Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

A Survey Paper on Differentially Private Frequent Item Mining

Ms. Chanchal Rathi
G. S. Moze College of Engineering, Balewadi, Pune-45.
University of Pune,
Pune, India.

Mr. J. Ratnaraj Kumar
G.S.Moze College of Engineering, Balewadi, pune-45.
University of Pune,
Pune, India.

Abstract **- In today's business world there is excess of available data & a great need to make good use of it. Data mining is art of extracting pattern and knowledge from large amount of data. Frequent itemset plays essential role in many Data Mining tasks that try to find out interesting patterns from database such as association rules. Association rule mining is a method of discovering interesting correlations between variables in large databases. Mining of frequent itemset is most popular problem in data mining. The discovery of frequent itemsets can be serve valuable economic & research purpose. But valuable discovered frequent itemsets should not only assure security but also achieve high data utility & offer time efficiency. The frequent itemsets are patterns or items like itemset, substructures or subsequences that comes out in data set frequently. There are several FIM algorithms for frequent item mining such as Apriori, FPgrowth, Eclat, UP growth algorithms.**

To provide security or privacy here we use differentially private FIM algorithm using UP- growth algorithm. It consist of Prepocessing & mining phase.In prepocessing phase, to enhance utility & privacy advance smart splitting method is proposed to transform database. For given Database preprocessing phase should be performed only once. In mining phase run time estimation & dynamic reduction performed. To cover the information loss by smart splitting, we contrive run time estimation to calculate actual support of itemsets in original database. For privacy we have added noise in the database, we put forward dynamic reduction method to reduce the noise dynamically which guarantees privacy during mining process. In this paper we proposed new algorithm for mining high utility itemsets called as UP growth which consider not only frequency of itemset but also utility associated with the itemset.

Keywords–Frequent Item Mining, ϵ- differential privacy, FP- growth, UP-growth.

1. INTRODUCTION

Data mining is called as uncovering hidden data in a database. In other word, it is called as also data analysis, data driven determination and deductive finding out. Among the areas of data mining, the problem of extracting associations from data has received a great deal of awareness. Association rules are used to identify correlations among a set of items in database. These correlations are not based on immanent properties of the data themselves (as with functional dependencies), but rather based on occurrence of the data items.

Association Rule Mining discovers application in market basket analysis. The market analysts would be focused in discovering frequently bought items by customers, so the organization can do effective arrangements of items according to their sales. Two strategically measures that command the association rule mining process are support and confidence. Support is the statistical importance of a rule while confidence is the degree of assuredly of the detective associations the whole association mining process is commanded by two variables, minimum support and confidence which are user defined.

Discovering useful patterns hidden in database plays an important role in different data mining jobs, such as frequent pattern mining, high utility pattern mining. Among them frequent pattern mining is a fundamental research topic, that has been used to different databases, such as transactional databases. It is used in the analysis of purchase of customer transactions in retail research where it is called as market basket analysis. It is used to identify the purchase patterns of the consumer. Given a database, where each transaction has a set of items, FIM tries to find itemsets that occur in transactions more often than a given threshold. The frequent itemsets detection can potentially provide, if the data is intuitive (e.g., web browsing history and medical records of patients), releasing the detected frequent itemsets might cause threats to individual privacy.

This paper addresses the frequent and weighted itemsets discovery, i.e., the frequent weighted itemsets, from transactional weighted data sets using UP tree and UP growth algorithm for high transaction itemsets.

2. RELATED WORK

Lots of studies have been proposed to solve the privacy preserving FIM problem from different aspects.

Main aim is to ensure that the resulted frequent itemsets itself does not leak private information and achieve differential privacy. Considering K- anonymity model for protecting privacy in [2], [12] propose an algorithm to publish annonymised frequent itemset. However these two studies do not satisfies differential privacy. And thus they cannot provide sufficient privacy protection against attackers with background knowledge. [3] A new novel and powerful privacy definition called l –diversity. [3] Show the weak points of k-anonymity; how it is weaken to protect information against attacker with background knowledge. Diversity framework introduced here to give strong privacy guarantee.

[4] Proposed fast algorithm for mining association rule i.e. Apriori & AprioriHybrid algorithms .These compared with previous algorithms and these algorithm gives excellent

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

performance for large database with transactions, but these generates candidate set.

[5] Introduces FP growth algorithm, with is nothing but mining frequent pattern without candidate generation, as we have seen in [4] apriori algorithm performs mining fastly with candidate set generation, which is costly. In [5] FP tree is used as data structure to store large database compressed in small data structure. Algorithm introduced in [5] is scalable and efficient than apriori algorithm.

[11] Present set of randomization operators to limit privacy beaches in FIM.

[13] Proposed new algorithm for discovering frequent patterns in sensitive data adopted exponential mechanism & Laplace noise-addition mechanism techniques which are efficient in context of frequent item mining.

[14] Proposes algorithm Privbasis with perform frequent itemset mining with differential privacy by using minimum support threshold. Any itemset that occurs in transaction often than minimum support threshold is subset of some basis with differential privacy guarantee.

But [6] [13] [14] addresses some issues performing frequent item mining with differential privacy.

Differential Privacy:-

For 2 databases D & D', they are neighboring databases if they differ by at most one record.

Є –Differential Privacy:

A private algorithm A satisfies Є –Differential Privacy Iff for any two neighboring databases D & D' and any subset of outputs S is subset of Range (A),

$\Pr [A (D) \in S] \leq e^{\epsilon} * \Pr [A (D') \in S]$

Where probability is taken over the randomness of A.

Frequent Itemset mining:

Itemset nothing but collection of one or more items & frequent items means itemsets whose support is greater than or equal to minimum support threshold. Finding such frequent itemset is called frequent item mining.

3. EXISTING SYSTEM

Differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth.

FP growth algorithm:

It is a partitioning based, depth first search algorithm.

It adopts divide and conquer manner to adopt to decompose the mining task into smaller tasks for finding frequent itemsets. In conditional pattern bases.

To efficiently generate conditional pattern bases, FP growth uses 2 data structures Header Table & FP- tree.

Header Table used to stores items and their support.

For FP- tree each branch represents an itemset and node has counter.

PFP -growth consists of

1. Preprocessing phase
2. Mining phase.

Preprocessing phase:-

Utility and privacy trade-off can be improved by using transaction splitting techniques. To improve privacy utility trade off transactions are splitted rather than truncated. Smart splitting is performed in this phase.

By extracting the information from original database smart splitting is performed and original database is transformed. For given Database preprocessing phase performed only once.

Mining phase:-

In this phase, given the transformed database and a user-specified threshold, it privately discovers frequent itemset. Run time estimation & dynamic reduction methods are used in this phase to improve quality of results.

In this phase we divide privacy budget ϵ in to 5 portions.

$Є_1$ is used to compute maximum length constraint

$Є_2$ is used to estimate maximal length of frequent itemsets.

$Є_3$ is used to reveal correlation between items in transaction.

$Є_4$ is used to compute vectors of itemsets.

$Є_5$ is used to compute support.

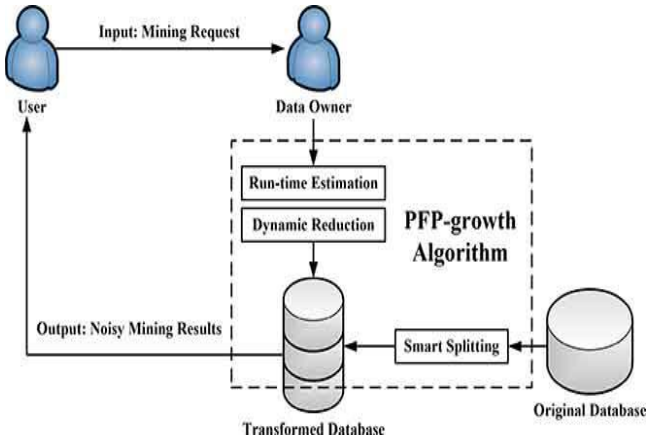

Figure 1: System Architecture.

Preprocessing Phase:-

Problem statement:

- The existing system does not deal with the high utility transactional itemsets.
- Existing methods require larger time to execute.
- Existing system gives comparatively large size output combination.

4. PROPOSED SYSTEM

Mining high utility itemsets from a database having large transactions refers to the discovery of itemsets with high utility like profits, sale. Although a number of relevant approaches have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. By referring paper [18], we propose an efficient algorithm, namely *UP-Growth* (*Utility Pattern Growth*), for mining high utility itemsets with a set of techniques for pruning candidate itemsets. The high utility itemsets information is preserved in a special data structure named *UP-Tree* (*Utility Pattern*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

Tree) such that the candidate itemsets can be created efficiently with only two scans of the database. The performance of UP-Growth was estimated in differentiation with the other algorithms on different types of datasets. Our algorithm outperforms in terms of execution time, especially when the database contains lots of long transactions. Also gives limited and accurate output.

Advantages:

- The proposed algorithm scans the database only limited no. Of times which requires less time complexity.
- Our proposed algorithm improves the accuracy as compare to existing algorithms for frequent itemsets.
- Gives limited and accurate item set.

5. CONCLUSION

In this paper we survey some frequent item mining with privacy methods. We studied about the method, which is useful for privacy such as K- anonymity,l-diversity,Privbasis.we have studied and analysed these methods observed drawbacks and benefits of these methods.We have studied different mining algorithms such as Apriori ,Apriorihybrid, FP –growth, UP –growth algorithm performed comparisons between Apriori and FP –growth & UP growth. Apriori is costly to perform and not time efficient than FP –growth algorithm. We studied existing system .We compared the FP- growth with UP-growth algorithm. We conclude that UP growth algorithm is time efficient and requires less memory as compared to FP growth especially when database contains lots of long transactions.

6. REFERENCES

[1] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Yang, “Differentially private frequent item mining via transaction splitting”,2015,In IEEE transactions on knowledge and data engineering vol. 27, No.7, pp.1875-1891

[2] C. Dwork, “Differential privacy,” in Proc. Int. Colloquium Automata, Languages Programm., 2006, pp. 1–12

[3] L. Sweeney, “k-anonymity: A model for protecting privacy,” Int. J.Uncertainity Fuzziness Knowl.-Base Syst., vol. 10, no. 5, pp. 557–570,2002.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity, “in Proc. 22nd Int. Conf. Data Eng., 2006, p. 24.

[5] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.

[6] 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.

[7] C. Zeng, J. F. Naughton, and J.-Y. Cai, “On differentially private frequent itemset mining,” Proc. VLDB Endowment, vol. 6, no. 1,pp. 25–36, 2012.

[8] 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.

[9] 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.

[10] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis,“Security in outsourcing of association rule mining,” in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 111–122.

[11] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis,“An audit environment for outsourcing of frequent itemset mining,”Proc. VLDB Endowment, vol. 2, no. 1, pp. 1162–1173, 2009.

[12] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 217–228.

[13] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi,“Anonymity preserving pattern discovery,” VLDB J., vol. 17,no. 4, pp. 703–727, 2008.

[14] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, “Discovering frequent patterns in sensitive data,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp. 503–512.

[15] N. Li, W. Qardaji, D. Su, and J. Cao, “Privbasis: Frequent itemset mining with differential privacy,” Proc. VLDB Endowment, vol. 5, no. 11, pp. 1340–1351, 2012.

[16] F. McSherry and K. Talwar, “Mechanism design via differential privacy,” in Proc. 48th Annu. IEEE Symp. Found. Comput. Sci., 2007,pp. 94–103.

[17] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Proc. 3rd Conf. Theory Cryptography, 2006, pp. 265–284.

[18] Vincent S. Tseng, Cheng Wei wu, Bai- En shei, Philip S. Yu“UP-Growth: An Efficient Algorithm for High Utility Itemset Mining”, Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA