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 11, November 2015

A Survey on Differentially Private Frequent Itemsets Mining

Varsha V. Dabhole, Prof. V. S. Nandedkar

ME Student, Dept. of Computer Engineering, P.V.P.I.T, Bavdhan, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, P.V.P.I.T, Bavdhan, Pune, Maharashtra, India

ABSTRACT: The differentially private frequent itemsets mining. We begin by exploring the theoretical difficulty of simultaneously providing good utility and good privacy in this task. While our analysis proves that in general this is very difficult, it leaves a glimmer of hope in that our proof of difficulty relies on the existence of long transactions (that is, transactions containing many items). 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 Itemsets 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 itemsets. Discovering frequent item set play an important role in mining association rules, clusters ,web long mining and many other interesting pattern among complex data Efficient algorithm for analyze frequent item set based on the memory utilization and performance at the run time .Differential private FIM to find high data utility and high degree of privacy in the database.

KEYWORDS: Frequent item set (FIM); cluster association rule; Differential Private frequent item set.

I. Introduction

The frequent item set play an essential role in many data mining task that try to find out interesting pattern from databases such as association rule, correlation ,sequences, classifiers and clusters .association rule helpful for analyzing customer behavior in retail trade, banking system etc. Frequent item set tries to find item set that occur in transaction more frequently than given threshold.FIM treat all the item having the same unit profit. Differential privacy offer strong privacy of released data without making assumption about an attacker background knowledge . Sequential pattern mining is define to finding statistically relevant pattern . The customer buying first a mobile phone, data cable and memory card if it occur frequently in a shopping history database is a sequential pattern.

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 [5]. 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 data mining, 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:

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

It will focus on the task of finding frequent itemsets in association rule 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. 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? 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. These compounding factors are exactly the ones addressed by differential privacy [8], which intuitively guarantees that the presence of an individual's data in a dataset does not reveal much about that individual. Accordingly, in this paper we 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. We quantify the utility of a differentially private frequent itemset mining algorithm by its likelihood to produce a complete and sound result. Intuitively speaking, "completeness" requires an algorithm to include all the sufficiently "frequent" itemsets, and "soundness" requires an algorithm to exclude all the sufficiently "infrequent" ones. We start by a theoretical investigation of the tradeoff between privacy and utility in frequent itemset mining. Our result unfortunately indicates that the problem is very hard that is, in general, one cannot simultaneously guarantee high utility and a high degree of privacy. 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.

II. Related work

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

A. 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 is used. Utility pattern growth algorithm for mining high utility item set.

B. 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.

C. 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. 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.

D. 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.

Efficient Algorithms For Mining The Concise and Lossless Representation of Closed+ High Utility Item sets [2] by Author Cheug-wei wu, Philippe Fournier viger, Philip S.Yu have proposed lossless and compact representation named high utility item set(frequent item set).To mine the representation three algorithm are proposed AprioriHC Apriority-based approach for mining high utility closed item set, Apriori HC-D. Apriori HC algorithm with discarding unpromising and isolated item and (CHUD) Closed High utility Item set. Apriori HC discard the global unpromising item and Isolated Item Discarding Strategy for finding itemset.AprioriHC perform breadthfirst search in horizontal database and CHUD perform depth first search in vertical database . Derive High Utility Item set (DAHU) for efficiently recovering all the High utility item set from the CHUD.

Anonymity preserving pattern discovery by Prof. Maurizio Atzori, F. Bonchi, F. Giannotti [3**]**, proposed 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.

The Differential privacy by author C. Dwork [4] has introduced 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.

PrivBasis: Frequent Itemset Mining with Differential Privacy by author N. Li, W. Qardaji, D. Su, and J. Cao [5] proposed 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.

On differentially private frequent itemset mining by author C. Zeng, J. F. Naughton, and J.-Y. Cai,[6], introduced 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.

Privacy preserving mining of association rules by author Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke [7], proposed 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.

An audit environment for outsourcing of frequent itemset mining by author W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis [8], proposed that 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.

III. CONCLUSION AND FUTURE WORK

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 made a significant contribution to the search of improving the efficiency of frequent item set mining algorithm and the analysis of efficient algorithm techniques, but these techniques have pros and cons, therefore there is necessity to develop such technique to overcome the entire disadvantage to find frequent item and to provide privacy accessing data from the database.

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] Cheug-wei wu, Philippe Fournier –viger, Philip S.Yu "Efficient Algorithms For Mining The Concise and Lossless Representation of Closed+ High Utility Item sets" pp,487-499 ,1994.
[3] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity preserving pattern discovery," VLDB Journal, 2008.
[4] C. Dwork, "Differential privacy," in ICALP, 2006.
[5] Ninghui Li, WahbehQardaji, Dong Su, Jianneng Cao,"PrivBasis: Frequent Itemset Mining with Differential Privacy.", in *VLDB*, 2012.
[6] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," in VLDB, 2012.
[7] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in KDD, 2002.
[8] 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.

BIOGRAPHY

Miss. Varsha V. Dabhole student of ME Computer Engineering second year from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune.

Prof. V. S. Nandedkar is a faculty in the Computer Engineering from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune, India.