IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

Understating Frequent Pattern Mining on Encrypted Cloud Data and its Security

Rupali Namdeo Bichitkar[1], **Prof. Vidya Jagtap**[2]

Dept., of Computer Engineering, G.H. Raisoni College of Engineering Management, Chas, Ahmednagar, Pune, India[1]

Dept., of Computer Engineering, G.H. Raisoni College of Engineering Management, Chas, Ahmednagar, Pune, India[2]

Abstract: Frequent Itemset mining is important part of data mining techniques, which focuses on looking at sequences of actions or event. In real world, a lot of exists on large datasets that are stored on cloud servers in recent years. By using data mining techniques on these large dataset increases the accuracy of in terms of data result and also the efficiency. But it also brings the possibility of data leakage of confidential private datasets. Therefore, in this proposed work we focus on the privacy and efficiency of frequent Itemset. Proposed system uses the FP growth algorithm, which is best performing algorithm for frequent pattern mining. Also to maintain the privacy of frequent item set patterns, encryption algorithm has been implemented.

Keywords: Data Mining, Cloud computing, frequent pattern mining

1 INTRODUCTION

Frequent pattern mining is one of the most important concept in data mining technique that helps in decision making. Frequent patterns are item sets, subsequences, or substructures that appear in a data set with higher frequency. Consider the example of set of items, such as milk and bread, mobile phone and its cover that appear frequently together in a transaction data set. Such item sets are the frequent item sets, as most of the time a person who buys milk also buys the bread, who buys mobile phone also buys cover for it.

A subsequence, such as buying first a computer or laptop, then a digital camera, and then a memory card is referred as frequent sequential pattern if it occurs frequently in a shopping history database.

As huge amount of data is available, the maintenance of this huge data creates the storage problem. That’s why user or client started using the cloud server to store their data. Public cloud server are provided with easy access and minimum cost which can help the data owner to reduce the mining cost on massive datasets. Accuracy of mining process increased as large-scale data is available on cloud server for mining. On other hand, this large-scale data can also contain the confidential and private data of client which should not be disclosed to anyone. For example, the insurance or financial information of a customer. Because of this the security of cloud data and the mining process became a big concern.

Many research papers are available and are going on to secure the privacy and confidentiality of frequent itemset mining process on cloud server. The rest of the paper is organized as follows- Section 2 is about some existing methodologies. In Section 3, the proposed Methodology. Finally we conclude in Section 4 and give future direction of our work.

2 RELATED WORKS

The sequential pattern mining problem was first introduced by Agrawal and Skrikant [2] and can be stated as if we are given a set of sequences, called data-sequences, consisting of list of transactions, where each transaction contains items, sequential pattern mining is to find all of the frequent subsequences whose ratios of appearance exceed the minimum support threshold. Many approaches [8] have been proposed to extract sequential patterns from sequence databases. Some methods focus on the efficient mining of sequential patterns in time-related data. There exists lots of algorithm to mine sequential patterns. These algorithms can be broadly categorized into classes such as below

(1) Apriori-based method supporting horizontal formatting, such as GSP [1].
(2) Apriori-based method supporting vertical formatting, such as SPADE.
(3) Apriori-based candidate generation and pruning using depth-first traversal, such as SPAM.
(4) Projection-based pattern growth method, such as PrefixSpan and FreeSpan.

Apriori based algorithm:-

The Apriori algorithm was first proposed by Agrawal in [9], for the discovery of frequent item sets. It is the most widely used algorithm for the discovery of frequent item sets and association rules. The Apriori property of sequences states that, if a sequence S is not frequent, then none of the super sequences of S can be frequent.

DOI 10.17148/IJARCCE.2017.61220

GSP: Generalized Sequential Pattern:-

GSP algorithm is similar to the Apriori algorithm. It makes multiple passes over the data. In the first pass it finds the frequent sequences i.e. it finds the sequences that have minimum support. These sequences are seed set for the next iteration. At each next iteration, each candidate sequence has one more item than the seed sequence. There are some drawbacks of GSP such as it generates large set of candidate sequences, it requires multiple scans of database and it is inefficient for mining long sequential patterns (as it needs to generate a large number of small candidates).

Apart from finding simple frequent patterns, GSP allows a user to specify time constraints (minimum and/or maximum time period between adjacent elements in a pattern). It relaxes the restriction that the items in an element of a sequential pattern must come from the same transaction, instead allowing the items to be present in a set of transactions whose transaction-times are within a user-specified time window. Given a user-defined taxonomy Mining Trajectory Patterns and its Application in Pattern Matching Query (is-a hierarchy) on items, it allows sequential patterns to include items across all levels of the taxonomy.

Vertical Format-Based Method (SPADE: Sequential Pattern Discovery using Equivalent Class):-

This is a vertical format sequential pattern mining method. SPADE first maps the sequence database to a vertical database format. It decomposes the original problem into smaller problems and solves the problems independently in memory using lattice search techniques. The important contribution of this algorithm is that it requires only three database scans to discover all sequences or only a single scan with some pre-processed information, thus minimizing the I/O costs. SPADE decouples the problem decomposition from the pattern search. Pattern search could be done in a BFS or DFS manner.

The SPAM and I-SPAM Algorithms:-

Sequential pattern mining technique that utilizes a bitmap representation called SPAM. The algorithm is the first sequential mining method that utilizes a depth-first approach to explore the search space. Combining this search strategy with an effective pruning technique that reduces the number of candidates makes the algorithm particularly suitable for very long sequential patterns. However, the algorithm requires that the whole database can be stored in main memory, which is the main drawback of the algorithm.

As sequences are generated traversing the tree, two types of children are generated from each node: sequence-extended sequences (sequence extension step or S-step) and itemset-extended sequences (item-extension step or I-step). Finally, an efficient representation of the data is used, which is a vertical bitmap representation. The bitmap is created for each item.

Pattern Growth Method:-

It comes up with solution of the problem of generate-and-test. It works on key features like avoid the candidate generation step and focus the search on a restricted portion of the initial database. These methods Scan DB once, find frequent

1. item set (single item pattern)
2. Order frequent items in frequency descending order
3. Scan DB again, construct FP-tree.

It works on projected database. It reduces candidate generation which is basic feature of FP-growth. It uses frequent items to recursively project sequence databases into a set of smaller projected databases and grows subsequence fragments in each projected database. Moreover, since a length-k subsequence may grow at any position, the search for length − (k + 1) candidate sequence will need to check every possible combination, which is costly.

J. Vaidya and C. Clifton [14] proposed the vertical partition of centralised data means each site contains some elements of a transaction. This paper creates a database as the primary, and is the initiator of the protocol. The other database is the responder and a join key present in both databases. It aims to find interesting association rules of attributes other than join key using secure scalar product protocol. This method is depends on the number of data values that the other party might know from some external source, since a dataset knows its own data and learns the resulting global association rules results in some disclosure.

J. Vaidya and C. Clifton [15] proposed the protocol for securely determining the size of set intersection. It presents a protocol for computing the size of the intersection of sets of items held by different parties and same can be used to compute association rules. For this they uses commutative hash function for encryption where each party encrypts its own items with its own key and then parties pass the set to their neighbour to be hashed. After this every party computes intersection of datasets. To mine the association rules it uses the vertically partitioned data. However this protocol discloses the some intersection information to other parties. It includes extensive interactions and are not feasible to the system framework.

C. Dong and L. Chen [16] focus on the efficiency data mining techniques and privacy of association rule mining. It introduces an efficient private set intersection protocol which is built on two well-defined cryptographic primitives: the

DOI 10.17148/IJARCCE.2017.61220

Goldwasser–Micali Encryption and the Oblivious Bloom Intersection. This protocol is two party protocol where each holds part of the transaction set that is vertically partitioned. The performance of protocol is enhanced by exploiting parallelization of Oblivious Bloom Intersection protocol. It is a two-party protocol which involve extensive interactions.
X. Yi, F.-Y. Rao, E. Bertino, and A. Bouguettaya [17] proposed a system where user encrypts the data before storing it to the cloud server and the association rule mining task is performed by the outsourced server. For this outsourced task n(≥ 2) aided servers are needed. To encrypt the datasets distributed ElGmal homomorphic encryption technique is used. The private key is spilt into n pieces and distributes them to the n servers. To prevent the background knowledge based attack, fake transactions added to the transaction database. As ElGamal encryption is probabilistic,so a item can have different encryptions. To identify these encryptions of same item a algorithm is proposed. This proposed system includes item privacy, transaction privacy and database privacy for privacy preserving association rule mining. Because of these n extra server the processing slows down the running time of frequent itemset mining, and introduces huge interactions and communication overheads.
S. R. Oliveira and O. R. Zaiane [18] proposed a framework to maintain the privacy of sensitive data. This paper focus on hiding sensitive item and pattern and disclosing non-sensitive data instead of encrpting all dataset. For this they use inverted file system (one for indexing the transactions per item and a second for indexing the sensitive transactions per restrictive pattern) and a search techniques on boolean queries against inverted file to find sensitive transactions and to sanitise them. To hide this sensitive items and pattern, then uses the Naive algorithm by removing them from the database.
M. Kantarcioglu and C. Clifton [19] proposed privacy-preserving mining of association rule methodology on horizontally partitioned database. For a global rule support threshold k, author compute the summation of support degree of each inter-site whose support is >k securely. So global frequent itemsets are acquired whose support is greater than threshold. For horizontal partioning of transaction data on different sites author uses the secure multi-party computation using encryption of each data on all sites. Because of this encryption of data and its support at each site it increases the cost of mining process.
W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis [20] proposed a encryption method to secure the outsourced association rule mining transactions from outside world. For this substitution cipher technique has been used to map the data item one-to-one and one-to-n design. To enhance the security of the association rule mining on service provider side, the non-deterministic one-to-n substitution scheme was proposed where random items were added to transaction to make more robust to attack. However based on background knowledge, the attacker can trace the information about the association rules and some itemsets.

3. PROPOSED METHODOLOGY

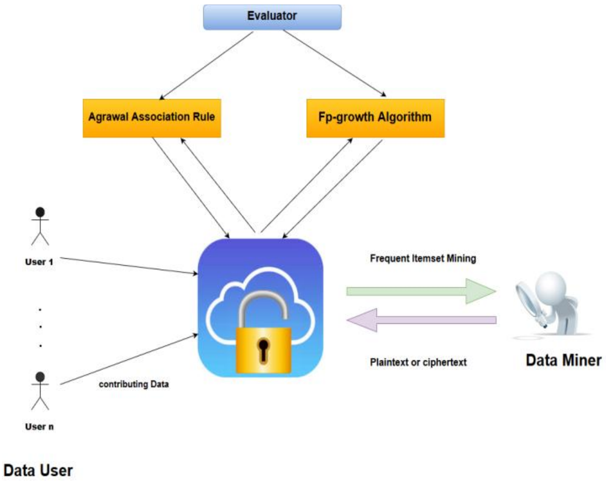

Fig 1: Proposed Architecture of System

There are mainly four modules required for securely generates frequent itemset which are follows,

1) Data user
2) Cloud service provider
3) Data miner
4) Evaluator

There is n number of users and their responsibility is to contribute data and those data store in cloud. Cloud provider maintains user data, collect mining request from data miner and generates frequent itemset by using evaluator.

3.1 Data User:-

Data user plays important role in database creation and user are continuously doing a transaction and those transactions are collected for frequent itemset mining. Before contributing its own data, the data is encrypted by using encryption algorithm.

For example user A go to shop and then buy milk, coffee, breads and then paid a bill. Those buying items we called as a transaction. Those transaction stored in cloud after that another user come in a shop then if user buy a milk then it may chance to user also buy coffee or bread or both. Our system suggests them to buy similar connected items.

3.2 Cloud Service Provider (CSP):-

Cloud service provider is responsible for maintaining user's transaction and stores them in cloud. There are so many open source cloud provider is available so it reduces the actual cost of resources and its maintenance. CSP hold mining request from data miner and then doing frequent itemset mining.

3.3 Data Miner:-

Data miner submits its query to cloud service provider either in plaintext or cipher text. If data miner wants to hide mining request then system encrypts query and then submits to the CSP. CSP evaluates mining request and generates frequent itemset by using Evaluator and then collects generated frequent itemset.

3.4 Evaluator:-

The role of evaluator is to generate frequent itemset and evaluator uses two independent algorithms and those algorithms are used for evaluating frequent items. Association rule mining is existing approach for mining transaction. We propose new protocol to identify frequent itemset. FP-growth is a proposed algorithm which is used for enhancing performance of the system.

4. CONCLUSION

In recent years, the need of data mining techniques and cloud computing growing due the requirements of businesses. Businesses are generating large amount of data and that data must be analyzed to understand the behavior of end user. Also due the frequent changes in hardware and software resources required for organizations, algorithm performance needs to be improved. This paper discusses issues, challenges, algorithms used in frequent pattern data mining.

ACKNOWLEDGEMENT

I express true sense of gratitude towards our project guide **Prof. Vidya Jagtap**, Assistant professor, Computer Department for her invaluable co-operation and guidance that she gave us. We specially thank our project coordinator Prof. xyz for inspiring us and for providing us all the lab facilities. We would also like to express our appreciation and thanks to HOD, Principal and all my friends who have assisted us throughout our hard work.

REFERENCES

[1] Mohammed J. Zaki," Scalable Algorithms for Association Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 12, NO. 3, MAY/JUNE 2000.

[2] R. Agrawal and R. Srikant, Mining Sequential Patterns, Proc. 11th International Conf. Data Eng., pp.3-14, Mar. 1995.

[3] J.Chang, H.Lee, New travel time prediction algorithms for intelligent transportation systems, Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology Volume 21 Issue 1, 2, April 2010.

[4] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.ACM, 2002, pp. 639–644.

[5] J. Vaidya and C. Clifton, "Secure set intersection cardinality with application to association rule mining," Journal of Computer Security,vol. 13, no. 4, pp. 593–622, 2005.

[6] C. Dong and L. Chen, "A fast secure dot product protocol with application to privacy preserving association rule mining," in Advances in Knowledge Discovery and Data Mining. Springer, 2014, pp. 606–617.

[7] X. Yi, F.-Y. Rao, E. Bertino, and A. Bouguettaya, "Privacy-preserving association rule mining in cloud computing," in Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security. ACM, 2015, pp. 439–450.

[8] S. R. Oliveira and O. R. Zaiane, "Privacy preserving frequent itemsetmining," in Proceedings of the IEEE international conference on Privacy,security and data mining-Volume 14. Australian Computer Society, Inc.,2002, pp. 43–54.

[9] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge & Data Engineering, no. 9, pp. 1026–1037, 2004.

[10] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining," in Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, 2007, pp. 111–122.

[11] https://aws.amazon.com/

[12] https://archive.ics.uci.edu/

DOI 10.17148/IJARCCE.2017.61220