SECURED ASSOCIATION RULE MINING IN CLOUD DATA

K. Neeraja[1] M. Tech (CSE), G Priyanka Jeeva Karunya[2] M. Tech (CSE)

[1, 2] Working as Lecturer (Ad Hoc) in Department of Computer Science and Engineering

JNTUH CES, Sulthanpur.

ABSTRACT

Frequent item set mining, which is the essential operation in association rule mining, is one of the most widely used data mining techniques on massive datasets nowadays. With the dramatic increase on the scale of datasets collected and stored with cloud services in recent years, it is promising to carry this computation-intensive mining process in the cloud. Amount of work also transferred the approximate mining computation into the exact computation, where such methods not only improve the accuracy also aim to enhance the efficiency. However, while mining data stored on public clouds, it inevitably introduces privacy concerns on sensitive datasets.

In this paper a framework is proposed, where data is stored in trusted public cloud in an encrypted format and performed mining process to extract the association rules. CSP (client service provider sends the extracted rules to data users. Here data is secured by applying homomorphic encryption schemes. Since to maintain privacy noise is added to the data. A Enhanced apriori algorithm is proposed which is applied on large dataset without losing its efficiency and accuracy. The overhead of fictitious data [35] is to reduce so that the algorithm works more efficient on large datasets.

Keywords: *Frequent item set mining, secured privacy preserving, cloud computing.*

1. INTRODUCTION

With the raise of big data, data mining has become one of the most emerging techniques for data analytics on massive datasets. As one of the fundamental operations in association rule mining, frequent itemset mining, plays a significant role in market prediction[1], intrusion detection[2], network traffic management [3]and web analysis[4]. Specifically, given a transaction database, where each tuple in this database is denoted as a transaction, frequent itemset mining is able to discover popular item sets and their potential interesting connections from this transaction database. For instance, Netflix or RedBox can perform frequent item set mining to discover whether people watched Fast and Furious will also watch Mission Impossible. In order to recommend new-release movies to costumers. The easy accesses and low prices of public cloud services can significantly save mining costs on massive datasets. More importantly, since large-scale data from multiple and variable data sources (e.g., millions of users and devices) are collected by cloud services, such as Google, mining on cloud data can also dramatically improve the accuracy and effectiveness of mining. On the other hand, since many datasets collected by cloud services are sensitive, such as locations, medical record sand financial data, mining these confidential data inevitably brings important privacy issues. For example, a curious service provider can easily reveal confidential data in a transaction database while users and miners would like to keep the data in private during the entire mining process. In order to preserve privacy of frequent itemset mining in public cloud services, some randomization-based approaches [5], [6], [7], [8], [9] have been proposed. Unfortunately, randomization based approaches dramatically scarify the accuracy and utility of frequent itemset mining with limited privacy guarantee1. Movinga step forward, Yi et al. [10] recently proposed a privacy-preserving frequent itemset mining protocol in public clouds, where all the encrypted transactions are centralized to the cloud and miners delegate all the mining tasks to the cloud. However, to enable frequent itemset mining, this work requires n (where $n \geq 2$) aided semi-honest servers(besides the cloud server storing data)to perform distributed

decryption during the evaluation on encrypted data. Moreover, the need of these additional n aided semi-honest servers slows down the running time of frequent itemset mining, and introduces huge interactions and communication overheads.

In this I propose a framework where association rule mining service is outsourced to cloud service provider. The data owner sends the encrypted data to store in the public cloud. Data owner also stores the information about the fiction records [18]. The cloud performs mining on the encrypted data and sends results to the data owner. A cloud-aided privacy-preserving frequent itemset mining solution for partitioned databases, which is then used to build a privacy-preserving association rule mining solution. Both solutions are designed for applications where data owners have a high level of privacy requirement. The solutions are also suitable for data owners looking to outsource data storage – i.e. data owners can outsource their encrypted data and mining task to a semi-trusted (i.e. curious but honest) cloud in a privacy preserving manner.

The contributions of this paper are

1. This paper proposes a homomorphic encryption scheme to facilitate secure outsourced computations of supports/ confidences, as well as a secure outsourced comparison scheme for comparing supports/confidences with thresholds. The scheme only requires modular additions and multiplications, and is more efficient.
2. Enhanced apriori algorithm is used to extract frequent item sets and association rules.
3. The data user receives the rules in encrypted format and decrypts them. The fictitious records are identified and removed by the authorized data user.

2. LITERATURE SURVEY

A. *Frequent itemset mining*

Apriori algorithm is the originality algorithm of Boolean association rules of mining frequent item sets, raised by R. Agrawa and R. Srikan in 1994. The core principles of this theory are the subsets of frequent item sets are frequent item sets and the supersets of infrequent item sets are infrequent item sets.

The algorithm is used to find out all the frequent item sets. In the first iteration, item set A directly constitutes the first candidate item set C1. Assume that A={a1,a2,…,am}, then C1={{a1},{a2},…,{am}}. In the Kth iteration, firstly, the candidate item set Ck of this iteration emerges according to the frequent item set Lk-1 found in the last iteration. (The candidate item set is the potential frequent item set and is the superset of the K-1th frequent item set. Item set with k candidate item sets is expressed as Ck, which was consisted by k frequent item sets Lk .) Then distribute a counter which has a initial value equals to zero to ever item set and scan affairs in database D in proper order. Make sure every affairs belongs to each item sets and the counter of these item sets will increase. When all the affairs have been scan, the support level can be gotten according to the actual value of |D| and the minimum support level of the certain Ck of the frequent item set. Repeat the process until no mew item occurs. [4] The algorithm includes two key processes: connecting step and pruning step. Connecting step: in order to get Lk , connect Lk-1 with itself. Set this candidate as Ck and assume L1 and L2 are the item sets of Lk-1. Li[j] is the jth item of Li . Assume the affairs and items of the item set are in the dictionary order. Execute the connection Lk-1><Lk-1, in which the elements of Lk-1, L1 and L1, are connectable, if they have the same first (k-2)th items. Research of an Improved Apriori Algorithm in Data Mining Association Rules Jiao Yabing Market basket analysis Static data association rules mining Agrawal,1993 Frequent closed itemsets mining Pasquier 1991 Maximu m frequent itemsets mining Bayardo, 1998 Fuzzy associati on rules mining Srikant, 1996 Uncertai n data mining based on Frequent Itemsets Mining Agarwal, 2009 International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013 25 That is, the elements of Lk-1, L1 and L1, are connectable, if (L1[1]=L2[1]) ∧ (L1[2]=L2[2]) ∧ ……(L1[k-2]=L2[k-2]) ∧ (L1 [k-

1]=L2[k-1]). The requirement of (L1[k-1] <L2[K-1]) simply assure no repetition.

B. Secure privacy preserving.

Paillier Encryption.

Paillier Encryption is an additively homomorphic encryption proposed by Paillier [33]. In this paper, we mainly focus on its additive homomorphism, which means it can compute a new ciphertext of plaintext (m1 + m2) based on the two ciphertexts of plaintext m1 and m2 without revealing the plaintexts. Specifically, • Given [m1] and [m2], we can compute [m1 + m2] = [m1]· [m2]. • Given [m1] and a plaintext α, we can compute [α·m1] = [m1]α. where [m] describes the Paillier ciphertext of a plaintext m. More details of construction of Paillier encryption can be found in [33].

BGN Encryption. Boneh-Goh-Nissim

(BGN) encryption is a somewhat homomorphic encryption, which was proposed by Boneh et al. [34]. With the homomorphic property, it can compute an arbitrary number of additions and also one multiplication on encrypted data. Specifically, we have • Given ‖m1‖ and ‖m2‖, compute ‖m1 + m2‖ = ‖m1‖⊙‖m2‖ • Given ‖m1‖ and ‖m2‖, compute ‖m1 ×m2‖ = ‖m1‖⊗‖m2‖where‖m‖representstheBGNciphertextofap laintextm,⊙and⊗ denote the corresponding operations in the ciphertext domain. More construction details of BGN can be found in [34]. We will use Paillier and BGN respectively for evaluating inner products with different privacy requirements, which essentially leads to secure association rule mining. Understanding the homomorphic properties of these two primitives will be sufficient to follow the design of our protocols. The details and security proofs of Paillier and BGN can be found in [33] and [34] respectively

C. Cloud computing

The cloud computing system provides the service for the user and has the character of high scalability and reliability. The resource in the cloud system is transparent for the application and the user do not know the place of the resource. The users can access your applications and data from anywhere. Resources in cloud systems can be shared among a large number of users. The cloud system could improve its capacity through adding more hardware to deal with the increased load effectively when the work load is growing. Cloud resources are provided as a service on an as needed basis. The cloud itself typically includes large numbers of commodity-grade servers, harnessed to deliver highly scalable and reliable on-demand services. The amount of resources provided in the cloud system for the users is increased when they need more and decrease when they need less. The resource can be the computing, storage and other specification service. The cloud computing is seen as the important change of information industry and will make more impact on the development of information technology for the society. The majority of cloud computing infrastructure currently consists of reliable services delivered through data centre that are built on servers with different levels of virtualization technologies. The services are accessible anywhere in the world, with The Cloud appearing as a single point of access for all the computing needs of consumers. The cloud computing changed the style of software. The data can be stored in the cloud system and the user can use the data in any time and in anywhere.

3. PROBLEM STATEMENT

a. System Model

As Shown in the Fig.1 the data owner can send data to CSP in encrypted form by some fictitious records added in order to maintain privacy from cloud. For CSP it will be difficult to identify the original records. The miner can send the query in plaintext **or** cipher which CSP will apply the enhanced Apriori algorithm to get association rules. These rules are sent to data owner/user who decrypt the results and removes noise i.e., fictitious records from the received dataset. The expected performance of the model will be improvised than [36] as the overhead of noise is reduced and the mining

algorithm applied will also give same performance on the larger dataset too.

b. Objectives

- The expected model maintains privacy and security among the data from third party users.
- Reduces overhead of noise
- Enhanced apriori algorithm works on larger dataset efficiently.

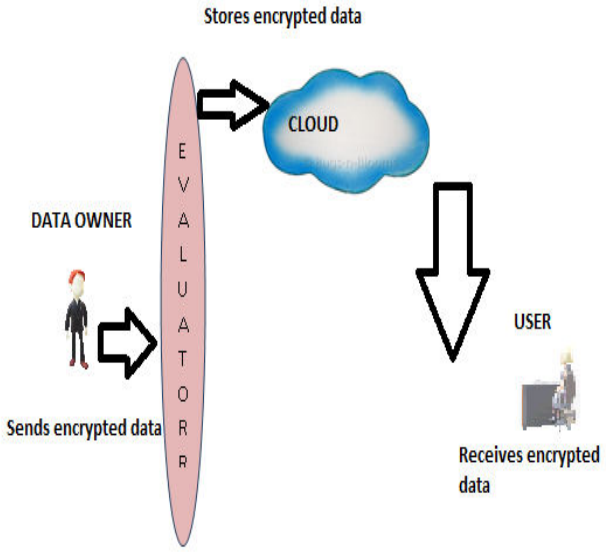
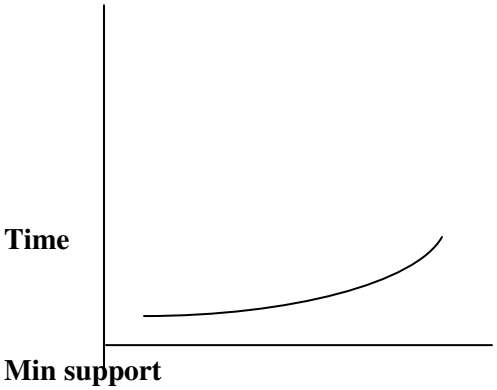


Fig 1. Secured mining on cloud storage

c. Implementation

As proposed in [37] the above mobel can be used by parallel processing. This is implemented by distributed the cloud storage among the systems and applying the apriori algorithm parallely and assembling the results from different systems and sending them to the client. This parallelism is because of increasing size in data and to improve the performance of apriori. The apriori algorithm basically suggestable for limited storage datasets. But in this case even though there is a increase in size parallelism can be applied. Parallelization of association rule mining algorithms is an important task in data mining to mine frequent patterns from transaction databases. These algorithms either distribute database horizontally or increase number of CPU to reduce execution time of frequent pattern mining.

d. Performance results:



This is the expected performance of proposed system where there is no difference between the [36] and the former one. Even though the size of the data base is increased the cost of parallelism is added with same efficiency.

4. CONCLUSION

In this research proposal, frequent itemset mining on encrypted data with reducing overhead of noise is expected to work better on larger dataset. Apriori algorithm is enhanced to work better on even more large datasets. When the data is outsourced the privacy also preserved by encrypting the data before sending it to cloud.

REFERENCES

[1] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic itemset counting and implication rules for market basket data,” in ACMSIGMOD Record, vol. 26, no. 2. ACM, 1997, pp. 255–264.

[2] W. Leeand S.J. Stolfo, “Data mining approaches for intrusion detection,” in Usenix Security, 1998.

[3] C. Estanand G. Varghese, New directions in traffic measurement and accounting. ACM, 2002, vol. 32, no. 4.

[4] B. Mobasher, N.Jain,E.-H.Han, and J.Srivastava, “Webmining: Pattern discovery from world wide web

transactions," Technical Report TR96050, Department of Computer Science, University of Minnesota, Tech. Rep., 1996.

[5] 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.

[6] 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.

[7] 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.

[8] T. Tassa, "Secure mining of association rules in horizontally distributed databases," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no. 4, pp. 970–983, 2014.

[9] 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.

[10] 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

[11] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, vol. 22, no. 2. ACM, 1993, pp. 207–216.

[12] R. Agrawal, R. Srikant et al., "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, 1994, pp. 487–499.

[13] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Information Systems, vol. 29, no. 4, pp. 343–364, 2004.

[14] S. R. Oliveira and O. R. Zaiane, "Privacy preserving frequent itemset mining,"in Proceedings of the IEEE international conferenceon Privacy, security and data mining-Volume14. Australian Computer Society, Inc., 2002, pp. 43–54.

[15] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining," in Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, 2002. RIDE-2EC 2002. Proceedings. Twelfth International Workshop on. IEEE, 2002, pp. 151– 158.

[16] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment, 2002, pp. 682–693.

[17] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Information Systems, vol. 29, no. 4, pp. 343–364, 2004.

[18] F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving mining of association rules from outsourced transaction databases," Systems Journal, IEEE, vol. 7, no. 3, pp. 385–395, 2013.

[19] K. Sathiyapriya and G. S. Sadasivam, "A survey on privacy preserving association rule mining,"International Journal of Data Mining & Knowledge Management Process, vol. 3, no. 2, pp. 119–131, 2013.

[20] R. Kharat, M. Kumbhar, and P. Bhamre, "Efficient privacy preserving distributed association rule mining protocol based on random number," in Intelligent Computing, Networking, and Informatics. Springer, 2014, pp. 827–836.

[21] 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.

[22] C.-H. Tai, P. S. Yu, and M.-S. Chen, "k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 473–482.

[23] M. Kantarcioglu, R. Nix, and J. Vaidya, "An efficient approximate protocol for privacy-preserving association rule mining," in Advances in Knowledge Discovery and Data Mining. Springer, 2009, pp. 515–524.

[24] L. Qiu, Y. Li, and X. Wu, "Preserving privacy in association rule mining with bloom filters," Journal of Intelligent Information Systems, vol. 29, no. 3, pp. 253–278, 2007.

[25] C. Dong, L. Chen, and Z. Wen, "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. 789–800.

[26] X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," in Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on. IEEE, 2010, pp. 345–350.

[27] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, "Towards semantically secure out sourcing of association rule mining on categorical data," Information Sciences, vol. 267, pp. 267–286, 2014.

[28] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft, "Privacy-Preserving Ridge Regression on Hundred of Millions of Records," in Proc. of IEEE S&P'13, 2013.

[29] A. Peter, E. Tews, and S. Katzenbeisser, "Efficiently Outsourcing Multiparty Computation Under Multiple Keys," IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, pp. 2046–2058, 2013.

[30] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data," Crypto ePrint Archive, 2014.

[31] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," in Proc. of ACM CCS'06, 2006. [32] O. Goldreich, Foundations of cryptography: volume 2, basic applications. Cambridge university press, 2004.

[33] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes." Springer, 1999, pp. 223–238.

[34] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in Theory of cryptography. Springer, 2005, pp. 325–341.

[35] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques:

[36] ShuoQui, Boyang Wang, Ming Li, Jiqiang Liu and Yanfengshi "Toward practical privacy preserving frequent itemset mining on Encrypted cloud data" IEEE transactions on cloud computing 2017.

[37] "Hp-Apriori: Horizontal Parallel-Apriori Algorithm For Frequent Itemset Mining from Big Data "Mohammad-Hossein Nadimi-Shahraki and Mehdi Mansouri in 2017 IEEE 2nd International Conference on Big Data Analysis.