A Survey on Discovering Frequent Item Set Mining using Private Frequent Pattern Algorithm

Khude Tejashree Vishnu[1] Dr.D.S.Bhosale[2]
[1,2]Department of Computer Science & Engineering
[1,2]Ashokrao Mane Group of Institutions, Vathar-416112, India

Abstract— Frequent Item sets Mining (FIM) is the most well-known techniques to extract knowledge from dataset. Many algorithms to analyze frequent item set are discovered. Previously Apriori and Frequent Pattern growth (FP-growth) algorithms were used for frequent item set mining but due to some disadvantages such as apriori needs candidate set generation and FP-Growth requires two database scans these algorithms failed to achieve accuracy in privacy and time utility. Private Frequent pattern growth algorithm is proposed to gain not only high data utility and high degree of privacy but also high time efficiency in the database using transaction splitting.

Key words: Frequent Pattern Algorithm, Frequent Item Set Mining

I. INTRODUCTION

In recent years the size of database has increased rapidly. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data. The term data mining or knowledge discovery in database has been adopted for a field of research dealing with the automatic discovery of implicit information or knowledge within the databases. Frequent item sets are appearing in a data set with frequency greater than a user-specified threshold. For example, a set of items, such as mobile phone and memory card that appear frequently together in a transaction data set is a frequent item set. But releasing discovered frequent item sets may hammer individual privacy if the data is sensitive example web browsing history and military records. Sequential pattern mining is define to find statistically relevant pattern. For example the customer buying first a computer, modem and pendrive if it occur frequently in a shopping history database then it is said to be sequential pattern. The existing system has problem of privacy threats and large time complexity. Existing system gives comparatively large size output combination. So, to solve this problem, this paper develops a time efficient differentially private FIM algorithm. Data mining, efficiently find valuable, non-obvious information from huge databases, which may be misused. The situation may become worse when there are lots of long transactions in database. In case of frequent item set mining (FIM) when database of transaction is given and each transaction contains a set of items, it tries to find item set that occur most frequently than given threshold. To perform utility on transaction with privacy consideration FP-Growth algorithm with differential privacy is used. The concept of transaction splitting (for long transaction) is used in Private-FP growth (PFP) algorithm. At first database of transaction is transformed so that the records of database get private. In transformed database long transactions are split into sub-transactions and splitting process is performed only once for given database. After that actual support of each item set is calculated in both original and transformed database. Finally Frequent Item set will mine from transformed database. As compared to Differential Privacy Apriori algorithm, the private Frequent Pattern algorithm achieves high privacy and high time efficiency. Private FP-Growth works in two phases. Phase one is pre-processing phase which include splitting of long transaction into multiple subset. Second phase is mining phase which works when transformed database is obtained from pre-processing phase and information loss is reduced.

Here algorithm takes a dataset consisting of the transactions by a group of individuals as an input and produces the frequent item sets as output. This may create considerable threat on privacy concern in publishing the frequent item sets in the dataset. These privacy concern factors ensures that the presence of an individual's data in a dataset does not reveal much about that individual. Here, in this paper we put forth the possibility of developing differentially private frequent item set mining algorithms. Here goal is to improve accuracy of differential privacy without destroying the utility of the algorithm. The utility of a differentially private frequent item set mining algorithm is improved by including all frequent item sets and excluding all infrequent item sets. In the pre-processing phase, some statistical information is extracted from the original database and long transactions are splitted into multiple subset. In the mining phase, to reduce the information loss caused by transaction splitting, it devise a run-time finding method to find the actual support of item sets in the original database.

II. LITERATURE REVIEW

Different algorithm in existing systems are used to find item set which are, 1) Apriori Algorithm-This algorithm uses the association rule to find out frequent item set in transactional database. It uses bottom up approach for finding frequent item set and breadth first search method. 2) UP-Growth- To reduce the number of item sets and produce only high utility item sets UP-Growth algorithm is used. The first step of UP-Growth algorithm is compute Transaction Unit (TU) of each transaction. And then Transaction Weight Unit (TWU) of each single item is also calculated. Then in next step infrequent items are discovered and eliminated from transaction. Further in next step remaining frequent and promising items are arranged in descending order and then those transaction are inserted into Utility Pattern. 3) FP-Growth-The FP-Growth algorithm do not generate the candidate item set and uses a FP- tree structure to store item set frequency information. It is top-down approach and depth-first search method. FP-Growth works like divide and conquer method. It requires two scans on the database. In first scan the frequent items are inserted into the Header Table (HT) and sorted in decreasing order of their supports. Then in the second database scan, FP-tree for the given database is constructed. For the frequent items in each transaction, they are arranged according to the order of HT

and inserted into FP-tree as a branch. 4) Frequent item set mining- A frequent item set mining algorithm takes as input a dataset consisting of the transactions by a group of individuals, and produces as output the frequent item sets. But privacy in case of not revealing private data or information of individual is utilised here.

Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu [1] outline study of two algorithms utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility item sets with a set of effective strategies for removing candidate item sets. Here tree-based data structure named as UP-tree is maintained which contains information about high utility item set. Although this algorithms improved runtime in case of databases containing many long transactions, but the problem of huge memory usage for constructing and visiting conditional trees is considerable.

J.Vaidya and C.Clifton [2] outlined problems of association rule mining. Here database is vertically partitioned and global frequent item sets is founded using secure computation of scalar products. This paper addresses problem of association rule mining where transactions are distributed across sources. Vertically partitioned, means that each site contains some elements of a transaction. For example, one site may contain electronic equipment purchases, and another has computer accessories purchases. Using a key such as credit card number and date, it can join these to identify relationships between purchases of electronic equipment and computer accessories. However, this discloses the individual purchases at each site, possibly violating consumer privacy agreements. Clifton and Kantarcioglu [12] consider the database is horizontally partitioned and proposed the problem as a secure multi-party computation.

J. Han, J. Pei, and Y. Yin,[13] outline existing mining algorithms. Previously Apriori which include candidate set generation approach was used. But in case of long patterns candidate-set-generation is costly. So to overcome this disadvantage frequent-pattern tree (FP-tree) structure is put forth where large database are compressed and only two database scan were done. Here frequent item sets are found without candidate-set-generation. FP growth algorithm use both horizontal and vertical database to store data set in main memory. FP growth algorithm is based on the depth first search method.

Prof. Maurizio Atzori, F. Bonchi, F. Giannotti [3] outlined algorithm to discover anonymised frequent item set based on k-anonymity concept. They characterized all possible threats to anonymity that will arise from revealing the set of patterns. So depending on this that threats to anonymity cannot be avoided they outlined the elimination of threats using pattern distortion.

C. Dwork [4] has introduced Differential Privacy which is considered as a standard notion of privacy in data analysis. Database A and B are said to be neighbouring databases if they differ by at most one record. On contrary new measure are suggested in, differential privacy, which, captures the increased risk to one's privacy generated by participating in a database.

N. Li, W. Qardaji, D. Su, and J. Cao [5] proposed an approach called PrivBasis to overcome challenges of transactional database with high dimensionality. They represented algorithms for privately constructing all basis set and then using it to find the most frequent item sets.

C. Zeng, J. F. Naughton, and J.-Y. Cai [6], introduced difficulties of finding good utilities and privacy and also they have proposed differentially private algorithm for the top-k item set mining. Here long transactions are truncated. All the item sets are founded whose supports exceed the given threshold. In general difficulties occur during processing of long transaction so transactions that contain more items are truncated, trading off errors introduced by the truncation with those introduced by the noise added to guarantee privacy.

Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke [7], proposed a work for mining association rules from transaction consisting of categorical items where the data has been randomized to maintain privacy of individual transactions. While it is possible to recover association rules and preserve privacy using a forward uniform randomization. Experimental analysis gives ensurity about the algorithm by applying it on real datasets.

W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis [8], proposed that finding frequent item sets in association rule mining is costly task. This task if outsourced then it may bring several benefits to the data owner such as cost relief and a less obligation to storage and computational resources. If the service provider makes error in the mining process and reduces costly computation, returning incomplete results then results of mining may lost. Thus audit environment for outsourcing is proposed by these authors which consist of dataset transformation and result verification methods. The main component of its audit environment is an artificial item set planting (AIP) technique which ensures appropriateness and accurate verification process. Some analytical and experimental studies represented that the technique is both effective and efficient.

Freddy Chong Tat Chua [11] proposed a social correlation framework. Two Generative models i.e. Sequential Generative model and Unified Generative Model. In this method efficient parameter estimation solution based on the expectation maximization is discussed. This paper focuses on item adoption predication based on the social links Differential private transaction splitting methods.

III. CONCLUSION

Frequent item set is very important to find out from the large data set. Here few algorithms contributes for achieving efficiency of frequent item sets mining ,but these algorithms have some pros and cons, therefore there is necessity to develop such technique to overcome the entire disadvantage to find frequent item and to provide privacy accessing data from the database. Private FP-growth (PFP-growth) algorithm tries to achieve better utility and privacy. This algorithm consists of a pre-processing phase and a mining phase. Pre-processing phase, converts original database into transformed database using transaction splitting. In the mining phase, a run-time estimation method is proposed to reduce the information loss due to transaction splitting. PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

REFERENCES

[1] Cheug-wei wu, Philippe Fournier–viger, Philip S.Yu "Efficient Algorithms For Mining The Concise and Lossless Representation of Closed+High Utility Item sets" pp,487-499 1994.
[2] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 639–644.
[3] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity preserving pattern discovery," VLDB Journal, 2008.
[4] C. Dwork, "Differential privacy," in Proc. Int. Colloquium Automata, Languages Programme., 2006.
[5] Ninghui Li, WahbehQardaji, Dong Su, Jianneng Cao,"PrivBasis: Frequent Itemset Mining with Differential Privacy.", in VLDB, 2012.
[6] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," in VLDB, 2012.
[7] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in KDD
[8] W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "An audit environment for outsourcing of frequent itemset mining," in VLDB,2009.
[9] 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.
[10] E. Shen and T. Yu, "Mining frequent graph patterns with differential privacy," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 545–553.
[11] Freddy Chong Tat Chua, Hady W.Lauw, Ee-peng Lim "Generative Models for Item Adoption using Social Correlation".
[12] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally Partitioned data" IEEE Trans. Knowl. Data Eng., vol. 16, no. 9, pp. 1026–1037, Sep.2004.
[13] 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.

