ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Privacy-Preserving Private Frequent Itemset Mining via Smart Splitting

Nandhini[1], Madhubala[2], Valampuri[3]

PG scholar, Department of CSE, Tagore Institute of Engineering Technology,Salem, Tamil Nadu, India[1]

Assistant Professor, Department of CSE, Tagore Institute of Engineering Technology,Salem, Tamil Nadu, India[2]

Assistant Professor, Department of CSE, Tagore Institute of Engineering Technology,Salem, Tamil Nadu, India[3]

ABSTRACT: FIMI(Frequent Itemset Mining) is one of the problem in data mining. Here we designed private FIM algorithm cannot be achieve high data utility, and high degree of privacy, and also offer high time of efficiency.Inexist,Apriori and FP growth are used. In this paper we proposed the PFP growth algorithm to the smart splitting method. In PFP growth can be consist of two phase (i.e. pre-processing and mining phase).In the pre-processing phase,we transform the database to limit the length of transactions. The pre-processing phase is irrelevant to user specified thresholds and needs to be performed only once for a given database. We argue, to enforce such a limit, long transactions should be split rather than truncated. 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. In the mining phase, given the transformed database and a user-specified threshold, we privately discover frequent itemsets. During the mining process, we dynamically estimate the number of support computations, so that we can gradually reduce the amount of noise required by differential privacy. 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 itemsets in the original database.

KEYWORDS: Frequent itemset mining, Differential privacy, Smart splitting.

I. INTRODUCTION

Frequent itemset mining (FIM) is one of the most fundamental problems in data mining. It has practical importance in a wide range of application areas such as decision support, Web usage mining, bioinformatics, etc. Given a database, where each transaction contains a set of items, FIM tries to find itemsets that occur in transactions more frequently than a given threshold. Despite valuable insights the discovery of frequent itemsets can potentially provide, if the data is sensitive (e.g., web browsing history and medical records), releasing the discovered frequent itemsets might pose considerable threats to individual privacy. Differential privacy has been proposed as a way to address such problem. Unlike the anonymization-based privacy models (e.g., k-anonymity and l-diversity), differential privacy offers strong theoretical guarantees on the privacy of released data without making assumptions about an attacker's background knowledge. In particular, by adding a carefully chosen amount of noise, differential privacy assures that the output of a computation is insensitive to changes in any individual's record, and thus restricting privacy leaks through the results. A variety of algorithms have been proposed for mining frequent itemsets. The Apriori and FP-growth are the two most prominent ones. In particular, Apriori is a breadth first search, candidate set generation-and-test algorithm. It needs l database scans if the maximal length of frequent itemsets is l. In contrast, FP-growth is a depth-first search algorithm, which requires no candidate generation. Compared with Apriori, 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, we argue that a practical 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, we are not aware of any existing studies that can satisfy all these requirements simultaneously. The resulting demands inevitably bring new challenges. It has been shown that the utility-privacy tradeoff can be improved by limiting the length of transactions. Existing work presents an Apriori-based differentially private FIM algorithm. It enforces the limit by truncating transactions (i.e., if a transaction has more items than the limit, deleting items until its length is under the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

limit). In particular, in each database scan, to preserve more frequency information, it leverages 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 proposed is not suitable for FP-growth. In addition, to avoid privacy breach, we add noise to the support of itemsets.

Given an i-itemset X (i.e., X contains i items), to satisfy differential privacy, the amount of noise added to the support of i-itemset X depends on the number of support computations of i-itemsets. Unlike Apriori, FPgrowth is a depth-first search algorithm. It is hard to obtain the exact number of support computations of i-itemsets during the mining process. A naive approach for computing the noisy support of i-itemset X is to use the number of all possible i-itemsets. However, it will definitely produce invalid results.

II. RELATED WORK

In [1] Authors 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 exista large number of patterns and/or long patterns. In [2] Authors In this paper, we 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, we propose a novel two-phase algorithm for mining both prefixes and substring patterns.In [3] Authorswe present a framework for mining association rules from transactions consisting of categorical 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.In [4] Authors outsourcing association rule mining to an outside service provider brings several important benefits to the data owner. These include (i) relief from the high mining cost, (ii) minimization of demands in resources, and (iii) effective centralized mining for multiple distributed owners. This paper proposes substitution cipher techniques in the encryption of transactional data for outsourcing association rule mining.In [5] Authors finding frequent itemsets is the most costly task in association rule mining. Outsourcing this task to a service provider brings several benefits to the data owner such as cost relief and a less commitment to storage and computational resources. Mining results, however, can be corrupted if the service provider (i) is honest but makes mistakes in the mining process, or (ii) is lazy and reduces costly computation, returning incomplete results, or (iii) is malicious and contaminates the mining results.

III. PROPOSED ALGORITHM

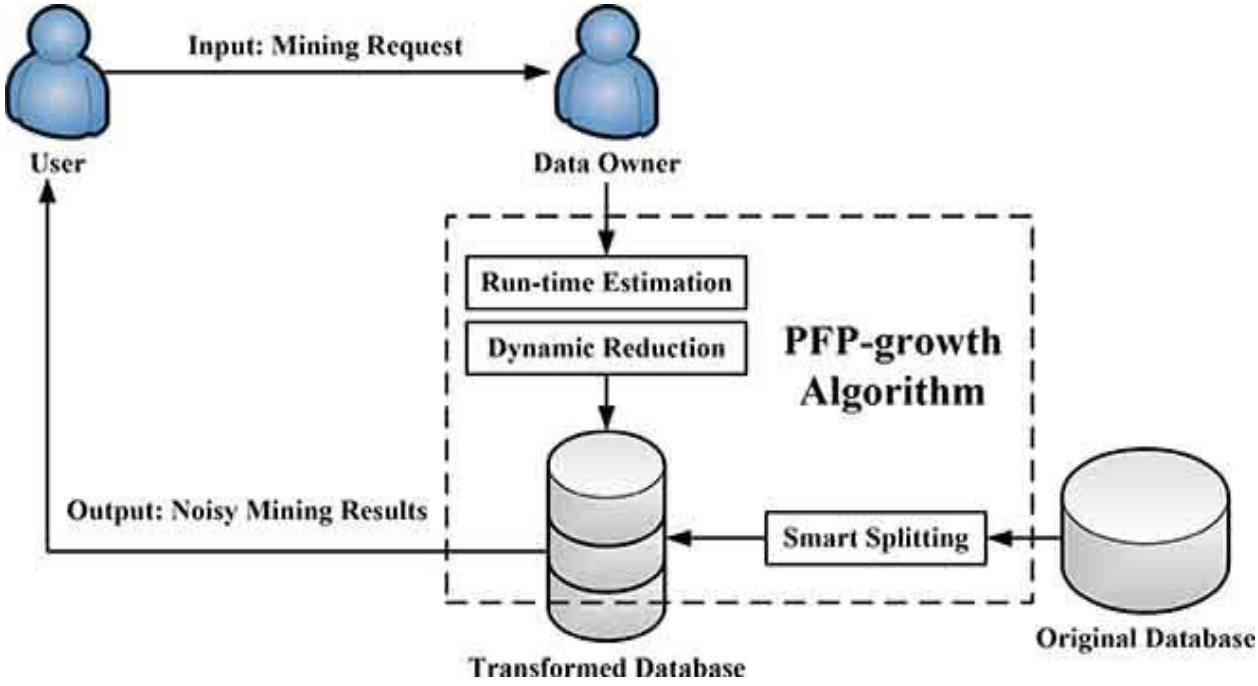
To address these challenges, we present our private FPgrowth (PFP-growth) algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, we transform the database to limit the length of transactions. The preprocessing phase is irrelevant to user specified thresholds and needs to be performed only once for a given database. We argue, to enforce such a limit, long transactions should be split rather than truncated. 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. In the mining phase, given the transformed database and a user-specified threshold, we privately discover frequent itemsets. During the mining process, we dynamically estimate the number of support computations, so that we can gradually reduce the amount of noise required by differential privacy. 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 itemsets in the original database. Runtime estimation method to quantify the information loss caused by transaction splitting Dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015



IV. PSEUDO CODE

Step 1:The PFP-growth algorithm consists of two phases.
In particular, in **the preprocessing phase**, we extract some statistical information from the original database and leverage the smart splitting method to transform the database.
Notice that, for a given database, the preprocessing phase is performed only once.
Step 2:In the **mining phase**,
for a given threshold, we privately find frequent itemsets.
The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.
Besides, we divide the total privacy budget ϵinto five portions:
ϵ1 is used to compute the maximal length constraint,
€2 is used to estimate the maximal length of frequent itemsets,
€3 is used to reveal the correlation of items within transactions,
€4 is used to compute μ-vectors of itemsets, and
ϵ5is used for the support computations.
PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

V. EXPERIMENTAL RESULTS

Login Application: Administrator Have to controlled all access in our application so he have unique password
Upload Products: Administrator Have to upload Product details.
View Details of Products: View Generated table graph by Grid view and Details view.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

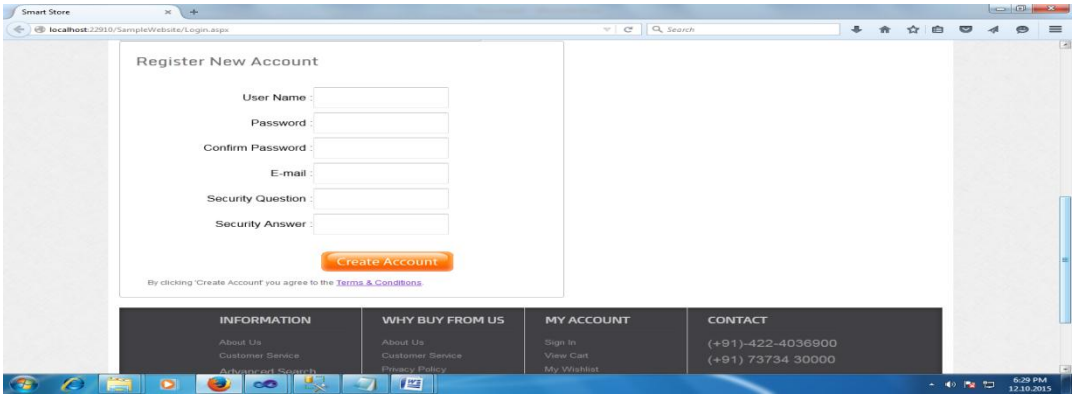
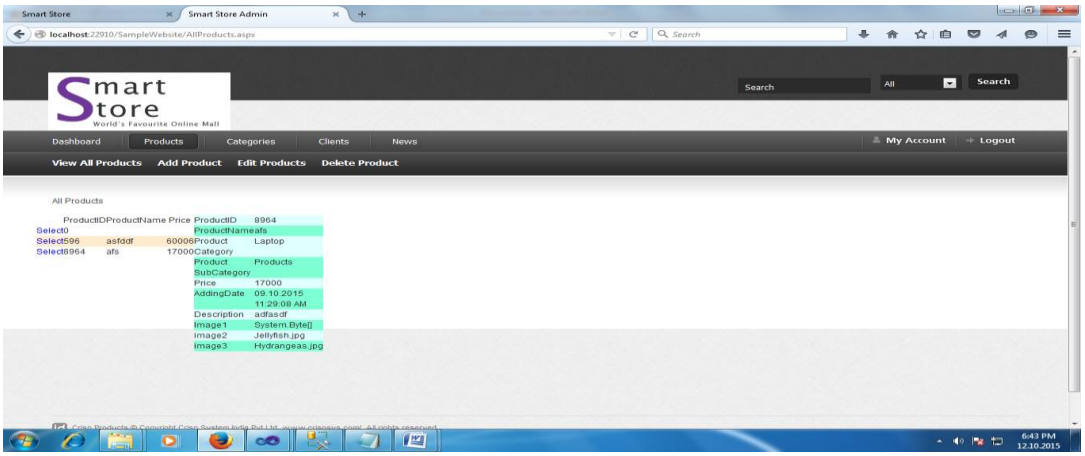


Fig 1:Register in to the new account

Fig 2:Frequently itemsets



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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

VI. CONCLUSION AND FUTURE WORK

In this project, we investigate the problem of designing a differentially private FIM algorithm. We propose our private FP-growth (PFP-growth) algorithm, which consists of a pre-processing phase and a mining phase. In the pre-processing phase, to better improve the utility-privacy trade-off, we devise a smart splitting method to transform the database. In the mining phase, a run-time estimation method is proposed to offset the information loss incurred by transaction splitting. Moreover, 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. Formal privacy analysis and the results of extensive experiments on real datasets show that our PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

REFERENCES

1. J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *KDD*, 2002.
2. W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis,"Security in outsourcing of association rule mining," in *VLDB*,2007.
3. 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.
4. L. Bonomi and L. Xiong, "A two-phase algorithm for mining sequential patterns with differential privacy," in *CIKM*, 2013.
5. M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *TKDE*, 2004.
6. C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," in *VLDB*, 2012.
7. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *VLDB*, 1994.
8. X. Zhang, X. Meng, and R. Chen, "Differentially private setvalued data release against incremental updates," in *DASFAA*,2013.

BIOGRAPHY

Nandhini.M is a Research Assistant in the Computer Science Department, College of Tagore Institute of Engineering and Technology, Salem, Tamil Nadu, India. Her research interests are Data Mining, Networks etc.

Madhubala .P is a Assistant Professor in Department of Computer Science, College of Tagore Institute of Engineering Technology, Salem, Tamil Nadu, India.

Valampuri.U is a Assistant Professor in Department of Computer Science, College of Tagore Institute of Engineering Technology, Salem, Tamil Nadu, India.