An Algorithm for Frequent Item-set Mining to incorporate Differential Privacy and to increase proficiency.

Akanksha Bhalerao-Kulkarni[1], Prof. Soumitra Das[2]

[1]Research Scholar, Department of Computer Engineering, Dr. D.Y.Patil School of Engineering, Pune, India
[2]Head off Department, Department of Computer Engineering, Dr. D.Y.Patil School of Engineering, Pune, India

---***---

Abstract – Today a number of E-commerce enterprises and Supermarkets depend upon surveys from their customers to reach out to formulate better business decisions. The mining of such examples is a prime duty of a data mining system which impacts the growth of its clients. A lot of research on this issue has prompted the excessive need of efficient and scalable algorithms for mining frequent patterns. The protection of personal data of the clients participating still is a major cause of worry. Thus, we examine this issue in a protection saving setting and propose an approach for differential private frequent item-set mining based on LCM algorithm; we refer it as P-LCM algorithm. P-LCM is extended version on PFP growth algorithm which basically works in two phases namely pre-processing and mining phase. The pre-processing phase is a onetime activity. To boost its utility and privacy transaction splitting is introduced to play. The Mining phase is responsible to limit the information loss caused and for noise reduction during the process. LCM is a polynomial time algorithm which finds all frequent item sets. The closed item-sets obtained do not occupy memory space. On analysis it is exhibited that our algorithms are faster and time effective simultaneously.

***Key Words*: Frequent Item-set Mining, Transaction Splitting, Differential Privacy, LCM.**

1.INTRODUCTION

Recently, almost all enterprises collect personal data from various users as surveys, feedbacks. This marks a threat to privacy and users apprehend from participating in public survey forums. Hence, in our paper, we focus on privacy issues that arise of finding frequent item-sets in "transactional" data. Frequent item-set mining is widely used in many applications especially in market basket analysis. It finds sets of items that are frequently bought together, and thus user can establish an association rule in them. This helps in formulation of various business decisions. These days FIM is being studied widely due to the above factors. However, the end user's privacy has received little attention. 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. This raises a privacy concern as this data should not reveal any private information about the participating individuals, but the enterprise cannot assure this. This problem is compounded by the fact that it is not even known what data the individuals would like to protect nor what background information might be possessed by an adversary. These compounding factors are exactly the ones addressed by differential privacy [2].

1.1LITERATURE SURVEY

A. Related Work

Many different algorithms have been proposed for frequent item-set mining. From that Apriori and FP-growth are the two most well-known ones.

- **Apriori:**

Apriori algorithm works as breadth-first search, along with candidate set generation-and-test algorithm. This algorithm needs only those number of database scans as per the length of frequent item-sets , if the maximal length of is one then scan would be single. Thus with the increase in number of frequent item-sets will promote increase in the number of scans as well. [1].

FP-growth algorithm is depth-first search algorithm, and does not require candidate generation. FP-growth only performs two database scans, which makes FP-growth faster in all cases.

- **FP-GROWTH**:

The promising features of FP-growth motivate us to design a differentially private FIM algorithm based on it. In this paper, we argue that a practical differentially private FIM algorithm should achieve high data utility and degree of privacy, and time efficiency all together. Some differentially private FIM calculations have been proposed, however any existing studies that can fulfill each criteria isn't found yet. It is not only time effective but has high degree of privacy. There are some limitations of these existing FIM algorithms such as FP-growth scans only two times hence cannot be used for longer transactions. [5][6]

1.2MOTIVATION

Enormous amount of research is going on Frequent item-set mining (FIM), But Differential privacy came as a break-thru as not much focus has been given to it. This made our inclination stronger towards this area. Data mining and Network Security being the opted elective courses provided the fundamental knowledge regarding the domain and thus embarked further interest into the topic. The known concepts helped in better understanding of the research papers. Also the business applications of FIM like supermarkets, healthcare centers, E-commerce etc. contributed to our objective.

1.3EXISTING ARCHITECTURES

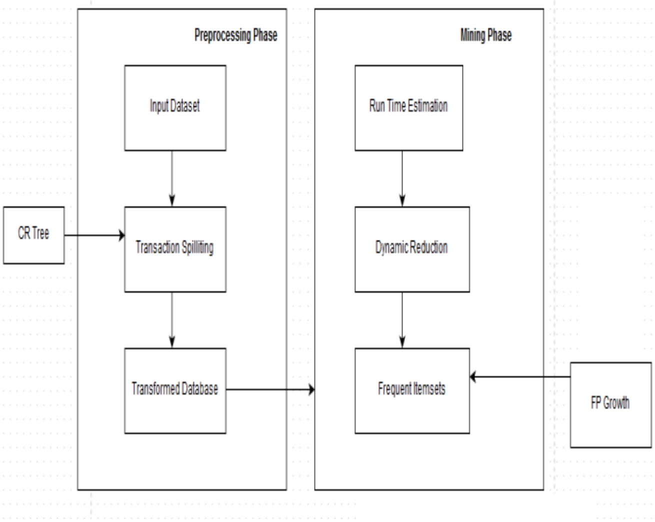


Fig. 1. Existing system architecture with both phases [This diagram is pictorial representation as studied by authors]

2.TAXONOMY CHART

The taxonomy chart denotes the comparison of various existing tools thus giving clarity on constraints and requirement parameters to be worked upon.

PARAMETERS/REFERENCES	TREE STRUCTURE	EFFICIENCY	TIME COMPLEXITY	PERFORMANCE
Frequent Pattern Growth (FP-Growth) Algorithm	YES	POOR	LOW	AVERAGE
UP-Growth: an efficient algorithm for high utility itemset mining	YES	AVERAGE	HIGH	AVERAGE
Mining Frequent Item-sets – Apriori Algorithm	NO	HIGH	HIGH	AVERAGE
Differentially Private Frequent	YES	HIGH	LOW	GOOD

3. PROPOSED FRAMEWORK AND DESIGN

A. PROBLEM DEFINITION

To develop a highly secure, efficient and more accurate system which provides mining strategy of Frequent Item-sets using P-LCM algorithms, this will ensure the reduction in data loss with optimum output.

B. MATHEMATICAL MODEL

Let S be the system which we use to find the private frequent item-sets. FP growth performs well in case of differential privacy for frequent item-set finding.

It consists of two phases:

I. Pre-processing
II. Mining phase

Mathematically it is as follows:

S = {P, M, FIM} where,

P = Pre-processing phase

M = Mining phase.

FIM = Frequent Item-sets.

Input: A transactional data set T= {t1, t2, t3,..., tn} is a set of transactions, where each transaction tq (q belongs to [1,n]) is a set of items in I and each is characterized by a transaction ID (tid) where,

I= {i1, i2,..., im} be a set of data items.

I.PRE-PROCESSING PHASE:

Assume that P= {D, N, ϵ1, ϵ2, ϵ3, TS}

Where, D= original database;

N= percentage,

ϵ1, ϵ2, ϵ3 are the privacy budgets,

TS = transaction splitting criteria.

For calculating privacy budgets we need following:

i. Sensitivity[1]:
Given p count queries Q, for any neighbouring databases D;

D' the sensitivity of Q is:

ΔQ = max ||Q (D) - Q (D')||.

The Laplace distribution with magnitude M, i.e., Lap (M), follows the probability density function as

$Pr[x|M] = 1 /2 M*e^{-|x|/M}$,

where M= ΔQ / ϵ

is determined by both the sensitivity ΔQ and the privacy budget ϵ.

ii. Threshold calculation[1]:
G(ϵ/Cn*Lf)

Where,

ϵ = privacy budget,

Cn is the length of transaction and

Lf is maximum transaction length.

iii. Smart splitting using Weighted Splitting Operation[1]:

Consider a transaction t whose length exceeds the maximal length constraint Lm.

A function f divides t into multiple subsets t1, ..., tk, where ti is assigned a weight wi and the length of ti is under the length constraint Lm.

Then, function f is said to be a weighted splitting operation iff:

$U^{k}_{i=1}$ ti and $\sum^{k}_{i=1}$ (wi≤1).

Given a transaction t of length p (p > Lm), we aim to partition the p items into q =[p=Lm] subsets t1, ..., tq, each of which satisfies the length constraint, so as to minimize

the within subset sum of shortest path lengths:

avg min $\sum^{q}_{i=1} \sum^{Iu,Iv\,\epsilon\, ti}_{i=1}$ dist (iu, iv)

II.MINING PHASE:

MI = {TD, T, PB, Z}

Where,

TD = transformed database,

T = threshold value,

PB = Privacy budget,

And Z = matrix.

Following are the processes from mining phase:

1. Estimate the actual support of transformed database.
2. Estimate the actual support of Original database

Output: FIM (frequently mined item-sets):

We have to perform algorithms i.e. Mining Phase algorithm for frequent item-set mining.

MI= {D, Lm, Lp, Dp, prefix, M, ϵ', upArray}

Where, D = the transformed dataset,

Lm = maximal length constraint,

LP = List,

DP= conditional pattern base,

Prefix= the prefix item-set,

Є' and M are the Privacy budget and threshold respectively, upArray is Up-Array.

Final Output: Frequent item-set F

Where F = {f1, f2,…,fn}

C. SOFTWARE ARCHITECTURE

The new architecture helps us understand how the proposed system functions. The FP-Growth algorithm is replaced with LCM algorithm [9] for better results.

LCM stands for ***L****inear time* ***C****losed item set* ***M****iner*. The algorithms which exist enlist the final output of frequent item sets with cutting off unnecessary item sets by pruning. Nonetheless, if pruning is not complete, they continue to function on unnecessary frequent item sets and may ultimately lead to data loss. In LCM, a parent-child relationship amongst frequent closed item sets comes to play. This relationship induces tree-shaped transversal routes consisting of all the frequent closed item sets only. Our algorithm traverses the routes in linear time of the number of frequent closed item sets. LCM is designed on the basis of reverse search technique. LCM by far has significantly outgrown its competitors as exhibited by the experimental results.[4][9]

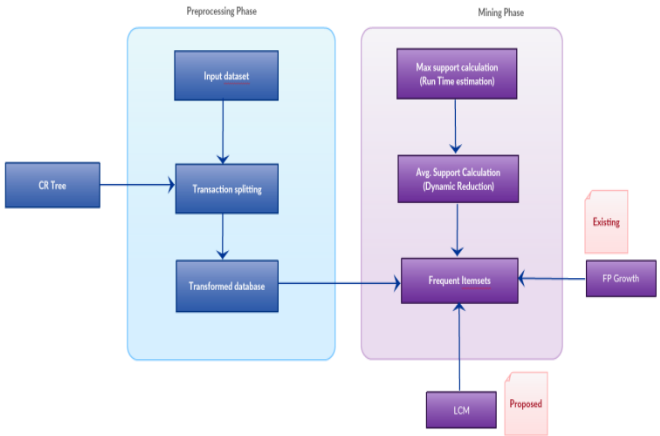


Fig. 2. The System Architecture using P-LCM Algorithm [This diagram is pictorial representation of proposed system designed by authors.]

4. PROJECT MODULES

Module 1: In this module we create Basic GUI of user side. User can insert input transaction dataset through this GUI and pass it for pre-processing steps.

- In this module user or we can say it as **admin**, who can browse input transaction dataset file and upload it for pre-processing operations.
- Then system will do pre-processing operations given in second algorithm of pre-processing. Such as assign privacy budgets, calculate **maximum threshold value** for transaction splitting, create CR tree etc.

- We will create different set of item-sets whose length is greater than calculated maximum threshold value.
- Then we will split long transactions for further mining phase.

Fig. 3. The Welcome Screen

Fig.4. The Login Screen

Fig.5. The Home Screen

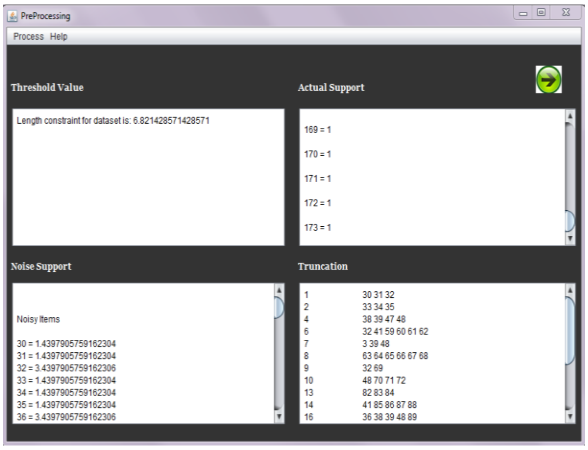


Fig.6. The Pre-processing Phase

Module 2: This module deals with implementation of Existing system.

- In this module we will implement mining phase using FP-growth algorithm.
- And generate and store its results.

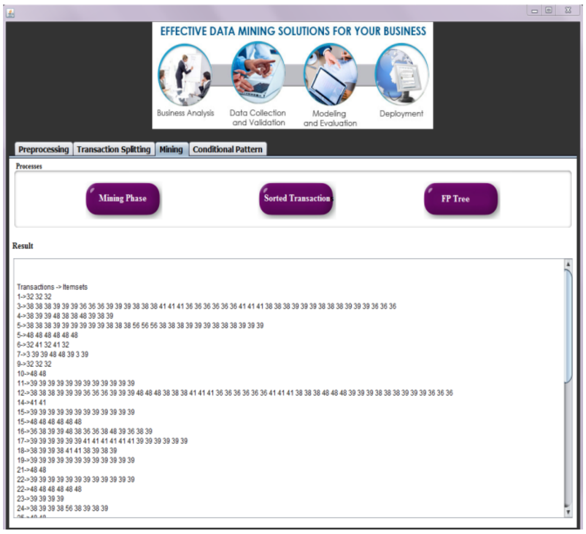


Fig. 7.The Mining Phase

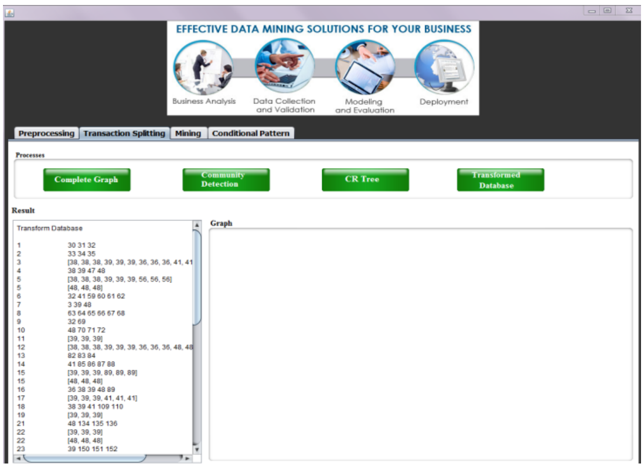


Fig.8.The Transaction-Splitting Algorithm

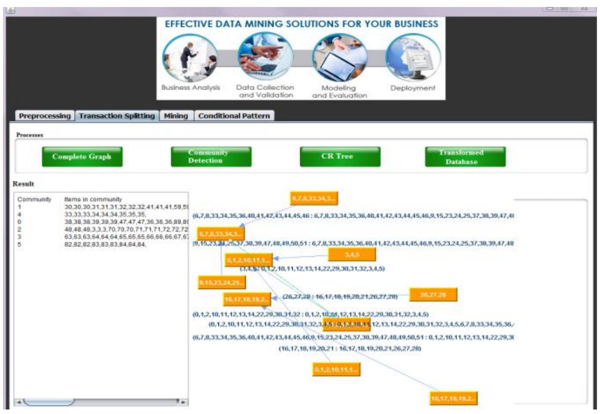


Fig.9.The Complete graph generation

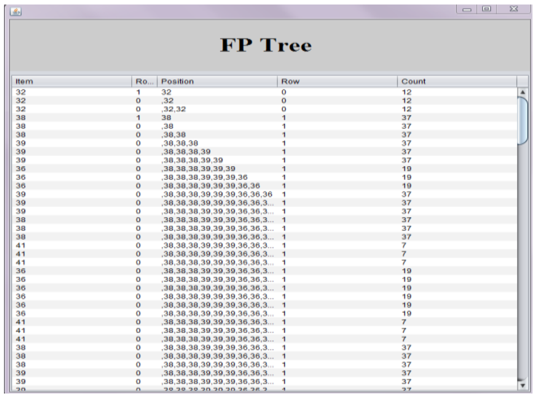


Fig.10. The FP- Tree

Module 3: In this module we experiment our proposed system and compare it with existing system for analysis.

- In this module our proposed P-LCM algorithm is replaced instead of FP growth for mining frequent closed item-sets.
- We plan to integrate LCM algorithm in existing mining phase algorithms and obtain results.
- The results obtained are stored for study of the comparative result of existing and proposed system.

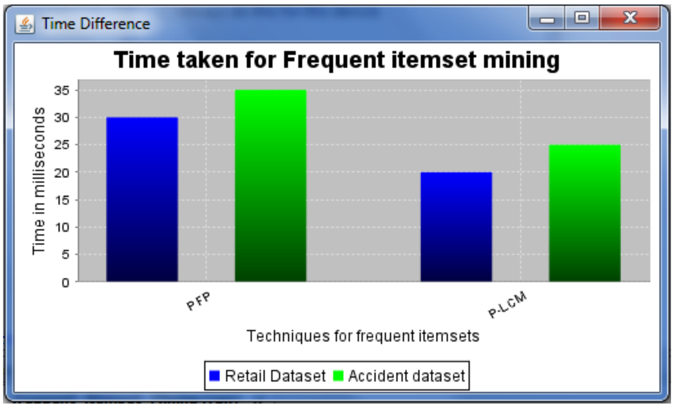
Module 4: In this module we test the new system for expected results.

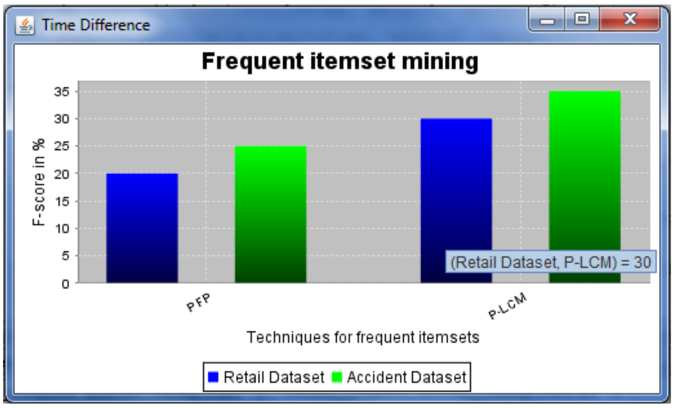
- In this module the analysis of obtained results is conducted with regards to expected ones.
- Thus statistics is drawn over system performance.

V. RESULTS AND DISCUSSION

The project is tested for Time efficiency and F-score and the following results were obtained. The Existing PFP growth and proposed P-LCM both are compared.The Retail and Accident Datasets were used as inputs.

LCM has evidently performed better on both parameters.





VI. CONCLUSION AND FUTURE ENHANCEMENT

The need for designing differentially private data mining algorithms has seen growth as for frequent item-set mining purposes. It is the backbone of Data Mining. The most traditional and not much effective algorithms have been the cause behind this development. Thus through this project we intend to provide better and time saving results of frequent item-set mining along with maintaining the security of long transactional datasets. An effort to considerably replace the traditional FP-growth algorithm with P-LCM algorithm is tested for results. The concept of Differential Privacy, Transaction splitting and Run Time Estimation are studied in depth.

Our future work extends to apply same techniques on higher dimensional dataset of transactions.

ACKNOWLEDGMENT

It gives me great pleasure and immense satisfaction to present this paper which is the result of unwavering support, expert guidance and focused direction of my guide and HOD, Prof. Soumitra Das and my M.E. Staff to whom I express my deep sense of gratitude and humble thanks, for their valuable guidance.

REFERENCES

[1] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Yang

'Differentially Private Frequent Itemset Mining via Transaction Splitting', 2015 IEEE Transactions on Knowledge and Data Engineering

[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]Takeaki Uno, Tatsuya Asai, Yuzo Uchida, Hiroki Arimura,'LCM: An Efficient Algorithm forEnumerating Frequent Closed Item Sets'.

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

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

[7] 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.

[8] 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.

[9] Takeaki Uno1, Tatsuya Asai2, Yuzo Uchida2, Hiroki Arimura2, "LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets", National Institute of Informatics.

[10] Akanksha Bhalerao-Kulkarni,Prof. Soumitra Das," A New Energy Efficient Vertical Handover Algorithm In Heterogeneous Networks." www.mjret.in/M61-2-4-10-2016

BIOGRAPHIES

Akanksha Bhalerao-Kulkarni,
Research Scholar, M.E. Computer Engineering, Dr. D.Y.Patil School of Engineering, Lohegaon, Pune.

Prof. Soumitra Das,
Head Of Department (Computer Engineering), Dr. D.Y.Patil School of Engineering, Lohegaon, Pune.