International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Special Issue 31(September, 2015), PP. 245-249

ASSOCIATION RULE MINING TO SECURE DATA IN DISTRIBUTED DATABASE

Sharda Darekar [1], Sandip Bankar[2], Prof.D.K.Chitre[3]
[1,2,3]Department Of Computer Engineering,Terna Engineering College,Nerul,Navi Mumbai.
[1]sharda.darekar@gmail.com, [2]sandip.bankar@gmail.com, [3]dkchitre1@rediffmail.com

Abstract— Data mining is that the most quick growing space these days that is employed to extract necessary data from massive knowledge collections however usually these collections area unit divided among many parties. With the fast development of knowledge mining analysis tools currently days it's penetration of knowledge mining and analysis at intervals completely different fields for disciplines, security and providing security at mining activities. The pattern mining in massive information provides introduction of recent and novel algorithms in data processing technology for providing secured pattern mining at outsourced or remote servers. The aim of paper is to produce security in data processing pattern analysis results and maintaining the principles that secures the personal data regarding organization or people corporations.

In older data processing analysis of patterns, quick Distributed Mining algorithmic program is employed. The quick Distributed Mining algorithmic program is invention of recent protocol with set of mining rules for secure mining of knowledge in horizontally distributed databases. However the protocol wasn't ready to secure distributed versions of information, as a result of quick distributed mining is extension of Apriori algorithmic program.

To beat the matter we tend to propose another technique known as secure computations with Multi party computations model on data. The goal of the projected system is to use mining and cryptography as joint technique for quicker and secure mining computations.

Keywords— ***Apriori Algorithm, Association Rule, Distributed Database, Fast Distributed Mining (FDM), secure mining.***

I. INTRODUCTION

Data mining is outlined because the methodology for extracting hidden predictive data from giant distributed databases. it's new technology that has emerged as a method of characteristic patterns and trends from giant quantities of information. the ultimate product of this method being the information, that means the numerous data provided by the unknown components [2]. This paper study the matter of association rules mining in horizontally distributed databases. within the distributed databases, there are many players that hold homogenized databases that share identical schema however hold data on totally different entities. The goal is to seek out all association rule with support s and confidence c to reduce the knowledge disclosed concerning the non-public databases command by those players [1].

Kantarcioglu and clifton studied the matter wherever additional appropriate security definitions that enable parties to settle on their desired level of security are required, to permit effective solutions that maintain the required security [2]. so that they devised a protocol for its resolution. the most a part of that protocol is sub protocol for secure computation of the union of personal subsets that ar command by the various players. It makes the protocol pricey and its implementation depends upon coding primitive's ways, oblivious transfer and hash operate additionally the escape of knowledge renders the protocol not absolutely secure [1].

This paper projected Associate in Nursing algorithmic rule, PPFDM, privacy protective quick distributed mining algorithmic rule for horizontally distributed knowledge sets and notice attention-grabbing association or correlation relationships among an outsized set of information things and to include cryptographically techniques to reduce the knowledge that goes to shared with others, whereas adding very little overhead to the mining task [1]. Within the projected theme, the inputs are the partial information and therefore the needed output is that the list of association rules that hold within the unified database with support and confidence no smaller than the given thresholds s and c, severally. the knowledge that may prefer to shield during this paper isn't solely individual dealing within the totally different databases, however additionally additional world or public data like what association rules are supported regionally in every of these databases. The projected protocol improves upon that in [2] in terms of simplicity and potency furthermore as privacy.

II. RELATED WORK

Data mining with security has been a crucial analysis space for last a few years. There square measure several things wherever data processing isn't done by the info house owners. Information house owners would possibly source the info mining task to another company. During this case, it's essential to expect secure data processing.

Anonymization is one amongst the techniques explored in [1] and [2] for secure data processing. Cryptographical measures are around for securing operations. ID3 [3] is employed for secure generation of call trees as a part of information discovery. Expectation Maximization [4] was conjointly employed by researchers to mine information from horizontally distributed information bases.

Association rule mining is one amongst the foremost helpful data processing techniques on the market as explored in [5], [6] and [7]. There are square measure instances

wherever associate rule mining is administrated in distributed surroundings. In [8] and [9] experiments square measure created with horizontally and horizontally distributed databases. Secure multi-party communications is one amongst the techniques for securing communications among multiple parties. It is accustomed have privacy protective distributed data processing. In [10], [11] and [12], this type of analysis was administrated for secure cooperative data processing. The thought of polynomials and privacy protective protocol were utilized in [10] and [11] severally. a form of secret writing called independent secret writing was utilized in [8]. In [12] a similar is administrated with less communication price. Several researchers experimented with 2 players for secure and distributed data processing as explored in [13]. Recently in [14] polynomial analysis is employed for addressing set inclusion drawback.

III. PROPOSED SYSTEM

The system includes a novel various protocol for providing secure computation with personal subsets in distributed information and it improves the potency and security of information mining. The planned and designed protocol provides full computing, parameterized computing and customized user computing, that we have a tendency to decision user threshold functions, within which the 2 extreme cases correspond to the issues of computing the union and intersection of personal subsets. This mechanism provides associate degree extension to Apriori rule with quick distributed and Secure Multiparty computations.

The planned system uses 2 algorithms, specifically Apriori and S-FDM for locating frequent item sets from horizontally distributed databases. Association Rules square measure generated from the frequent things sets and classified whose confidence is larger than the minimum threshold confidence referred to as sturdy Association Rules. The sturdy associations rules square measure classified during this manner square measure presented the user.

While extracting data from distributed database system more number of irrelevant data will occur. Irrelevant data is avoided by using the Apriori algorithm. Data leakage is more in Apriori algorithm. Encryption is done at the time of retrieving data from the database.

A. Advantages

- As a rising subject, data processing is taking part in associate degree more and more vital role within the call support activity of each walk of life.
- Get economical item set result supported the client request.
- Provides associate degree increased cryptography theme that allows protractile formal privacy guarantees in large-scale and real-life dealing information.

B. System Design

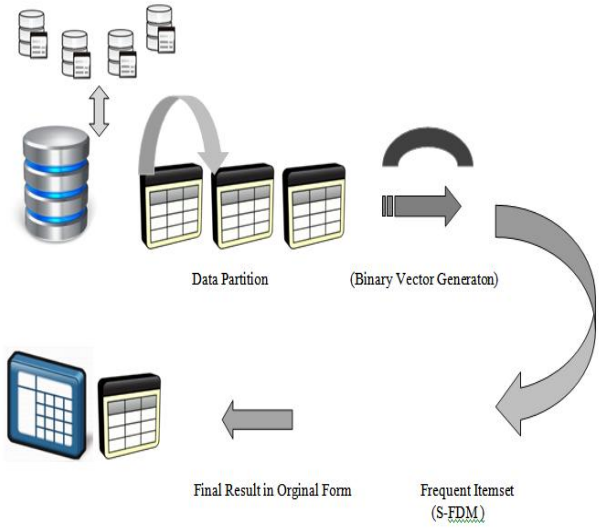

Figure 1: System Flow Diagram

The projected System flow is shown in figure1. Association Rule mining is one among the foremost necessary data processing tools employed in several world applications [2]. This paper, presents the matter of computing association rules at intervals a horizontally partitioned off info. we have a tendency to assume consistent databases. To mine the association rules the primary task is to come up with the frequent item sets. Second task is to mine the association rules from the frequent item sets.

1. Generation of Frequent Itemsets

Frequent item sets from completely different information return to a world database. Since there square measure such a large amount of information bases through that frequent data goes to the worldwide info thus this will increase the quantity of messages that require to be passed thus on notice frequent k item set. the main downside with frequent set mining strategies is that the explosion of the quantity of results then it's tough to seek out the foremost attention-grabbing frequent item sets. therefore the idea of finding frequent itemsets from the info that is at completely different Distributions, and mine the association rules has been highlighted during this paper.

2. Mining associations Rules

Following the first definition by Agrawal the matter of association rule mining is outlined as[5] : Let I = be a group of n binary attributes referred to as things. Let D = be a group of transactions referred to as the info. every dealing in D contains a distinctive dealing ID and contains a set of the things in I. A association rule is outlined as Associate in Nursing implication of the shape X->Y wherever X, Y ⊆ I and X ∩ Yb = ∅. The sets of things (for short itemsets) X and Y square measure referred to as antecedent (left-hand-side or

LHS) and resulting (right-hand-side or RHS) of the rule severally. Associate in Nursing example of Associate in Nursing association rule would be "If a client buys a bread, he's eightieth possible to conjointly purchase milk."

Given a minimum confidence threshold minconf and a minimum support threshold minsup, the matter is to come up with all association rules that have support and confidence bigger than the user-specified minimum support and minimum confidence. within the initial pass, the support of every individual item is counted, and therefore the giant ones square measure determined. In every sequent pass, the big itemsets determined within the previous pass is employed to come up with new itemsets referred to as candidate itemsets. The support of every candidate itemset is counted, and therefore the giant ones square measure determined. This method continues till no new giant itemsets square measure found.

The overall method of Association Rule Mining consists of following modules:

i. User Module: -

In this module, privacy protective data processing has thought of 2 connected settings. One, within which {the information} owner and therefore the data mineworker square measure 2 completely different entities, and another, within which the info is distributed among many parties World Health Organization aim to conjointly perform data processing on the unified corpus of knowledge that they hold.

In the first setting**,** the goal is to protect the data records from the data miner. Hence, the data owner aims at anonymizing the data prior to its release. The main approach in this context is to apply data perturbation. He perturbed data can be used to infer general trends in the data, without revealing original record information.

In the second setting**,** the goal is to perform data mining while protecting the data records of each of the data owners from the other data owners.

ii. Admin Module:

This module is used to view user details. Admin to view the item set based on the user processing details using association role with Apriori algorithm.

iii. Association Rule:-

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

Association rules are created by analyzing data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

iv. Apriori Algorithm :-

Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.

The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Number of transaction is present in each set of data. Initial scan/pass of algorithm counts occurrence of each item in order to determine the frequent items set. Next scan K consists two phases.

1) In first phase, Candidate item set C_K is generated using frequent item set L_k-1 found in $(K-1)_{th}$ pass. This is candidate generation process in Apriori Algorithm.

2) In second phase database is scanned to find support for Candidates C_K. In next step, it prunes the candidates which have an infrequent sub pattern and keep only subset of candidate sets which are already identified as frequent items sets. Output of Apriori algorithm generates sets of rules which determine how often items are brought together in single set.

v. Fast Distributed Mining(FDM) :

Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the Apriori algorithm. Its main idea is that any s-frequent itemset must be also locally s-frequent in at least one of the sites. Hence, in order to find all globally - frequent itemsets, each player reveals his locally s-frequent itemsets and then the players check each of them to see if they are s-frequent also globally.

The FDM algorithm proceeds as follows:

- Initialization
- Candidate Sets Generation
- Local Pruning
- Unifying the candidate item sets
- Computing local supports
- Broadcast Mining Results

vi. S-FDM:

In this paper we have a tendency to discuss concerning the protection of knowledge whereas mining. victimization Advanced cryptography customary (AES) and DES (Data cryptography Standard) technique, information are encrypted and decrypted whereas inserting and retrieving the info. Advanced cryptography customary (AES) takes less quantity of your time to encode and rewrite the info.

The S-FDM algorithm proceeds as follows:

- Cryptographic Primitive Selection

- All item sets Encryption
- Item set Merging
- Decryption
- View final Result

IV. RESULTS

There are several sites that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities. The system is designed to find all association rules with support at least *s* and confidence at least *c*, for some given minimal support size *s* and confidence level *c*, that hold in the unified database, while minimizing the information disclosed about the private databases held by those sites. The protected context is not only individual transactions in the different databases, but also more global information such as what association rules are supported locally in each of those database.

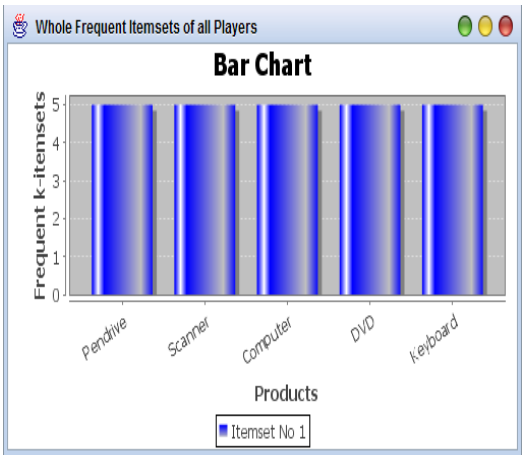

Figure 2: Frequent Item Sets

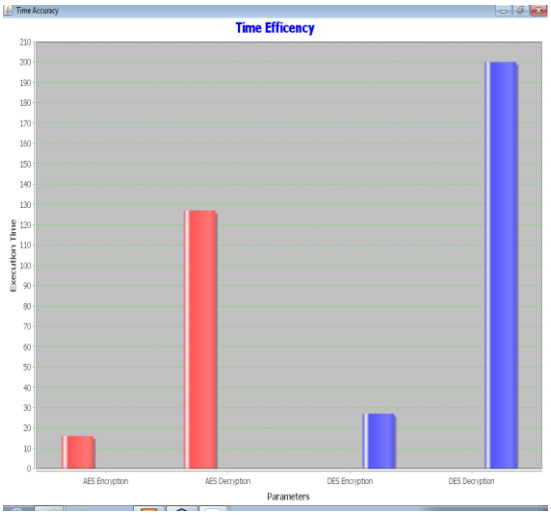

Figure 3: Time Efficiency Graph

Figure2 shows the frequent item sets & figure3 shows the analysis of time efficiency for cryptographic methods (i.e. AES and DES). AES(Advanced Encryption Standard) is a symmetric-key block cipher and DES(Data Encryption Standard) is a symmetric key block cipher. Figure2 shows the time efficiency graph which gives time required for encryption and decryption process. AES takes less amount of time to encrypt and decrypt the data than DES.

V. CONCLUSION

Mining association rules is one among the information mining techniques that are terribly helpful for creating well sophisticated choices. This work is distributed on horizontally distributed info in secure setting. Support and confidence are the applied math measures used for mining association rules. Thus, the applied math measures will be wont to shrewdness the foundations are helpful. The lot of in support and confidence, the lot of in utility of the foundations. Frequent item sets are generated through apriori and remainder of the mechanisms are distributed by the planned algorithmic program. Knowledge accesses from distributed information are a lot of secured and extremely economical once this mechanism is applied.

REFERENCES

[1] Tamir Tassa,"Secure mining of association rule in horizontally distributed databases" ,IEEE trans. Konwledege and Data Engg.,Vol.26, no.2, April 2014.

[2] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.

[3] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *KDD*, pages 217–228, 2002.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD Conference*, pages 439–450, 2000.

[5] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Crypto*, pages 36–54, 2000.

[6] X. Lin, C. Clifton, and M.Y. Zhu. Privacy-preserving clustering, with distributed EM mixture modeling. *Knowl. Inf. Syst.*, 8:68–81,2005.

[7] J. Zhan, S. Matwin, and L. Chang. Privacy preserving *Security*, collaborative association rule mining. In *Data and Applications* pages 153– 165, 2005.

[8] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *KDD*, pages 639–644, 2002.

[9] M. Kantarcioglu, R. Nix, and J. Vaidya. An efficient approximate protocol for privacy-preserving association rule mining. In *PAKDD*,pages 515–524, 2009.

[10] A. Schuster, R. Wolff, and B. Gilburd. Privacy-preserving association rule mining in large-scale distributed systems. In *CCGRID*, pages 411– 418, 2004.

[11] L. Kissner and D.X. Song. Privacy-preserving set operations. In *CRYPTO*, pages 241–257, 2005.

[12] M.J. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *EUROCRYPT*, pages 1–19, 2004.

[13] J. Brickell and V. Shmatikov. Privacy-preserving graph 252,

algorithms in the semi-honest model. In *ASIACRYPT*, pages 236–2005.

[14] M. Freedman, Y. Ishai, B. Pinkas, and O. Reingold. Keyword search and oblivious pseudorandom functions. In *TCC*, pages 303–324, 2005.