# Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach

*Lalanthika Vasudevan , S.E. Deepa Sukanya, N. Aarthi\**

*Abstract*—**In our era, Knowledge is not "just" information anymore, it is an asset. Data mining is thus extensively used for knowledge discovery from large databases. The problem with data mining is that with the availability of non-sensitive information, one is able to infer sensitive information that is not to be disclosed. Thus privacy is becoming an increasingly important issue in many data mining applications. This has led to the development of privacy preserving data mining. Two main approaches to privacy preserving data mining have emerged in recent years. The first approach protects the privacy of the data by using an extended role based access control approach where sensitive objects identification is used to protect an individual's privacy. The second approach uses cryptographic techniques. We propose a new solution by integrating the advantages of both these techniques with the view of minimizing information loss and privacy loss. By making use of cryptographic techniques to store sensitive data and providing access to the stored data based on an individual's role, we ensure that the data is safe from privacy breaches.**

*Index Terms*— Cryptographic technique,Extended Role based access control,Privacy preserving data mining.

## I. INTRODUCTION

Explosive progress in networking, storage, and processor technology has led to the creation of ultra large databases that record unprecedented amount of transactional information. The main problem is that with the availability of non-sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Despite its benefits in various areas such as marketing, business, medical analysis, bioinformatics and others, data mining can also pose a threat to privacy in database security if not done or used properly. Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. [1].

In privacy preserving data mining (PPDM), the goal is to perform data mining operations on sets of data without disclosing the contents of the sensitive data. Since the results of the mining tell us something about the data, some information about the original data is leaked to the mining results. This leads to privacy loss. If the data is perturbed on the other hand for privacy concerns, it leads to information loss, which typically refers to the amount of critical information preserved about the datasets after the perturbation [3].Thus, we need to work towards minimizing both privacy loss and information loss.

Manuscript received Dec 22, 2007.
Lalanthika Vasudevan is with the Computer science department of Sri Venkateswara College of Engineering, Sriperumbudhur, TamilNadu, India ( phone: 044-24618868; e-mail: lalanthikav@gmail.com).
SE. Deepa Sukanya is with the Computer science department of Sri Venkateswara College of Engineering, Sriperumbudhur, TamilNadu, India ( phone: 044-23631095; e-mail: deepa.sukan@gmail.com).
N.Aarthi is a lecturer with is with the Computer science department of Sri Venkateswara College of Engineering, Sriperumbudhur, TamilNadu, India

Many approaches emerged for privacy preserving data mining. The first approach involved perturbing the input before mining. Though it has the benefit of simplicity it does not provide a formal framework for proving how much privacy is guaranteed. Secure Computation technique [2] has the advantage of providing a well defined model for privacy using cryptographic techniques and is also accurate. However it is a slower method.

In PRBAC technique, access to sensitive objects(SOBS) is based on roles [1].But it has the drawback of space complexity. i.e, as all data are stored in the Database Server ,it leads to a large memory requirement. Also, risk of illegal access of data is not completely ruled out as the entire data is stored at one site.In our paper,we address these problems by applying vertical fragmentation and cryptographic techniques for data storage.

Here, we propose a new approach to privacy preserving data mining based on cryptographic role based access control approach (PCRBAC) where we have 2 sets of object: Sensitive objects (SOBS) and Non sensitive objects (NSOBS). Using the data mining technique, users are allowed to mine different sets of data based on their roles. The data is first classified as sensitive objects and non sensitive objects. Sensitive objects are encrypted and stored. The permitted user can access the sensitive objects only after decryption ensuring privacy.

## II. RELATED WORK

Over the past few years, several approaches have been proposed in the context of privacy preserving data mining. Some of the main approaches include heuristic based approach, reconstruction based approach, and cryptographic approach. The underlying concept of the heuristic based approach technique is: how to hide sensitive rules that can be mined from the original data while maximizing the utility of the released data. In the reconstruction based approach [4,5], we first use some methods to distort the values of the original data and then release these distorted data. The third approach is Cryptography based approach [6, 7] which has been developed to solve the following problem: Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. This problem is referred to as the Secure Multiparty Computation (SMC) problem, which requires that no more information be revealed to a participant in the computation than that participant's input and output. It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one measures the confidential data protection, while the second measures the loss of functionality. Another important approach to Privacy Preserving Data Mining is the Access control based approach .In this approach, the groundwork to build an access control model over existing technologies was proposed called Multi-relational association rules (MRAR). This model is composed of three layers notably Authenticator, checker and the database server. In MRAR, the type of policy is Mandatory access control where the users are associated to mining levels. The addressed problem in MRAR is multilevel association rules. The major disadvantage of MRAR is that it is not always possible to

assign clearances to users of commercial information systems and not always possible to assign sensitivity levels to data in case level contains another level. This problem was overcome using the PRBAC model [1] which falls into the category of access control based approach; In Role based concept, the type of policy is Role based and the target system is Privacy preservation in data mining in the context of databases which can be built over existing database technologies.

The idea of Cryptographic approach and PRBAC (Privacy Preserving Role based access control approach) has motivated us to provide a more secure approach to privacy preserving data mining by combining the benefits of these two techniques along with the idea of vertical fragmentation of the data for distributed storage. We illustrate this idea by identifying data as sensitive and non-sensitive objects and using cryptographic and vertical partitioning techniques to securely store the data and taking into account the flexibility of role based access control models to access the stored data.

## III. PRIVACY PRESERVATION IN PCRBAC

There are two different views on privacy: Individual privacy: protecting individual data, and corporate privacy- the release of information about a collection of data rather than an individual data item. In PCRBAC focus is on individual's privacy.

### A Preliminaries & Requirements for PCRBAC:

Type of Policy: Policy in access control models can be classified into mandatory, discretionary, role- based. In PCRBAC the policy type is Cryptographic Role based policy since users are assigned to roles and roles have certain permission to access the encrypted sensitive objects.

Type of Subject: is a user.

Type of Object: In this model we have two object types: Sensitive objects (SOBS) and Non-sensitive objects. Sensitive object (SOBS) types could be Database table, Columns, Rows, in specific tables predetermined by System Administrator and accessed only according to role permission. Non-sensitive objects include Text document, Audio file, video file, web board, power point presentation file (PPT), Adobe acrobat file (PDF), etc

Target System: PCRBAC targets Privacy preservation in data mining, in the context of databases.

Security Aspects: The focus is mainly on securing SOBS from privacy breaches..

The requirements of PCRBAC model include:

- Role based concept.

- Decryption based on user access permission.

- Users are granted limited access i.e., parts of the data that they need to perform their mining tasks.

- Multiple authorized users can mine data concurrently, but the same user cannot have multiple active sessions.

- The Server maintains a SDB database containing SOBS name, corresponding identifier name and site of storage of the SOB.It also contains the decryption algorithms needed for decryption.

- The Encrypter maintains a EDB database containing the identifier name of the SOB and the corresponding encryption: decryption algorithm pair.

- The Encrypter also contains the encryption algorithms needed for encryption.

## IV. PRIVACY PRESERVING DATA MINING USING CRYPTOGRAPHIC ROLE BASED APPROACH (PCRBAC)

The preliminary concept of PCRBAC is PRBAC and is presented in this section followed by our PCRBAC technique.

### A Privacy Preserving Role based access control (PRBAC)

PRBAC is one of the known approaches to protect information in the context of relational databases which prevents users from obtaining sufficiently large and varied samples of a database .It also resolves the problem of tracing patterns that are not supposed to be disclosed .It does so by classifying objects as sensitive objects (SOBS) And Non-sensitive objects (NSOBS). PRBAC model is defined as follows:
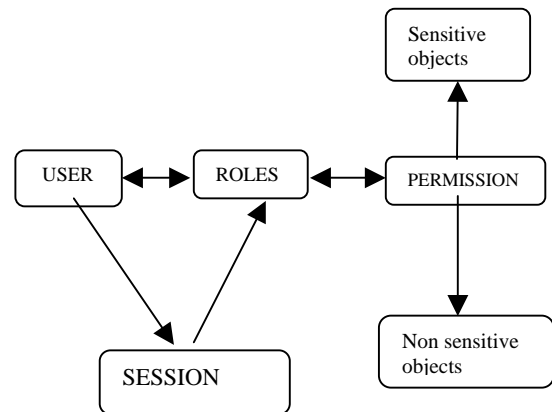


*Fig 1: PRBAC MODEL*

User: User is a human being

Role R and Role hierarchy: A role is a job function or job title within the company

Permission P: Permission is an approval for a particular mode of access.

Session S: A user establishes a session during which he activates some subset of roles that he or she is a member of.

Sensitive objects (SOBS): Sensitive objects are predetermined by the database administrator (DBA) or system administrator and it differs according to role permission. Examples for SOBS include Database Table, columns, Rows, in specific tables).

Non Sensitive Objects: An entity that contains or receives information, or has exhaustible system resources that can be accessed without permission. [1].

### B. Cryptographic Approach

**Standard Encryption Algorithm:**

Get the SOB;

Convert each character of the attribute
value into  its corresponding ASCII value;

Convert it into binary value;

Perform NOT operation;

Add flag bits between two binary values;

**Standard Decryption Algorithm**:

Get the encrypted SOB;

Retrieve the binary values;

Perform NOT operation;

Convert it into corresponding ASCII value;

Retrieve the characters corresponding to the
ASCII value;

### C. Vertical partitioning of data

Vertical partitioning of data refers to the method of partitioning data in which each site holds a subset of the attributes for all entities ensuring secrecy.
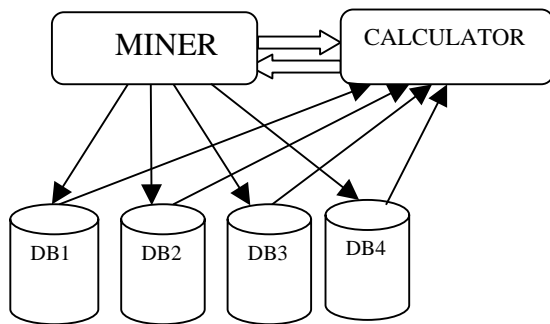


Fig 2: Vertical Partitioning

The architecture in the above figure shows the participating databases:

- A Miner which decides what computation is to be done

- A Calculator that does the computation. The Calculator is unaware of what data it computes.

-  It is important to note that, only the Miner and the participants get the mining results while the Calculator performs only auxiliary computations, without knowing their meaning.

### D. Proposed Work

a) PCRBAC  Model

Improving  the  above  two  approaches  the  PCRBAC  architecture has been developed.
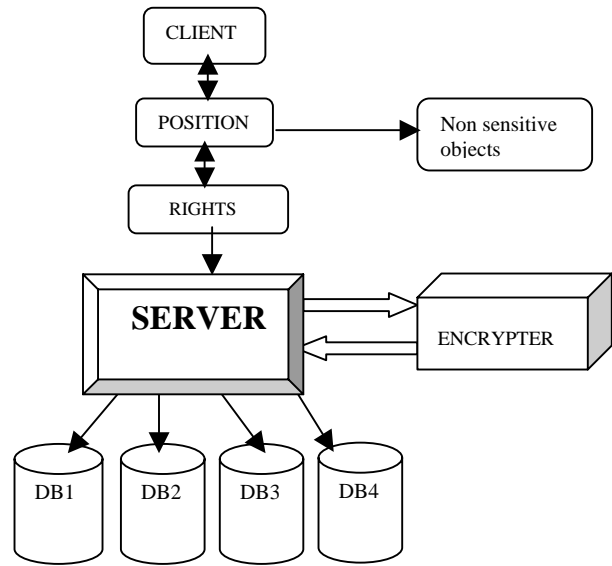


Fig 3: PCRBAC Architecture

Our proposed PCRBAC model is defined as follows:

Client: Client is a human or a computer who wants to access the entity.

Position:  Position is a rank of an individual within the organization based on which rights are assigned.

Rights:  Privileges or the access rights of the client enjoyed within an organization to access one or more objects in the database system.

Non Sensitive Object: An entity that contains or receives information, or has exhaustible system resources and can be accessed without permission.

Server:  Server identifies the sensitive objects, decides how the data is to be partitioned and where it is stored in the databases after encryption.  It also verifies the permission rights of the user.

Encrypter: Encrypter encrypts the given partitioned data unaware of the data it is encrypting.

Databases: Databases are the sites where the vertically partitioned data is stored after the double encryption.

b) PCRBAC Algorithm

The following algorithm describes the working of the PCRBAC model.

If (client == active)   //already has an active session
    then disconnect;

else if (client != active)
        get the (id  and password);

check the position of the client;

if (object required == Non sensitive)

     then Create session;
     display the object in active session;

else

  get position of the user and the SOB required;
  verify the input with SDB //permission rights

  if permission denied
    Logout;

  else
    Create session
      get the location of the SOB from the SDB;

      retrieve the encrypted SOB from the site;

      get the decryption algorithm name

      corresponding to the SOB  from EDB;

      double decrypt the SOB;

      display the SOB to the user;

      close all completed tasks;

      close the session;

c) PCRBAC Process:

In the PCRBAC process model privacy can be achieved using encryption technique where the server first splits the database attributes into sensitive and non sensitive objects. Non sensitive objects can be accessed by all the clients using the operation $<c,p,op>$ where c is the client; p is the position; op is the output. Sensitive objects can be accessed by the operation $<c,p,r,SOB,op>$ where c is the client; p is the position; r is the rights; s is the SOB required ;op is the output. The server using vertical partitioning approach first partitions sensitive object. It then encrypts the vertically partitioned data using a simple encryption technique (Standard encryption technique) and assigns an identifier name for the SOB. It then sends the encrypted SOB and its identifier name to the Encrypter.  The Encrypter then encrypts the data (i.e., the encrypted SOB) choosing randomly from any of the encryption techniques stored in it. It then stores the identifier name of the sensitive object and the $<encryption, decryption>$ algorithm pair name used in the Encrypterdatabase(EDB).The encrypted sensitive object is sent to the server which is stored in the respective site(Database) allocated by the vertical partitioning approach. Server maintains a database (SDB) which contains the name of the SOB, the corresponding identifier name, and the site (Database) where the double encrypted version of the SOB is stored. For decrypting the encrypted SOB, the Server finds the decryption algorithm corresponding to the SOB from the Encrypter module and decrypts the encrypted data in two steps: 1.Decryption using the decryption algorithm specified by the Encrypter module.2.Decryption using the decryption algorithm corresponding to the standard encryption technique [used for the first round of encryption by the Server].

d) Object access in PCRBAC model:

When an object has to be accessed by the client, the following sequence of steps are followed:

1.The client makes a request to the server for accessing the object.

2.If the requested object is a non sensitive object it is displayed without checking for permission.

3.If the requested object is a sensitive object, the client's position in the organization is checked. If permission is denied, the client is prevented from accessing the SOB.

4.If the client has the necessary permission to view the SOB, the server checks its database (SDB) to find the location where the requested sensitive object is stored. The server then gets information from the Encrypter database (EDB) about the decryption algorithm needed for decrypting that SOB (by passing the identifier name of the SOB to it). Once the required detail is obtained from the Encrypter, the Server decrypts the encrypted SOB using the corresponding decryption algorithm and does the second round of decryption using the standard decryption algorithm for retrieval of the actual data.

5.Then the Sensitive object is displayed to the user.

e) Advantages

In PCRBAC, the retrieval of data occurs through a secure path. The Server does not have a prior knowledge of the encryption techniques used and the Encrypter on the other hand is not aware of the actual data that is being encrypted by it. Thus, illegal access to SOB's is prevented. .

V.  EXAMPLE APPLICATION

Consider a hospital management system.The permission rights for object access is based on the policy:
<Role, permission, type of objects>

The Access policies are given by:

DEAN policy:
<DEAN,{R,W,U,D},NSOBS,SOBS>;

Chief Doctor policy:
<CHIEF DOCTOR,{R,W,U},NSOBS,SOBS>;

House surgeon policy:
<HOUSE SURGEON,{R},NSOBS,SOBS >;

Nurse policy:
<NURSE,{R},{all}, - >

An object identification module splits the objects into sensitive and non-sensitive data. The SOB's are then double encrypted and stored by vertical partitioning approach in several databases.

Let the ID names assigned to the sensitive objects be as follows:Address—I1; Symptoms—I2;Diseases—I3; Treatment—I4. Let I1,I2,I3,I4 be stored in databases DB1, DB2 ,DB3 and DB4 respectively.

TABLE I: SERVER DATABASE (SDB)

| SOB NAME | IDENTIFIER NAME | DATABASE |
|---|---|---|
| Address | I1 | DB 1 |
| Symptoms | I2 | DB2 |
| Diseases | I3 | DB3 |
| Treatment | I4 | DB4 |

TABLE II :ENCRYPTER DATABASE(EDB)

| IDENTIFIER NAME | ENCRPTION-DECRYPTION PAIR |
|---|---|
| I1 | <E1,D1> |
| I2 | <E2,D2> |
| I3 | <E3,D3> |
| I4 | <E4,D4> |

CASE 1: NORMAL ACCESS

If the dean makes a request for the symptoms of a patient 'A', the information is retrieved as follows:
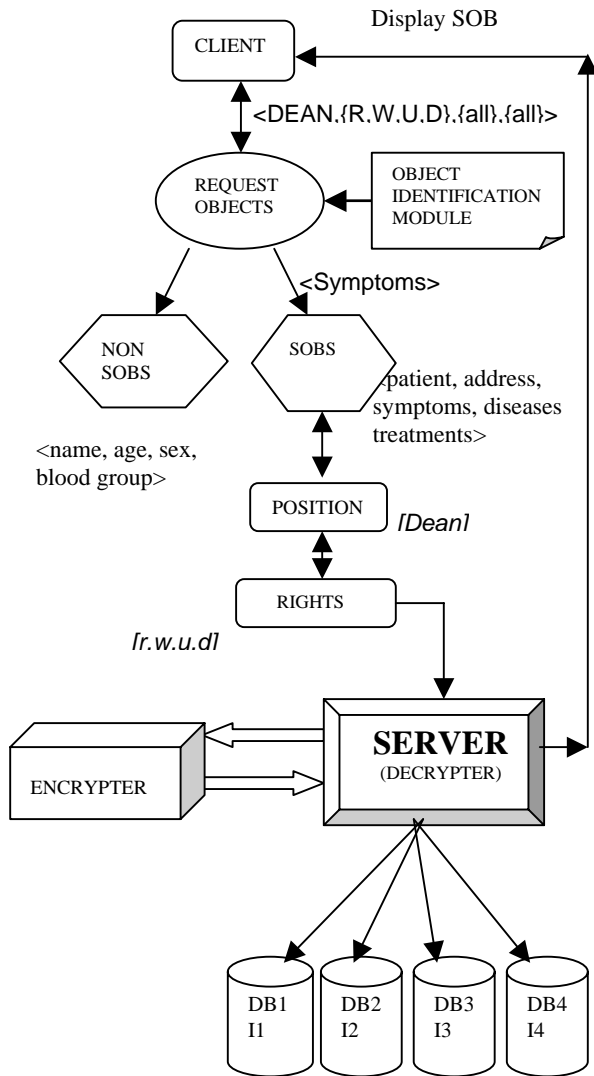
The Object Identification module identifies symptoms as a sensitive object.The client's position is checked and his role permission is found out : <DEAN,{R,W,U,D},{all},{all}>;where {R,W,U,D} represents the access rights;{all} represents the access rights for NSOBS;{all} represents the access rights for SOBS. And R represents Read, W represents Write, U represents Update, D represents Delete.The Server Database passes the Id name of the SOB to the Encrypter database to determine the Decryption algorithm to be used. Once the Encrypter identifies and sends the decryption algorithm name to the Server, the double encrypted SOB is decrypted first by using the decryption algorithm specified by the Encrypter and then by the Standard Decryption algorithm.The requested object is displayed to the user.

CASE 2: CRUCIAL TIME ACCESS:

In real time applications, where the time constraint is very crucial, the sensitive objects are treated like non-sensitive objects. For example,when the sensitive object "symptoms" is requested by the nurse during a crucial time, access permission is granted.
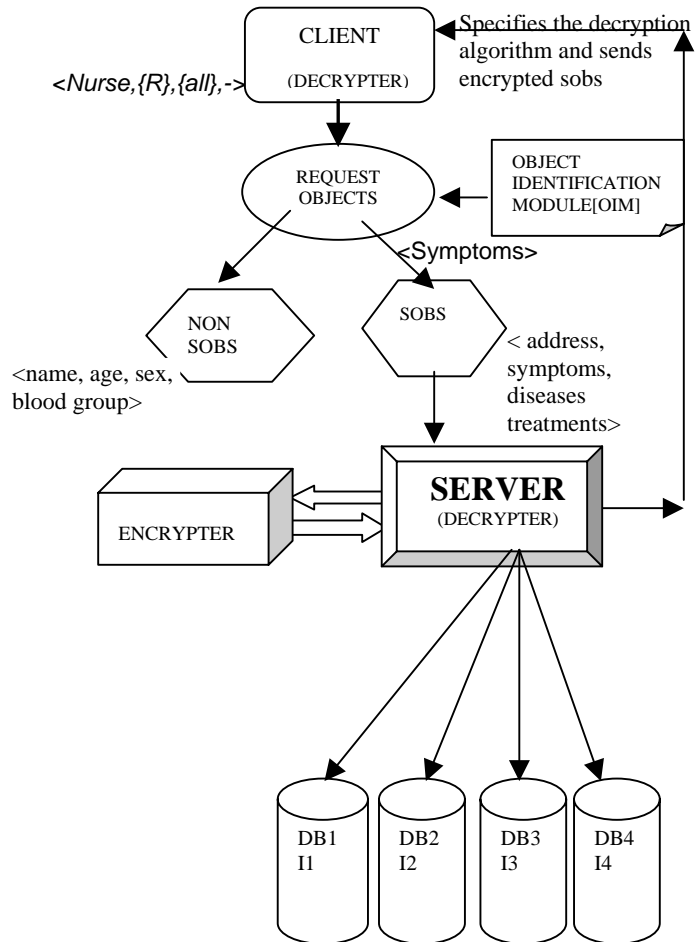


Fig.4: Normal access



Fig 5:Crucial access

## VI. CONCLUSION AND FUTURE WORK

In our approach, we have implemented privacy preservation in data mining by using the cryptographic role based access control. Here, we have assumed that the decryption occurs entirely at the Server. For real time applications with crucial time-constraints like Bio-medical applications, the keys for decryption can be distributed to the user for faster decryption and retrieval of data. For frequently accessed critical data, (for example, maintaining records of critical patients,) horizontal fragmentation can be used for quick retrieval of information.

Our approach can be extended by incorporating the concept of Role delegation into our model. Also, determining optimized algorithms for identification and classification of sensitive and non sensitive data will make our PCRBAC , a complete efficient system for data mining that will work towards mining useful information with minimal privacy breaches.

## VII. REFERENCES

1. Anor F.A. Dafa-Alla, Eun Hee Kim, Keun Ho Ryu, *Yong Jun Heo "PRBAC: An Extended Role Based Access Control for Privacy preserving Data mining" In Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05) of IEEE, 2005 .

2. Alex Gurevich, Ehud Gudes "Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation" In the proceedings of the 10th international Database Engineering and Applications Symposium (IDEAS'06) IEEE , 2006.

3. Chai Wah Wu IBM T. J. Watson Research Center "Privacy preserving data mining with unidirectional interaction" In the proceedings of the international conference of IEEE, 2005.

4. Yucel Saygin, Vassilios S.Verykios and Ahmed Elmagarmid K. Privacy preserving association rule mining, In Proceedings of the 12th International Workshop on Research Issues in Data Engineering pages 151.158.

5. Rakesh Agrawal, Srikant. Privacy Preserving Data Mining. ACM SIGMOD, 2000.

6. C. Clifton,M. Kantarcioglu, J. Vaidya, X. Lin and M. Y. Zhu. "Tools for Privacy Preserving Distributed Data Mining". In SIGKDD Explorations, 4(2): 28-34 December 2002.

7.Murat Kantarcioglu, Chris Clifton. "Privacy preserving Distributed Mining of association Rules on Horizontally partitioned Data.IEEE transactions on knowledge and data engineering, 2003.