Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 2 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Privacy Preserving Access of Outsourced Data in Heterogeneous Databases

By J.Bama & Dr. M.S.Thanabal
PSNA College of Engineering and Technology

Abstract- Privacy is main concern in the world, among present technological phase. Information security has become a dangerous issue since the information sharing has a common need. Recently, privacy issues have been increased enormously when internet is flourishing with forums, social media, blogs and e-commerce, etc. Hence research area is retaining privacy in data mining. The sensitive data of the data owners should not be known to the third parties and other data owners. To make it efficient, the horizontal partitioning is done on the heterogeneous databases is introduced to improve privacy and efficiency. we address the major issues of privacy preservation in information mining. In particular, we consider to provide protection between different data owners and to give privacy between them by partitioning the databases horizontally and the data's are available in the heterogeneous databases. Our proposed work is to center around the study of security saving on unknown databases and conceiving private refresh methods to database frameworks that backings thoughts of obscurity assorted than k-secrecy.

Keywords: *privacy preserving, homomorphic encryption, third parties.*

GJCST-C Classification: *H.2.5*

Strictly as per the compliance and regulations of:

Privacy Preserving Access of Outsourced Data in Heterogeneous Databases

J.Bama [α] & Dr. M.S.Thanabal [σ]

Abstract- Privacy is main concern in the world, among present technological phase. Information security has become a dangerous issue since the information sharing has a common need. Recently, privacy issues have been increased enormously when internet is flourishing with forums, social media, blogs and e-commerce, etc. Hence research area is retaining privacy in data mining. The sensitive data of the data owners should not be known to the third parties and other data owners. To make it efficient, the horizontal partitioning is done on the heterogeneous databases is introduced to improve privacy and efficiency. we address the major issues of privacy preservation in information mining. In particular, we consider to provide protection between different data owners and to give privacy between them by partitioning the databases horizontally and the data's are available in the heterogeneous databases. Our proposed work is to center around the study of security saving on unknown databases and conceiving private refresh methods to database frameworks that backings thoughts of obscurity assorted than k-secrecy. Symmetric homomorphic encryption scheme, which is significantly more efficient than the asymmetric schemes. Our proposed work helps the valid user can extract with key issue in partition data in automated approach and the data's are partitioned horizontally.

Keywords: privacy preserving, homomorphic encryption, third parties.

I. Introduction

Now a days, data's are the biggest assets. We can see that increasing number of organizations that collect data very often concerning about the individuals and used them for various purposes such as scientific research, medical data, marketing etc. Organization may also give access to the data they own or even release such data to third parties. Data once released are no longer under the control of the organization owning them. So, the organization owners cannot prevent the modification of the data. The main problem is addressed as preserving the privacy of the data being stored in the databases.

Data mining is the process of extracting the knowledge from the enormous set of databases. The data mining has various applications such as Market Analysis and Management, Corporate Analysis and Risk Management, Fraud Detection, Intrusion Detection (Intrusion Detection means any kind of action that threatens integrity, confidentiality or the availability of the network resources),Retail Industry, Biological Data Analysis, Financial Data Analysis, Telecommunication Industry. There are some major disadvantages of data mining are their privacy issues, security issues and others as the misuse of information.

So, data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as Clustering, Classification and Association Rule Mining [10]. Data mining deals with the mining of kinds of patterns. The kinds of patterns can be done in two ways either descriptive or classification and prediction. Our project deals with the descriptive way (ie) Mining of Associations.

The preview of data mining which falls within the problem of finding association rules is also called as Knowledge discovery in databases. In the Retail stores the Associations are used to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules [3]. The main objective of data privacy is to protect the personally identifiable information. In case, if the general information is considered then it should be linked either directly or indirectly to an individual person. The personal data subjected to mining and then the attribute values related with the individuals should be kept private and must be protected from disclosure. Data miners can be able to learn from the global models instead from the characteristics of a particular individual. The objective is not only to protect the personally known data but also to identify the trends and patterns that are not supposed to be discovered. Some information requires special care and handling. Incase if the information is handled inappropriately then it results in penalties, identify theft, financial loss, invasion of privacy or unauthorized access by one or more individuals. The confidential data's sensitivity is high. For example, confidential data's are research details, library transactions, personal information, information covered by non-disclosure agreements, contracts, facilities, management information. The concept of partitioning is used in our project for ensuring the privacy of data's available in the databases. Partitioning is the process of physically or logically partitioning data into segments that are more easily accessed or maintained and also

Author α: PG SCHOLAR, Dept. of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu.
e-mail: Bamavani1993@gmail.com
Author σ: ASP/CSE, Dept. of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu.
e-mail: Ms_thanabal@yahoo.com

the arrangement of allocating data to data sites which occurs within the same common data architecture. Partitioning can be done in two ways. They are vertical partitioning and horizontal partitioning .If the databases are partitioned in the column wise then it is called vertical partitioning of databases. If the databases are partitioned in the row wise then the partitioning is known to be horizontal partitioning of databases.

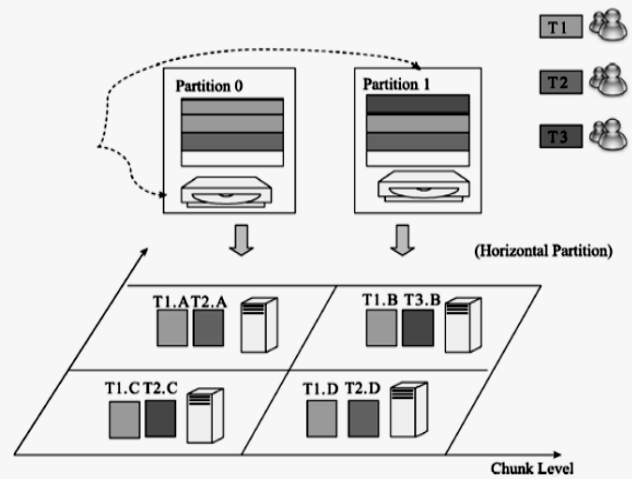


Figure 1

II. Background

Sensitive information is defined as data that is protected against unwanted disclosure. Access to sensitive information should be safeguarded. Sensitive information includes all data, in its original and duplicate form, for which there is either legal, ethical or contractual requirement that it be protected or access restricted. Also includes the unauthorized access of any data that is protected by the university policy. This information must be restricted to those with a legitimate business need for access. For example, public safety information, financial donor information, information security records, information file encryption keys. In order to find the association rules and their causal relationship between the set of items can be found using the association rule mining. So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. It is machine learning - rule based approach for determining the relationship between the variables used in large databases. Association rules are largely used between the products in large scale transaction data being recorded by the point-of-sale (POS) systems in supermarkets which was done according to the concept of Rakesh Agarwal, Tomasz and Arun Swami. Association Rules are used in market analysis, intrusion detection, mining, bioinformatics, and continuous production . In spite of the sequence mining, they don't consider the order of items either across the transaction or within the transaction.

a) Association Rule Mining (Arm)

The popular research method used in data mining for discovering the interesting relations between variable in large databases. Association Rules (AR) are useful for analyzing and predicting the customer behavior. The IF-THEN statements are the association rules (AR) that help to uncover the relationship between unrelated data available in a relational database or other information Repository. Example If a customer buys bread then 80% of people are expected to buy butter. These association rules expresses about how the items or objects are related to each other and how they tend to group together.

The Association strength can be measured using the support, confidence values. Support is the ratio of the number of itemsets satisfying both antecedent and consequent to the total number of transactions. Confidence is derived from subset of transactions in which the two entities are related. Association Rule Mining can be done using three algorithms. They are Apriori algorithm, FP Growth algorithm, Éclat algorithm. In my work I have used FP Growth algorithm for discovering the frequent patterns in the database.

- Support(X) =No. of transactions contains X/total number of transactions.
- Confidence= (support(X union Y))/Support(X) .

b) FP Growth Algorithm

FP Growth algorithm means Frequent Pattern Growth Algorithm which is a scalable technique for mining the frequent pattern in the database. Frequent item set mining is possible without candidate generation so that the FP Growth algorithm is more efficient and a biggest improved over the Apriori Algorithm. This algorithm consumes less memory and a linear running time.

Procedure:

1. Build a compact data structure called the FP tree.
2. Extracts Frequent Itemsets directly from the FP tree.

III. Proposed Work

The proposed system introduced a privacy-preserving outsourced frequent itemset mining solution for horizontal partitioned from heterogeneous databases. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. Based on this solution, we built a privacy-preserving outsourced association rule mining solution for horizontal partitioned databases for the unknown database and conceiving private refresh methods to database frameworks that backings thoughts of obscurity assorted than K-secrecy. Our solutions protect data owner's raw data from other data owners and the server. Our solutions also ensure the privacy of the mining results from the server is shown in Figure 3.1

Compared with most existing solutions, our solutions cannot leak less information about the data owners' raw data. Therefore, our solutions are suitable to be used by data owners wishing to outsource their databases to the cloud but require a high level of privacy without compromising on performance. Other than the settings of vertically partitioned databases and cloud/third-party-aided mining, privacy-preserving frequent itemset mining and association rule mining have been studied in the settings of horizontally partitioned databases data publishing and differential privacy.

The proposed system introduce a symmetric homomorphic encryption scheme using mediate Certificate less algorithm (using only modular additions and multiplications), which is significantly more efficient than asymmetric schemes. The scheme supports many homomorphic additions and limited number of homomorphic multiplications, and comprises the following three algorithms Key generation algorithm, Encryption algorithm, Decryption algorithm.

Advantage

The advantage of the proposed system is that the valid user can extract with key issue in partition data in automated approach.

a) Features of Proposed System

(i) The feature of the proposed system is managing data in horizontal partitioning.
(ii) The partition data is converted sensitive format by using Mediated Certificate less Algorithm.
(iii) If any valid user wants to review their original sources, they must submit valid attribute to extract heterogeneous databases.

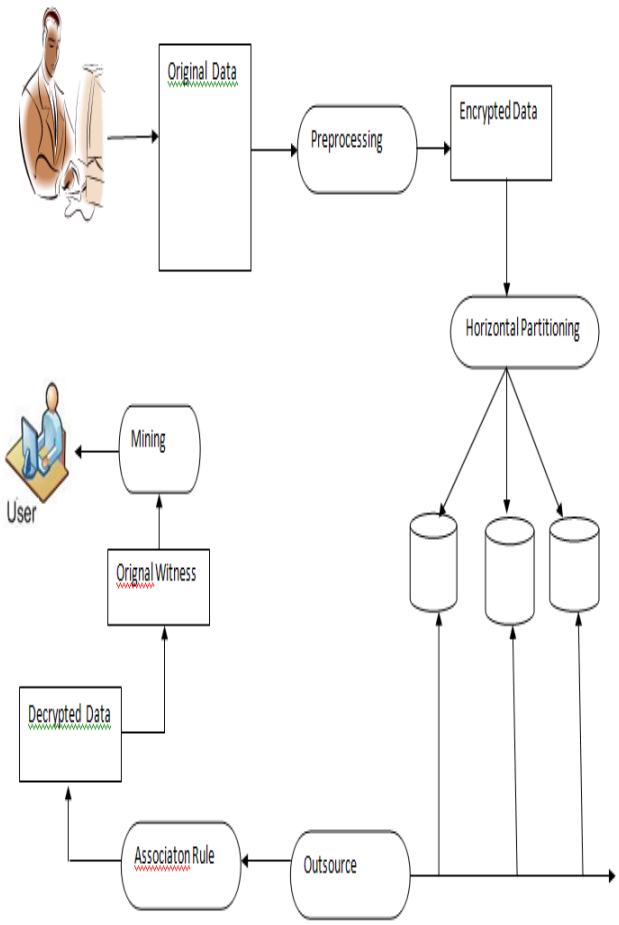


Figure 3.1: Architecture of proposed system

IV. Performance Evaluation

The horizontal partitioning algorithms used for comparison
As follow:

1. *Hash partition:* The data is evenly distributed to the predefined individual partitions, which ensures that the data of each partition has the same amount roughly;
2. *Schism partition:* Duplicate overlapped nodes; generate Partition according to the spanning graph.

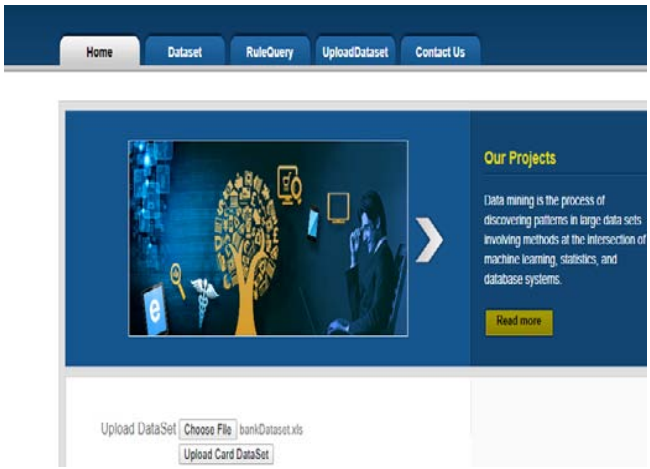
a) Partition Quality Evaluating

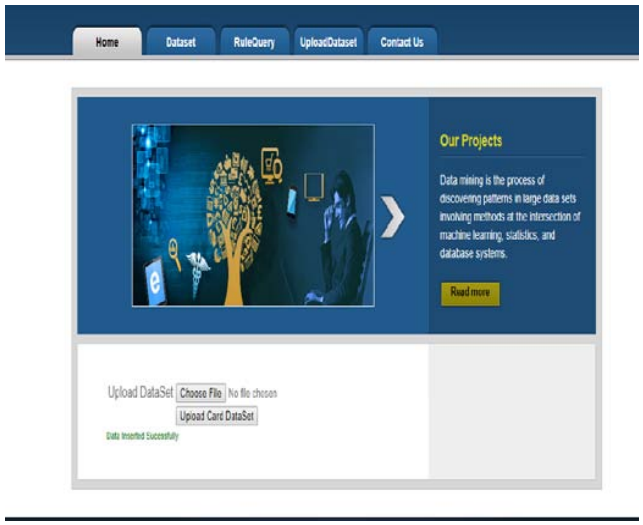
When a query accessed attributes on different partitions we call it a distributed query. The distributed queries cost more resources, so we regard the proportion of non distributed queries as the quality of the partitioning scheme.

The result shows that the proportion of non-distributed queries of larger than Hash and Schism, which indicates that with a certain number of partitions, the efficiency is better than Hash and Schism. In addition, Hash and Schism must know it before. The proposed Scheme can dynamically adapt to the coming

workload and predict the trend of the workload to give a better partition scheme, so proposed Scheme has better partition quality than Hash and Schism.

V. Result





VI. Conclusion

We proposed a privacy-preserving outsourced frequent itemset mining solution for horizontal partitioned databases. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. Based on this solution, we built a privacy-preserving outsourced association rule mining solution for horizontal partitioned databases. Our solutions also ensure the privacy of the mining results from the cloud. Compared with most existing solutions, our solutions leak less information about the data owners' raw data. Our evaluation has also demonstrated that our solutions are very efficient; therefore, our solutions are suitable to be used by data owners wishing to outsource their databases to the cloud but require a high level of privacy without compromising on performance.

References Références Referencias

1. Arun, V., Gowthami, S., & Padma, S. K. (2015, December)' Securely mining transactional databases for association rules using FDM' In Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on (pp. 340-345) IEEE.
2. Adhvaryu, R. V., & Domadiya, N. H. (2014)'Privacy Preserving in Association Rule Mining On Horizontally Partitioned Database' International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 3(5).
3. Agrawal, R., & Srikant, R. (1994, September)' Fast algorithms for mining association rules' In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).
4. Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., & Moser, S. A. (1998)'Association rules and data mining in hospital infection control and public health surveillance' Journal of the American medical informatics association, 5(4), 373-381.
5. Creighton, C., & Hanash, S. (2003)' Mining gene expression databases for association rules' Bioinformatics, 19(1), 79-86.
6. Dong, X., & Li, X. (2015, November)'A Novel Distributed Database Solution Based on MySQL' In Information Technology in Medicine and Education (ITME), 2015 7th International Conference on (pp. 329-333) IEEE.
7. Han, J., Pei, J., & Yin, Y. (2000, May)'Mining frequent patterns without candidate generation' In ACM sigmod record (Vol. 29, No. 2, pp. 1-12). ACM.
8. Kantarcioglu, M., & Clifton, C. (2004)'Privacy-preserving distributed mining of association rules on

horizontally partitioned data' IEEE transactions on knowledge and data engineering, 16(9), 1026-1037.
9. Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001, November)'Effective personalization based on association rule discovery from web usage data'In Proceedings of the 3rd international workshop on Web information and data management (pp. 9-15). ACM.
10. Rozenberg, B., & Gudes, E. (2006)'Association rules mining in vertically partitioned databases' Data & Knowledge Engineering, 59(2), 378-396.
11. Sheikhalishahi, M., and Martinelli, F. (2017, July)'Privacy preserving clustering over horizontal and vertical partitioned data' In Computers and Communications (ISCC), 2017 IEEE Symposium on (pp. 1237-1244) IEEE.
12. Yin, X., & Han, J. (2003, May) 'CPAR: Classification based on predictive association rules' In Proceedings of the 2003 SIAM International Conference on Data Mining (pp. 331-335). Society for Industrial and Applied Mathematics.
13. Zhan, J., Matwin, S., & Chang, L. (2005, August)'Privacy-preserving collaborative association rule mining' In DBSec (pp. 153-165).
14. Zaki, M.J. (2000)'Scalable algorithms for association mining'IEEE Transactions on Knowledge and Data Engineering, 12(3), 372-390.