Journal of Theoretical and Applied Information Technology
31st October 2014. Vol. 68 No.3
© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRIVACY PRESERVING COLLABORATIVE DATA MINING USING STEGANOGRAPHY AND ENCRYPTION

[1]HARE RAM SAH AND[2]G.GUNASEKARAN

[1]Research Scholar, Faculty of CSE, Sathyabama University, Chennai, INDIA
[2]Professor and Principal, MeenakshiCollege of Engineering, Chennai, INDIA
Email: [1]ramaayu@gmail.com

ABSTRACT

Data collection is an essential step in data miningprocess. Collecting data of varying nature and still preserving privacy is essential for many applications. Privacy concerns maypreventdirect sharing of data andhow multipleparties collaboratively conduct data mining withoutbreaching data privacy presents a challenge.Cryptography involves converting a message text into anunreadable cipher and steganography embedsmessage into a cover media and hides its existence.Both these schemes are effectivelyimplemented in image files. In this paper an advanced system of encrypting datathat combines the features of cryptography and steganography along with privacy preservation is presented.A case study with medical image is also presented. The details of the person's medical image along with clinical interpretation is encrypted, stegano and stored in the database. Key based retrieval technique is used to recover the hidden details. The advantage of this scheme is that the stegano can operate on encrypted texts also and hence offers a double layer data protection.

Keyword: *Privacy Preserving, Data mining,Steganography,Image Encryption*

1. INTRODUCTION

Privacy is an issue when used data involve individual sensitive information as the increasing useof data mining, large volumes of personaldata are regularly collected and analyzed. These data are importantasset to business organizations and governments both todecision making processes (provide social benefit,such as medical research, crime reduction, national security,etc.) However, data owners are becomingincreasingly concerned about their privacy, since the datacontains some personal information about individuals, medical records, health insurance data etc. Privacy preserving aims to prevent informationdisclosure and ensure legitimate access to the data. Thus, privacypreserving is different from conventional data security,access control and encryption technology that tries to preventinformation disclosure against illegitimate means. The privacy and security of individuals should be taken care to solve ethical, legal and social issues.The mainconsideration in privacy preserving data mining is,

(i) Sensitive raw data (like identifiers, names, addresses) should be modified from the originaldatabase.

(ii) Sensitiveknowledge which can be mined from a database by usingdata mining algorithms should also be excluded.

Privacy and security has ability to communicate and share data; an omniscient data source carries value to research and building accurate data analysismodels. Such an ambitious task requiresthe collaboration of geographically distributed industries,etc. It needs toshare their private data for building data analysis models tounderstand the underlying physical phenomena.

1.1Problem Statement

Records are encrypted or watermarked and provided with hidden information i.e. steganography. The encryption or watermarking or stegano based schemes serve two purpose (i) indexing (ii) privacy preserving (particularly for medical images). However, when a particular text and its associated record has to be accessed and related with similar hidden text or encrypted text of other records, special models are required to perform efficient data mining. Also, the metrics which define the efficiency of the data mining scheme shall be different than the conventional ones. In this research, it is planned to perform data mining for three emerging record storage approaches namely (i) encrypted record: Decryption models are to be designed to coexist with the data mining scheme and retrieve the query (ii) Watermarked records: In the case of watermarked record, relational data base structure has to be studied as the watermark text or data of similar records may be the relational query and the mined data is the dewatermarked record. (iii) Stegano records: This area first requires retrieving the hidden data and then performing mining. More important in this work, to improve the speed of mining certain preemptive algorithms are to be studied and implemented and hardware implementation of the mining work is planned to use threading and concurrent search tasks more effectively.

1.2 Steganograghyand Encryption Process

Steganography and encryption processes arecounter parts in digital security the obvious advantage ofsteganography over encryption that messages do notattract attention to themselves, to messengers or torecipients.To make a steganographic communication even moresecure the message can be compressed and encryptedbeforebeing hidden in the carrier. The types of steganography are,

(i) **Linguistic Steganography:** It uses language in the cover that is categories as,

(a) **Open codes**: Openly readable text is mostly well constructed.

(b) **Text semagrams**: works with graphical modification of text.

(ii) **Technical Steganography:** Technical steganography is a method where a tool, adevice or a method is used to conceal the message.

Steganography techniques strive to hide the very presenceof the message itself from an observer. It hides information in digital content has a widerclass of applications that go beyond steganography, Figure 1.

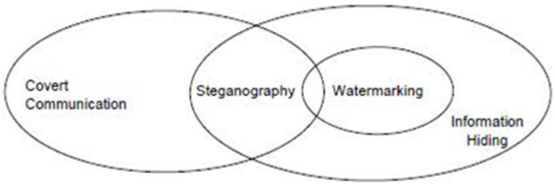


Fig. 1 Relationship of steganography to related fields

The modern steganography is a term of the prisoner's problem where Alice and Bob are two inmates who wish tocommunicate in order tohatch an escape plan. The data hiding information into a media requires the cover medium that holds the hidden data,secret message, cipher text, the stego function and an optional stego-key that is used to hide and unhide the message is shown in Figure 2.

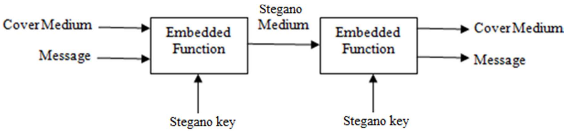


Fig.2 Steganography Flow

1.3 Privacy-Preserving Data Analysis

Many privacy-preserving data analysis protocols have beendesigned using cryptographic techniques, where data are generallyassumed to be either vertically or horizontally partitioned.In the case of horizontally partitioned data, differentsites collect the same set of information about different entities. It distributed protocols for horizontally partitioned data for many differentdata mining tasks such as building decision trees, mining association rules, and generating k-meansclusters and K-NN classifiers.Again, privacy-preserving protocols for the verticallypartitioned case have many differentdata mining tasks such as association rules,buildingdecision trees and k-means clusters.

2. PREVIOUS WORK

Tipawan, et al., [2012] proposed knowledge discovery in databases process and is considered as significant subfield in knowledge management in data mining. Berson, et al., [1999] presented to discover valuable information hidden in the data by transforming these data into useful knowledge.Lalitha, et al., [2014] presented an encryption with high capacity and low distortion that can be achieved efficiently and easy for the data hider to reversibly embed data in the encrypted image. Puech, W., et al., [2012] presented an increasing number of image and video cryptographic techniques are used to enforce content access control, identity verification, authentication and privacy protection. Lavrac., et al., [2007] presented the combination of the data mining and decision support approaches in planning of the regional health-care system.Rahman, et al., [2012] proposed knowledge refinement through a use of the technique on the construction industry dataset.Abhishek, et al., [2014] presented steganography method with text media in a picture or image format towards steganography techniques.Benny, et al., [2002] proposed non-trusting parties can jointly compute functions of the different inputs while ensuring that no party learns anything. Akansha and Virendra [2014] presented to discover and study the approaches for securing video files. The videos send securely and data will be protected from any unauthorized access. Murat and Wei [2013] presented to develop key theorems and base on the theorems, analyze certain important privacy-preserving data analysis tasks. Kantarcioglu, et al., [2009] proposed privacy-preserving data mining in the malicious modelthat consider curious modelagainst malicious adversaries.Sweeney [2002] presented the k-anonymity privacy requirement, for each record in an anonymzed table to be indistinguishable with at least k-1 other records within the dataset. Savitaand Lata[2012] proposed protect private data with better accuracy and reconstruct original data and provide data with no information loss and makes usability of data.Vijayalakshmi, V., et al., [2014] proposed how to embed secret message into anon standard cover file.txt and encryption using DES algorithm and shifting cipher method where both uses randomly generated key. Jian, et al., [2009] presented topreserve the sensitive data and establish the extent and depth of existing techniques to preserve sensitive data.

3. ENCRYPTION TECHNIQUE FOR DATA MINING USING BLOWFISH ALGORITHM

Blowfish Algorithm is used for encryption and decryption data transmission process, which has symmetric block cipher that can be effectively used for encryption andsafeguarding of data. It takes a variable-length key, from 32 bits to 448 bitsand ideal for securing data. Althougha complex initialization phase required an encryption of data, which is very efficient on large microprocessors. A variable-length key block cipher is suitable for applicationthat does not change a communications link. It issignificantly faster than other encryption algorithms, when implemented on 32-bitmicroprocessors with large data caches. The blowfish algorithm has,

(i) Manipulates data in large blocks
(ii) Has a 64-bit block size.
(iii) Has a scalable key, from 32 bits to at least 256 bits.
(iv) Uses simple operations that are efficient on microprocessors.

The blowfish consists of a variable number of iterations, where the applications are with a small key size, the trade-off between the complexity ofa brute-force attack and a differential attack make a large number of iterationssuperfluous. Hence, it reduces the number of iterationswithout loss of security (beyond that of the reduced key size).

The privacy protected data storage module has ‘n’ number of raw data, (variable size) that has to be encrypted as shown in Figure 3.

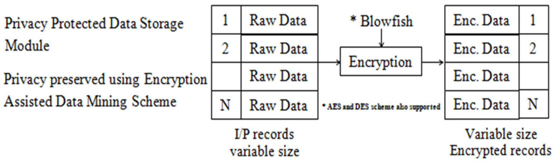

Privacy Protected Data Storage Module
Privacy preserved using Encryption Assisted Data Mining Scheme
1
Raw Data
2
Raw Data
Raw Data
N
Raw Data
* Blowfish
Encryption
* AES and DES scheme also supported
Enc. Data
1
Enc. Data
2
Enc. Data
Enc. Data
N
I/P records variable size
Variable size Encrypted records

Fig. 3 Privacy protection using data encryption (Blowfish encryption used)

The retrieved data variables decrypt from the specific data base is shown in Figure 4 and Figure 5.

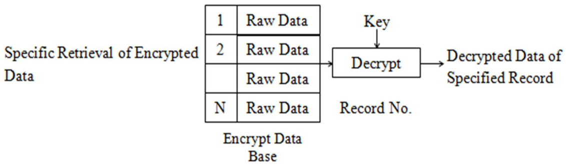


Fig. 4Specific retrievalof encrypted data

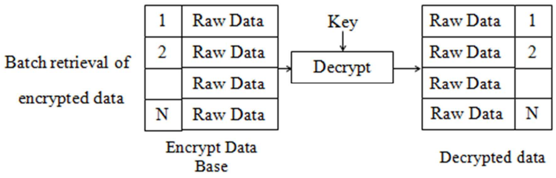


Fig. 5Decryption and batch retrieval from encrypted data

The encrypted data module decrypt the data module using query process methodology and matches specific records (those are decrypted) as shown in Figure 6.

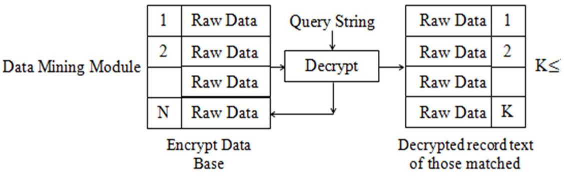


Fig. 6 Decryption of data record using querymethod

3.1 DataMining Task

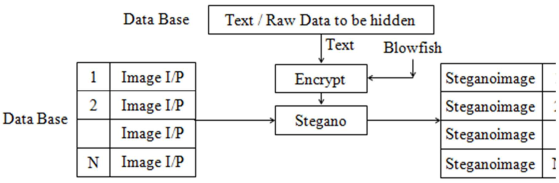


Fig. 7 Privacy protection for medical images

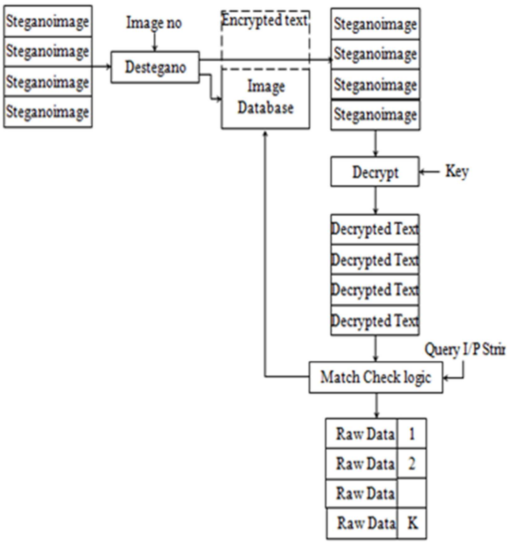
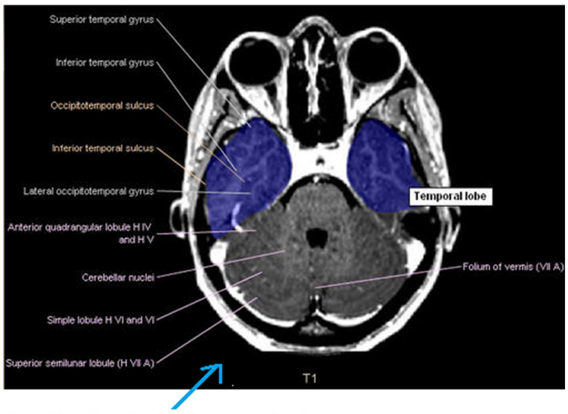


Fig. 8Retrieved specific images matched with given query

4. RESULTS AND DISCUSSION

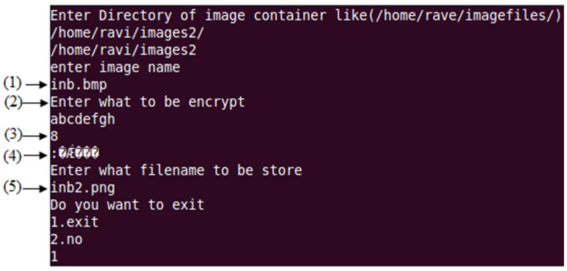
In this work, an image as input file to be encrypted with the data and to be stegano data is shown in Figure 9.



Input files where data is to be encrypted and stegano

Fig. 9 Input data for encrypted and stegano

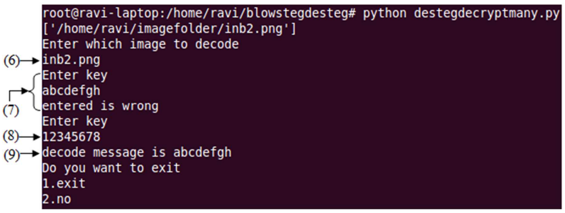
The user is presented with a menu to input the image in which the text is to be encrypted and stegano. A sample text is shown in figure 10 along with the encrypted output. Figure 11 shows the recovereddata.



```
Enter Directory of image container like(/home/rave/imagefiles/)
/home/ravi/images2/
/home/ravi/images2
enter image name
inb.bmp
Enter what to be encrypt
abcdefgh
8
:�Æ���
Enter what filename to be store
inb2.png
Do you want to exit
1.exit
2.no
1
```

Note: (1)Interactive menu to select input file
(2) Data to be encrypted and hidden (stegano) in Image selected
(3) Size of the text to be encrypted
(4) Encrypted output
(5) Encrypted text stored in file_2

Fig. 10 Selected input image data with Encrypted output



```
root@ravi-laptop:/home/ravi/blowstegdesteg# python destegdecryptmany.py
['/home/ravi/imagefolder/inb2.png']
Enter which image to decode
inb2.png
Enter key
abcdefgh
entered is wrong
Enter key
12345678
decode message is abcdefgh
Do you want to exit
1.exit
2.no
```

Note: (6) File from where text is to be decrypted after destegano
(7) Incorrect key match and data not retrieved
(8) Matching key
(9) Correctly decrypted message

Fig. 11 Decrypted text for the matched key

5. CONCLUSION

In this paper, a data hiding and retrieval technique suited for even watermarked records and images is presented. Hardware implementation on ARM core is also discussed. The work shall assist in privacy preserving for medical images, image based identification, etc. Future direction of work shall focus on video cryptography and use of thread based search for data retrieval.

REFERENCES:

[1] TipawanSilwattananusarn and KulthidaTuamsuk, "Data Mining and Its Applications for Knowledge Management", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.2, No.5, September 2012.

[2] Berson, A., Smith, S.J. &Thearling, K., "Building Data Mining Applications for CRM", NewYork: McGraw-Hill, 1999.

[3] Lalitha P., Vidhushavarshini S., "Retrieving Information using Reversible Data Hiding", International Journal of scientific research and management, Vol. 2, No. 5, pp. 802-808, 2014.

[4] Puech, W., Erkin, Z., Barni, M., Rane, S., Lagendijk, R.L., "Emerging cryptographic challenges in image and video processing", Image Processing (ICIP), 19th IEEE International Conference, pp. 2629-2632, Sept. 2012.

[5] Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. &Kobler, A., "Data mining and visualization for decision support and modeling of public health-care resources", Journal of Biomedical Informatics, 40, 438-447, 2007.

[6] Rahman, N. & Harding, J.A., "Textual data mining for industrial knowledge management and text classification: A business oriented approach", Expert Systems with Applications, 39, 4729-4739, 2012.

[7] Abhishek Koluguri, Sheikh Gouse, P. Bhaskara Reddy, "Text Steganography Methods and its Tools", International Journal of Advanced Scientific and Technical Research, Vo. 2, No. 4, pp. 888-902, April 2014.

[8] Benny Pinkas, "Cryptographic techniques for privacy preserving data mining", SIGKDD Explorations, Vol.4, Issue 2, pp. 12-19, 2002.

[9] AkanshaAgrawal, Virendra Singh, "Securing Video Data: A Critical Review", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2014.

[10] Murat Kantarcioglu and Wei Jiang, "Incentive Compatible Privacy-Preserving Data Analysis", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 6, June 2013.

[11] M. Kantarcioglu and O. Kardes, "Privacy-Preserving Data Mining in the Malicious Model", International Journal of Information and Computer Security, Vol. 2, pp. 353-375, Jan. 2009.

[12] L. Sweeney, "K-anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570, 2002.

[13] SavitaLohiya, LataRagha, "Privacy Preserving in Data Mining Using Hybrid Approach", Fourth International Conference on Computational Intelligence and Communication Networks, 2012.

[14] V. Vijayalakshmi, Mahalakshmi, Thamizharasan, "Data Encryption hiding technique in non-standard cover files", International Journal of Advanced Research in Computer Science and Technology, Vol. 2, No. 1, March 2014.

[15] Jian Wang, Yongcheng Lou, Yen Zhao, Jiajin Le, "A Survey on Privacy Preserving Data Mining", International Workshop on Database Technology and Applications, pp.111-114, 2009.