ISSN (Online) : 2454-4159

International Journal of Advanced Research in Science Management and Technology

IJARSMT

Volume 2, Issue 3, March 2016

A Novel Method for Association Rules in Horizontal Distributed Database in Secure Mining

Patil Nikhil[1], BalsaneAbhijeet[2], Bhoye Kishor[3], Pagar Pramod [4], Prof. J. A. Dandge[5]

UG Student, Department Of I.T., PVG's College Of Engineering, Nashik, Maharashtra, India[1,2,3,4]

Associate Professor, Department Of I.T., PVG's College Of Engineering, Nashik, Maharashtra, India[5]

ABSTRACT- We propose a process for an innovative way for association guidelines in horizontally distributed data source in secure mining. The existing existed protocol is certainly that of Kantarcioglu and Clifton. Our protocol, is founded on the Fast Distributed Mining (FDM) algorithm of Cheung et al. that is a distributed variant of the Apriori algorithm. The primary contents in our process happen to be two novel secure multi-party algorithms the one that computes the union of individual subsets that all of the interacting players maintain, and another that exams the inclusion of an factor held by one person in a subset kept by another. Our protocol offers increased privacy and security with respect to the protocol. In addition, it is simpler and is more efficient with regards to communication rounds significantly, communication cost and computational cost.

KEYWORDS – Privacy preserving data mining, distributed computation, frequent item sets, association rules.

I. INTRODUCTION

We propose a process for an innovative way for association guidelines in horizontally distributed data source in secure mining. The existing existed protocol is certainly that of Kantarcioglu and Clifton. Our protocol, is founded on the Fast Distributed Mining (FDM) algorithm of Cheung et al. that is a distributed variant of the Apriori algorithm. The primary contents in our process happen to be two novel secure multi-party algorithms the one that computes the union of individual subsets that all of the interacting players maintain, and another that exams the inclusion of an factor held by one person in a subset kept by another. Our protocol offers increased privacy and security with respect to the protocol. In addition, it is simpler and is more efficient with regards to communication rounds significantly, communication cost and computational cost.

We study here the condition of a novel method of association rules in horizontally distributed database in secure mining. In that setting, there are many sites (or players) that hold homogeneous databases, i.e., databases that show the same schema but hold information on several entities. The target is to find all association guidelines with support at least s and self-confidence at least c, for some given little support size sand self confidence level c, that carry in the unified data source, while minimizing the granted details disclosed about the individual databases kept by those players. The information that people want to protect in this context isn't only individual transactions in the several databases, but also more global information such as for example what association rules are supported locally in each of these databases. That target defines a issue of secure multi-party computation. In such problems, there are M players that hold private inputs, x1; . . . ; xM, plus they desire to securely compute 1/4 f?x1; . . . ; xM? for a few general public function f. If there existed a reliable alternative party, the players could surrender to him their inputs and he'd perform the function analysis and mail to them the resulting end result. In the lack of such a trusted alternative party, it is had a need to devise a process that the players can operate on their own as a way to arrive at the mandatory result y. Such a process is considered correctly secure if no person can study from his viewpoint of the protocol a lot more than what he would have got learnt in the idealized setting up where in fact the computation is completed by a trusted alternative party. Yao was the first ever to propose a generic remedy because of this nagging problem regarding two players. Other generic solutions, for the multi-party case, were proposed Inside our problem later, the inputs will be the partial databases, and the mandatory output is the set of association rules that holding the unified database with support and confidence no smaller compared to the given thresholds s and c, respectively. As all these generic solutions rely after a information of the function f as a Boolean circuit, they might be applied and then small features and inputs which happen to be realizable by straightforward circuits. In more technical settings, such as for example ours, other methods are necessary for undertaking this computation. In such instances, some relaxations of the idea of perfect security could be unavoidable when looking for functional protocols, so long as the excess information is regarded as benign Kantarcioglu and Clifton studied that trouble in devised a process because of its solution. The main portion of the process is a sub-process for the safe and sound computation of the union of exclusive subsets that are kept by the several players.(The non-public subset of confirmed player, as we below explain, includes the item models that are s-repeated in his partial data source.) This is the most costly portion of the protocol and its own implementation relies after cryptographic primitives such as for example commutative encryption, oblivious transfer, and hash functions. That is also the only component in the protocol where the players may extract from their look at of the protocol data on different databases, beyond what I simplied by the ultimate output and their unique input. While such leakage of information renders the protocol not secure perfectly, the perimeter of the surplus information is normally explicitly bounded in in fact it is argued presently there that such info leakage is normally innocuous, when suitable from a practical perspective.

II. LITERATURE SURVEY

1] Cost With Finish Time-Based Algorithm:[BolluJyothi, K. VenkateswaraRao]

In this paper, The CwFT algorithm is a workflow scheduling algorithm extended from the HEFT algorithm for distributed environments with multiple heterogeneous processing nodes. Instead of optimizing only the workflow make span as usual, CwFT algorithm also considers reducing the monetary cost that CCs need to pay in a computing framework with the combination between numerous Cloud nodes and a local system. Similar to HEFFT, the CwFT algorithm is comprised of two phases: Task Prioritizing to mark the priority level for all tasks and Node Selection to select tasks in a descending order by the priority level and then schedule each selected task on an appropriate processing node to optimize the value of the utility function.

2] Horizontal Aggregations Used In SQL:[Rajkumar.S ,V.Elavarasi]

This paper introduced a new class of aggregations that have similar behavior to SQL standard aggregations, but which produce tables with a horizontal layout. In contrast, we call standard SQL aggregations vertical aggregations since they produce tables with a vertical layout. Horizontal aggregations just require a small syntax extension to aggregate functions called in a SELECT statement. Alternatively, horizontal aggregations can be used to generate SQL code from a data mining tool to build data sets for data mining analysis.

3] Enhanced Kantarcioglu And Clifton's Scheme (ekcs):[chin-chenchang,jieh-shanyeh]
This study has proposed EKCS to reduce the communication overhead of KCS in the first phase. KCS requires k round of communication to transmit all local frequent itemsetsduring the distributed mining operation. According to the downward closure property of Apriori, a frequent kitemsetcontains 2k-1 frequent sub-item sets. A frequent item set has no frequent superset is called a maximum frequent item set (MFI). In database DB, the set of all MFIs by deleting the redundant frequent sub-item sets from Fcan represent all frequent item sets.

4] Synthetic database generation:[Ms.Manali Rajeev Raut, Ms.HemlataDakhore]
The generation of synthetic transactions is to evaluate the performance of the algorithms over a large range of data characteristics. The creation of synthetic data is an involved process of data anonymization; that is to say that synthetic data is a subset of anonym zed data. This data is used in a variety of fields as a filter for information that would otherwise compromise the confidentiality of particular aspects of the dat.

III. PROBLEM DEFINITION

The job is that the individual language can be interpreted with Temporal Data source and produce appropriate results. The machine is definitely for an English sentence to become interpreted by the pc and appropriate actions taken. Asking problems to databases in normal language is an extremely convenient and easy approach to data access, especially for everyday users who don't realize complicated data source query languages such as for example SQL.
In this the organic language query is used English Language, any kind of statement (WH type issues [2], word, any kind of statement etc.). Then your query in English vocabulary mapped regarding to syntax of SQL query that delivers user the accurate info from data source after execution of mapped SQL query. The reliability in mapped Query is targeted here.

The objective of the system is as follows:

- To get the accurate result with any database one should know the association rules of that particular database software (Microsoft SQL, Oracle, etc.).
- To provide better performance we use c# as platform.
- To use English language as natural language for particular data extraction.
- To reduce the human efforts.
- To make user-friendly we use C# as front end and SQL 2008 as backend because it easy to learn.

IV. PROPOSED SOLUTION

FREQUENT ITEM SETS

Frequent sets play an important role in lots of Data Mining responsibilities that look for interesting habits from databases, such as for example association guidelines, correlations, sequences, episodes, clusters and classifiers. The mining of association rules is probably the most popular problems of most these. The identification of units of items, products, characteristics and symptoms, which occur mutually in the given data source often, can be seen among the most elementary tasks in Data Mining.

PRIVACY PRESERVING

Privacy is one of the main properties of an information system must satisfy, in which systems the necessity to share information among unique, certainly not trusted entities, the safeguard of sensible information includes a relevant role. Thus privacy is becoming a crucial issue in many data mining applications increasingly. For that privacy secure distributed computation that was done within a more substantial body of research in the idea of cryptography has achieved exceptional results. These outcomes were proven using generic constructions which can be put on any function which has a competent representation as a circuit. A comparatively new trend implies that classical access control methods are not adequate to ensure privacy when info mining techniques are being used in a malicious method. Privacy preserving data mining algorithms have already been recently introduced with the purpose of stopping the discovery of practical information.

We describe here outcomes of a body system of cryptographic research that presents how separate get-togethers scan jointly compute any function of their inputs, without revealing any various other information. As we above argued, these effects achieve maximal personal privacy that hides all facts aside from the designated productivity of the function.

DECRYPTION

Decryption is the procedure for transforming data that is rendered unreadable through encryption back again to its unencrypted kind. In decryption, the machine extracts and converts the garbled info and transforms it to texts and photos that are often understandable not merely by the reader but also by the machine. Decryption may immediately be achieved manually or. It could also be performed with a couple of keys or passwords.

GLOBAL ITEMSET

Association guidelines find the romantic relationships between the several items in a data source of product sales transactions. Such guidelines track the buying habits in buyer behaviour eg. finding how the occurrence of one item in the occurrence is damaged by the transaction of another and so forth. The issue of association rule generation has gained considerable prominence in the info mining community as a result of the ability of its being used as a significant tool for knowledge discovery

ODD DATA

Data mining technology provides emerged as a way of identifying tendencies and patterns from large levels of data. Data mining and data warehousing go hand-in-hand: most tools operate by gathering all data right into a central site, running an algorithm against that info then. However, privacy concerns can prevent creating a centralized warehouse - data distributed among several custodians maybe, none which are permitted to transfer their data to some other site.

ARCHITECTURE

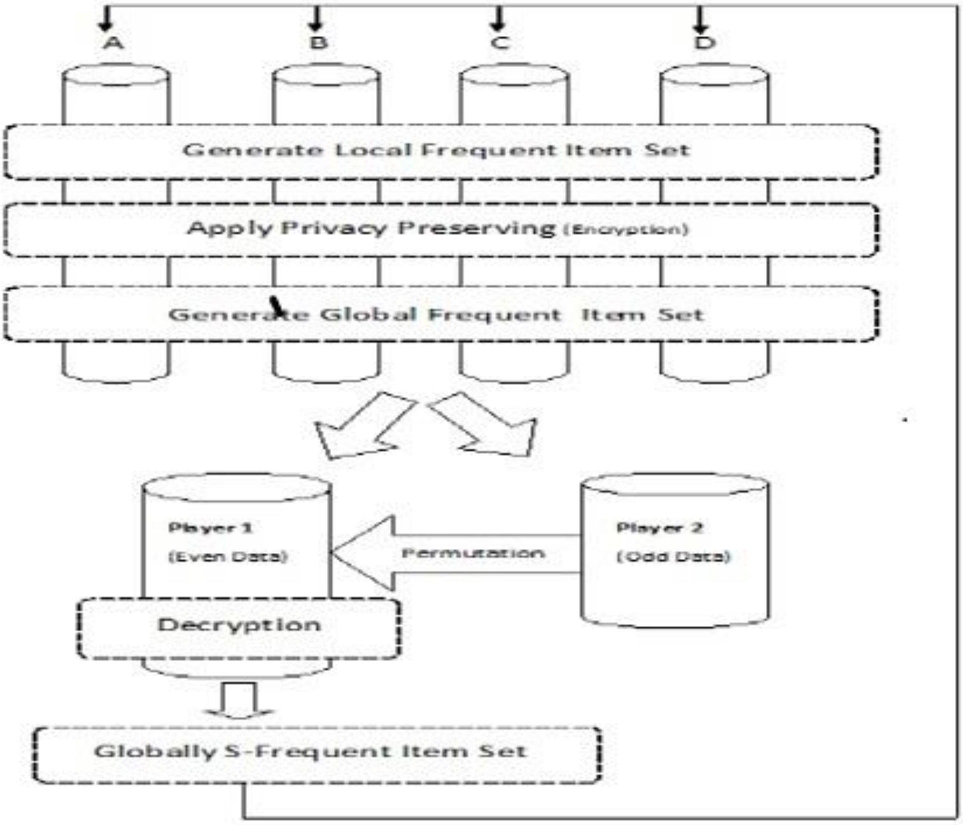
A
B
C
D
Generate Local Frequent Item Set
Apply Privacy Preserving (Encryption)
Generate Global Frequent Item Set
Player 1
(Even Data)
Permutation
Player 2
(Odd Data)
Decryption
Globally S-Frequent Item Set

Fig: 1. System Architecture

PERMUTATION

We check out the framework of permutation for assessing the effectiveness of classifiers. We review two basic permutation tests. The initial test evaluate if the classifier has found a genuine class structure in the info; the corresponding null distribution can be estimated by permuting labels in the data. This evaluation has been found in classification concerns in computational biology extensively. The next test studies if the classifier is exploiting the dependency between your features in classification; the corresponding null distribution can be approximated by permuting the features within classes, motivated by restricted randomization methods used in statistics. This new test can serve to recognize descriptive features that can be valuable information in increasing the classifier performance. We study the properties of these tests and present an comprehensive empirical analysis on real and synthetic data. Our analysis demonstrates studying the classifier performance via permutation tests works well. Specifically, the restricted permutation check obviously reveals if the classifier exploits the interdependency between your features in the info.

GLOBALLY s-FREQUENT ITEMSETS

Protocols UNIFI-KC and UNIFI yield the place Cks that contains all item units that happen to be locally s-frequent in at least one webpage. Those are the k-item sets which may have potential to be also globally s-frequent. To be able to reveal which of these item sets is globally s-frequent you will find a have to securely compute the support of every of these item sets. That computation should never reveal the neighbourhood support in virtually any of the sites.

PROBLEM SOLVINGAPPROACH:-

The system includes the following modules to solve the problem:

- GUI: Designing the front end or the user interface where the user will access the particular data using association rules.
- Parsing: Derives the Semantics of the Natural Query given by the user and parses it in its technical form.
- Query Generation: After the successful parsing of the statement given by the user, the system generates a query against the user statement in SQL and further gives it to the back end database.
- Data Collection: This module collects throughput of the SQL statement and places it in the User Interface Screen as a result form.

V. CONCLUSION

We partially proposed a process for protected mining of association guidelines in horizontally distributed databases that increases considerably after the existing leading protocol with regards to privacy and efficiency. Among the primary ingredients inside our proposed process is for processing the union (or intersection) of private subsets that every of the interacting players keep. Another element is a process that testing the adding of an aspect held by one person in a subset kept by another. Those protocols exploit the actual fact that the underlying issue is of interest only once the amount of players is higher than two.

REFERENCES

[1] International Journal of Research Studies in Science, Engineering and Technology Volume 1, Issue 9, December 2014, PP 10-13 ISSN 2349-4751 (Print) & ISSN 2349-476X (Online)

[2] IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.11, November 2006.

[3]International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 12 4091 – 4094

[4]Manali Rajeev Raut et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7540-7544.

[5] International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)Volume 3, Issue 3, May-June 2014 ISSN 2278-6856

[6] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc 20th Int’l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.

[7] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” Proc. ACM SIGMOD Conf., pp. 439-450, 2000.

[8] D. Beaver, S. Micali, and P. Rogaway, “The Round Complexity of Secure Protocols,” Proc. 22nd Ann. ACM Symp. Theory of Computing (STOC), pp. 503-513, 1990.

[9] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, “Efficient Mining of Association Rules in Distributed Databases,” IEEE Trans.Knowledge and Data Eng., vol. 8, no. 6, Dec. 1996.

[10] T. ElGamal, “A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms,” IEEE Trans. Information Theory, vol. IT-31, no. 4, July 1985.