A Study of Differentially Private Frequent Itemset Mining

Trupti Kenekar[1], A. R. Dani[2]

[1, 2]GHRIET, Department of Computer Engineering & Technology, Savitribai Phule Pune University, Pune ,India

Abstract*: Frequent sets play an important role in many Data Mining tasks that try to search interesting patterns from databases, such as association rules, sequences, correlations, episodes, classifiers and clusters. FrequentItemsets Mining (FIM) is the most well-known techniques to extract knowledge from dataset. In this paper differential privacy aims to get means to increase the accuracy of queries from statistical databases while minimizing the chances of identifying its records and itemset. We studied algorithm consists of a preprocessing phase as well as a mining phase. We under seek the applicability of FIM techniques on the MapReduce platform, transaction splitting. We analyzed how differentially private frequent itemset mining of existing system as well.*

Keywords: Frequent itemset mining, Differential Privacy, Transaction Splitting.

1. Introduction

Frequent sets play an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters. The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database, can be seen as one of the most basic tasks in Data Mining. The original motivation for searching frequent sets came from the need to analyze so called supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. Frequent sets of products describe how often items are purchased together.

The existing system has problem of tradeoff between utility and privacy in designing a differentially private FIM algorithm. The existing system does not deal with the high utility transactional itemsets. Existing methods has large time complexity. Existing system gives comparatively large size output combination. To solve this problem, this project develops a time efficient differentially private FIM algorithm. With communication, data storage technology, a huge amount of information is being collected and stored in the Internet. Data mining, with its promise to efficiently find valuable, non-obvious information from huge databases, is particularly vulnerable to misuse. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. To solve this, we propose an efficient algorithm, namely used hadoop, for parallel processing on high utility item sets. Frequent itemset mining (FIM) is one of the most basic problems in data mining. We present a framework for mining association rules from transactions consisting of different items where the datahas been randomized to preserve privacy of individual transactions [2]. We continue the investigation of the data mining by following:

- categorical data instead of numerical data, and
- Association rule mining instead of classification.

It will focus on the task of finding frequent itemsets in association rule mining.

Definition: Suppose it have a set I of n items:I={a1,a2,a3….an}
Let T be a sequence of N transactions T={t1,t2…tn}
Where each transactions is a subset of I
Given an itemset A϶ I;
Its supp$^{(T)}$(A) is defined as

$$\mathrm{supp}^{T}(A) := \frac{\#\{t \in T \mid A \subseteq t\}}{N}.$$

An itemset A ϵ I is called frequent in T if supp$^{(T)}$(A)≥ t; where t is a user-defined parameter.

In the mining phase, to offset the information loss caused by transaction splitting, It devise a run-time finding method to find the actual support of itemsets in the original database. Here, we search the applicability of FIM techniques on the MapReduce platform. It is a parallel distributed programming framework introduced in [4, 6], which can process large amounts of data in a massively parallel way using simple commodity machines. We use MapReduce to implement the parallelization of algorithm, thereby improving the overall performance of frequent itemsets mining.

2. Existing System

- **C. Dwork [5]** The author give a general impossibility result showing that a formalization of Dalenius‟ goal along the lines of semantic security cannot be achieved. Contrary to intuition, a variant of the result threatens the privacy even of someone not in the database. This state of affairs suggests a new measure, differential privacy, which, intuitively, captures the increased risk to one‟s privacy incurred by participating in a database. The techniques developed in a sequence of papers, culminating in those described in, can achieve any desired level of privacy under this measure. In many cases, extremely accurate information about the database can be provided while simultaneously ensuring very high levels of privacy.
- C. Zeng, J. F. Naughton, and J.-Y. Cai,[22],In this author elaborates difficulties of finding good utilities and privacy and also they have proposed differentially private algorithm for the top-k item set mining. In general it is

difficulties occur during processing of long transaction so they had investigate an approach that begins by truncating transactions that contains more items, trading off errors introduced by the truncation with those introduced by the noise added to guarantee privacy. their algorithm solves the frequent item set mining problem in which they find all item set whose support exceeds a threshold. The advantage of this algorithm is it achieves better F-score unless k is small.

- N. Li, W. Qardaji, D. Su, and J. Cao [8], In this paper, they searched the problem of how to perform frequent itemset mining on transaction databases while satisfying no of privacy. They propose an approach, called PrivBasis, which leverages a novel notion called basis sets. A θ-basis set has the property that any itemset with frequency highest than θ is a subset of some basis. They represented algorithms for privately constructing all basis set and then using it to find the most frequent itemsets. Experiments show that our approach greatly outperforms the state of the art.
- Maurizio Atzori, F. Bonchi, F. Giannotti [18**],** In this paper author show that this belief is ill-founded. By concept of *k-anonymity* from the source data to the extracted patterns, they formally characterize the notion of a threat to anonymity in the context of pattern, and gives a methodology to efficiently and effectively show all such possible threats that arise from the disclosure of the set of patterns. On this basis, they gain a formal notion of privacy protection that allows the disclosure of the extracted knowledge while protecting the anonymity of the individuals in the source database. Rather in order to handle the cases where the threats to anonymity cannot be avoided, they study how to eliminate such threats by means of pattern distortion performed in a dataset.
- Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke [19], Author present 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, the searched rules can unfortunately be exploited to gain privacy. They analyze the nature of privacy and propose a class of operators that are much more effective than uniform randomization in limiting the breaches. They prove formulae for an unbiased support estimator and its variance, which allow us to get backitem set supports from randomized database, and show how to incorporate these formulae into mining algorithms. At last, they present experimental analysis that validates the algorithm by applying it on real datasets.
- W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis [16], They found frequent item sets is the most costly task in association rule mining. This task to a service provider brings several benefits to the data owner such as cost relief and a less obligation to storage and computational resources. Mining results, can be loss if the service provider (i) is honest but makes error in the mining process, or (ii) is lazy and reduces costly computation, returning incomplete results, or (iii) is malicious and infected the mining results. They show the integrity issue in the outsourcing process, i.e., how the data owner verified the accuracy of the mining results. For this purpose, we propose and develop an audit environment, which consists of a dataset transformation method and a result verification method. The main component of its audit environment is an artificial itemset planting (AIP) technique. They provide a theoretical base on our method by showing its appropriateness and showing probabilistic guarantees about the correctness of the verification process. Through analytical and experimental studies, they represented that their technique is both effective and efficient.
- E. Shen and T. Yu[23],author discovered frequent graph patterns in a graph database .If graph database contains sensitive data of individuals, liberating discovered frequent patterns may present a threat to privacy of individuals. so they proposed the first differentially private algorithm for mining frequent graph patterns. Their proposed solution incorporates the process of graph mining and privacy protection into an MCMC(Markov Chain Monte Carlo) sampling framework.
- SheelaGole and Bharat Tidke [24], implement FIM algorithm based on MapReduce programming model. They introduced ClustBigFIM algorithm that works on large datasets with increased execution efficiency using pre-processing. Their Result of ClustBigFIM shows that it works on BigData very efficiently and with higher speed.

3. Related Work

In differentially private frequent mining it uses different algorithm to find itemset as follows :

- ***UP-Growth:*** The basic method to generate high utility item sets is the FP-Growth [3] algorithm. However, it produces huge number of item sets. In order to reduce the number of item sets and produce only high utility item sets UP-Growth algorithm [21] is used. Utility pattern growth algorithm for mining high utility item set .
- **FP-Growth:** The FP-Growth algorithm skips the candidate itemset generation process by using a compact tree structure to store itemset frequency information. FP-Growth works in a divide and conquers way. It requires two scans on the database. FP-Growth first computes a list of frequent items sorted by frequency in descending order (F-List) during its first database scan [15].
- **Frequent itemset mining**: A frequent itemset mining algorithm takes as input a dataset consisting of the transactions by a group of individuals, and produces as output the frequent itemsets [22]. This immediately creates a privacy concern how can we be confident that publishing the frequent itemsets in the dataset does not reveal private information about the individuals whose data is being studied.
- **PFP-growth:** We devise partitioning strategies at different stages of the mining process to achieve balance between processors and adopt some data structure to reduce the information transportation between processors. The experiments on national high performance parallel computer show that the PFP-growth is an efficient parallel algorithm for mining frequent itemset.
- **Apriori**: Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets

(item combinations). The set of possible itemsets is the power set over I and has size 2n − 1 (excluding the empty set which is not a valid itemset). Although the size of the power set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity)

The below table shows comparison between each algorithm used in frequent itemset mining.

Table 1: Comparison between above algorithm

Algorithm/ Parameter	FP-growth	PFP-growth	UP-growth	Apriori	FIM
Based on	Bottom up approach	Breadth-first search	Tree structure	Tree structure	transaction splitting
Efficiency	Low	Average	Average, High for dataset	Average	High
Time Complexity	High	High	Average	Average	low
Performance	Low	Medium	Medium	average	high

4. System Model

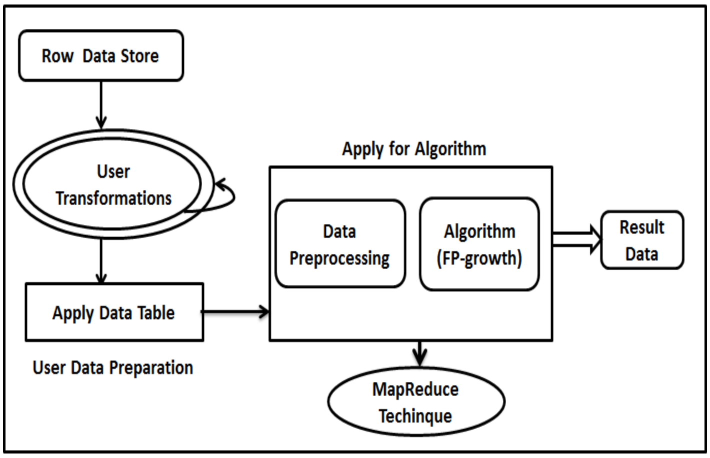

Figure 1: System Diagram

Figure 1 shows system model Frequent ItemsetMining. It takes data from row data store then by using user transformation it apply data processing and algorithm on data table. It uses Map reduce to find frequent itemset and gives result. Main aim of using mapreduce is to handle big data. As we know transaction file may contain sensitive and huge data. We are proposing private frequent itemset mining based on mapreduce so that long dataset will get splits into multiple parts and these files will be parallel handle, which in turns reduce space and time complexity. It was proposed which is used to obtain frequent itemsets from the dataset.

MINimal Infrequent Itemsets (MINIT) is the algorithm designed specifically for mining minimal infrequent itemsets [9]. MINIT computes both minimal (weighted) and non-minimal (unweighted) infrequent itemset mining from unweighted data which is based on algorithm and also proved that the minimal infrequent itemset problem is NP-complete problem. Different from [10], Clifton and Kantarcioglu [11] consider the dataset is horizontally partitioned and model the problem as a secure multi-party computation. Evfimievski et al. [12] present a set of randomization operators the privacy breaches in FIM. Based on k-anonymity [13], Atzorietal.The most relevant work from the statistical database literature is the work by Warner [4], where he represented the „„randomized response‟‟ method for survey results. Through formal privacy analysis, heshow that our PFP-growth algorithm is differentially private. Extensive experiments on real database illustrate that our PFP-growth algorithm and its outperforms the state-of-the-art techniques. The problem of outsourcing the task of data mining with accurate result was introduced in our previous work [16]. Frequent itemsets, as name suggest, are the sets of items often occurring frequently in transactional dataset. It leads to discovery of the association rules from the datasets. Frequent itemsets are appearing with frequency more than a user-specified threshold. This task to a service provider brings several bents to the data owner such as cost relief and a less commitment to storage and computational resources. The following section shows algorithm and mining long patterns.

4.1 Algorithm

Construct data Set: Construct a Basis Set Using Frequent Items and Pairs
Input: F, frequent items, and P, frequent pairs.
Output: B, a basis set covering all maximal cliques in the graph (F, P).

```
1: function Construct Basis Set(F, P)
2: B1 ← all maximal cliques of size at least 2 in the graph given by P
3: B2 ← items in F but not in P, divided into the smallest number of itemsets such that each contains at most 3 items
4: Repeatedly find Bi, Bj ∈ B1 such that merging Bi and Bj results in the huge reduction of average-case error variance (EV) when using B = B1 ∪B2 to obtain no. of itemsets in F
```

```
and P; and update B1 by merging Bi, Bj; stop when no merging reduces EV
5: Again find Bi ∈ B2 such that deleting Bi and moving items in Bi to bases in B1 ∪ B2 with smallest sizes results in the largest EV-reduction; update B when Bi is found; stop when no such Bi can be found
6: return B = B1 ∪ B2
7: end.
```

5. Conclusion and Future Work

In this paper, we investigate the problem of designing a differentially private FIM algorithm. We use differential privacy to stop the potential information exposure about individual recordset during the data mining process. Here we studied system model of Frequent Itemset Mining using Map-reduce. We put forward algorithm and mining long patterns.. It minimizes time required for large dataset. As we are using map reduce here, can also handle huge size dataset without any problem. We represented comparative table between different algorithms used in FIM. As our future work we plan to design more effective differentially private FIM on big data.

6. Acknowledgment

We would like to thank all the authors of different research papers referred during writing this paper. It was very knowledge gaining and helpful for the further research to be done in future.

References

[1] Sen Su, ShengzhiXu, Xiang Cheng, Zhengyi Li, and Fangchun Yang ,"Differentially Private Frequent Itemset Mining via Transaction Splitting",,IEEE Trans. On Knowl. And Data Engg., Vol. 27, NO. 7, Jul 2015
[2] AlexandreEvfimievskia, RamakrishnanSrikantb, RakeshAgrawalb,JohannesGehrkea Privacy preserving mining of association rules.
[3] Office of the Information and Privacy Commissioner, Ontario, Data Mining: Staking a Claim on Your Privacy , Jan 1998.
[4] S. Warner , Randomized response: a survey technique for eliminating evasive answer bias, J. Am. Stat. Assoc.,1965.
[5] C. Dwork, "Differential privacy," in ICALP, 2006.
[6] An Audit Environment for Outsourcing of Frequent Itemset Mining
[7] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis. Security in outsourcing of association rule mining. ,in VLDB, 2007.
[8] Ninghui Li, WahbehQardaji, Dong Su, Jianneng Cao,"PrivBasis: Frequent Itemset Mining with Differential Privacy.", in *VLDB*, 2012.
[9] R. Agrawal, T. Imielinski, and A. Swami," Mining association rules between sets of items in large databases", in Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93), , pp. 207–216, May 1993.
[10] JyothiPillai, O.P.Vyas, "Overview of Itemset Utility Mining and its Applications ",International Journal of Computer Applications (0975 – 8887) Volume 5– No.11, Aug 2010
[11] 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, , pp. 639–6442002.
[12] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributedmining of association rules on horizontally partitioned data,",IEEE Trans. Knowl. Data Eng., Vol. 16, no. 9, pp. 1026–1037, Sep.2004.
[13] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacypreserving mining of association rules," in Proc. 8th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, , pp. 217–228, 2002.
[14] L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J.Uncertainity Fuzziness Knowl.-Base Syst., vol. 10, no. 5, pp. 557–570,2002.
[15] Revealing Information while Preserving PrivacyIritDinurKobbiNissim
[16] 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.
[17] PFP: Parallel FP-Growth for Query Recommendation Haoyuan Li Google Beijing Research, Beijing, 100084, China
[18] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity preserving pattern discovery," VLDB Journal, 2008.
[19] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in KDD, 2002.
[20] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in FOCS, 2007.
[21] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases",IEEE Trans. On Knowledge And Data Engg., VOL. 25, NO. 8, AUG 2013.
[22] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," in VLDB, 2012.
[23] E. Shen and T. Yu, "Mining frequent graph patterns with differentialprivacy," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. DiscoveryData Mining, , pp. 545–553, 2013.
[24] SheelaGole and Bharat Tidke,"ClustBigFIM-Frequent itemset mining of bigdata using pre-processing based on map reduce framework"