ENCRYPTION AND DECRYPTION BASED INFORMATION SECURITY IN BIG DATA

G.VIHARI
Department of IT
Sir C.R.R College of Engineering, Eluru.

P.RAMAIAH CHOWDARY
Department of IT
Sir C.R.R College of Engineering, Eluru

ABSTRACT

The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each type of user, we discuss his privacy concerns and the methods that can be adopted to protect sensitive information. We briefly introduce the basics of related research topics, review state-of-the-art approaches, and present some preliminary thoughts on future research directions. Besides exploring the privacy-preserving a approaches for each type of user, we also review the game theoretical approaches, which are proposed for analyzing the interactions among different users in a data mining scenario, each of whom has his own valuation on the sensitive information. By differentiating the responsibilities of different users with respect to security of sensitive information, we would like to provide some useful insights into the study of PPDM.

Keywords: *Data mining, sensitive information, privacy-preserving data mining, anonymization, provenance, game theory, and privacy auction, anti-tracking.*

1. INTRODUCTION

Data mining has attracted more and more attention in recent years, probably because of the popularity of the ``big data" concept. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. As a highly application-driven discipline, data mining has been successfully applied to many domains, such as business intelligence, Web search, scientific discovery, digital libraries, etc.

The term data mining" is often treated as a synonym for another term ``knowledge discovery from data" (KDD) which highlights the goal of the mining process. To obtain useful knowledge from data, the following steps are performed in an iterative way (see Fig. 1.1):

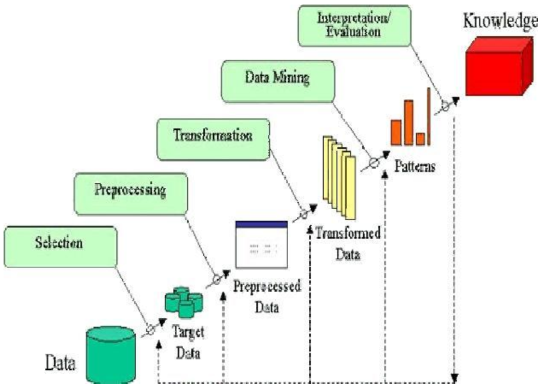


FIGURE 1.1 *An overview of the KDD process.*

We can identify four different types of users, namely four *user roles*, in a typical data mining scenario (see Fig. 1.2)

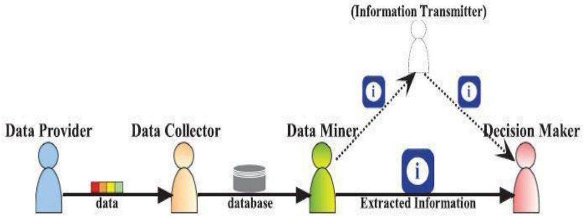


FIGURE 1.2 *A simple illustration of the application scenario with data mining at the core.*

Data Provider: The user who owns some data that are desired by the data mining task.
Data Collector: The users who collects data from data providers and then publish the data to the data miner.
Data Miner: The user who perfo0rms data mining tasks on the data.
Decision Maker: The user who makes decisions based on the data mining results in order to achieve certain goals.

The data that is provided by the user can be breached or can be get by other users of the database since there is less security in data base the data provided by the user is not safe and sensitive data is not fully secured we have to developed an application that encrypts the data and then stores the data in database so that other unauthorized user cannot get the data and does not know the data that is been hidden in the encrypted data.

2. SYSTEM ANALYSIS

PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:

Identifier (ID): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile number. Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code. Sensitive Attribute (SA): Attributes that an individual wants to conceal, such as disease and salary. Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA.

Before being published to others, the table is anonym zed, that is, identifiers are removed and quasi-identifiers are modified. As a result, individual's identity and sensitive attribute values can be hidden from adversaries.

The standard security techniques in database management system, such as username and password or access control mechanisms, does not provide full security to the data that is been provided by the data provider.

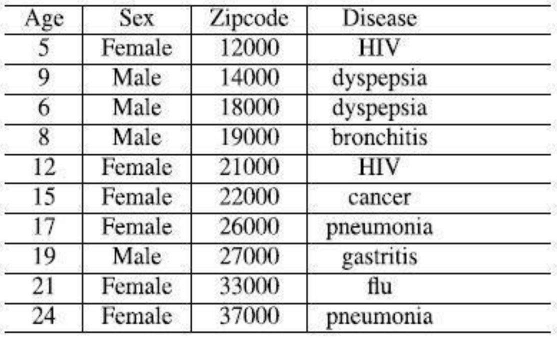
DISADVANTAGE

- Security computations in distributed programming frameworks
- Security best practices for non-relational data stores
- Secure data storage and transactions logs
- End-point input validation/filtering
- Real-time security monitoring
- Scalable and compostable privacy-preserving data mining and analytics
- Granular access control
- Granular audits
- Data provenance

IMPLEMENTATION

The basic idea of this project is that the data is to be secured with the help of encryption and decryption technique. Since the data is been encrypted and then stored in the database system the unauthorized users cannot know the data if the data is breached. Here I have used SHA3 algorithm for the implementation of encryption and decryption technique. For example: How the data table should be anonymized mainly depends on how much privacy we want to preserve in the anonymized data. Different privacy models have been proposed to quantify the preservation of privacy. Based on the attack model which describes the ability of the adversary in terms of identifying a target individual, privacy models can be roughly classified into two categories.

The first category considers that the adversary is able to identify the record of a target individual by linking the record to data from other sources, such as liking the record to a record in a published data table (called *record linkage*), to a sensitive attribute in a published data table (called *attribute linkage*), or to the published data table itself (called *table linkage*). The second category considers that the adversary has enough background knowledge to carry out a *probabilistic attack*, that is, the adversary is able to make a confident inference about whether the target's record exist in the table or which value the target's sensitive attribute would take. Typical privacy models. Includes *k*-anonymity (for preventing record linkage), *l*-diversity (for preventing record linkage and attribute linkage), *t*-closeness (for preventing attribute linkage and probabilistic attack), *epsilon*-differential privacy (for preventing table linkage and probabilistic attack), etc.



Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	flu
24	Female	37000	pneumonia

TABLE 2.1: Table for sensitive data.

ADVANTAGE'S

Privacy-preserving publishing of social network data Data utility

Privacy-preserving publishing of trajectory data

3. MODULES

The system consists of the following module encrypt and decrypt

ENCRYPTION AND DECRYPTION:

Encryption: a process of encoding a message so that it's meaning is not obvious Decryption: the reverse process (encipher) vs. decode (decipher) Encoding: the process of translating entire words or phrases to other words or phrases Enciphering: translating letters or symbols individually Encryption: the group term that covers both encoding and enciphering P(plaintext): the original form of a message

C(ciphertext): the encrypted form Basic operations plaintext to ciphertext: encryption: C = E(P) ciphertext to plaintext:

decryption: P = D(C) requirement: P = D(E(P))

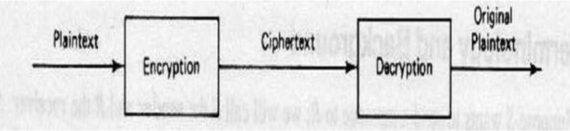
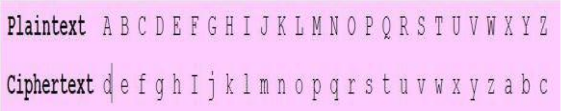


FIGURE 3.1: *Encryption and decryption*

A cipher is more cryptographically secure would display a rather flat distribution, which gives no information to a cryptanalyst



Plaintext	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Ciphertext	d	e	f	g	h	I	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	a	b	c

TABLE 3.2: *Table for plaintext and cipher text.*

4. SYSTEM ARCHITECTURE

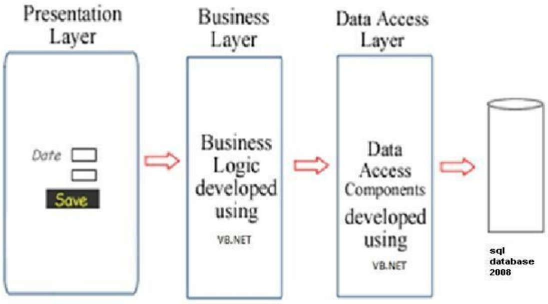


FIGURE 4.1: *System architecture for proposed application.*

Presentation layer is the end user layer where the data collector will collect the data and enter the data into the system. Business layer is the software where it is used as an end application layer used by the data collector. Data Access layer is the layer where the data is been accessed from the database to the application layer. SQL Database is the database where it is used to store the data it act as a data source

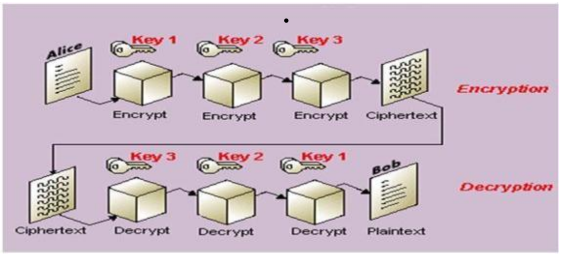


FIGURE 4.2: SHA3 Encryption and decryption

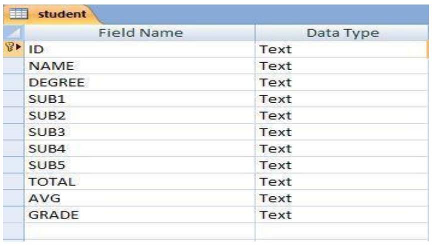
Encryption is the process of adding the key to the plain text and hiding the information stored and showing different kind of data. Decryption is the reverse process of encryption removing the key that is added to the plain text. Plain text that is added with the key is called chipper text. Plain text is the original text where the data without the key added. In cryptography encryption is the process of encoding message or information in such a way that only authorized parties can read it. Encryption does not of itself prevent interception, but denies the message content to the interceptor. There are two types of encryption available they are symmetric key encryption and public key encryption.

5. IMPLEMENTATION RESULT

Here we are going to create an application which will provide more security for the data in the database. I have used the SHA3 encryption and decryption algorithm for encrypt the data and then store it into the database.

The system implementation I have used VB.NET and SQL DATABASE software's to handle the data. The user is provided two types of options to secure the data they are "ENCRYPTION" and "DECRYPTION" The user also provided more operation to manipulate the data available in the database they are ADD NEW, UPDATE, DELETE, EDIT, AND SEARCH. Thus this Application provides more security for the data and if any security breach occurs then the attacker will not get any data from the database. This application is less cost and easy to access the data.

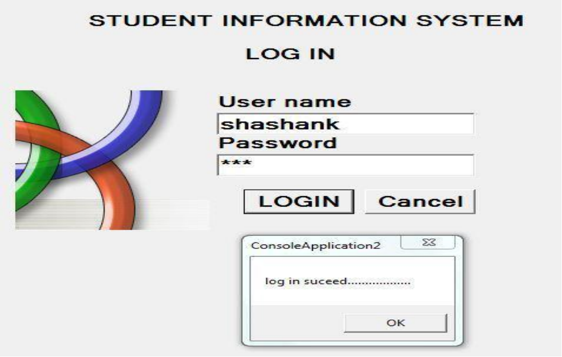
DATABASE DIAGRAM



student

Field Name	Data Type
ID	Text
NAME	Text
DEGREE	Text
SUB1	Text
SUB2	Text
SUB3	Text
SUB4	Text
SUB5	Text
TOTAL	Text
AVG	Text
GRADE	Text

TABLE 5.1*: STUDENT DATABASE SYSTEM*



STUDENT INFORMATION SYSTEM

LOG IN

User name
shashank
Password
***

LOGIN Cancel

ConsoleApplication2

log in suceed.................

OK

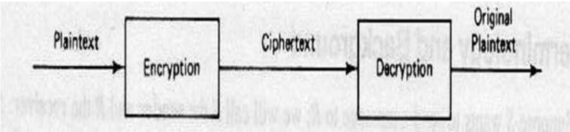
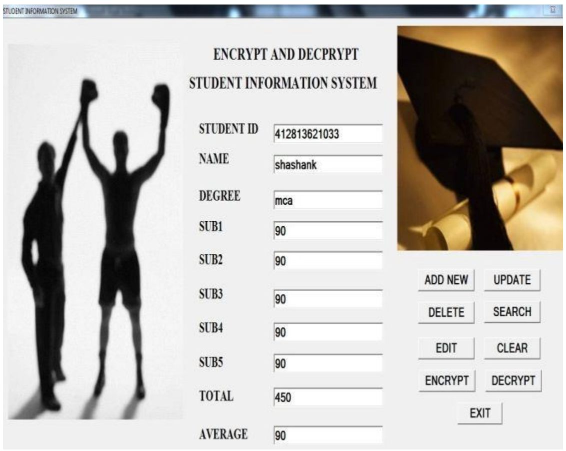


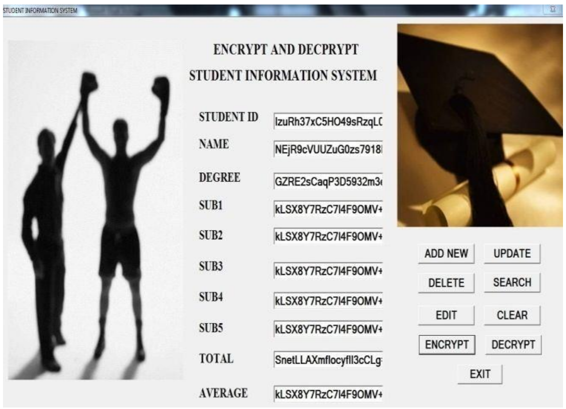
FIGURE 5.2: *log in page for student data base*

management system



STUDENT ID	412813621033
NAME	shashank
DEGREE	mca
SUB1	90
SUB2	90
SUB3	90
SUB4	90
SUB5	90
TOTAL	450
AVERAGE	90

FIGURE 5.3: *Record after searching in database*



STUDENT ID	IzuRh37xC5HO49sRzqL(
NAME	NEjR9cVUUZuG0zs7918I
DEGREE	GZRE2sCaqP3D5932m3(
SUB1	kLSX8Y7RzC7I4F9OMV+
SUB2	kLSX8Y7RzC7I4F9OMV+
SUB3	kLSX8Y7RzC7I4F9OMV+
SUB4	kLSX8Y7RzC7I4F9OMV+
SUB5	kLSX8Y7RzC7I4F9OMV+
TOTAL	SnetLLAXmflocyfIl3cCLg:
AVERAGE	kLSX8Y7RzC7I4F9OMV+

FIGURE 5.4: *Records after encryption*

6. CONCLUSION

A dynamic secret-based encryption scheme is designed to secure the data that is been stored in the database to reduce its complexity, the retransmission sequence is proposed to update dynamic encryption key, here we are using the SHA3 encryption and decryption algorithm. It provides more security to the data.

A demo system is developed to show the performance and security of SHA3 encryption and decryption algorithm. An application is been developed with the help of using the application VB.NET and SQL database. To implement the big data security we have taken the student database management system for the implement of encryption and decryption technique.

This application will provide many operations and flexibility of the data that is been stored in the data base. The user will have the permission for adding, deleting, editing, updating and searching of data. The data is encrypted and then stored in the database and in search operation the data is decrypted and then searched in the database. While adding new data the data is provided as a normal data by the user and the data is encrypted and then stored in the database.

If the user retrieve the data from the database the data is been decrypted and then shown to the end user. The user is also provided options to encrypt and decrypt the data.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[2] L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining,'' in *Proc. Austral. Inst. Comput. Ethics Conf.*, 1999,

[3] R. Agrawal and R. Srikant, ``Privacy-preserving data mining,'' *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 439_450, 2000.

[4] Y. Lindell and B. Pinkas, ``Privacy preserving data mining,'' in *Advances in Cryptology*. Berlin, Germany: Springer-Verlag, 2000, pp. 36_54

[5] C. C. Aggarwal and S. Y. Philip, *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*. New York, NY, USA: Springer-Verlag, 2008.

Mr. G.Vihari is working as an Asst Professor, in I.T Department, Sir C. R Reddy College of Engg, Eluru, A.P., India. He has received his B.Tech (CSE) from Al-Ameer College of Engineering from Al-Ameer College of Engineering, Visakhapatnam and M.Tech (I.T) from Gitam Institute of Technology, GITAM University Visakhapatnam A.P., INDIA. His research interests include, Cloud Computing, Networks security, Web security, Software Engineering and Computer Networks.

Mr.P.Ramaiah Chowdary is working as an Asst.Professor, in Department of I.T , Sir C. R. Reddy College of Engg, Eluru , A.P.,India. He has received his B.Tech(I.T) from SSIET, Nuzvid, A.P and M.Tech(S.E.) from Gitam Institute of Technology, GITAM University, Visakhapatnam, A.P., INDIA. His research interests include Cloud Computing, Software Engineering, Neural Networks, Fuzzy logic, and Computer Networks.