A Survey on Frequent Itemset Mining using Differential Private with Transaction Splitting

Bhagyashree R. Vhatkar¹, Prof. (Dr.). S.A. Itkar²
¹Computer Department, P.E.S. Modern College of Engineering
²Computer Department, P.E.S. Modern College of Engineering

Abstract- In recent years, individuals are interested in designing differentially private data mining algorithms. Many researchers are now working on designing of data mining algorithms which provides differential privacy. In this paper, to explore the possibility of designing a differentially private FIM algorithm, can not only achieve high data utility and a high degree of privacy, but also offer high time efficiency. To this end, the propose a differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. The Private FP-growth algorithm consists of a preprocessing phase and a mining phase. In the preprocessing phase, to improve the utility and privacy tradeoff, a novel smart splitting method is proposed to transform the database. A frequent itemset mining with differential privacy is important which will follow two phase process of preprocessing and mining. Through formal privacy analysis, show that our Private FP-growth algorithm is \(\epsilon\)-differentially private. Extensive experiments on real datasets illustrate that our PFP-growth algorithm substantially outperforms the state-of-the-art techniques.

Keywords-Frequent itemset mining, Differentially private, Preprocessing, Mining, Private FP-Growth, Transaction splitting.

I. INTRODUCTION

In the database, where each transaction contains a set of items, FIM tries to find itemsets that occur in transactions more frequently than a given threshold. A variety of algorithms have been proposed for mining frequent itemsets. The Apriori and FP-growth are the two most important ones. In particular, Apriori is a breadth first search, candidate set generation-and-test algorithm. It needs one database scan if the maximal length of frequent itemsets is one. In contrast, FP-growth is a depth-first search algorithm, which requires no candidate generation. In FP-growth only performs two database scans, which makes FP-growth an order of magnitude faster than Apriori. The appealing features of FP-growth motivate us to design a differentially private FIM algorithm based on the FP-growth algorithm. In this paper, the differentially private FIM algorithm should not only achieve high data utility and a high degree of privacy, but also offer high time efficiency. Although several differentially private FIM algorithms have been proposed, they are not aware of any existing studies that can satisfy all these requirements simultaneously. The resulting demands necessarily bring new challenges. It has been shown that the utility-privacy trade-off can be improved by limiting the length of transactions.

In previous work presents an Apriori-based differentially private FIM algorithm. It enforces the limit by truncating. In particular, in each database scan, to preserve more frequency information, it advantage to discovered frequent itemsets to re-truncate transactions. However, FP-growth only performs two database scans. There is no opportunity to re-truncate transactions during the mining process. Thus, the transaction truncating approach is not suitable for FP-growth. In addition, to avoid privacy breach, the add noise to the support of itemsets. FP-growth is a depth-first search algorithm not similar to Apriori. It
is hard to obtain the exact number of support computations of i-itemsets during the mining process. A naive approach to calculate the noisy support of i-itemset is to use the number of all possible i-itemsets. However, it will definitely produce invalid results. Apriori-based algorithm in is significantly improved by adopting our transaction splitting techniques:

1) The revisit the trade-off between utility and privacy in designing a differentially private FIM algorithm. The demonstrate that the trade-off can be increased by our novel transaction splitting techniques. Such techniques are not only suitable for FP-growth, but also can be utilized to design other differentially private FIM algorithms.
2) To generate a time-efficient differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. In particular, by leveraging the downward closure property, a dynamic reduction method is proposed to dynamically reduce the amount of noise added to guarantee privacy during the mining process.
3) Through formal privacy analysis, the show that our PFP-growth algorithm is \( \epsilon \)-differentially private.

The paper is organized into four sections: Section II gives brief review of the differentially private frequent itemset mining. Section III describes the performance parameters considered to compare these approaches and finally, Section IV summarizes and presents the conclusions.

II. RELATED WORK

Differentially Private Frequent Itemset Mining via Transaction Splitting, Sen Su, Shengzhi Xu[1]. In this paper the features of FP-growth motivate us to design a differentially private FIM algorithm based on ther FP-growth algorithm. We argue that a practical differentially private FIM algorithm should not only achieve high data utility and high degree of privacy, but also offer high time efficiency. FP-growth only performs two database scan. There is no opportunity to re-truncate transaction during mining process. Private FP-growth (PFP-growth) algorithms, which consist of preprocessing phase and mining phase. In preprocessing phase we transform the database to limit the length of transactions. The preprocessing phase is irrelevant to userspecified thresholds and needs to be performed only once for a given database. That is, if a transaction has more items than the limit, we divide it into multiple subsets (i.e., sub-transactions) and guarantee each subset is under the limit. We devise a novel smart splitting method to transform the database. In particular, to ensure applying \( \epsilon \)-differentially private algorithm on the transformed database still satisfies \( \epsilon \)-differential privacy for the original database, we propose a weighted splitting operation. Moreover, to preserve more frequency information insubsets, we propose a graph-based approach to reveal the correlation of items within transactions and utilize such correlation to guide the splitting process.

On differentially to design private frequent itemset mining, Zeng C[2]. In this paper, we focus on privacy issues that arise in the context of finding frequent itemsets in transactional data. It can explore the possibility of developing differentially private frequent itemset mining algorithms. Our goal is to guarantee differential privacy without obliterating the utility of the algorithm. A closer investigation of this negative result reveals that it relies on the possibility of very long transactions. This raises the possibility of improving the utility-privacy trade-off by limiting transactions cardinality. Of course, one cannot in general impose such a limit; so instead, we explore enforcing the limit by truncating transactions [2]. That is, if a transaction has more than a specified number of items, we delete items until the transaction is under the limit. Of course, this deletion must be done in a deferentially private way; perhaps equally important, while it reduces the error due to the noise required to enforce privacy. Idea of limiting the maximal cardinality of transactions is simple we truncate a transaction whose cardinality
violates that constraint by only keeping a subset of that transaction. Of course, that truncating approach incurs certain information loss. However, if the cardinality of transactions in a dataset follows a distribution in which most are short and a few are long, then these few long transactions, while having little impact on which itemsets are frequent, have a major effect on the sensitivity.[2]

PrivBasis: Frequent Itemset Mining with Differential Privacy by Ninghui Li[3]. In this paper we propose a novel approach that avoids the selection of top k itemsets from a very large candidate set. More specially, we introduce the notion of basis sets. A 0-basis set $B = B_1; B_2; ...; B_w$; where each $B_i$ is a set of items, has the property that any itemset with frequency higher than _ is a subset of some basis $B_i$. A good basis set is one where $w$ is small and the lengths of all $B_i$ are also small. Given a good basis set $B$, one can reconstruct the frequencies of all subsets of $B_i$ with good accuracy. One can then select the most frequent itemsets from these. We also introduce techniques to construct good basis sets while satisfying differential privacy. It meets the challenge of high dimensionality by projecting the input data set onto a small number of selected dimensions that one cares about. In fact, PrivBasis often uses several sets of dimension $s$ for such projections, to avoid any one set containing too many dimensions. Each basis in $B$ corresponds to one such set of dimensions for projection. Our techniques enable one to select which sets of dimensions are most helpful for the purpose of finding the most frequent itemsets. A key concept introduced in this approach is the notion of Truncated Frequencies (TF). The TF technique tries to address the running time challenge by pruning the search space, but it does not address the accuracy challenge.

J.han, J.pei[4] Mining frequent patterns in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist a large number of patterns and/or long patterns. In this study, to propose a novel frequent-pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. Efficiency of mining is achieved with three techniques: (1) a large database is compressed into a condensed, smaller data structure, FP-tree which avoids costly, repeated database scans, (2) our FP-tree-based mining adopts a pattern-fragment growth method to avoid the costly generation of a large number of candidate sets, and (3) a partitioning-based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods.

J.Vaidya and C.Clifton[5] This paper addresses the problem of association rule mining where transactions are distributed across sources. Each site holds some attributes of each transaction, and the sites wish to collaborate to identify globally valid association rules. However, the sites must not reveal individual transaction data. We present a two-party algorithm for efficiently dis-covering frequent itemsets with minimum support levels, without either site revealing individual transaction values. To present a framework for mining association rules from transactions consisting ofcategorical items where the data has been randomized to preserve privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward “uniform” randomization, the discovered rules can unfortunately be exploited to and privacy breaches analyze the nature of privacy breaches and propose a class of randomization operators that are much more effective than uniform randomization in limiting the breaches. The derive formulae for an unbiased support estimator and its variance, which allow us to recover itemset supports from randomized datasets, and show how to
incorporate these formulae into mining algorithms. Finally, to present experimental results that validates the algorithm by applying it on real datasets.

By vertically partitioned, the mean that each site contains some elements of a transaction. Using the traditional market basket" example, one site may contain grocery purchases, while another has clothing purchases. Using a key such as credit card number and date, it can join these to identify relationships between purchases of clothing and groceries. However, this discloses the individual purchases at each site, possibly violating consumer privacy agreements. There are more realistic examples. In the sub-assembly manufacturing process, different manufacturers provide components of the finished product. Cars incorporate several subcomponents; tires, electrical equipment, etc. made by independent producers. Discovering frequent patterns in sensitive data by Bhaskar R[6]. In this paper present two efficient algorithms for discovering the K most frequent patterns in a data set of sensitive records. Our algorithms satisfy differential privacy, a recently introduced definition that provides meaningful privacy guarantees in the presence of arbitrary external information. Differentially private algorithms require a degree of uncertainty in their out-put to preserve privacy. Our algorithms handle this by returning noisy lists of patterns that are close to the actual list of K most frequent patterns in the data. We define a new notion of utility that quantifies the output accuracy of private top-K pattern mining algorithms.[6]

L.Bonomi[16] Frequent sequential pattern mining is a central task in many fields such as biology and finance. However, release of these patterns is raising increasing concerns on individual privacy. In this paper, study the sequential pattern mining problem under the differential privacy framework which provides formal and provable guarantees of privacy. Due to the nature of the differential privacy mechanism which perturbs the frequency results with noise, and the high dimensionality of the pattern space, this mining problem is particularly challenging. In this work, the propose a novel two-phase algorithm for mining both prefixes and substring patterns. In the first phase, our approach takes advantage of the statistical properties of the data to construct a model-based prefix tree which is used to mine prefixes and a candidate set of substring patterns. The frequency of the substring patterns is further refined in the successive phase where the employ a novel transformation of the original data to reduce the perturbation noise.

III. PERFORMANCE PARAMETER

To evaluate the performance of algorithm, we employ the widely used standard metrics.

In an information retrieval scenario, the instances are documents and the task is to return a set of relevant documents given a search term; or equivalently, to assign each document to one of two categories, "relevant" and "not relevant". In this case, the "relevant" documents are simply those that belong to the "relevant" category.

1) Precision: Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. Precision is the probability that retrieved document is relevant.

2) Recall: Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents. Recall is probability that a relevant document is retrieved in search.

3) F-score: It measures the utility of generated frequent itemset.

4) Relative error: To measure the error with respect to the actual support of itemset.
IV. CONCLUSION

In this paper we examine the problem of designing a private PFP with differential privacy algorithm, which consist of preprocessing phase and mining phase. In first phase, to better improve utility trade-off, using smart splitting method. In mining phase, a run time estimation method is proposed to offset the information loss incurred by transaction splitting. By using dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. The PFP-growth algorithm is time efficient and can achieve both utility and good privacy. The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.

REFERENCES

4. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in SIGMOD, 2000.
8. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000.