International Journal of Academic Research and Development
ISSN: 2455-4197
Impact Factor: RJIF 5.22
www.academicsjournal.com
Volume 3; Special Issue 2; March 2018; Page No. 136-140

Privacy preserving frequent pattern mining technique on encrypted cloud data

Rupali Bichitkar[1], VV Jagtap[2]
[1] Department of Computer Engineering, GH Raisoni College of Engineering and Management, Pune, Maharashtra, India
[2] Professor, Department of Computer Engineering, GH Raisoni College of Engineering and Management, Pune, Maharashtra, India

Abstract
Privacy preservation in data mining has gained significant recognition because of the increased concerns to ensure privacy of sensitive information. It enables multiple parties to conduct collaborative data mining while preserving the privacy of their data. Through the use of data mining techniques on this large data set, accuracy increases in terms of data result and efficiency. But it also involves the possibility of data leakage of confidential private data sets. Techniques for data mining, in particular sequential pattern mining, can be used to extract frequent patterns. Traditional cryptographic methods use encryption techniques or secure multiparty computation (SMC) to ensure privacy of data. But privacy in these techniques is at the expense of additional communication cost, which limits their use in practical applications. Therefore, in this proposed work we focus on the privacy and efficiency of frequent Item set. The proposed system uses the FP growth algorithm, which is the best performing algorithm for frequent pattern mining. To maintain the privacy of common item set patterns, the encryption algorithm has also been implemented.

Keywords: cloud computing, data mining, frequent pat-tern 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.
As a current trend in information technology (IT), it is being used in market analysis and a wide variety of other fields. Complementing this, it is anticipated that software as-a-service (SaaS) services that analyze customers data on cloud servers will become widely used. A recognized problem when having analysis performed on a third-party cloud server, however, is the risk of information leaks due to unauthorized data access or criminal activity within the service provider, and therefore the challenge is to develop secure ways of performing this data mining. 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.

Existing System
The existing system implemented Agrawal association rule mining algorithm for finding frequent itemset from a dataset. This algorithm takes more time to execute and find frequent item sets from the dataset.

2. Review of literature
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 sub-sequences 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.

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 [1] 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 [2] 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 [3] 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 Gold was ser Mi-cali 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 Obliv-ious Bloom Intersection protocol. It is a two-party protocol which involve extensive interactions.
X. Yi, F.-Y. Rao, E. Bertino, and A. Bouguettaya [4] 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 dis-tributed 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 [5] 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 [6] proposed privacy-preserving mining of association rule methodology on horizon-tally 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 portioning 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 [7] 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.

M. Kantarcioglu, R. Nix, and J. Vaidya [8] proposed methods to efficiently compute the approximate dot product for privacy-preserving data mining. To compute the scalar product of two different vectors of two different parties uses the bloom filter. The intersection size between two itemsets is approximated using the bloom filter. For this, each party has to create its own bloom filter values and then participate in dot product and multiplication computation. Here the author assumes that dot product, multiplications and comparison protocols which are being used are secure. However the bloom filter may decrease the accuracy of mining result.

L. Qiu, Y. Li, and X. Wu [9] proposed to use the keyed Bloom filters to represent data items and transactions to maintain the privacy of association rule mining. For this author uses the hash function with secret key to concatenate with item and then inserting it in bloom filter set. To mine the association rule, three phases have been used counting, pruning and candidate generation phase. In candidate generation phase multiple interactions need to conduct in client and server to prevent the servers from decrypting the partial sensitive data using the share key.

C. Dong, L. Chen, and Z. Wen [10] proposed to enhanced privacy set intersection protocol based on oblivious Bloom intersection approach means the using the Oblivious transfer technique between two parties. For this each party represents it own itemset using bloom filter independent uniform hash functions mapping. Author also uses the new version of bloom filter called garbled Bloom filters which has different data structure (includes security parameter in basic structure). Then intersection are produced between the bloom filters of two parties. The protocol uses the symmetric key operations in parallelized manner. They enhance the protocol utilization to malicious model. However because of the oblivious transfer between parties involve extensive interactions and increased the communication overhead.

J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan [11] proposed outsourced association rule mining along with semantic secure solution. This semantic secure solution in-cludes both data privacy and mining result privacy. In this paper categorical data is assumed. To provide the semantic security, symmetric-key predicate-only encryption for inner products have been used. To enhance the accuracy of the data mining result, author introduces the soundness so that the data owner can identify the false data. To retrieve the frequent itemset is 2-way interaction happens between data owner and cloud server for mining request and encrypted mining result response. However the solution is efficient for practical use.

3. System Architecture / System Overview

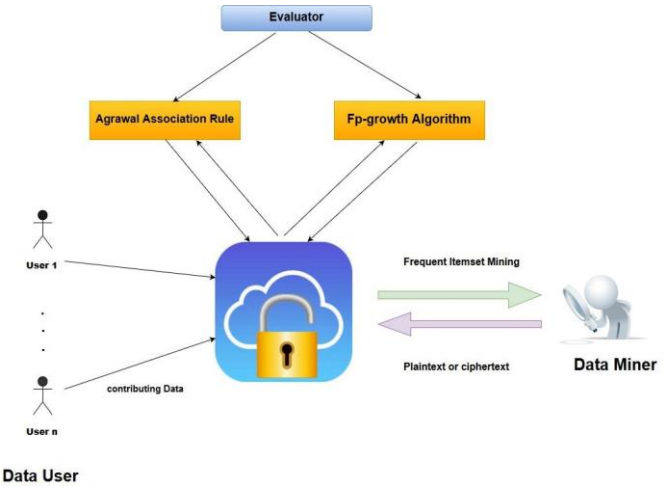

Fig 1: System Architecture

There are mainly four modules required for securely gener-ates 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

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 users

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 and those algorithm is 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. System Analysis

4.1 Algorithm

1) FP Growth Algorithm: Input: A database DB, represented by FP-tree constructed according to Algorithm 1, and a minimum support threshold.
Output: The complete set of frequent patterns.
Method: Call FP-growth (FP-tree, null). Procedure FP-growth (Tree, a)
(01) if Tree contains a single prefix path then // Mining single prefix-path FP-tree
(02) let P be the single prefix-path part of Tree;
(03) let Q be the multipath part with the top branching node replaced by a null root;
(04) for each combination (denoted as) of the nodes in the path P do
(05) generate pattern ł a with support = minimum support of nodes in ;
(06) let freq pattern set(P) be the set of patterns so generated;
(07) else let Q be Tree;
(08) for each item ai in Q do // Mining multipath FP-tree
(09) generate pattern = ai ł a with support = ai. support;
(10) construct s conditional pattern-base and then s conditional FP-tree Tree ;
(11) if Tree then
(12) call FP-growth (Tree ,);
(13) let freq pattern set(Q) be the set of patterns so generated;
(14) return(freq pattern set(P) ł freq pattern set(Q) ł (freq pattern set(P) freq pattern set(Q)))

2) BGN: Gen ()

1. Choose large primes q, r and set n = qr.
2. Find a super singular elliptic curve E/Fp with a point P of order n as described above, and let G = hPi.
3. Choose Q0 R G and set Q = [r] Q0; then Q has order q. Let e: G G n Ł Fp 2 be the modified Weil pairing (constructed from the Weil pairing using a distortion map)

Output the public key pk = (E, e, n, P, Q) and the secret key sk = q.

Encryption

Enc (pk, m): Choose t R [1, n] and output C = [m]P + [t]Q.
Decryption Dec (sk, C): Compute P = [q]P and C = [q]C, and output m0 = logP C.

3) Paillier: Encryption

1. Randomly take two prime numbers p and q
2. Generate public key and private key PKeva, SKeva KeyGeneration
3. Encrypt data by using PKeva (public key) $Encrypt_data$ = (M; P Keva)

Decryption

1. Decrypt data by using SKeva (Secret key)
2. $Decrypt_data$ = (M; SKeva)

4.2 Mathematical Model

Set theory: Let S = I, P, R, O
Where,
S: Frequent Itemset Mining system.
I: Set of inputs.
P: Set of processes.
R: Rules or constraints.
O: Set of outputs/Final output I = i1, i2,.....,in

Where,
i1,i2,...,in = data sets. P=p1, p2, p3, p4, p5, p6, p7 Where,
p1: Dataset collection
p2: Working with cloud provider
p3: Data Mining
p4: Security R = r1 Where,
r1: No one should be able access others data. O= o1
Where,
O1:Valid user cloud access any file.

5. Conclusion

Proposed system sufficiently shows the better precision for the extraction of interesting patterns form the dataset. System efficiently extracts the data and parses them too to get rid of the redundant data. Then the parsed data is being preprocessed to get the most important data using BGN algorithm. System successfully identifies the all the possible frequent item sets with their candidates sets and this horizontal data is being converted into vertical data for FP growth mining algorithm. Above research can be extended by researchers for distributed information retrieval and finding new better interesting pat-terns.

6. Acknowledgment

Miss. Rupali Bichitkar currently pursuing M.E (Computer) from Department of Computer Engineering, G.H.RAISONI COLLEGE OF ENGINEERING AND MANAGEMENT, CHAS, AHMEDNAGAR -414 008. My area of interest is data mining and cloud computing.
Prof. V.V. Jagtap currently working as M.E coordinator and HOD in Department of Computer Engineering, G.H.RAISONI COLLEGE OF ENGINEERING AND MANAGEMENT, CHAS, AHMEDNAGAR -414 008.His area of interest is data mining and cloud computing.

7. References

1. Vaidya J, Clifton C. 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.639644.
2. Vaidya and C. Clifton, Secure set intersection cardinality with applica-tion to association rule mining, Journal of Computer Security, 2005; 13(4):593622.
3. Dong C, Chen L. A fast secure dot product protocol with application to privacy preserving association rule mining, in Advances in Knowledge Discovery and Data Mining. Springer, 2014, 606617.
4. Yi X, Rao F.-Y, Bertino E, Bouguettaya A. Privacy-preserving association rule mining in cloud computing, in Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security. ACM, 2015, pp.439450.
5. Oliveira SR, Zaiane OR. Privacy preserving frequent itemset mining, in Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14. Australian Computer Society, Inc., 2002, pp.4354.
6. Kantarcioglu M, Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data, IEEE Transactions on Knowledge Data Engineering, 2004; 9, pp.10261037.
7. Wong WK, Cheung DW, Hung E, Kao B, 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.111122.
8. Wong WK, Cheung DW, Hung E, Kao B, Mamoulis N. Security in outsourcing of association rule mining, in Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, 2007, pp.111122.
9. Kantarcioglu M, Nix R, Vaidya J. An efficient approximate protocol for privacy-preserving association rule mining, in Advances in Knowledge Discovery and Data Mining. Springer, 2009, pp. 515524.
10. Qiu L, Li Y, Wu X. Preserving privacy in association rule mining with bloom filters, Journal of Intelligent Information Systems. 2007; 29(3):253278.
11. Dong C, Chen L, Wen Z. When private set intersection meets big data: an efficient and scalable protocol, in Proceedings of the 2013 ACM SIGSAC conference on Computer communications security. ACM, 2013, pp.789800.
12. Lai J, Li Y, Deng RH, Weng J, Guan C, Yan Q. Towards semantically secure outsourcing of association rule mining on categorical data, Information Sciences, 2014; 267, pp.267286.