Secure and Distributed Approach for Mining Association Rules

G. Chandana
M. Tech (C.S.E),Student,
S.K.D Engineering College,
Gooty, Ananthapuramu District.
chandanagowd@gmail.com

C. Vasu Murthy
M.C.A, M.Tech, Assistant Professor
S.K.D Engineering College
Gooty, Ananthapuramu District.
vasumurthy.chintala@gmail.com

S. G. Nawaz
M.Tech, Associate Professor
S.K.D Engineering College
Gooty, Ananthapuramu District.
sngnawaz@gmail.com

Abstract--Data mining is the process of extracting trends from data sources. Domain exerts can make use of the trends to derive business intelligence. Big organizations store data in multiple server and often data is horizontally distributed. Mining such database provides useful and actionable knowledge which can help in making well informed decisions. However, secure mining of extracting association rules can provide interesting information that can help enterprises to make expert decisions. In this paper, we propose an algorithm and have a secure mechanism in order to mine association rules for deriving knowledge. We also incorporated auditing of data in the proposed system. We built a prototype application that demonstrates the secure mining of association rules with support and confidence. The statistical measures such as support and confidence help in knowing the usefulness of the rules. The empirical results are encouraging.

Index Terms *– Data mining, distributed databases, association rule mining, security*

*****

I.INTRODUCTION

Data mining has plethora of algorithms or techniques that can be used to mine useful knowledge from databases for growth of organizations. However, due to the rapid growth of data and the business expansion, organizations are maintaining data in multiple servers and data is horizontally distributed among the servers. Mining such data as a whole can provide useful knowledge that can be used to make intelligent decisions. Data mining has become indispensable for enterprises as manual analysis of data is not feasible. Businesses of all fields can grow faster by utilizing data mining domain knowledge. Data mining can bring about trends or patterns that are useful. Data mining has become de facto standard for organizations to obtain trends that can provide customer behavior for better decision making. Of late many companies are involving data mining making it as collaborative data mining. In the same fashion a single company can have data stored in distributed fashion. In this paper we considered the horizontally distributed data bases for mining association rules which provide actionable knowledge.

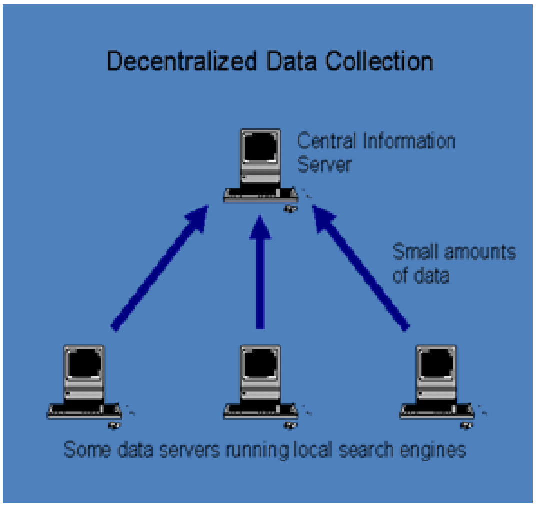

Decentralized Data Collection
Central Information Server
Small amounts of data
Some data servers running local search engines

Figure 1 – Data is distributed across many servers

As shown in Figure 1, it is visible that data is stored in multiple locations. In other words, data is collected from multiple places and mining such data as a whole can provide more useful and accurate business intelligence. When data is stored in multiple servers, it is required to have mechanism to mine the whole data so as to get comprehensive knowledge.

One of the important mechanisms of data mining is known as association rule mining. Some item sets are frequently stored and mining such frequent item sets and brings about their association with support and confidence measures will result in association rules. Apriori is one of the well known algorithms for mining frequent item sets. The general flow of this algorithm is as shown below.

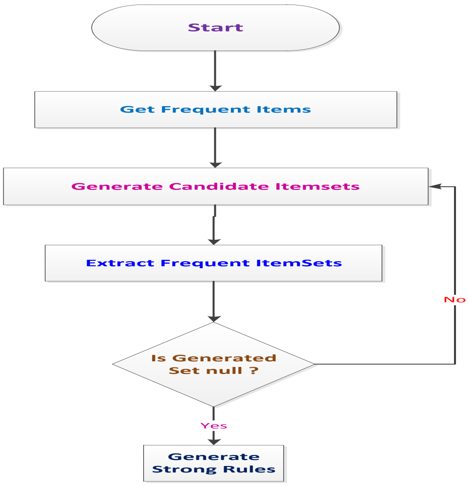
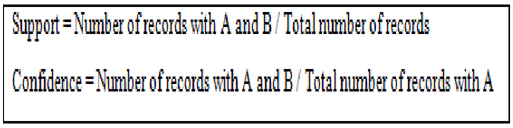


Figure 2 – Overview of association rule mining using Apriori

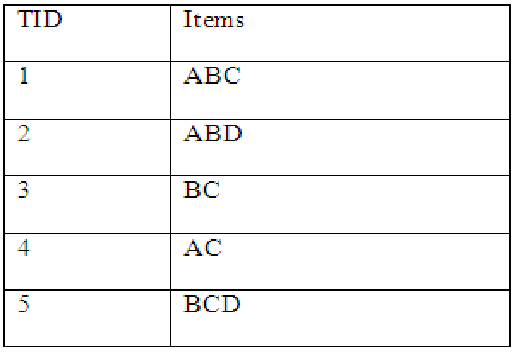
The flow presented in Figure 2, shows that Apriori follows an iterative process that generates strong association rules. These rules can help domain experts to derive business intelligence. Support and confidence are the statistical measures that can be used to know the usefulness of the rules generated. The support and confidence are computed as follows.



Support = Number of records with A and B / Total number of records

Confidence = Number of records with A and B / Total number of records with A

Consider two item sets namely A and B. Their frequency in data source provides details about their frequency and association from which support and confidence can be computed. These measures can be used further to gain actionable knowledge.



TID	Items
1	ABC
2	ABD
3	BC
4	AC
5	BCD

Figure 2 – Sample dataset

According to the computations of support and confidence the transactions present in Figure 2 can be analyzed. The item sets are A, B, C and D. Their frequency is different. AB item set has 40% support, ABC 20% and BC 60% support. With respect to confidence 66%, 50% and 75% are the values computed respectively. These statistics can help in quantitative data analysis to make expert decisions. These figures are useful to domain experts in making well informed decisions that lead to organizational growth.

In this paper we proposed an algorithm for mining association rules from horizontally distributed databases in secure fashion. We also build a prototype application that demonstrates the proof of concept. The remainder of the paper is structured as follows. Section II provides review of literature pertaining to secure association rule mining. Section III presents proposed architecture and also the algorithm. Section IV presents experimental results while section V concludes the paper.

II.RELATED WORKS

Data mining with security has been an important research area for last many years. There are many situations where data mining is not done by the data owners. Data owners might outsource the data mining task to some other company. In this case, it is essential to expect secure data mining. Anonymization is one of the techniques explored in [1] and [2] for secure data mining. Other useful ideas include perturbing data and involvement of multiple parties in the data mining process. Cryptographic measures have been around for securing operations. ID3 [3] is used for secure generation of decision trees as part of knowledge discovery. Expectation Maximization [4] was also used by researchers to mine knowledge from horizontally distributed data bases. Association rule mining is one of the most useful data mining techniques available as explored in [5], [6] and [7]. There are instances where associate rule mining is carried out in distributed environment. In [8] and [9] experiments are made with horizontally and vertically distributed databases.

Secure multi-party communications is one of the techniques for securing communications among multiple parties. It can be used to have privacy preserving distributed data mining. In [10], [11] and [12], this kind of research was carried out for secure collaborative data mining. The concept of polynomials and privacy preserving protocol were used in [10] and [11] respectively. A kind of encryption known as commutative encryption was used in [8]. In [12] the same is carried out with less communication cost. Many researchers experimented with two players for secure and distributed dada mining as explored in [13]. Recently in [14] polynomial evaluation is used for addressing set inclusion problem.

III. Proposed Architecture and Algorithm For Secure Arm

Architecture is proposed in this paper for secure association rule mining on horizontally distributed databases. The overview of the architecture is as shown in Figure 3. Many sites are found in the architecture in distributed fashion. From each site association rules are mined in sure fashion. Then all the rules collected from all the sites are merged together. This has paved the way for obtaining association rules with support and confidence to form business intelligence. The cryptographic mechanisms are also used in order to perform operations in secure environment.

Architectural Overview

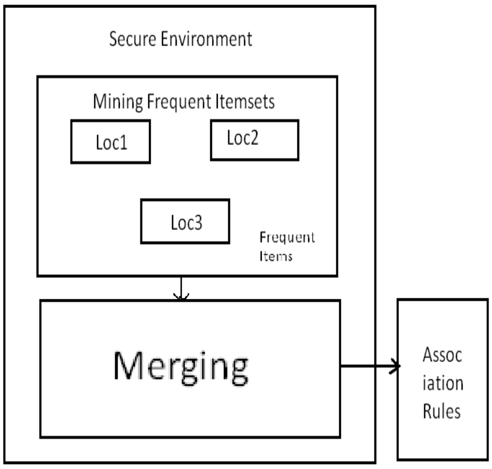


Figure 3 – Overview of the architecture

The architecture shows merging process that is used to obtain frequent item sets from all sites and then combine them so as to form association rules for the whole database that is distributed horizontally. This kind of architecture when implemented can become a tool that can mine useful association rules that can be used further for deriving knowledge. Later on this architecture is further extended to provide built in auditing mechanism for data consistency.

IV. Proposed Algorithm

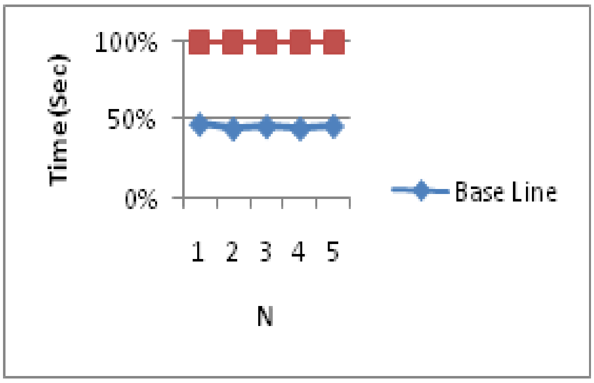
In order to realize the proposed architecture as shown in Figure 3, we proposed an algorithm that can help in achieving the intended purpose of the architecture. The aim of the algorithm is to mine association rules from horizontally distributed databases.

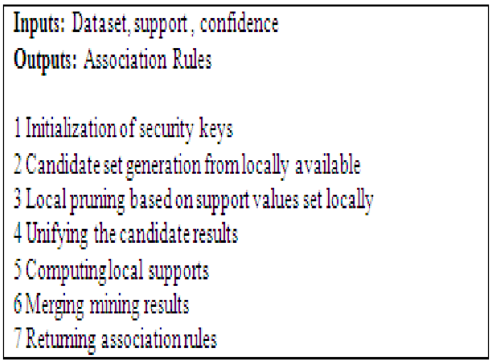
Listing 1 – Outline of the proposed algorithm

As can be viewed in listing 1, it is evident that the algorithm takes dataset, support and confidence as input and returns association rules. Initialization of security keys, candidate set generation, local pruning, unifying candidate results, computing local supports, merging mining results and retuning association rules are the important steps carried out.

V. EXPERIMENTAL RESULTS

We built a prototype application that demonstrates the proof of concept. The application has been built in Visual Studio 2012 which is the IDE for .NET environment. The experiments are made in a PC with 2GB RAM, Dual core processing running Windows 7 operating system. The experiments are made in terms of total time taken for mining task.





```
Inputs: Dataset, support , confidence
Outputs: Association Rules

1 Initialization of security keys
2 Candidate set generation from locally available
3 Local pruning based on support values set locally
4 Unifying the candidate results
5 Computing local supports
6 Merging mining results
7 Returning association rules
```

Figure 4 – Computational time comparison

As seen in Figure 4, it is evident that the algorithm has been tested with number of rows and columns represented by N which is plotted in horizontal axis while the time taken in seconds is represented in vertical axis. The proposed Secure ARM algorithm is compared with a baseline algorithm. The performance of the proposed algorithm is far better than the baseline algorithm.

VI. CONCLUSION

Mining association rules is one of the data mining techniques which are very useful for making well informed decisions. In this paper we study secure mining of association rules. Our work is carried out on horizontally distributed database in secure environment. Support and confidence are the statistical measures used for mining association rules. Thus the statistical measures can be used to know how the rules are useful. The more in support and confidence, the more in usefulness of the rules. We used Apriori algorithm along with our algorithm in order to achieve this. Frequent item sets are generated through Apriori and rest of the mechanisms are carried out by the proposed algorithm. We built a prototype application that demonstrates the proof of concept. Later on this architecture is further extended to provide built in auditing mechanism for data consistency. The empirical results are encouraging. In future we improve the prototype and make it a useful tool for mining business intelligence using other data mining algorithms that can be employed to various domains. Thus the tool become useful for acquiring business intelligence for making well informed decisions.

REFERENCES

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

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

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

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

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

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

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

[8] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, 16:1026–1037, 2004.

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

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

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

[12] J. Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model. In *ASIACRYPT*, pages 236–252, 2005.

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

[14] T. Tassa, A. Jarrous, and J. Ben-Ya'akov. Oblivious evaluation of multivariate polynomials. *Submitted*.