Privacy Preserving Association Rule Mining Using Horizontally Partition Data

Arpita B. Modh[1]

[1]Department of Computer Science and Engineering, L.J. Institute of Technology, Ahmedabad -382210.Gujarat, India

ABSTRACT

In this paper we propose algorithm to mine association rule using modified RSA Cryptography technique over horizontally partition data. Here we consider unsecured distributed environment. Our proposed algorithm provides privacy and security against involving parties and intruder.

Keywords - Privacy preserving association rule mining, Modified RSA Cryptography**.**

1. INTRODUCTION

During the data mining process use a sensitive or personal data. It may be of mutual benefit for two parties or multiple parties to share their data for an analysis task. However, they would like to ensure their own data remains private. Means, there is a need to protect sensitive knowledge during a data mining process. This problem is called Privacy-Preserving Data Mining (PPDM). So maintaining privacy is challenging issue in data mining. An association rule is a rule shows certain relationship among set of attribute in database. Finding association rule may useful for finding interesting pattern in marketing, statistical analysis, medical diagnosis etc.. Mining association rule require s iterative scanning of database which is quite costly in processing. These techniques can be demonstrated in centralize [1,2] as well as distributed environment [3, 4] where data can be distributed among the different sites. Distributed database scenario can be classified in horizontally partitioned data and vertically partitioned data. In horizontally partitioned data each site contained same set of attributes with different number of transaction. In vertically partitioned data each site contains different number of attributes with same number of transaction. Figure 1 shows different between horizontally partition database and vertically partitioned database

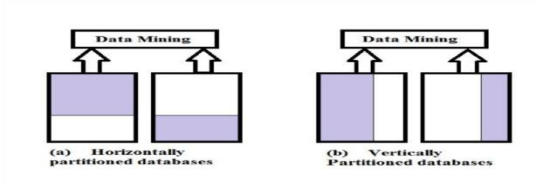

Fig-1 a) Horizontally Partiton Data b) Vertically Partition Data

1.1 Motivation

In commercial application organizations are very much concerned with privacy issue. Most organization collect their information from individual party ad their specific need. They find it essential to share information to each other. In such cases each unit want to be sure that privacy of individual party must not violated.

If Government wants to survey about some medical disease. Individual hospital don't want to revel personal informationof it's patient, It is against law. This information is regarded private and we want to avoid exposing confidential information of patient

In this paper we propose cryptography approach to mine association rules. We have used modified RAS algorithm. This provides privacy against involving parties

The remaining paper is organized as follows. In section 2, theoretical background and related work are discussed. In section 3, proposed method for finding association rule. The

analysis of proposed method and conclusion are in presented in section 4 and 5.

2. RELETED WORK

2.1 Data Mining of Association Rule in Distributed Environment

In a distributed environment Let *DB* be a database with *D* transactions. Assume that there are *n* sites, *S*1, *S*2, ・・*Sn*, in a distributed system, and the database *DB* is partitioned over the *n* sites into *{DB*1,*DB*2, ・・・*DBn}*, respectively.

Let the size of the partitions *DBi* be *Di*, for *i* = 1, ・・ *n*. Let *X.sup* be the global support count, and *X.supi be* the local Support count of *X* at site *Si*. For a given minimum support Threshold *s*, *X* is globally frequent if *X.sup* ≥ *s* × *D*; correspondingly, *X* is locally frequent at site *Si*, if *X.supi* ≥ *s*×*Di*. In the following, *L* denotes the globally frequent itemsets in *DB*, and *L*(*k*) the globally frequent *k*-itemsets in *L*. The aim of mining association rules in distributed database is to find all rules whose global support and global confidence are higher then the user specified minimum support and confidence.[5]

$$(X \rightarrow Y).\sup = \frac{X.\sup}{|DB|} = \frac{\sum_{i=1}^{n} X.\sup_i}{\sum_{i=1}^{n} |DB_i|} \geq \sup\%$$

$$(X \rightarrow Y).conf = \frac{\{X \cup Y\}.\sup}{X.\sup} = \frac{\sum_{i=1}^{n} XY.\sup_i}{\sum_{i=1}^{n} X.\sup_i} \geq conf\%$$

2.2 RSA Algorithm

The key feature of public-key cryptosystem is that the encryption and decryption procedure are done with two different keys - public key and private key, and the private key cannot be derived from the public key, that enables the publication of the encryption key without the risk of leaking the secrets. The most significant approach of public key cryptography algorithm is RSA, which can resist almost all the known passwords attacks so far. RSA algorithm, which is named after the inventors, is the first algorithm that can be used both for data encryption and digital signatures [6]. RSA algorithm security depends on the difficulty of decomposition of large numbers. In the algorithm, two large prime numbers are used for constructing the public key and the private-key. It is estimated that the difficulty of guessing the plaintext from signal key and the cipher text equals to that Decomposition of the product of two large prime numbers.

Key Generation:

- Select p and q both prime number, p is not equal to q.
- Calculate n = p x q.
- Calculate ø (n) = (p -1) x (q-1).
- Select integer e whose gcd (Ø (n), e) = 1; 1 < e <Ø (n).
- Calculate private key d = e-1 (mod Ø (n)).
- Public key PU = {e, n}.
- Private Key PR = {d, n}

Encryption:

- $C = M^e \bmod n$.

Decryption:

- $M = C^d \bmod n$.

Where, C - Cipher text

M - Message

p and q - Prime Numbers,

N - Common Modulus,

e and d - Public and Private Keys

2.3 Disadvantage of RSA Algorithm

- Loss Of Private Key May Break The Security.
- Attacks on RSA can break security fore. factorization
- Problem, low decryption exponent, common modulus, short message, cyclic attack etc.
- High Computational Cost.

- As n is transmitted in public key, thus its factors can be found out by hit and trial, due to which the security quotient of RSA algorithm gets reduced.

3. PROPOSED METHOD

3.1 Problem definition

RSA algorithm security can be compromised over the network. To increase security of computation of RSA algorithm we need to modify the RSA algorithm which can be done by third prime number and using a new variable for Encryption and Decryption.

3.2 Flow chart of proposed method.

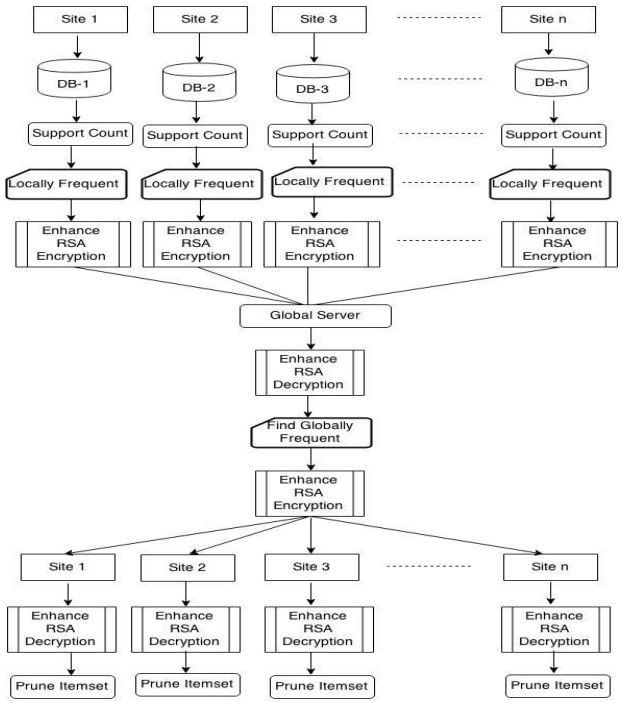

Fig-2 Proposed Diagram

1. Websites stored on web server have database that resides at the database server and connected with webserver. Every website have there on database.

2. Database have different attribute (column). If there are site-n then there are database-n.

3. Using Association Rule Mining technique we can find the support count for the different attributes different values.

4. According to the minimum support criteria, itemset which are below minimum support will be prune and above minimum support itemset are called the locally frequent itemset. This procedure will be followed in all the n participating database.

5. Locally frequent itemset will be send to global server to check that this item set will be frequent or not, but before sending the data to global server RSA encryption algorithm will be applied for the privacy preserving.

6. Global server will decrypt all the itemsets using RSA Decryption algorithm w

7. hich have been received from different web sites and combine all together.

8. Association Rule mining has been applied on combine itemsets to find global support.

9. Combine Global support divide to all the sites according to their itemset and RSA Encryption Algorithm apply on it.

10. Itemset with global support will be send to all sites.

11. Every site will apply RSA Decryption algorithm for decrypt their itemset.

12. Item set will be prune using the pruning algorithm.

3.3 Modified RSA Algorithm

Key Generation	
Select three number p, q, r	p, q, r are prime number
Calculate $n = p \times q \times r$	
Calculate $\phi(n) = (p-1)(q-1)(r-1)$	
Select integer e such that co-prime to n	gcd $(\phi(n), e) = 1$; $1 < e < \phi(n)$

	e is public key exponent
Compute d: If p > q Then n-p < d < n If p < q Then n-q < d < n General Formula to compute d: (e x k) mod $\phi(n)$ =1 k is found by the formula: k x e = 1 x mod (d)	
Public key	$PU = \{ e, d \}$
Private key	$PR = \{ k, d \}$
Encryption:	$c = m^e \bmod (d)$
Decryption	$m = [c^k \bmod (d)]$

3.4 Advantages of Proposed Method

- The strength of large prime number depend on three variables p, q and r. It is difficult to break the large prime number into three.
- Eliminate the use of common modulus n by generating a new variable from value of n and the prime numbers.
- Using new variable for encryption and decryption gives more security for data transfer.

4. RESULT ANALYSIS

TABLE-4.1

File Name	Plain text(in Kilobyte)	RSA Encryption time(in Second)	Modified RSA Encryption Time
Dataset1	66	0.0239963	0.0160066
Dataset2	75	0.0160049	0.0160058
Dataset3	91	0.0159980	0.0120068

Table 4.1 Comparison of Encryption Time: RSA and Modified RSA

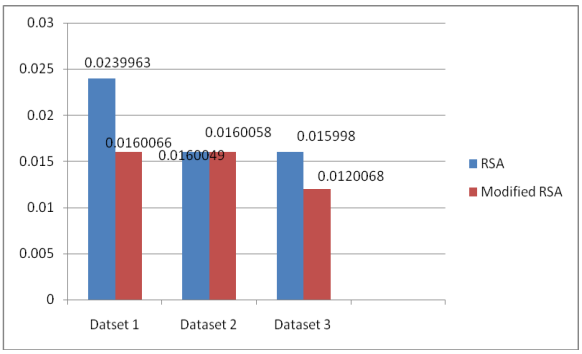

Fig-3 Encryption Time of RSA & Modified RSA

TABLE-4.2

Existing RSA system	**Proposed RSA system**
Two prime numbers are selected to generate Common modulus n.	Three prime are selected to generate the common modulus n.
Common modulus n is used for Encryption and Decryption.	A new variable d is generated used for Encryption and decryption.
Encryption and Decryption Required more time	Encryption and decryption Require less time
Provide Less Security	Provide More Security Compare to RSA.
Process Speed is Fast	Process Speed is Slow

Table 4.2 Comparison of Existing and proposed method

5. CONCLUSION

In Privacy Preserving Association rule mining over horizontally partition database to find a global association rule from the local frequent itemset using Modified RSA Cryptography technique. This Method Provides Decryption with no information loss. So this approach can be used for transmission of private data efficiency and securely through

unsecured channel without loss of any information. It's providing good security and less computational time.

REFERENCES

[1] Elena Dasseni, Vassilion S. Verykios, Ahmed K. Elmagarmid and Elisa Bertino, "Hiding Association Rules by using Confidence and support," In Proceedings of the 4th information Hiding workshop,pp 369-383,2001.

[2] Vassilios S. Verykios Ahmed K. Elmagarmid, Bertino Elisa yucel saygin and Dasseni Elena, "Association Rule hiding", IEEE Transactions on knowledge and Data Engineering, 2003.

[3] Vaidya J. & Clifton, C.W, "Privacy preserving association rule mining in vertically partitioned data," In Proceedings of the eight ACM SIGKDD international conference on knowledge discovery and data mining Edmonton Canada 2002

[4] Marut Kantarcioglou and charis Clifton, " Privacy-Preserving distributed mining of association rules on horizontally partitioned data," In proceedings of the ACM SIGMOD workshop on Reserch Isuues in Data Mining and knowledge discovery, PP 24-31,2002.

[5] Liu j., Piao X.,& Huang, S, " A privacy-Perserving mining algorithm of association rules in distributed database" proceeding of the international mutti-symposium on computer and computational Science, pp 740-750, 2006.

[6] XinZhou and Xiaofei Tang, "Research and Implementation of RSA Algorithm for Encryption and Decryption",The 6th International Forum on Strategic Technology ISSN: 978-1-4577-0399-7/11, pp. 1118-1121, August 2011.

