IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.2, Jan 2019 E-ISSN: 2347-2693

Effective Algorithm to Find The Frequency Item Sets Using Datamining.

R. Pandiammal[1*], N. Kavitha[2]

[1]Department of Computer Applications, Nehru Arts and Science College, Coimbatore, Tamilnadu, India
[2]Department of Computer Science, Nehru Arts and Science College, Coimbatore, Tamilnad, India

[*]*Corresponding Author: panbhavya@gmail.com,*

Available online at: www.ijcseonline.org

***Abstract*—** Now a day when we transfer the data, we are using private frequency item sets mining algorithm. It has 2 phases that are pre-processing and mining phase. This algorithm is used for utility, privacy and efficiency the frequency item set is planned which is based on frequency pattern growth algorithm. In pre processing phase consists to improve the privacy, utility and novel smart splitting to transpose database. The mining phase consists to offset the information lost during the transformation splitting information and calculate the runtime estimate for actual support of item set in a given data base . Further the dynamic noise reduction technique is used to reduce the noise at the time of mining phase.

***Keywords*—** item set, frequent item set mining, differential privacy.

I. INTRODUCTION

Differentially private frequency item set mining algorithms is more interested in data mining because data item mining facing more problems in data mining. Here two types of data mining algorithms are used. The FIM algorithm find out the number of item set transformed more time than a given occurrences. The Apriori and FP-growth algorithm are used for this.

Apriori algorithm is as breadth first search and FP algorithm is used as depth first search. It needs 1 database scan if maximum number of frequency item set is 1 and FIM algorithm is used to scan two transaction databases. In Apriori algorithm is used to truncate truncations. Truncate transaction means if the transaction has more items than the limitations then delete that item until its length under the limit. In FIM, we can't do the truncation transaction (TT) between the databases and to avoid the breach, noise is added to the support of data item sets.

II. RELATED WORK

Shailza Chaudhary, Pardeep Kumar, Abhilasha Sharma, Ravideep Singh, in this study, Mining information from the database is the main aim of data mining. Most relevant information as the result of data mining is getting relation on various items. Many algorithms discussed in IEEE require multiple scan of a database to get the information by various steps of algorithm which becomes difficult.

Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren, [2] in this study, 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). The basic idea of PPDM is to modify the data to perform data mining algorithms effectively with security of 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 there are data miner, data provider, data collector, and decision maker. Each user, discuss his time alone 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 differentiate 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 study, Data mining is used for mining useful data from huge datasets and finding out

meaningful sequences from the dataset. More institutes/Business centres are now using data mining techniques in a day to life. Frequent pattern mining is an important in the field of research. Frequent sequences are patterns that appear in a data set most commonly. This study provides the preliminaries of basic concepts about frequent sequence tree(fp-tree) and present a survey of the developments. An experimental result shows better result than Apriori. So 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.

Feng Gui, Yunlong Ma, Feng Zhang, Min Liu, Fei Li, Weiming Shen, Hua Bai, [4] in this, Frequent itemset mining is dine by 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 item set 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.

III. PROPOSED SYSTEM

A. Problem Definition

To design PFP-growth algorithm, which is divided into two phases namely Preprocessing phase and Mining phase. The pre-processing 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 item set in a given database.

B. Proposed System Architecture

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

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.

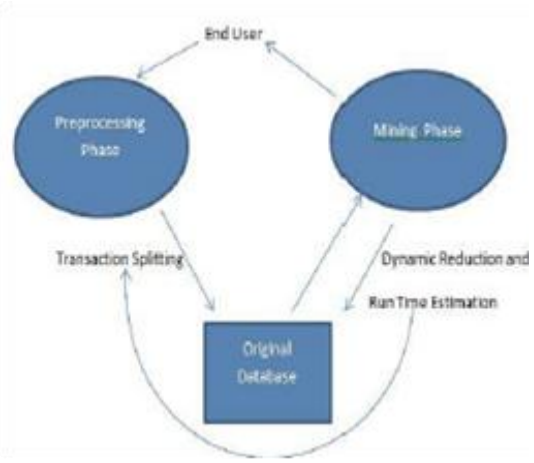

Figure 1: Proposed System

The three key methods are as follows:

- ***Smart Splitting***

The smart splitting 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 & t2, t1= {a; b; c} and t2 = { d; e; f}. And doing so results in to the support of item sets {a; b; c}, {d; e; f} and their subsets will not be affected.

- ***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) actual support in the transformed database, and 2) then compute its actual support in the original database.

- ***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 item set 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.

IV. CONCLUSION AND FUTURE SCOPE

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 usefulness, 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 item set 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.

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

