ISSN 2395-1621

Private Frequent Itemset Mining Using PFP Growth Algorithm(Transaction Splitting)

[#1]Nikita C. Khandare, [#2]Shrikant Nagure

[1]nikitakhandare7@gmail.com

[#1]M.E.,Dept. of Computer,RMD Sinhgad School of Engineering,Pune,India
[#2]Assistant Professor,Dept. of Computer,RMD Sinhgad School of Engineering,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. Frequent Item sets Mining (FIM) is the most well-known techniques to extract knowledge from dataset. Accordingly, we investigate an approach that begins by truncating long transactions, trading off errors introduced by the truncation with those introduced by the noise added to guarantee privacy. Experimental results over standard benchmark databases show that truncating is indeed effective. We studied algorithm consists of a pre-processing phase as well as a mining phase. We under seek the applicability of FIM techniques on the Map Reduce platform, transaction splitting. We analysed how differentially private frequent item set mining of existing system as well.Related work has proposed differentially private algorithms for the top-k itemset mining problem ("find the k most frequent item-sets".) An experimental comparison with those algorithms show that our algorithm achieves better F-score unless k is small.

Keywords— Frequent itemset mining,differential privacy,transaction -splitting,dynamic reduction.

ARTICLE INFO

Article History

Received:19th December 2015

Received in revised form :

21st December 2015

Accepted:23rd December, 2015

Published online :

28th December 2015

I. INTRODUCTION

Recently, concomitant with the increasing ability to collect personal data, privacy has become a major concern. In this paper, we focus on privacy issues that arise in the context of finding frequent item-sets in "transactional" data. Frequent item-sets mining is widely used in many applications, perhaps the best known of which is market basket analysis. The goal of frequent item-sets mining in market basket analysis is to find sets of items that are frequently bought together, which is helpful in applications ranging from product placement to marketing and beyond. The problem of developing efficient algorithms for frequent item-sets mining has been widely studied by our community. However,the privacy issues arising in frequent item-sets mining have received little attention. A frequent item-sets 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 immediately creates a privacy concern—how can we be confident that publishing the frequent item-sets in the dataset does not reveal private information about the individuals whose data is being studied? This problem is compounded by the fact that we may not even know what data the individuals would like to protect nor what background information might be possessed by an adversary. Fortunately, a possible answer to that challenge is presented by differential privacy, which intuitively guarantees that the presence of an individual's data in a dataset does not reveal much about that individual. In this paper, we explore the possibility of developing differentially private frequent item-sets mining algorithms. Our goal is to guarantee differential privacy while still finding useful frequent item-sets. To the best of our knowledge, ours is the first paper to explore differential privacy in this context. An obvious but important observation is that privacy is just one aspect of the problem; utility also matters. In this paper, we quantify

the utility of a differentially private frequent item-sets mining by its likelihood to produce a complete and sound result. Intuitively speaking, "completeness" requires an algorithm to include the sufficiently "frequent" item-sets, and "soundness" needs an algorithm to exclude the sufficiently "infrequent" ones. We start by showing the tradeoff between privacy and utility in frequent item-sets mining. Our result indicates that no matter how sophisticated a differentially private frequent item-sets mining algorithm is, it has to suffer from a huge risk of revealing an individual's information in order to satisfy a non-trivial utility requirement. In view of that result, we utilize the constraint on each transaction's cardinality in databases to promote the utility of a frequent item-sets mining algorithm while ensuring differential privacy. We first consider a constrained frequent item-sets mining problem —

frequent 1-item-sets mining in which we are only interested in the item-sets of cardinality 1. We prove that by limiting the maximal cardinality of transactions, we can significantly promote the utility of a specific frequent 1-item-sets mining algorithm while still guaranteeing differential privacy. At the most abstract level, that algorithm works by adding "noise" in computing frequent item-sets. As we will prove, if this is done properly, that algorithm is differentially private. Motivated by that theoretical result, we impose the cardinality constraint to a database by truncating transactions. However, that truncating approach has privacy implications since it needs to access the database. Fortunately, we can prove that as long as that truncating approach is "local", which takes as input a single transaction, and produces as an output only depending on the input, then applying any differentially private algorithm on the truncated database users, by truncating transactions, we are adding a new kind of error by "throwing away" information that was in the original dataset. In an experimental study using three benchmark datasets, we find that this trade-off is worth making. In particular, we also evaluate our frequent 1-item-sets mining algorithm using an intuitive metric F-score, which measures the percentage of correct frequent 1-item-sets, and our results indicate that by truncating transactions, our algorithm gets much better F-score than the non-truncating algorithm. Encouraged by that success, we generalize our idea of truncating transactions to frequent k-item-sets mining in which we are interested in item-sets of cardinality not exceeding k. However, as described in the body of the paper, that generalization is far from trivial since

(a) the total number of item-sets is too large.
(b) the precise computation of the required "noise" to guarantee differential privacy is hard.
(c) Specifically, we first propose a naive frequent k item-sets mining algorithm which attacks by exploiting the "a priori" property of frequent item-sets to reduce the number of computations, and
(d) by computing the upper bound of the required "noise." However, we find the performance of our naive algorithm is not satisfying in practice.

To remedy that problem, we propose two heuristic methods to improve our naive algorithm. We further extend our frequent k-item-sets mining algorithm to frequent item-sets mining by estimating the maximal cardinality of frequent item-sets in a differentially private way, and then applying our frequent k-item-sets mining algorithm by setting k to be that estimated maximal cardinality. In an experimental study using the benchmark datasets, we find that our differentially private frequent item-sets mining algorithm is able to generate reasonable accuracy on these test datasets.

II. LITERATURE SURVEY

1. We achieve good privacy and utility may prove elusive we investigate the problem of signing a differentially private FIM algorithm. We use differential privacy to stop the potential information exposure about individual record set during the data mining process. Here we studied system model of Frequent Item set 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.

2. We have proposed a differentially private frequent item set mining algorithm. We have precisely quantified the trade-off between privacy and utility in frequent item set mining, and our results indicate that in order to satisfy a non-trivial utility requirement, a frequent item set mining algorithm incurs a huge risk of privacy breach. However, we find that we can greatly promote the utility of a differentially private frequent item-set mining algorithm by limiting the maximal cardinality of transactions. Motivated by that observation, we have proposed a new differentially private frequent item set mining algorithm. Our results on benchmark datasets indicate that in comparison to the latest algorithm on publishing transactional data in a differentially private way , our algorithm improves the F score of frequent item sets by more than 200% in one dataset, and by an order of magnitude on the other two datasets. Our results also show that our algorithm significantly improves the quality of top-k frequent item sets comparing to the differentially private top-k frequent item set mining algorithm proposed in except when k is small. There are many potential opportunities for future work. One such direction would be to explore other more sophisticated truncating algorithms. Another direction would be to explore alternative methods to limit the information loss due the truncating. Finally, the success of our algorithm relies on the assumption that the "short" transactions dominate the datasets. How to deal with datasets dominated by long transactions is an open problem, although with no constraints on the database our theoretical results on privacy/ utility tradeoffs suggest that algorithms that simultaneously.
3. we investigate the problem of designing a differentially private FIM algorithm. We use differential privacy to stop the potential information exposure about individual record set during the data mining process. Here we studied system model of Frequent Item-set 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.

4. Frequent item set is very important to find out from the large data set. Online transaction has increases need to find out which item has frequently access. Privacy mechanism discussed adds an amount of noise in the data set. Summary of the in-depth analysis of few algorithm s is done which of improving the efficiency of frequent itemset mining algorithm.

III.SYSTEM ARCHITECTURE

In this paper, we explore the possibility of designing a differentially private FIM algorithm which can not only achieve high data utility and a high degree of privacy, but also offer high time efficiency. To this end, we propose a differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. The PFP-growth algorithm consists of a pre-processing phase and a mining phase. In the pre-processing phase, to improve the utility and privacy tradeoff, a novel smart splitting method is proposed to transform the database. For a given database, the pre-processing phase needs to be performed only once. In the mining phase, to offset the information loss caused by transaction splitting, we devise a run-time estimation method to estimate the actual support of item-sets in the original database. In addition, by leveraging the downward closure property, we put forward a dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. Through formal privacy analysis, we show that our PFP-growth algorithm is ϵ-differentially private. Extensive experiments on real datasets illustrate that our PFP-growth algorithm substantially outperforms the state-of-the-art techniques.

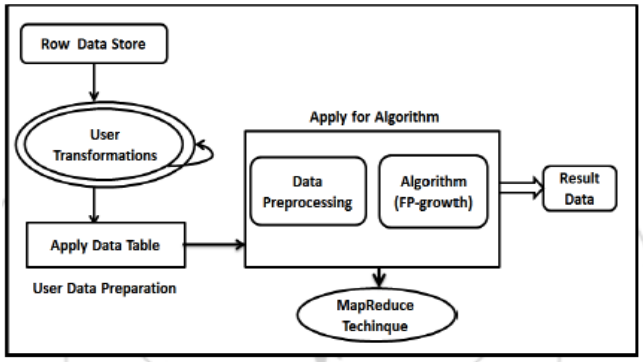

Figure 1: System Diagram

Fig shows system model Frequent Item-set Mining. 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 item-set and gives result. Main aim of using map-reduce is to handle big data. As we know transaction file may contain sensitive and huge data. We are proposing private frequent item-set mining based on map-reduce 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 item-sets from the dataset. Minimal Infrequent Item-sets (MINIT) is the algorithm designed specifically for mining minimal infrequent item-sets. MINIT computes both minimal (weighted) and non-minimal (un-weighted) infrequent item-set mining from un-weighted data which is based on algorithm and also proved that the minimal infrequent item-set problem is NP-complete problem. Different from , Clifton and Kantarcioglu consider the dataset is horizontally partitioned and model the problem as a secure multi-party computation. Evfimievski present a set of randomization operators the privacy breaches in FIM. Based on k-anonymity .The most relevant work from the statistical database literature is the work by Warner, where he represented the randomized response method for survey results. Through formal privacy analysis, he showed that our PFP-growth algorithm is differentially private. Extensive experiments on real database illustrate that our PFP-growth algorithm and its out performs the state-of-the-art techniques. The problem of outsourcing the task of data mining with accurate result was introduced in our previous work, Frequent item-sets, 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 item-sets 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.

SYSTEM MODULES AND DESCRIPTION

Frequent item-set 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 item-sets. This immediately creates a privacy concern how can we be confident that publishing the frequent item-sets in the dataset does not reveal private information about the individuals whose data is being studied.

Differential Privacy :

Differential privacy has gradually emerged as the de factor standard notion of privacy in data analysis.

For two databases D and D', they are neighbouring databases if they differ by at most one record. Formally, the differential privacy is defined as follows.

Definition 1:

(ϵ-differential privacy). A private algorithm A satisfies ϵ-differential privacy iff for any two neighboring databases D and D', and any subset of outputs

$$S \subseteq \text{Range(A)},$$

$$\Pr[A(D) \in S] \leq e\epsilon \times \Pr[A(D) \epsilon\ S],$$

where the probability is taken over the randomness of A. A fundamental concept in differential privacy is the sensitivity. The amount of injected noise is carefully calibrated to the sensitivity. The sensitivity of count queries is used to measure the maximum possible change in the outputs over any two neighbouring databases.

Transaction Splitting:

To better understand the benefit of transaction splitting, we apply it to Apriori by modifying TT. In particular, in the first database scan, we find frequent 1-item-sets from the database which is transformed by our smart splitting method. In each subsequent database scan, to preserve more information, we re-transform the database in the following manner. For each long transaction, we divide it into subsets

by recursively using TT's smart truncating method. The weights of resulting subsets are evenly assigned. In addition, in the mining process, we use our run-time estimation method to quantify the information loss caused by transaction splitting. The results are shown in the performance of TT is significantly improved by adopting our transaction splitting techniques.

IV. CONCLUSION

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 record set during the data mining process. So the proposed PFP algorithm which consists of two phases; pre-processing to better improve the privacy tradeoff and mining phase in which runtime estimation is proposed to offset the information loss due to transaction splitting . 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.

REFRENCES

[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.

