ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Privacy preserving association rule mining in vertically partitioned data

1 T.Nandhini, 2 D. Vanathi, 3 Dr.P.Sengottuvelan

1M.E.Scholar, 2Associate Professor& 3Professor

Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052,

3Associate, Department Of Computer Science, Periyar University PG Extension Cente, Darmapuri

nandynithu912@gmail.com, vanathi.d@nandhaengg.org

Abstract: Privacy concerns typically constrain data processing technique. This paper addresses the matter of association rule mining wherever transactions square measure distributed across sources. Every web site holds some attributes of every dealings and also the sites would like to collaborate to globally valid association rules. However, the sites should not reveal individual dealings knowledge. This paper addresses the matter of association rule mining wherever transactions are distributed across sources. This paper have a tendency to specialize in privacy-preserving mining on vertically divided knowledge.This paper addresses a replacement model is planned to seek out association rules by satisfying the privacy constraints for vertically divided databases at n range of web sites in conjunction with data laborer. This model adopts cryptography techniques like encryption, decoding techniques and real number technique to seek out association rules expeditiously and firmly for vertically divided databases.

Keywords: Privacy Preserving Data Mining, Distributed Data Mining, Information Security, Association Rule Mining, Secure Multiparty Computation

1. INTRODUCTION

The main aim of data mining technology is to explore hidden information from massive databases. Several data mining techniques square measure exist like association rule mining, clustering, classification and wide applications within the world. In recent years, several organizations are showing interest to share the information with alternative parties to urge mutual edges however at a similar time no organization is willing to produce their non-public knowledge. To realize this, new space of analysis that's privacy conserving data mining has evolved. The most aim of privacy conserving data mining is to get the uncovered data from massive info whereas protective the sensitive data/information of people.

The issue of privacy arises in 2 things particularly centralized and distributed setting. In centralized setting, info is offered in single location and also the multiple users square measure allowed to access the info .The main aim of privacy conserving data {processing} during this scenario is to perform the mining process by activity sensitive data/information from users. In distributed setting, the info is offered across multiple sites and also the main aim of privacy conserving data mining during this setting is to seek out the worldwide mining

results by conserving the individual sites non-public data/information. Each web site will access the worldwide results that are helpful for analysis.

In recent years, several researchers are specializing in privacy conserving data mining in distributed setting because it has ton of applications in numerous fields. In distributed info setting, the info among totally different sites are often divided as horizontally, vertically and mixed mode. Several privacy conserving data mining algorithms are planned for various partitioning ways so as to seek out the worldwide mining results by satisfying the privacy constraints. In horizontally divided info, every {site|website|web web site} possess totally different set of tuples for a similar set of attributes wherever as within the case of vertically divided databases every site possess the common set of transactions for distinct set of attributes. In mixed partitioning methodology, knowledge is divided horizontally so every horizontally divided info is more divided into vertical and contrariwise.

Among several data mining techniques, association rule mining is receiving additional attention from the researchers to get the associations between item sets. Once several users have an interest to grasp the worldwide mining results while not revealing their non-public knowledge, the difficulty of privacy arises in distributed setting. The difficulty of privacy additionally arises even in centralized setting wherever sensitive data/information exist and that should be protected against the users. during this case sensitive rules are to be hidden. during this paper, privacy conserving association rule mining for n range of vertically divided databases at n web sites in conjunction with data mine wherever no site are often treated as trusty party is taken into account and is mentioned within the next section..

2. RELATED WORK

Data mining algorithms that partition the information into subsets are developed by [1]. particularly, add parallel data mining which will be relevant [2, 3]. though the goal of parallelizing data processing algorithms is performance, the communication value between nodes is a problem. Parallel data mining algorithms could function a place to begin for naive solutions. Algorithms are planned for distributed data mining. Cheung et al. planned a technique for horizontally divided data[4]. Distributed classification has additionally been addressed . A meta-learning approach has been developed that uses classifiers trained at individual to develop a worldwide classifier [5, 6]. This might defend the individual entities, however it remains to be shown that the individual classifiers don't unharness personal data. Recent work has addressed classification using Bayesian Networks in vertically partitioned data [7], and things wherever the distribution is itself fascinating with reference to what's learned [8]. However, none of this work addresses privacy issues.

There has been analysis considering what proportion data may be inferred, calculated or unconcealed from the information created accessible through data processing algorithms, and the way to reduce the outpouring of data [9]. However, this has been restricted to classification. The matter has been treated with an “all or nothing” approach. We tend to need quantification of the protection of the method. Firms might not need temperature data protocols (that leak no data at all) as long as They will keep the data shared among strict (though presumably adjustable) bounds. [10] Uses information perturbation techniques to safeguard individual privacy for classification (this is completed by adding random values from a normal/Gaussian distribution of mean zero to the particular information values).

Secure 2 party computation was 1st investigated by Yao [11] and was later generalized to multiparty computation. The seminal paper by Goldreich proves that there exists a secure answer for any functionality[12]. The approach used is as follows: the perform F to be computed is 1st described as a combinatorial circuit, then the parties run a brief protocol for each gate within the circuit. each participant gets corresponding shares of the input wires and therefore the output wires for each gate.

This approach, tho' appealing in its generality and ease, implies that the dimensions of the protocol depends on the dimensions of the circuit, that depends on the dimensions of the input. This can be extremely inefficient for giant inputs, as in data mining. though this shows that secure solutions exist, achieving economical secure solutions for distributed data mining remains open. [13] Examine many issues within the data mining domain within the Secure Multiparty Computation framework.

3.PROPOSED METHODOLOGY

The system model is comprised of 2 or a lot of information homeowners and a cloud. Every information owner contains a non-public information, and also the information homeowners cipher their non-public databases before outsourcing the encrypted databases to the cloud. Information homeowners may request the cloud to mine association rules or frequent itemsets from the joint information on their behalf. The (honest however curious) cloud is tasked with the aggregation and storing of informationbases received from completely different data homeowners, the mining of association rules or frequent itemsets for information homeowners, and also the causing of the mining result to relevant information homeowners. Fig.1 shows the design diagram.

We think about an equivalent heterogeneous information situation thought-about. Specifically, we tend to think about a vertically partitioning of the information, divided between 2 parties A and B. The subsequent may be a formal statement of the association rule mining drawback : Let I = be a group of literals, referred to as things. Let D be a group of transactions, wherever every transaction T may be a set of things specified T ⊆ I. Related to every transaction may be a distinctive symbol, referred to as its TID. We are saying that a transaction T contains X, a group of some things in I, if X ⊆ T.

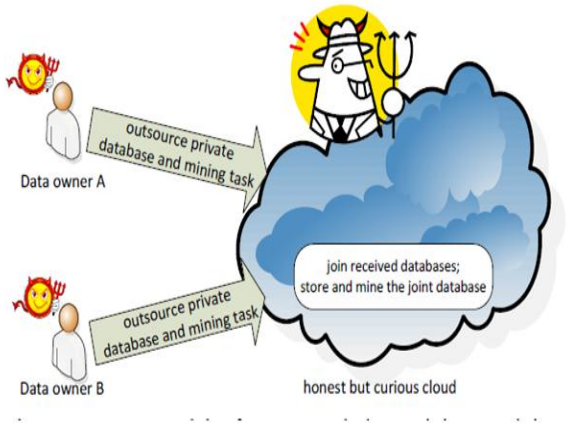

Fig.1. System Architecture Diagram

An association rule is associate degree implication of the shape, X =⇒ Y , wherever X ⊂ I, Y ⊂ I, and X ∩ Y = φ. The rule X =⇒ Y holds within the transaction set D confidently c if recall to mind transactions in D that contain X conjointly contain Y . The rule X =⇒ Y has support s within the transaction set D if that is so of transactions in D contain X ∪ Y . Inside this framework, we have a tendency to think about mining of mathematician association rules. The absence November twenty seven, 2001 DRAFT eight or presence of an attribute is delineated as a zero or one severally. So transactions seem like strings of zero and one and also the entire information may be delineated as a matrix of . to seek out out if a selected itemset is frequent, we have a tendency to simply have to be compelled to count the

amount of records within which the values for all the attributes within the itemset are one. This may translate into an easy mathematical drawback, given the subsequent definitions: Let the overall range of attributes be l + m, wherever A has l attributes, and B has the remaining m attributes. so A has the values for the attributes A1 through Al , and B has the values for the m attributes, B1 through Bm. k is that the support threshold needed. n is that the total range of transaction/records. Transactions/records are a sequence of l + m 1s or 0s. Let its represent the vectors of processed information values at every website, A and B severally. We have a tendency to describe the way to cipher the and vectors somewhat whereas later. The dot product (or dot product) of 2 vectors and of cardinality n is outlined as . = Now, if we would like to cipher whether or not a 2-itemset is frequent, wherever one among the attributes is understood to a short time the opposite is understood to B, the vector is precisely an equivalent because the attribute vector with A and also the vector is precisely an equivalent because the attribute vector with B. Computing $\sum_{i=1}^{n} x_i * y_i$ whether or not an itemset is frequent interprets to checking if the amount of transactions within which all attributes provide within the itemset are gift is bigger than the support threshold, k. Now, since we have a tendency to represent absence or presence of an attribute as zero and one, this interprets to computing wherever the xi and yi are the attribute values. Thus, finding the (in)frequency of the itemset interprets to conniving the scalar product of the two attributes. We have a tendency to provide an economical thanks to do that firmly once each of the parties possess one among the attributes and want to limit the data unconcealed within the section on the component algorithm.

The Component Algorithm

Secure computation of dot product is that the key to our protocol. Dot product protocols are projected within the Secure Multiparty Computation literature[10], but these scientific discipline solutions don't scale well to the present data processing drawback. We have a tendency to provide an algebraic answer that hides true values by inserting them in equations covert with random values. The data disclosed by these equations solely permits computation {of non-public|of personal} values if one facet learns a considerable range of the private values from an out of doors supply. (A completely different algebraic technique has recently been projected [14], but it needs a minimum of doubly the bitwise communication price of the tactic bestowed in [15].)

4.RESULTS AND DISCUSSION

In The component algorithm

In the projected model, every site’s information is represented in TID type that facilitates simple computations of native frequent item sets for its information by victimisation dot product technique. This TID type conjointly helps to seek out the dot product between the forerunner site´s computed results with its own leads to order to get all the frequent item sets for all possible mixtures of attributes associated with all the sites databases that area unit processed thus far (all forerunner sites and its own).

By adopting encryption, cryptography cryptography technique within the projected model, no successor website will predict its forerunner site’s data/information once it receives processed results from forerunner site. By adopting dot product technique within the projected model, each successor website will with efficiency determines the frequent item sets between its own frequent item sets and every one predecessors sites frequent item sets. The

dot product technique helps to explore all attainable combination of forerunner site's frequent item sets with successor site's frequent item sets.

This technique conjointly helps to work out that frequent item sets area unit by tally the amount of one's within the computed matrix and if the worth of count is bigger than or adequate to MinSup then the item set is said to be frequent for additional process. Though each site appends its computed results to the received results (consists of processed results of all predecessor sites) sent by its forerunner site to find globally frequent item sets, no site will predict any predecessor site's non-public data/information like attributes, native frequent item sets, support values as frequent item sets are in encrypted type within the received results.

DM cannot predict any site's non-public data/information even once DM has sure privileges like initiation of the mining method, cryptography of frequent item sets, finding international frequent item sets and their supports, generation of association rules. The DM receives processed results from sites that contains native frequent item sets of all attainable mixtures of attributes of all sites and connected supporting transactions. These transactions are obtained once completion of method the least bit sites and supported these data, DM cannot guess a person site's non-public data/information. The information transfers between sites and last site to miner is performed as a bulk information transfer rather than single data transfer for every frequent item set. Within the projected model, just one information transfer is needed for causing processed results from every forerunner website to its successor site. Thus solely n range of information transfers are required to get all sites processed leads to order to seek out the world frequent item sets.

As every website has distinct set of attributes for an equivalent set of transactions, the projected model with efficiency finds international frequent item sets by looking out all attainable mixtures of attributes of all sites. From the on top of discussion, the projected model is well, with efficiency and with minimum range of information transfers, finds the global association rules for vertically divided databases while not revealing any sites non-public data/information to any site and DM.

5.CONCLUSION

In this paper, a brand new model that utilizes the idea of inner product is projected to seek out international association rules once the information is divided vertically among n number of websites. within the projected model, DM has privileges to initiate the mining method, finding international association rules. Secured computations for association rules is achieved with this model by conserving the privacy of the individual sites data. The functioning of the projected model is illustrated with sample databases. With the projected model, association rules may be generated simply, with efficiency with minimum range of computations and communications by satisfying privacy constraints. The performance of this model is analyzed in terms of privacy and communications.

REFERENCES

[1]. Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In Proceedings of 21th International Conference on Very Large Data Bases, pages 432–444. VLDB, September 11-15 1995.

[2]. Mohammed J. Zaki. Parallel and distributed association mining: A survey. IEEE

Concurrency, special issue on Parallel Mechanisms for Data Mining, 7(4):14–25, December 1999.

[3]. H. Kargupta and P. Chan, editors. Advances in Distributed and Parallel Knowledge Discovery. AAAI/MIT Press, 2000

[4]. David Wai-Lok Cheung, Vincent Ng, Ada Wai-Chee Fu, and Yongjian Fu. Efficient mining of association rules in distributed databases. Transactions on Knowledge and Data Engineering, 8(6):911–922, December 1996.

[5]. Philip Chan. On the accuracy of meta-learning for scalable data mining. Journal of Intelligent Information Systems, 8:5–28, 1997.

[6]. Andreas Prodromidis, Philip Chan, and Salvatore Stolfo. Meta-learning in distributed data mining systems: Issues and approaches, chapter 3. AAAI/MIT Press, 2000.

[7]. Rong Chen, Krishnamoorthy Sivakumar, and Hillol Kargupta. Distributed web mining using bayesian networks from multiple data streams. In The 2001 IEEE International Conference on Data Mining. IEEE, November 29 - December 2 2001.

[8]. Rudiger ¨ Wirth, Michael Borth, and Jochen Hipp. When distribution is part of the semantics: A new problem class for distributed knowledge discovery. In Ubiquitous Data Mining for Mobile and Distributed Environments workshop associated with the Joint 12th European Conference on Machine Learning (ECML'01) and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany, September3-7 2001.

[9]. Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In Advances in Cryptology – CRYPTO 2000, pages 36–54. Springer-Verlag, August 20-24 2000.

[10].Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In Proceedings of the 1997 ACM SIGMOD Conference on Management of Data, Dallas, TX, May 14-19 2000. ACM.

[11].Andrew C. Yao. How to generate and exchange secrets. In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pages 162–167. IEEE, 1986

[12].O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game - a completeness theorem for protocols with honest majority. In 19th ACM Symposium on the Theory of Computing, pages 218–229, 1987

[13].Wenliang Du and Mikhail J. Atallah. Secure multi-party computation problems and their applications: A review and open problems. In Proceedings of the 2001 New Security Paradigms Workshop, Cloudcroft, New Mexico, September 11-13 2001.

[14].I. Ioannidis, A. Grama, and M. Atallah. A secure protocol for computing dot products in clustered and distributed environments. In The International Conference on Parallel Processing, Vancouver, Canada, Aug. 18-21 2002.

[15]. Vaidya, J., & Clifton, C. (2002, July). Privacy preserving association rule mining in vertically partitioned data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 639-644). ACM.