, the fundamental problem is to find frequent item set in the large database [14].

Frequent item set mining is importance in the wide range of application fields such as bioinformatics, web usage mining etc. A variety of different algorithm has been proposed for finding frequent itemset. The Apriori[2] and FP- Growth algorithm[3],[11] are the most famous algorithm in association rule mining. The Apriori algorithm is bottom-up approach algorithm or level-wise search algorithm. A subset of frequent itemset must also be a frequent itemset i.e if {AB} is a requent itemset, both A and B should be a frequent itemsets, called as Apriori property. The FP-growth algorithm is one of the approaches that eliminates the generation of a large number of candidate itemsets invented by Han *et al* [3]. It performs like a depth-first search algorithm.

In data mining, association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. When data is distributed among different sites, finding the global association rules is a challenging task as the privacy of the individual site's data is to be preserved. In this paper, a model is proposed to find global association rules by preserving the privacy of individual sites data when the data is partitioned horizontally among n number of sites[4].

Advanced Encryption Standard (AES) algorithm is not only used for security but also for great speed. Both hardware and software implementation are faster still [5]. AES algorithm is used to encrypt and decrypt the frequent itemsets using data blocks of 128 bits in 10, 12 and 14 round depending on key size. It can be implemented on various platforms specially in small devices and also large companies.

II. RELATED WORKS

A large number of researchers had proposed a private FP growth algorithm for producing accurate frequent items from the large database. Sen su et al.[6] proposed a differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. In this paper, PFP-growth algorithm consists of a preprocessing phase and a mining phase. In the preprocessing phase, a novel smart splitting method is proposed to transform the database and also

it improve the utility and privacy tradeoff. For a given database, the preprocessing phase needs to be performed only once. In the mining phase, to offset the information loss caused by transaction splitting, we devise a run-time estimation method to estimate the actual support of itemsets in the original database. There is another set of researchers studies on mining frequent patterns with differential privacy in sequence and graph databases based on rule mining. Shen and Yu [8] address the problem of mining frequent graphs under differential privacy. They integrate the process of the privacy protection and frequent graph mining into a Markov Chain Monte Carlo framework. Besides, two studies have been proposed for publishing sequence databases under differential privacy. Bonomi and Xiong [7] proposed a two-phase algorithm for finding frequent itemset. It privately mining both prefixes and substring patterns in the sequence database. Chen et al. [9] leverage a hybrid-granularity prefix tree and conduct constrained inferences. In [10], the authors utilize variable length n-grams to represent the necessary information in the sequence database. They use a tree to group grams and adaptively allocate privacy budget to compute the noisy counts of nodes in the tree. Different from [10], we use FPtree to generate conditional pattern bases and add noise to the support of itemsets to avoid privacy breach. Based on the above paper, we tried a private FP growth for finding frequent items in the large database. Also add cryptography technique to secure that frequent items from the third party.

III. ASSOCIATION RULE MINING

Association rule mining is a very important research topic in the data mining field. The concept of Association Rules was given by Agrawal, Imielinski, and Swami in 1993 [11]. It is a simple rule for data application and an interdependence relationship among the data objects. It is well studied and most widely used in a variety of data mining applications. The original problem addressed by association rule mining was to find a correlation among sales of different products from the analysis of a large set of supermarket data. However Association Rule Mining can also applied for other data structures, gene sequences, protein sequences. A typical association rule will be of the structure $X \rightarrow Y$ which states that, for every instance of X is true, Y is also true, where X and Y are sets of items. The level of significance of an association rule is generally measured by two indicators, support and confidence [12].

The goal of association rule mining is to discover all rules that have support and confidence greater than some user-defined minimum support and minimum confidence thresholds, respectively. Two parameters as following used to describe the properties of association rules:

1.Support: The rule $A \Rightarrow B$ has a support (denoted as supp) s in database if s% of the transactions in database contains $A \cup B$.

$$supp(A \Rightarrow B) = supp(A \cup B) = P(A \cup B)$$

2.Confidence: The rule $A \Rightarrow B$ has a measure of its strength called confidence(denoted as conf) i.e. the transactions in database that contain A also contain B.

$$conf(A \Rightarrow B) = P(B \mid A) = supp(A \cup B) / supp(A)$$

The normally followed scheme for mining association rules consists of two stages [11]:

1. The discovery of frequent itemsets
2. The generation of association rules

A. Defintion of Private FP growth Algorithm

The PFP-growth algorithm consists of preprocessing phase and mining phase. In the preprocessing phase, we extract some statistical information from the original database and leverage the smart splitting method to transform the database shown in Fig.1. Notice that, for a given database, the preprocessing phase is performed only once. In the mining phase, for a given threshold, we privately find frequent itemsets. The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.

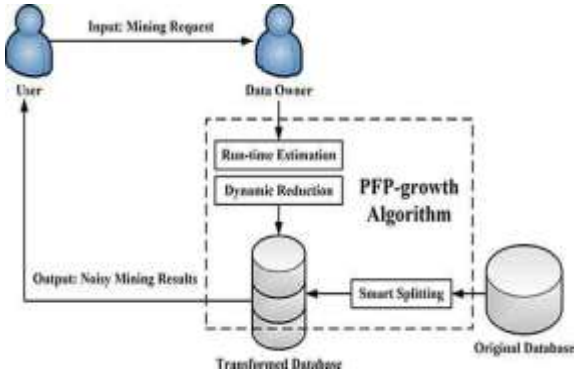

Fig.1 PFP algorithm process

IV.CRYPTOGRAPHY TECHNIQUE

Cryptography technique is used to secure the data from third parties. It contains different types of algorithm such as AES, DES and RSA. Encryption is a well known technology for protecting sensitive data in data mining. It is used to encrypt the plain text into cipher text. Use of the combination of Public and Private Key encryption to hide the sensitive data of users, and cipher text retrieval [13].

A. Advanced Encryption Standard Algorithm

Advanced Encryption Standard (AES) algorithm not only for security but also for great speed. Both

hardware and software implementation are faster still. Encrypts data blocks of 128 bits in 10, 12 and 14 round depending on key size as shown in Fig 2. It can be implemented on various platforms specially in small devices. It is carefully tested for many security applications.

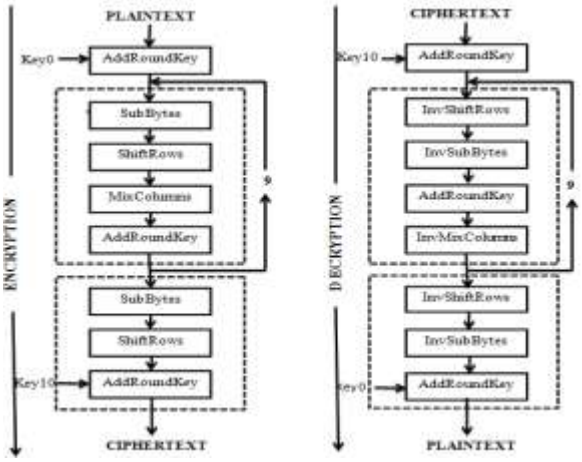

Fig. 2 AES algorithm process

V. PROPOSED METHODOLOGY

In the proposed methodology, we present our private FPgrowth (PFP-growth) algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, we transform the original database to limit the length of transactions. It also irrelevant to user specified thresholds and needs to be performed only once for a given database[6]. The long transactions should be split rather than truncated. That is, if a transaction has more items than the limit of database, we divide it into multiple subsets (i.e., sub-transactions) and guarantee each subset is under the limit. In the mining phase, given the transformed database and a user-specified threshold, to find frequent itemsets. During the mining process, we dynamically estimate the number of support computations, so that we can gradually reduce the amount of noise required by differential privacy. Runtime estimation method to quantify the information loss caused by transaction splitting Dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. Finally we generate a association rule based on FP growth algorithm. Then, the frequent item should encrypt by AES algorithm.

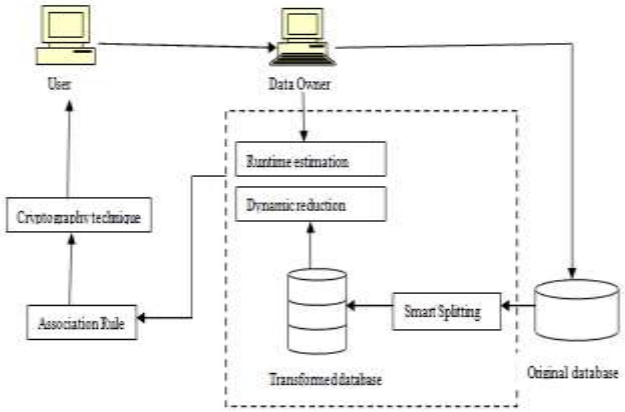

Fig. 3 Proposed method

A.Pseudo code

The PFP-growth algorithm consists of two phases.

Step 1: Preprocessing phase

Extract some statistical information from the original database and leverage the smart splitting method to transform the database.

Notice that, for a given database, the preprocessing phase is performed only once.

Step 2:Mining phase

Given threshold value, we privately find frequent itemsets.The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.

Besides, we divide the total privacy budget ϵ into five portions:

$\epsilon 1$ is used to compute the maximal length constraint,

$\epsilon 2$ is used to estimate the maximal length of frequent itemsets,

$\epsilon 3$ is used to reveal the correlation of items within transactions,

$\epsilon 4$ is used to compute μ-vectors of itemsets, and

$\epsilon 5$is used for the support computations.

PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy

Step 3: Rule generation

Two parameters as following used to describe the properties of association rules:

1.Support: $supp(A \Rightarrow B) = supp(A \cup B) = P(A \cup B)$

2.Confidence:

$conf(A \Rightarrow B) = P(B \mid A) = supp(A \cup B) / supp(A)$

These steps are used to find the association rule for frequent items.

Step 4: Cryptography technique(AES algorithm)

Association rule are encrypted using AES algorithm

These steps used to encrypt 128-bit block

1. The set of round keys from the cipher key.
2. Initialize state array and add the initial round key to the starting state array.
3. Perform round = 1 to 9 : Execute Usual Round (1.Sub Bytes 2. Shift Rows 3. Mix Columns 4. Add Round Key , using K(round))

4. Execute Final Round(1.Sub Bytes, 2. Shift Rows 3. Add Round Key, using K(10))
5. Corresponding cipher text chunk output of Final Round Step

Encryption :

1.Sub Bytes : The first transformation, Sub Bytes, is used at the encryption site. To substitute a byte, we interpret the byte as two hexadecimal digits.

2. Shift Rows : In the encryption, the transformation is called Shift Rows.

3. Mix Columns : The Mix Columns transformation operates at the column level; it transforms each column of the state to a new column.

4. Add Round Key : Add Round Key proceeds one column at a time. Add Round Key adds a round key word with each state column matrix; the operation in Add Round Key is matrix addition.

Decryption:

Decryption involves reversing all the steps taken in encryption using inverse functions like a) Inverse shift rows, b) Inverse substitute bytes, c) Add round key, and d) Inverse mix columns.

VI.EXPERIMENTAL RESULTS

In these experimental results, we collect sample data from the supermarket and performed it by proposed methods. The below figures shows the proposed method result.

Fig.4 Transaction data

Fig. 5 Split the transaction

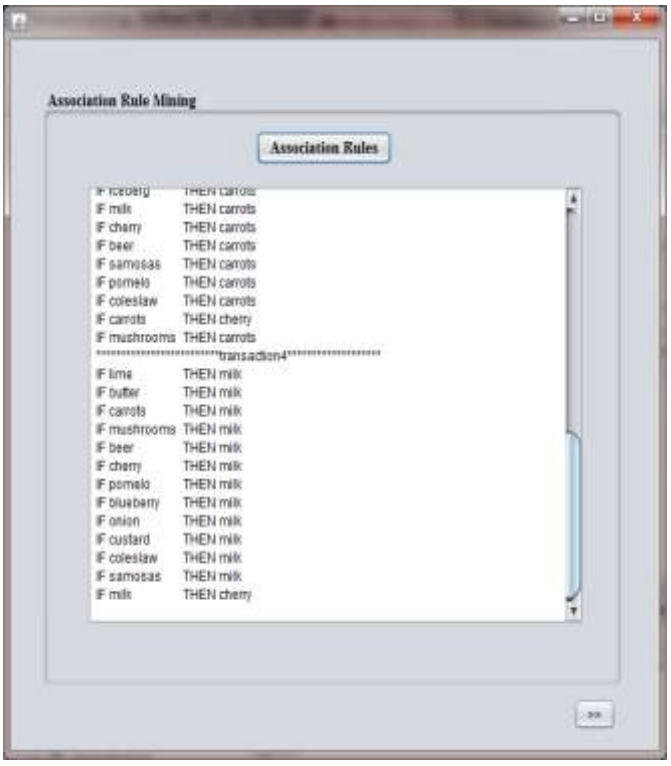

Fig 6 Association Rule generation

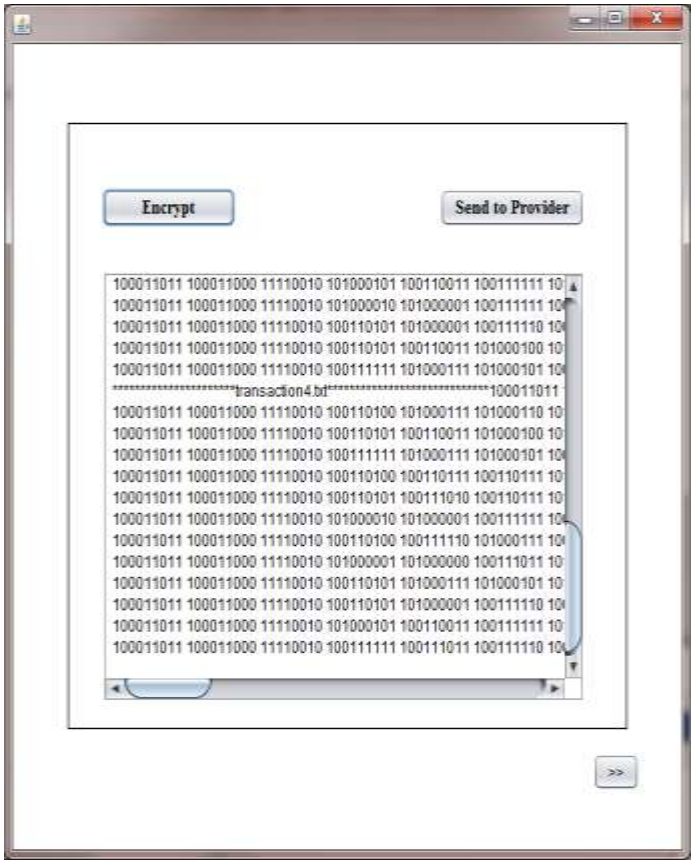

Fig. 7 Encrypt the association rule

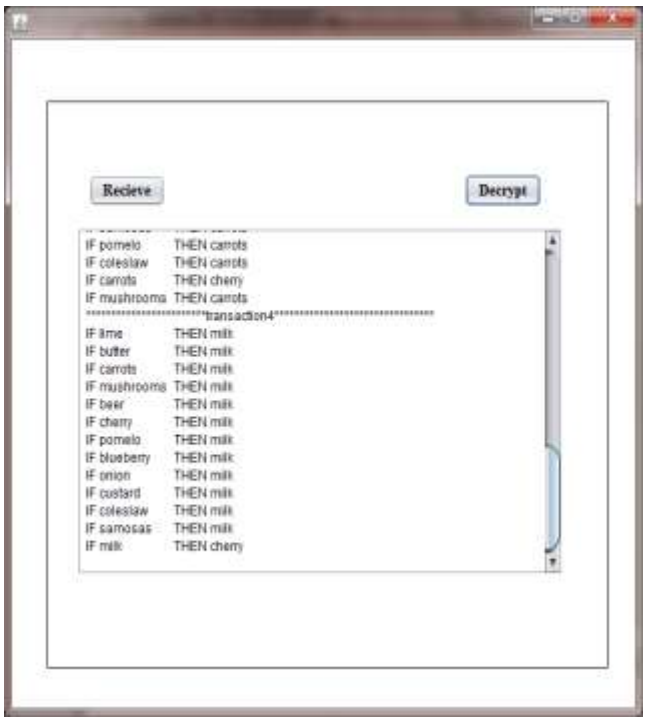

Fig. 8 Decryption process

In this process, we have to select the transaction data for processing. Based on PFP growth algorithm we find the frequent item. Association rules are generated based for frequent items. AES algorithm used to encrypt and decrypt that association rule.

7.CONCLUSION

In this paper, we examine the problem of designing a private FP-growth algorithm and finding association rule mining. Security is also a major problem in data mining, so we used cryptography algorithm to secure the association rule. We propose private FP-growth algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, to better improve the utility-privacy tradeoff. Smart splitting method used to transform the database. In the mining phase, a run-time estimation method is proposed to offset the information loss incurred by transaction splitting. Dynamic reduction method used to dynamically reduce the amount of noise added to guarantee privacy. We add cryptography technique (AES - Advanced Encryption Standard algorithm) to secure the frequent item set. Trusted party should preserve the privacy of individual data while the data is distributed among different sites. Formal privacy analysis and the results of extensive experiments on real datasets show that our PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

REFERENCES

1. Frawley,W.,Piatetsky-Shapiro,G.,Matheus,C.(1992) Knowledge Discovery in Databases: An Overview.AI Magazine, Fall 1992,pp.213-228.
2. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
3. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD Int. Conf. Manage.Data, 2000, pp. 1–12.
4. N.V.Muthu Lakshmi, and K.sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies, Vol. 3 (1), pp. 3176 – 3182, 2012.
5. Prerna Mahajan and Abhishek Sachdeva, "A Study of Encryption Algorithms AES, DES and RSA for Security", Global Journal of Computer Science and Technology Network, Web & Security Vol.13(15), Version 1.0, pp.14-22,2013.
6. S.Sen , X. Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Yang, "Differentially Private Frequent Itemset Mining via Transaction Splitting", IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 7, pp.1875-1891, 2015.
7. L. Bonomi and L. Xiong, "A two-phase algorithm for mining sequential patterns with differential privacy," in Proc. 22nd ACM Conf. Inf. Knowl. Manage., 2013, pp. 269–278.
8. E. Shen and T. Yu, "Mining frequent graph patterns with differential privacy," in Proc. 12th CM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 545–553.
9. R. Chen, B. C. M. Fung, and B. C. Desai, "Differentially private transit data publication: A case study on the montreal

transportation system," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 213–221.

10. R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in Proc. ACM Conf. Comput. Commun. Security, 2012, pp. 638–649.
11. Agrawal.R and Srikant.R. Fast algorithms for mining association rules in large databases. In Proc. 20th VLDB,Sept. 1994.
12. Lin D.I and Kedem Z.M. "Pincer-Search : An Efficient Algorithm for Discovering the Maximal Frequent Set",Knowledge and Data Engineering IEEE, pp: 553-566, 1999.
13. Padmapriya, Dr.A, Subhasri, P. "Cloud Computing: Security Challenges & Encryption Practices". International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 3, pp. 257, March 2013.
14. M.A.Santhi, "Application of Data Mining Using Snort rule for intrusion detection", SSRG International Journal of Computer Science and Engineering, Volume 1, Issue 8, 2014.
15. B.Muruganantham and Ankita Dubey, "Outlier Detection Using Distributed Mining Technology In Large Database", SSRG International Journal of Computer Science and Engineering, Volume 2, Issue 2, 2015.