IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 3, March 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

A Survey of Frequent Itemset Mining Using Different Techniques

Miss.Pooja Purohit[1], Prof.Sonal Patil[2]

[1]PG Student,Computer Engineering Department,G.H. Raisoni College of Engineering, Jalgaon,Maharashtra,India

[2] Head of Computer Engineering Department,G.H. Raisoni College of Engineering Jalgaon,Maharashtra,India

Abstract

Nowdays, businesses are evolving. For having business people needs to deal with much amount of data and this data needs to be delicate and confidential. So, to secure and preserve our data there are plenty of technologies used one of them is Data Mining. Data Mining is the technique in which it tries to find out interesting patterns or knowledge from database such as association or correlation etc. Frequent Itemset Mining is the critical problem in data mining. The frequent can contains valuable and research purpose. Frequent itemsets are items or patterns like itemset, substructures or subsequences that occurs frequently in transaction. To find out frequent itemset there are many Frequent Itemset Mining Algorithms used such as Apriori, FP-growth, Elcat. The famous two most important algorithm, to find Frequent itemset are Apriori and FP-growth. Apriori is a candidate set generation and-test algorithm. It needs multiple database scans.It have two steps: It find all itemsets which have minimum support frequent itemsets(candidate list) and generate frequent itemsets list . FP-growth which does not generates candidate set. Compared with Apriori, FP-growth is less time consuming algorithm. But it enforcing the limit by truncating transactions, if a transaction our has more items than the given limit, then deleting items until its length is under the limit. FP-growth is more faster than apriori algorithm. But this all only consider frequent itemset from large transaction

Keywords: *Frequent Item Mining, Apriori algorithm, FP-growth.*

1. Introduction

Data Mining is the method which find out the hidden data in a database. It is also called as data determination or data analysis and deductive finding out. There is a great deal of knowledge to extract the associations from data. To identify the correlated set of item in database association rules are used. Data Mining uses different kinds of techniques which are combined from database technologies and many more.

Market basket analysis is the one of the most famous example of association rule mining. In this, market analysts focused on discovering frequently purchased items by the consumer. So it is easy for arrangement of items according to their sales by organization. Association rule mining is the method of finding interesting or similar relations between the variables or items in large transaction of database. Association rule is using two measures support and confidence to identify the most important relationships. Support is an sign of how frequently the itemset appear in the database and it is also set of preconditions.

Confidence is an mark of finding how frequently the rule has been found to be true. It uses minimum support and confidence which are user defined. Association rule used in many application areas including web usage mining,intrusion detection , continuous production and bioinformatics.

Discovering useful patterns hidden in database plays an critical role in different data mining jobs such as frequent pattern mining , high utility pattern mining .Among the all ,frequent pattern mining is a basic research topic that has been used to different database having long transactions. It is used in the analysis of customer transactions in retail research where it is marked as market basket analysis and also used to identify purchase items of the consumer. Given a database, in which each of transaction has a set of items where Frequent Itemset Mining used to find out itemsets that occurs in transactions more than a given user specified threshold.

The data is perceptive (e.g web browsing history and medical records of patients) the frequent itemsets detection can provide. Releasing that detected frequent itemsets can cause threats to individual privacy. But limitations of frequent itemset mining are that only consider frequently occurring items in a transaction database above the user specified frequency threshold, without considering the quantity and profit of items. The quantity and utility are important for real world decision problem.

This paper marks the frequent and weighted itemsets discovery. Most of the methods in finding frequent itemsets which designed for traditional databases they are apriori and FP-growth algorithm.

2. Related Work

Many of researchers have been proposed to solve the privacy preserving FIM problem from different ways. Mining the frequent patterns is mine in many different kinds of databases such as transaction database, time-series databases, and many other. These databases have been investigated in data mining research. Many of the

previous examination accept an *Apriori* as like candidate set which is generation-and-test method. The candidate set generation is costly. In this study, the FP-growth, it structure an extended prefix-tree structure which is used for storing compressed and important information about frequent patterns

Main purpose is that the resulted frequent itemsets itself does not leak private information and achieve differential privacy. The k-anonymity model for protecting privacy in [2] and in [12] which propose an algorithm to publish annonymised frequent itemset. Both studies don't satisfy differential privacy as well as they cannot provide sufficient privacy protection from attackers having background knowledge. Attribute [3] introduced l-diversity, a framework which gives stronger privacy guarantees and shows the weak points of k-anonymity.

In [4] proposed Apriori & AprioriHybrid algorithms which are fast algorithm for mining association rule. These both compared with previous algorithms and gives excellent performance for large database with transactions, but it generates candidate set which is costly to handle these which is costly to handle these.

In [5] Introduces FP growth algorithm, which is mining frequent pattern without candidate generation and as we seen in [4] apriori algorithm performs mining fastly with candidate set generation, which is costly. In [5] FP-growth it uses FP-tree as a data structure to store large database which compressed in small data structure. Here, in [5] the algorithm is scalable and efficient than apriori algorithm.

In this [7], algorithm solves the frequent item set mining problem that they find all item set whose support exceeds a threshold. The benefit of this algorithm is that it achieves better F-score unless k is small. C. Zeng, J. F. Naughton, and J.-Y. Cai,[7] proposed an apriori algorithm for large transaction. Major problem is for long transaction which contains many items. Its truncating the long transactions means it limiting the transaction. Deleting the items until the transaction is under the limit if transaction has more than a specified number of items. It must be done in a differentially private way. It also discarding items from transactions gives a new source of error. In transaction truncating, the more frequent subsets are kept and other items which are not frequent are truncated.

In this [8] it solve the problem of association rule mining algorithm which gives a privacy preserving scalar product protocol as well as gives an efficient protocol for computing scalar product which preserve privacy of the individual transactions. In [9] Clifton and Kantarcioglu , which addresses the problem as a secure multi-party computation and consider the database as a horizontally partitioned.

L.Bonomi [6] proposed Frequent sequential pattern mining. It is a central task in many fields as like biology and finance. In this paper, it provides the provable and formal guarantees of privacy by studying the sequential pattern mining problem under the differential privacy framework. In this paper, proposed a two-phase algorithm. This algorithm mining both prefixes and substring patterns. In the first phase, construct a model-based prefix tree. This is used to mine a candidate set of substring patterns and its prefixes. In next phase, substring patterns is refined. Here it transformed the original data to reduce the perturbation noise.

W.K.Wong [10] proposed Outsourcing association rule mining to an outside service provider which gives many important benefits to the data owner. These include (i) release from the high mining cost, (ii) minimize the demands in resources, and (iii) for multiple distributed owners effectively centralized the mining and also provide security. In this paper, it develop an effective and efficient encryption algorithm and performs a single pass over the database. It is applicable for the application which sends the streams of transactions to the service provider.

In [11] it present the set of randomization operators to limit privacy which margin the FIM. Proposed new algorithm which discovers the frequent patterns in sensitive data and adopted two mechanism techniques i.e. exponential & Laplace noise-addition mechanism. These are efficient in context of frequent item mining.

[12] Proposes algorithm Privbasis which perform frequent itemset mining with differential privacy with the help of minimum support threshold. An item set that found in transaction is frequently than minimum support threshold and subset of some basis with differential privacy guarantee.

3. Methodology

Apriori is a breadth first search algorithm. It generates the candidate set. So, it also called as generation-and-test algorithm. It needs only one database scans if the maximal length of frequent itemsets is l. It Support count is expensive, due to generation of candidate set and requires multiple database scans (I/O). It consist of two steps: Find all itemsets which have minimum support frequent itemsets. Use that frequent itemsets to generate rules. The Apriori property follows a two step process:

- Join step: *Ck* is generated by joining *Lk*-1 with itself.
- Prune step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

The Apriori Algorithm

Ck: Candidate itemset of size k

Lk: frequent itemset of size k

L1= {frequent items};
For (k= 1; *Lk*!=∅; k++) do begin
Ck+1= candidates generated from *Lk*;
for each transaction tin database do
increment the count of all candidates in *Ck*+1that are contained in t.
Lk+1= candidates in *Ck*+1 with min_support
End
Return ∪ *k Lk*

Limitation

• Needs so many iterations of the data and uses uniform minimum support threshold. Problem in finding rarely occurring events
• Due to generation of large number of candidates it requires large memory space. Execution is more as time is wasted in producing candidate every time
• Execution time is more as the time is wasted in producing candidate every time and computational cost is also more.

FP-growth

FP-growth algorithm is used for Frequent Itemset Mining. It is a depth-first search algorithm, which requires no candidate generation. FP-growth is faster than Apriori. In the mining process of FP-growth, there is not a single chance to re-truncate transactions. So, the transaction truncating approach is not suitable for FP-growth. In FP-growth during mining process, it is hard to obtain the exact number of support computations of i-itemsets. In FP-growth, the structure is an extended prefix-tree structure which is used for storing compressed and important information about frequent patterns.

FP-growth uses two data structures i.e. header table and FP-tree. Branch in FP-tree represents an itemset. Each node in FP-tree contains of a counter. In Header table, it stores items and their supports.

- Construction of the FP-Tree.
- Extracts frequent itemsets directly from the FP-Tree.

The FP-growth also referred as Private FP-growth (PFPgrowth) Algorithm. It consist of two modules preprocessing and mining. The preprocessing module is used to improve privacy as well as smart splitting method to transform the database. In the mining module a run time estimation method is used to estimate the support of item and dynamic reduction
method to dynamically reduce the amount of noise. In between the formal privacy analysis which shows that the PFP-growth algorithm is ϵ-differentially private.

Problem statement:-

- The existing system have many problem one of is to find with the high utility transactional itemsets.
- Existing methods takes more times for mining.
- Existing system gives many more combinations of output.
- Deleting the items until its length is under the limit, if a transaction has more items than the limit means its enforcing the limit by truncating the transactions.

4. Conclusions

In this paper, we survey so many methods for frequent item mining with privacy such as K-anonymity-diversity, Privbasis. Also we have studied different mining algorithms such as Apriori , Apriorihybrid , FP –growth. Apriori is costly and its time efficiency is less than FP – growth algorithm. We have studied existing system. In this paper, there is problem in designing a differentially FIM algorithm.Frequent itemset mining only consider the frequently occurring itemset and is challenged in many areas such as retail, marketing etc. It has been seen that in many real application domains that itemsets which share the most are not well enough the frequent itemset.

References

[1] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Yang, “Differentially private frequent item mining via transaction splitting”,2015,In IEEE transactions on knowledge and data engineering vol. 27, No.7, pp.1875-1891

[2] C.Dwork, “Differential privacy,” in Proc. Int. Colloquium Automata, Languages Programm., 2006, pp. 1–12

[3] L. Sweeney, “k-anonymity: A model for protecting privacy,” Int. J.Uncertainity Fuzziness Knowl.-Base Syst., vol. 10, no. 5, pp. 557–570,2002.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M.Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity, “in Proc. 22nd Int. Conf. Data Eng., 2006, p. 24.

[5] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large dataBases,1994,pp.487–499.

[6] L. Bonomi and L. Xiong, “A two-phase algorithm for mining sequential patterns with differential privacy,” in CIKM, 2013.

[7] C. Zeng, J. F. Naughton, and J.-Y. Cai, “On differentially private frequent itemset mining,” Proc. VLDB Endowment, vol. 6, no. 1,pp. 25–36, 2012.

[8] J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 639–644.

[9] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,”IEEE Trans. Knowl. Data Eng., vol. 16, no. 9, pp. 1026–1037, Sep.2004.

[10] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis,"Security in outsourcing of association rule mining," in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 111–122.

[11] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis,"An audit environment for outsourcing of frequent itemset mining,"Proc. VLDB Endowment, vol. 2, no. 1, pp. 1162–1173, 2009.

[12] N. Li, W. Qardaji, D. Su, and J. Cao, "Privbasis: Frequent itemset mining with differential privacy," Proc. VLDB Endowment, vol. 5, no. 11, pp. 1340–1351, 2012.