Results and Discussions on Transaction Splitting Technique for Mining Differential Private Frequent Itemsets

Sheetal K. Labade
Computer Engineering Dept.,
JSCOE, Hadapsar
Pune, India

Srinivasa Narasimha Kini
Computer Engineering. Dept.,
JSCOE, Hadapsar
Pune, India

ABSTRACT
Many researchers are now working on designing of data mining algorithms which also provides differential privacy. Especially so, in mining of frequent itemsets. Individual privacy may get affected by revealing frequent itemsets. Therefore, a frequent itemset mining algorithm with differential privacy is important which will follow two phase process of preprocessing and mining. This paper discusses diagonal splitting of transactions in splitting mechanism. As proposed mechanism, diagonally splits each transaction then size of transaction reduces, resulting in complexity and processing time reduction. By splitting the transaction diagonally, it divides the transaction in two subparts. This paper demonstrated the performance of diagonal algorithm through experiments on real datasets. Result has been taken on various threshold values and calculated f-score measure for output frequent itemsets. Time taken for frequent itemset mining also studied. An experimental comparison with existing algorithms shows that diagonal splitting algorithm achieves better F-score measure and is about an order of magnitude faster for various top k frequent item mining.

General Terms
Data mining, Frequent item set mining, Algorithms.

Keywords
Differential, Privacy, Transaction, Splitting, Diagonally.

1. INTRODUCTION
With the World Wide Web, there is now a large amount of information about individuals is available, which can be gain within seconds. This information could be obtained through mining or just from information retrieval. Data mining is the process of users giving queries and mining knowledge which is previously unknown.

Now, data mining is an important technology for many applications. Mining of association rule is most substantial data mining application. This application find-out customer purchasing behavior. Association rule can be well defined as {X, Y} => {Z}. If for example, customer buys X, Y he may buy Z also. For this rule forming minimum support and confidence value is used.

However data mining also origins privacy worries, as users can now put pieces of information together and extract knowledge that is sensitive or private. Therefore, one needs to apply some control strategy on databases and data mining tools. That is, while data mining is an vital tool for several applications, we do not want the knowledge mined to be used in an improper way. For example, based on information about a person, an insurance company could reject insurance or a loan agency could reject loans. In many cases these rejections may not be genuine. Therefore, information providers have to be very alert in what they issue. Also, data mining researchers have to make sure that privacy features are addressed.

Differential privacy offers strong theoretical guarantees on the privacy of released data without creating any assumptions about an attacker's background knowledge. A detail literature survey related to frequent item set mining methods and privacy preservation was carried out in [2].Other existing systems related to frequent item set mining also considered in literature review unit of [1].In this paper the focus is on experimental setup, results and conclusion.

The rest of this paper is organized as follows: Section II provides the problem definition. Section III discusses how differential privacy is achieved in frequent item set mining via splitting the transaction diagonally. Section IV presents Experimental Setup and Results, followed by conclusions in Section V.

2. PROBLEM DEFINITION
Private FP growth algorithm is used for differential private frequent item set mining. Many algorithms are used for above stated purpose. Base of this notion is FP-growth algorithm. FP-growth suffers from some limitations like only two time scanning was available in this algorithm. And due to this limitation researchers cannot re-truncate their transactions many times [1]. Instead of using transaction truncating approach, splitting of transaction can be used [1]. But splitting technique might cause loss of information. To avoid this problem this paper proposes diagonal splitting of transactions [1].

2.1 Problem Statement
To achieve the privacy and reduce the loss of information during transaction splitting in differentially private frequent item set mining algorithm.

3. WORKING OF SYSTEM AND ARCHITECTURE

3.1 Overview
As discussed in problem definition section, some challenges are present while designing frequent item set mining (FIM) algorithm with differential privacy. To overcome those challenges three key methods was studied in [1]. These methods are smart splitting, run-time estimation and dynamic reduction and these methods were based on FP-growth algorithm. Smart splitting mechanism was used for splitting the long transactions into many subsets.

In diagonal spitting system, splitting algorithm was modified and which works according to the pseudo code given in [1]. In this approach transactions are splitted diagonally in splitting mechanism.

Main drawback of existing system was data loss while splitting long transaction [1]. Diagonal splitting directly splits the transaction in two sub-parts and therefore gives fewer probabilities for data loss.

3.2 System Architecture

Fig. 1. shows System Architecture. Pre-processing and mining are two main processes are here. Solid line from pre-processing to mining indicates existing system flow and dotted line indicates diagonal splitting system flow. In existing system which is explained in [1], run-time estimation was done after getting transformed database to balance the loss of information [1]. Finally, output is in the form of frequent itemsets by applying FP growth algorithm.

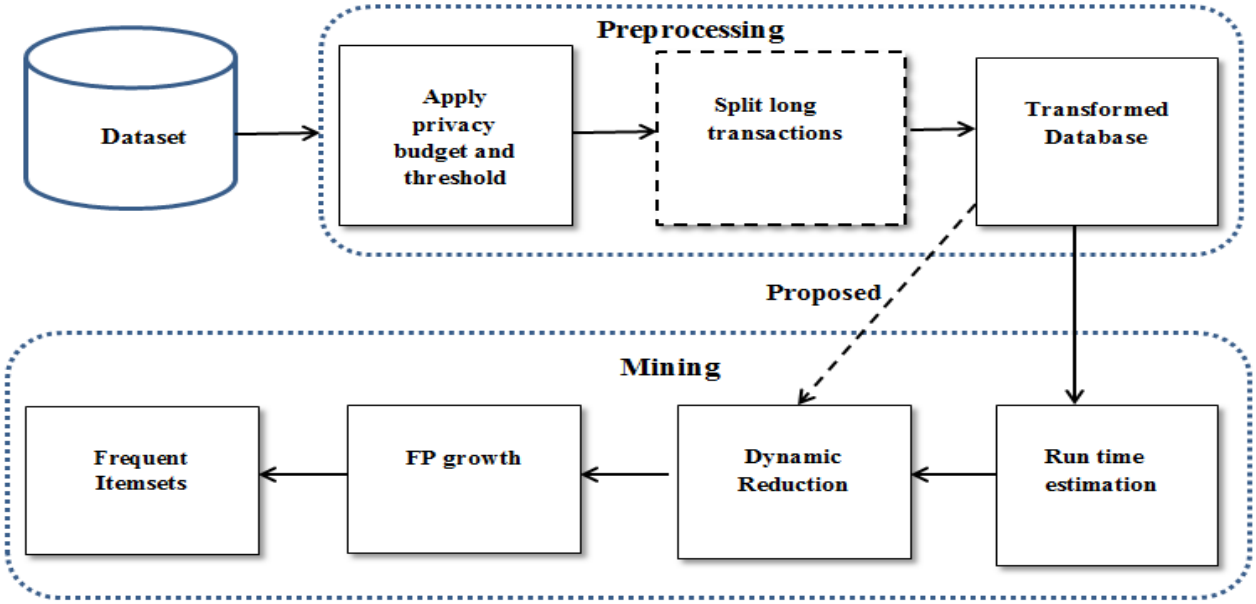

Fig 1: System Architecture

3.3 Data flow Diagram (DFD)

Fig.2. Shows data flow diagram of system explained in this paper. This is DFD level 2 diagram.

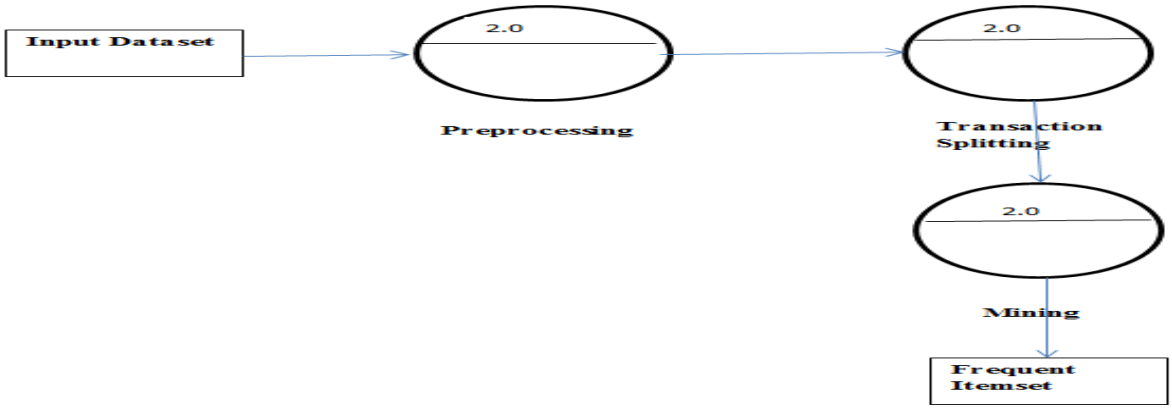

Fig 2. DFD diagram

4. EXPERIMENTAL SETUP AND RESULTS

4.1 Experiment Setup in Smart Transaction Splitting

The System proposed in [3] compared their PFP-growth algorithm with the following two algorithms.

1) the Apriori-based algorithm proposed in [4].
2) the "PrivBasis" algorithm proposed in [5]. They use PFP to denote their algorithm, while TT and PB to denote the algorithms in [4] and [5], respectively.

They implemented algorithms mentioned in [3], [4] and [5] in JAVA and conduct all experiments on a PC with Intel Core2 Duo E8400 CPU (3.0 GHz) and used 4 GB RAM. Since these algorithms was based on randomization, they [3] report average result by running algorithms 10 times. In their experiments, they used the relative threshold.

In PFP[3], since the number of transactions is increased by transaction splitting, they used the relative threshold with respect to the original dataset. They set privacy budget ε to be 1.0. The parameter η used in the preprocessing phase was set to be 0.85.

- **Datasets**: In the experiments, they used four publicly available real datasets. [A] Dense datasets: Pumsb-star (PUMSB) [6] and Accidents [6]. [B] Sparse datasets: BMS-POS (POS) [7] and Retail [6].
- **Utility Measures**: For performance evaluation of their algorithm, they used some standard metrics like F-score [4] to measure the utility of generated frequent itemsets, Relative error parameter (RE) and running time. Mining High Utility itemsets from a transaction database is to find itemsets that have utility above a user-specified threshold.To measure the error by considering actual support of itemsets they used RE parameter.

4.2 Experiment Setup in Diagonal Splitting

The Diagonal Splitting System proposed in these paper is compared with PFP-growth algorithm [3] which proposed smart transaction splitting approach.

This system can run with operating system like Windows XP/7 and upper versions. Programming language is Java, Tools used is Netbeans with minimum requirement of Pentium iv 2.6 ghz Processor, 512 MB ddr RAM and 20 GB Hard Disk.

- **Datasets**: In the experiments, two publicly available real datasets was used. Datasets used: Retail [6] and Accidents [6].
- **Utility Measures**: For performance evaluation of diagonal splitting algorithm mentioned in[1] ,some standard metrics like F-score [4] and running time was used.

4.3 Results

4.3.1 Time Comparison between Existing and Proposed System

Table 1 shows comparison between time consumed in Smart Transaction Splitting and Diagonal Splitting respectively.

Table 1.Time Comparison

Dataset	Smart Transaction splitting	Diagonal splitting
Retail	41,750 milliseconds	2,953 milliseconds
Accidents	12,265 milliseconds	2,484 milliseconds

4.3.2 Time graph

Graph 1: Fig. 3. Shows results for Time graph of frequent item set mining.

Dataset: Retail dataset [6] and Accidents dataset [6] was considered in graph 1.

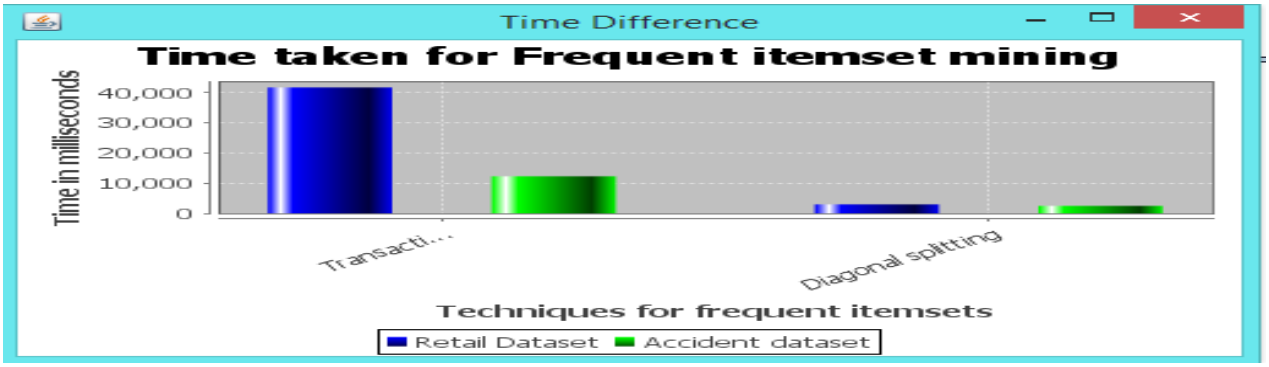

Fig 3. Time graph (Retail and Accidents Dataset)

Screen shots captured below in Fig.4 and Fig. 5 shows result analysis of Time taken for frequent itemset mining on retail dataset[6] for smart splitting and diagonal splitting approach respectively.

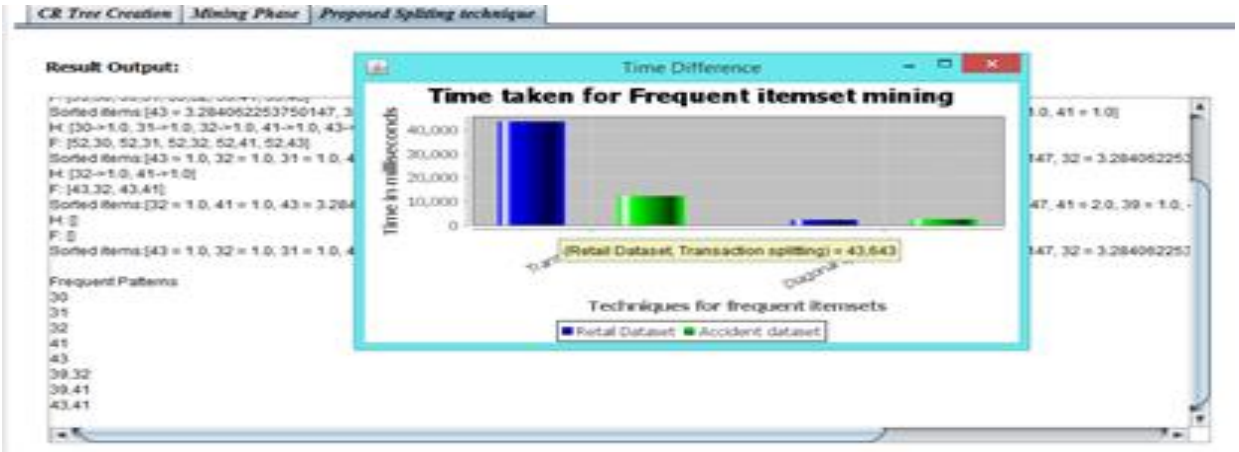

Fig. 4 retail dataset (Smart splitting)

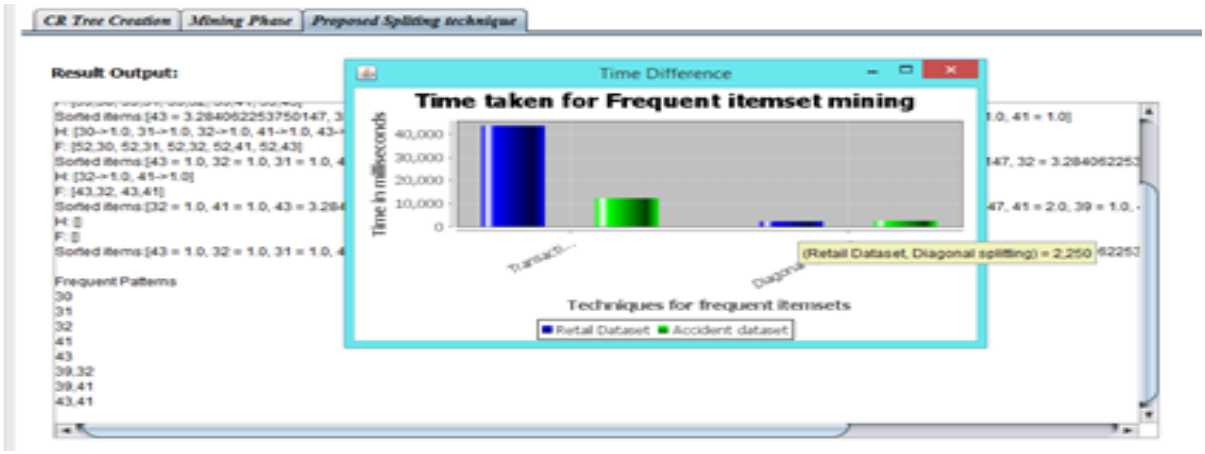

Fig. 5 retail dataset (Diagonal splitting)

4.3.3 F-Measure

Table 2 shows comparison between F-Measure of Smart Transaction Splitting and Diagonal Splitting respectively.

Table 2. F-Measure

Dataset	Smart Transaction splitting	Diagonal splitting
Retail	70.219	76.171
Accidents	70.109	76.055

4.3.4 F-Measure graph

Graph 2: Fig. 6. Shows results for frequent item set mining for F-Measure parameter. Retail [6] and Accident [6] dataset used for experimental evaluation.

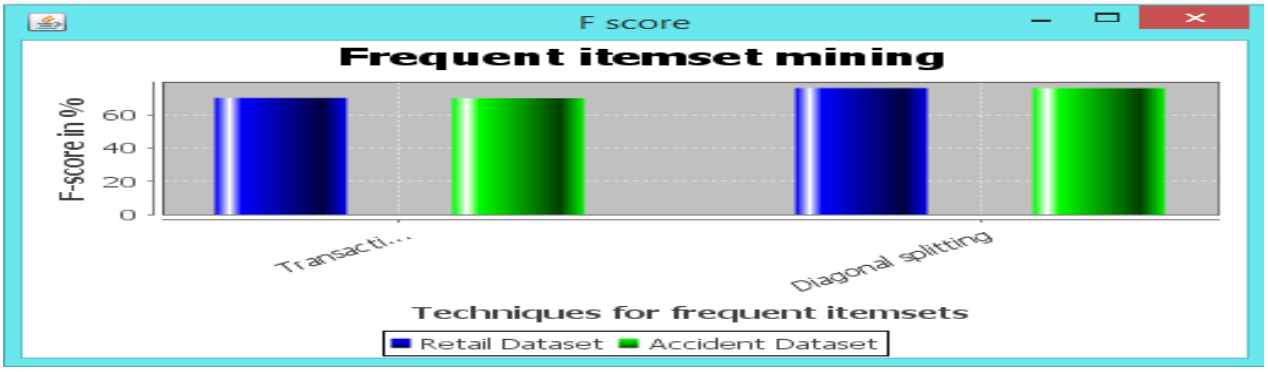

Fig. 6 F-Measure graph

Screen shots captured below in Fig. 7 and Fig. 8 shows result analysis of F-Measure parameter for Accident Dataset [6].Graph shows that diagonal splitting gives better performance for F-Measure on Retail[6] and Accident Datasets[6] than existing smart splitting approach..

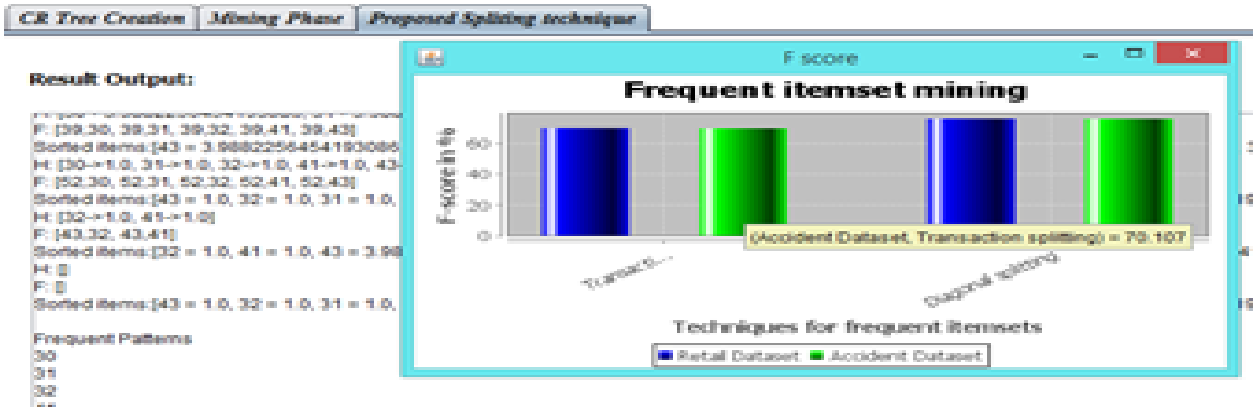

Fig. 7 F-Measure (Accident dataset)-Smart transaction splitting

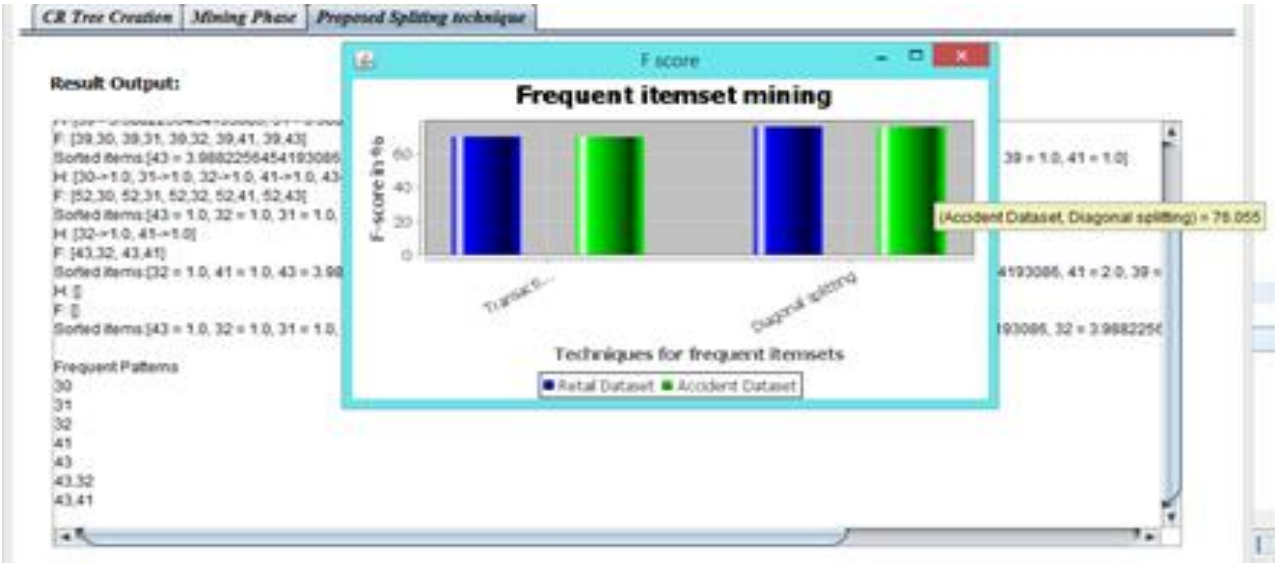

Fig 8. F-Measure (Accident Dataset)-Diagonal Splitting

4.3.5 Time evaluation graph:

Graph 3: Fig.9 .shows the time difference between existing and proposed system for various top k frequent items.

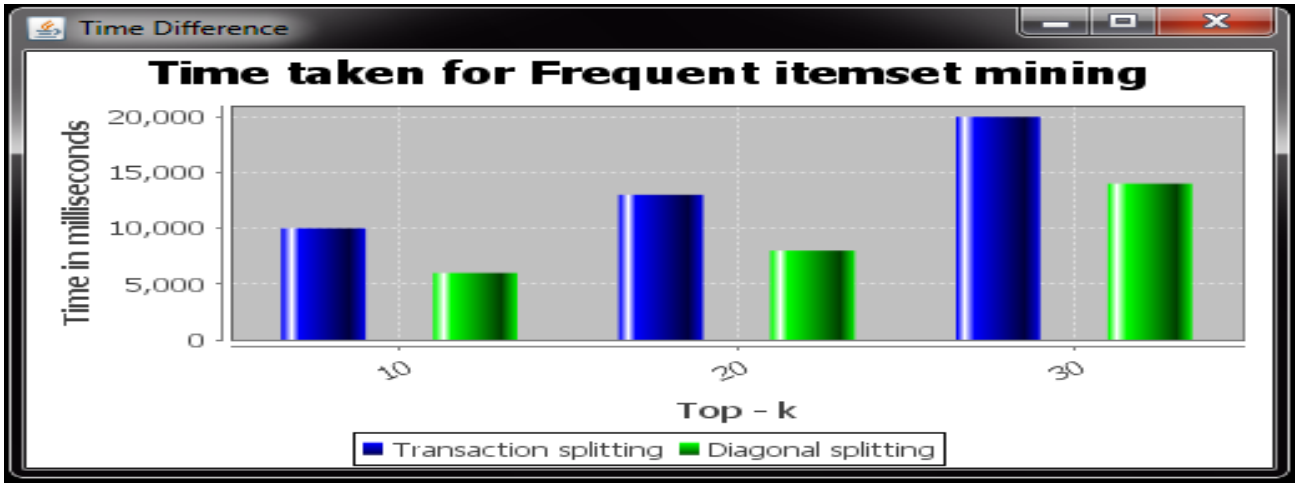

Fig 9. Time difference

4.3.6 Time consumed in preprocessing phase:

Table 3 Shows the time consumed in the preprocessing phase of smart transaction splitting and diagonal splitting. From table it is observed that, the preprocessing phase does not consume too much time in diagonal splitting.

Table 3.Time Consumed in the Preprocessing Phase

Dataset	Smart Transaction splitting	Diagonal splitting
Accidents	12893 milliseconds	2341 milliseconds
Retail	32034 milliseconds	7371 milliseconds

5. CONCLUSION AND FUTURE SCOPE

This paper examined the problem of designing frequent itemset mining algorithm with differential privacy. This paper discussed diagonal splitting of transactions in splitting mechanism. As transactions are splitted diagonally, then size of transaction reduces, resulting in complexity and processing time reduction. Also this splitting divides the transaction in two subparts. For performance evaluation of diagonal splitting algorithm two different real datasets was used. Result has been taken on various threshold values and calculated f-score measure parameter for output frequent itemsets. Time taken for frequent itemset mining also studied. An experimental comparison with existing algorithms mentioned in [3] shows that diagonal splitting achieves better F-score measure and is about an order of magnitude faster for various top k frequent item mining.

In future this system can be useful for high utility itemsets, which is another growing area of research [1].

6. REFERENCES

[1] Sheetal Labade, Srinivasa Narasimha Kini," A Novel Approach towards Transaction Splitting for Differential Private Frequent Itemset Mining,"fifth post graduate conference of computer engineering, cpgcon,25-26 march,2016.

[2] Sheetal Labade, Srinivasa Narasimha Kini," A survey paper on frequent itemset mining methods and

techniques" in International Journal of Science and Research (IJSR), Volume 4 Issue 12,Paper ID: NOV151884 , Dec.2015.

[3] Sen Su, 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,July 2015.

[4] 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.

[5] 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.

[6] Z. Zheng, R. Kohavi, and L. Mason, "Real world performance of association rule algorithms," in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2001, pp. 401–406.

[7] Frequent itemset mining dataset repository [Online]. Available: http:// fimi.ua.ac.be/data, 2004.