International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

Effieient Algorithms to find Frequent Itemset Using Data Mining

Sagar Bhise[1], Prof. Sweta Kale[2]

[1]*Department of Information Technology, RMD Sinhgad School of Engineering,Warje, Pune, MH,India*
[2] *Professor ,Department of Information Technology, RMD Sinhgad School of Engineering,Warje, Pune, MH,India*

Abstract - *Now a days, designing differentially private data mining algorithm shows more interest because item mining is most facing problem in data mining. During this study the possibility of designing a private Frequent Itemset Mining algorithm obtains high degree of privacy, data utility and high time efficiency. To achieve privacy, utility and efficiency Frequent Itemset Mining algorithm is proposed which is based on the Frequent Pattern growth algorithm. Private Frequent Pattern -growth algorithm is divided into two phases namely preprocessing phase and Mining phase. The preprocessing phase consists to improve utility, privacy and novel smart splitting method to transform the database; the preprocessing phase is performed only once. The mining phase consists to offset the information lost during the transaction splitting and calculates a run time estimation method to find the actual support of itemset in a given database. Further dynamic reduction method is used dynamically to reduce the noise added to guarantee privacy during the mining process of an itemset.*

***Key Words*: Itemset, Frequent Itemset Mining, Data mining, differential privacy.**

1. INTRODUCTION

Differentially private data mining algorithms shows more interest because data item mining is most facing problem in data mining. It is useful in most applications like decision support, Web usage mining, bioinformatics, etc. In a given a database, each transaction consists a set of items, FIM tries to find itemset that occur in transactions multiple times than a given occurrences.

The data may be private which may cause threats to personal privacy. This problem is solved by proposing differential privacy, It gives much guarantees on the privacy of released data without making assumptions about an attacker's information. By adding noise it assures that the resulted data itemset of an estimation is insensitive to changes in any personal record, and thus limiting privacy leaks through the results.

A variety of algorithms are already implemented for mining sequence itemsets. The Apriori and FP-growth algorithm are the two most prominent ones. In particular, Apriori algorithm is a breadth-first search algorithm. It needs l database scans if the maximal length of frequent itemsets is l. In contrast, FP-growth algorithm is a depth-first search algorithm, which requires no candidate generation. While FP-growth only performs two database scans,which makes FP-growth algorithm an order of magnitude faster than Apriori. The features of FP-growth inspire to design a differentially private FIM algorithm based on the FP-growth algorithm. During this study, a practical differentially private FIM gains high data utility, a high degree of privacy and high time efficiency. It has been shown that utility privacy can be improved by reducing the length of transactions. Existing work shows an Apriori based private FIM algorithm. It reduces the length of transactions by truncating transactions (It means if a transaction has more items than the limitations then delete items until its length is under the limit). In each database scan, to preserve more frequent items, it influences discovered frequent itemsets to re-truncate the transactions. However, FP-growth only performs two database scans. Due to this it is not possible to re-truncate transactions during the data mining process. Thus, the transaction truncating (TT) approach proposed in is not suitable for FP-growth. In addition, to avoid privacy breach, noise is added to the support of data itemsets. Given an data itemset in X to satisfy differential privacy, the amount of noise added to the support of i data itemset X depends on the number of support computations of i-itemsets. Unlike Apriori, FP-growth is a depth-first search algorithm. It is hard to obtain the perfect number of support computations of i-itemsets during the mining process of a transaction. A native approach for computing the noisy support of ith item is to use the number of all possible ith item. However, it will definitely produce invalid results.

2. Related Work

Shailza Chaudhary, Pardeep Kumar, Abhilasha Sharma, Ravideep Singh, [1] in this, Mining information from a database is the main aim of data mining. The most relevant information as a result of data mining is getting relations among various items. More preciously mining frequent itemset is the most significant step in the mining of different data itemsets. Many algorithms discussed in the IEEE require multiple scan of the database to get the information on various sub steps of the algorithm which becomes difficult. Here Author is proposing an algorithm Lexicographic Frequent Itemset Generation (LFIG), It should extract maximum data from a database only in one scan. It use Lexicographic ordering of data item values and arrange

itemsets in multiple hashes which are linked to their logical predecessor.

Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren, [2] in this, The growing popularity and development of data mining technologies bring serious threat to the security of personal sensitive useful data. The latest research topic in data mining called privacy preserving data mining (PPDM), was studied in past few years. The basic idea of PPDM is to modify the data in such a way that to perform data mining algorithms effectively with security of personal information contained in the data. Now a day studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, but in fact, unwanted disclosure of personal data may also happen in the process of data information, data publishing, and collecting (i.e., the data mining results) delivering. This study concentrate on the privacy issues correlated to data mining from a wider perspective and study various approaches that can help to protect personal data. In particular, find four different types of users involved in data mining applications, namely, data miner, data provider, data collector, and decision maker. Each user, discuss his privacy concerns and the methods that can be adopted to protect sensitive information. Then briefly present the basics of related research topics, review state of the art approaches, and present some thoughts on future research directions. Besides exploring the privacy preserving approaches for each type of user, review the game theoretical approaches, which are proposed for analyzing the communications among different users in a data mining scenario, each of whom has his own valuation on the personal data. By differentiating the responsibilities of different users with respect to security of personal data, it will provide some useful insights into the study of PPDM.

O.Jamsheela, Raju.G, [3] in this, Data mining is used for mining useful data from huge datasets and finding out meaningful sequences from the data. More institutes are now using data mining techniques in a day to life. Frequent pattern mining has become an important in the field of research. Frequent sequences are patterns that appear in a data set most commonly. Various technologies have been implemented to improve the performance of frequent sequence mining algorithms. This study provides the preliminaries of basic concepts about frequent sequence tree(fp-tree) and present a survey of the developments. Experimental results shows better performance than Apriori. So here concentrate on recent fp-tree modifications new algorithms than Apriori algorithm. A single paper cannot be a complete review of all the algorithms, here relevant papers which are recent and directly using the basic concept of fp tree. The detailed literature survey is a pilot to the proposed research which is to be further carried on.

Feng Gui, Yunlong Ma, Feng Zhang, Min Liu, Fei Li, Weiming Shen, Hua Bai, [4] in this, Frequent itemset mining is an significant step of association rules mining. Almost all Frequent itemset mining algorithms have few drawbacks. For example Apriori algorithm has to scan the input data repeatedly, which leads to high load, low performance, and the FP-Growth algorithm is incomplete by the capacity of storage device since it needs to build a FP-tree and it mine frequent data itemset on the basis of the FP-tree in storage device. In the coming of the Big Data, these limitations are becoming more bulging when confronted with mining large data. Distributed matrix-based pruning algorithm depend on Spark, is proposed to deal with sequence of item. DPBM can greatly decrease the amount of candidate item by introducing a novel pruning technique for matrix-based frequent itemset DllRlng algorithm, an better-quality Apriori algorithm which only needs to scan the input data items at only one time. In addition, each computer node reduces greatly the memory usage by applying DPBM. The experimental results show that DPBM gives better performance than MapReduce-based algorithms for frequent item set mining in terms of speed and scalability.

Hongjian Qiu, Rong Gu, Chunfeng Yuan, Yihua Huang , [5] in this, The frequent itemset mining (FIM) is , more important techniques to extract knowledge from data in many daily used applications. The Apriori algorithm is used for mining frequent itemsets from a dataset, and FIM process is both data intensive and computing-intensive. However, the large scale data sets are usually accepted in data mining now a days; on the other side, in order to generate valid data, the algorithm needs to scan the datasets frequently for many times. It makes the FIM algorithm more time-consuming over big data itemset mining. Computing is effective and mostly-used policy for speeding up large scale dataset algorithms. The existing parallel Apriori algorithms executed with the MapReduce model are not effective enough for iterative computation. This study, proposed YAFIM (Yet another Frequent Itemset Mining), a parallel Apriori algorithm based on the Spark RDD framework and specially-designed in-memory parallel computing model which support iterative algorithms and also supports interactive data mining. Experimental results show that, compared with the algorithms executed with Mapreduce, YAFIM attained 18 times speedup in average for various benchmarks. Especially, apply YAFIM in a real-world medical application to explore the associations in medicine. It outperforms the MapReduce method around 25 times.

3. PROPOSED SYSTEM

3.1 Problem Definition

To design PFP-growth algorithm, which is divided into two phases namely Preprocessing phase and Mining phase. The preprocessing phase consists to improve utility, privacy and novel smart spliting method to transform the database, it is perform only one time. The mining phase consists to offset the information lost during the transaction splitting and

calculates a run time estimation method to find the actual support of itemset in a given database.

3.2 Proposed System Architecture

The PFP-growth algorithm consists of a preprocessing phase consists to improve utility, privacy and novel smart spliting method to transform the database, it is perform only one time. The mining phase consists to offset the information lost during the transaction splitting and calculates a run time estimation method to find the actual support of itemset in a given database. Further dynamic reduction method is used dynamically to reduce the noise added to guarantee privacy during the mining process of a itemset.

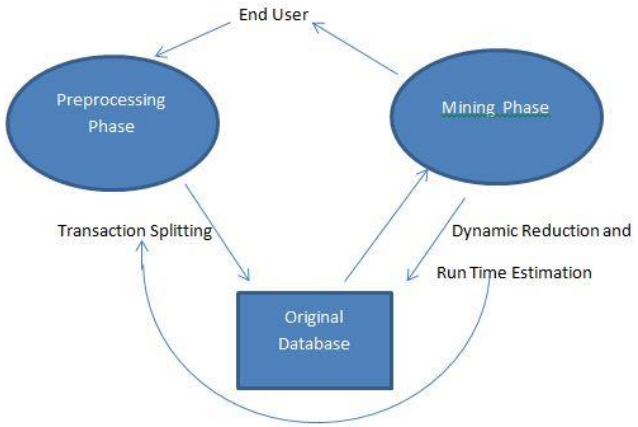

Fig -1: Proposed System Architecture

In Proposed system, three key methods to address the challenges in designing a differentially private FIM algorithm is based on the FP-growth algorithm are proposed.
The three key methods are as follows:

1. Smart Splitting

In a smart splitting the long transactions are splitted rather than truncated. It is nothing but dividing long running database transactions into more than one subset.
Given a transaction t = {a; b; c; d; e; f}
Instead of processing transaction t solely, divide t into t1 = {a; b; c} and t2 = { d; e; f}. Doing so results in to the support of itemsets {a; b; c}, {d; e; f} and their subsets will not be affected.

2. Run-time Estimation

This method finds weights of the sub transactions. While splitting the transactions there is data loss. To overcome this problem, a run-time estimation method is proposed. It consist of two steps: based on the noisy support of an itemset in the transformed database, 1) first estimate its actual support in the transformed database, and 2) then compute its actual support in the original database.

3. Dynamic Reduction

Dynamic reduction is the proposed lightweight method. This method would not introduce much computational overhead. The main idea is to leverage the downward closure property (i.e., the supersets of an infrequent itemset are infrequent), and dynamically reduce the sensitivity of support computations by decreasing the upper bound on the number of support computations.

To achieve both good utility and good privacy, PFP-Growth algorithm is developed which consists of two phases i.e. Pre-processing and Mining phase. In pre-processing phase it compute the maximal length constraint enforced in the database. Also compute maximal support of ith item, after computing smart splitting, transform the database by using smart splitting.

In mining phase given the threshold first estimate the maximal length of frequent itemsets based on maximal support.

4. CONCLUSIONS

The main focus of this work is to study PFP-growth algorithm, which is divided into two phases namely Pre-processing phase and Mining phase. The pre-processing phase consists to improve utility, privacy and novel smart splitting method to transform the database; it is perform only one time. The mining phase consists to offset the information lost during the transaction splitting and calculates a run time estimation method to find the actual support of itemset in a given database. Moreover, by leveraging the downward closure property, put forward a dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the data mining process. The formal privacy analysis and the results of extensive experiments on real datasets show that PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

REFERENCES

[1] Shailza Chaudhary, Pardeep Kumar, Abhilasha Sharma, Ravideep Singh, "Lexicographic Logical Multi-Hashing For Frequent Itemset Mining", International Conference on Computing, Communication and Automation (ICCCA2015)
[2] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren,"Information Security in Big Data: Privacy and Data Mining", 2014 VOLUME 2, IEEE 29th International Conference on Information Security in Big Data
[3] O.Jamsheela, Raju.G, "Frequent Itemset Mining Algorithms :A Literature Survey", 2015 IEEE International Advance Computing Conference (IACC)
[4] Feng Gui, Yunlong Ma, Feng Zhang, Min Liu, Fei Li, Weiming Shen, Hua Bai, "A Distributed Frequent

Itemset Mining Algorithm Based on Spark", Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)

[5] Hongjian Qiu, Yihua Huang, Rong Gu, Chunfeng Yuan, "YAFIM: A Parallel Frequent Itemset Mining Algorithm with Spark", 2014 IEEE 28th International Parallel & Distributed Processing Symposium Workshops

BIOGRAPHIES

Sagar Bhise
B.E.Computer Science and Engineering
M.E. Information Technology