Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study of Privacy Preserving Frequent Itemset Mining Via Smart Splitting

A. Jayakumari*
PG Scholar, Department of CSE
Vivekanandha College of Engineering for Women,
Namakkal, Tamilnadu, India

R. Rohini
Assistant Professor, Department of CSE,
Vivekanandha College of Engineering for Women,
Namakkal, Tamilnadu, India

B. Anitha
Assistant Professor, Department of CSE,
Vivekanandha College of Engineering for Women,
Namakkal, Tamilnadu, India

Abstract----A variety of algorithms have been proposed for mining frequent itemsets. 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. In this paper, to 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, the 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 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. For a given database, the preprocessing phase needs to be performed only once. In the mining phase, to offset the information loss caused by transaction splitting, the devise a run-time estimation method to estimate the actual support of itemsets in the original database. In addition, by leveraging the downward closure property, the 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, the 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.

Keywords----Frequent itemset mining, differential privacy, transaction splitting,preprocessing,mining process.

I. INTRODUCTION

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. 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, the 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, the 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 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, the add noise to the support of 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.Apriori-based algorithm in is significantly improved by adopting our transaction splitting techniques:

1). The revisit the tradeoff between utility and privacy in designing a differentially private FIM algorithm. the demonstrate that the tradeoff can be improved 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). The develop 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 ϵ-differentially private.

II. LITERATURE SURVEY

J.han,J.pei[5]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.

The PFP-growth algorithm consists of two phases.**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.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 ϵ into five portions:$\epsilon 1$ is used to compute the maximal length constraint, $\epsilon 2$ is used to estimate the maximal length of frequent itemsets,$\epsilon 3$ is used to reveal the correlation of items within transactions,$\epsilon 4$ is used to compute μ-vectors of itemsets, and $\epsilon 5$ is used for the support computations.PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

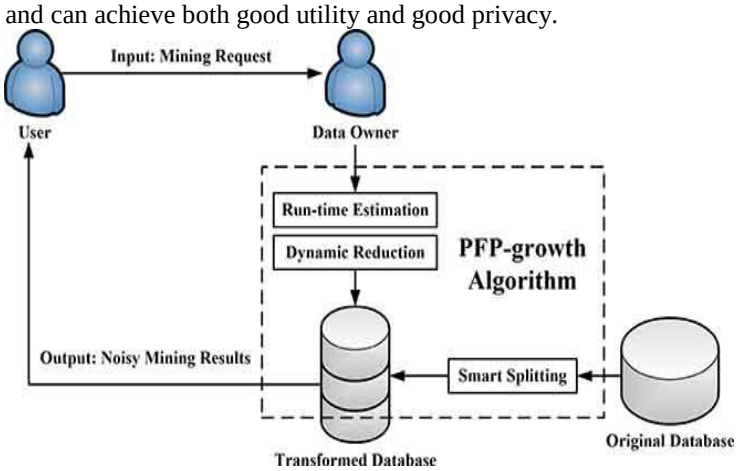

Fig.2.1.Mining Process

L.Bonomi[19]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.

Extensive experiment results using real datasets showed that our approach is effective for mining both substring and prefix patterns in comparison to the state-of-the art solutions. In this paper to propose a novel approach that avoids the selection of top k itemsets from a very large candidate set.More specifically, to introduce the notion of basis sets. A θ-basis set B = {B1,B2, . . . ,Bw}, where each Bi is a set of items, has the property that any itemset with frequency higher than θ is a subset of some basis Bi. A good basis set is one where w is small and the lengths of all Bi's are also small. Given a good basis set B, one can reconstruct the frequencies of all subsets of Bi's with good accuracy. One can then select the most frequent itemsets from these. thes also introduce techniques to construct good basis sets while satisfying differential privacy. Finally,the conducted extensive experiments, and the results show that our approach greatly outperforms the existing approach.

J.Vaidya and C.Clifton[7] 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 e_ciently 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 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 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 validate 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 con- sumer privacy agreements.There are more realistic examples. In the sub-assembly manufacturing process, di_erent manufacturers provide components of the _nished product. Cars incorporate several subcomponents; tires, electrical equipment, etc.; made by independent producers.

W.K.Wong,[9]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. On the other hand, security is an issue; the service provider should be prevented from accessing the actual data since (i) the data may be associated with private information, (ii) the frequency analysis is meant to be used solely by the owner. This paper proposes substitution cipher techniques in the encryption of transactional data for outsourcing association rule mining. After identifying the non-trivial threats to a straightforward one-to-one item mapping substitution cipher, to propose a more secure encryption scheme based on a one-to-n item mapping that transforms transactions non-deterministically, yet guarantees correct decryption. to develop an effective and efficient encryption algorithm based on this method. Our algorithm performs a single pass over the database and thus is suitable for applications in which data owners send streams of transactions to the service provider.

W.K.Wong and D.W.Cheung [10] 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. to address the integrity issue in the outsourcing process, i.e., how the data owner verifies the correctness of the mining results. For this purpose, the propose and develop an audit environment, which consists of a database transformation method and a result verification method. The main component of our audit environment is an artificial itemset planting (AIP) technique.Through analytical and experimental studies, to show that our technique is both effective.

III. CONCLUSION

To investigate the problem of designing a differentially private FIM algorithm. the propose our private FP-growth (PFP-growth) algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, to better improve the utility-privacy tradeoff, to 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, to 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. Further reduce the sensitivity while avoiding too much overhead is an interesting direction for future work. The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.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.

REFERENCES

[1] C. Dwork, "*Differential privacy*," in *ICALP*, 2006.
[2] L. Sweeney, "k-anonymity: *A model for protecting privacy*,"*Int. J. Uncertain. Fuzziness Knowl.-Base Syst*,2002.
[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkita subramaniam,"l-diversity: *Privacy beyond k-anonymity*," in *ICDE*,2006.
[4] R. Agrawal and R. Srikant, "*Fast algorithms for mining association rules,"* in *VLDB*, 1994.
[5] J. Han, J. Pei, and Y. Yin, *"Mining frequent patterns without candidate generation,"* in *SIGMOD*, 2000.
[6] C. Zeng, J. F. Naughton, and J.-Y. Cai, *"On differentially private frequent itemset mining,"* in *VLDB*, 2012.
[7] J. Vaidya and C. Clifton, *"Privacy preserving association rule mining in vertically partitioned data*," in *KDD*,2002.
[8] M. Kantarcioglu and C. Clifton, "*Privacy-preserving distributed mining of association rules on horizontally partitioned data*," *TKDE*, 2004.
[9] W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis,"*Security in outsourcing of association rule mining,*" in *VLDB*,2007.
[10] 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.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "*Privacy preserving mining of association rules*," in *KDD*, 2002.
[12] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, *"Anonymity preserving pattern discovery," VLDB Journal*,2008.
[13] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, *"Discovering frequent patterns in sensitive data,"* in *KDD*, 2010.
[14] N. Li, W. Qardaji, D. Su, and J. Cao, "*Privbasis: frequent itemset mining with differential privacy,"* in *VLDB*, 2012.
[15] F. McSherry and K. Talwar, "*Mechanism design via differential privacy*," in *FOCS*, 2007.
[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, *"Calibrating noise to sensitivity in private data analysis,"* in *TCC*, 2006.
[17] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "*Publishing set-valued data via differential privacy*," in *VLDB*, 2011.
[18] X. Zhang, X. Meng, and R. Chen, "*Differentially private setvalued data release against incremental updates*," in *DASFAA*, 2013.
[19] L. Bonomi and L. Xiong, *"A two-phase algorithm for mining sequential patterns with differential privacy*," in *CIKM*, 2013.
[20] E. Shen and T. Yu, *"Mining frequent graph patterns with differential privacy,"* in *KDD*, 2013.
[21] R. Chen, B. C. M. Fung, and B. C. Desai, *"Differentially private transit data publication: A case study on the montreal transportation system*," in *KDD*, 2012.
[22] R. Chen, G. Acs, and C. Castelluccia, *"Differentially private sequential data publication via variable-length n grams,"* in *CCS*, 2012.
[23] A. Ghosh, T. Roughgarden, and M. Sundararajan, "*Universally utility-maximizing privacy mechanisms," SIAM Journal on Computing*, 2012.
[24] L. Parsons, E. Haque, and H. Liu, *"Subspace clustering for high dimensional data:* a review," *SIGKDD Explorations*, 2004.
[25] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre*,"Fast unfolding of communities in large networks," J. Statistical Mechanics: Theory and Experiment*, 2008.
[26] N. Guttmann-Beck and R. Hassin, *"Approximation algorithms for minimum sum p-clustering," Discrete Applied Mathematics*,1998.
[27] Z. Zheng, R. Kohavi, and L. Mason, "*Real world performance of association rule algorithms,"* in *KDD*, 2001.
[28] "*Frequent itemset mining dataset repository*," http:// fimi. ua.ac.be/data.

