An Efficient Model for Secure Data Publishing

[1] Gaurav Kumar Ameta, [2] Divya Bhatnagar

[1, 2] Department of Computer Science and Engineering,
Sir Padampat Singhania University,
Bhatewar, Udaipur, Rajasthan, India

Abstract - Data Mining is the field of extracting and analyzing the data from large datasets. Exchange of databases is very important to get financial benefits now a day. To review business strategies and to get maximum benefit data analytics is needed. Data stored at distributed sites are integrated and published by data publisher. Data Publishing is the technique in which the data is released to others for the use. In addition to privacy preserving data size and security is also a challenge while publishing and transmitting the database. So there is a requirement of a technique which can reduce the size of database efficiently and transfer it in a secure manner. This paper proposes an efficient model for secure data publishing by using both compression and color encryption thus introducing a new approach. A new algorithm is designed and a tool is developed for implementing the proposed work.

Keywords - Data Publishing, Data Mining, Distributed Databases, Tool for Data Publishing.

1. Introduction

Data Mining is the field which has multiple disciplines. Data Publishing is an emerging field in present perspective. Various organizations rely on data analytics to know information about the details of business of different companies or within different branches of the same company to check behavior of data in their sectors. Now days Data Analytics is part of almost every sector like healthcare, automobile, transportation, telecommunication, finance and retail etc. and after analyzing company owners want to get benefits from it. There are various challenges before data publishing. Some of them are related to collection of database at the centralized place for analysis. Another challenge is that databases are kept in heterogeneous formats on each and every site from which data is to be collected which will be used for collaborative analysis. Security of the database must be retained during transmission. Data Size is also a major issue; it should be less in size for efficient transmission from sender to receiver. All of the above discussed challenges are faced by Data Publisher**.** In this paper we are considering several data providers that are situated at distributed locations and associated to centralized data publisher. Sender sites are responsible to send data intermittently to data publisher for analysis. Therefore, it is required to maintain uniformity in all database formats so that the burden of making all the received databases uniform is taken off from the data publisher's end. Data Size is another parameter on which efficiency depends. Data size should be small. In proposed technique transactional database is converted to very small thus reducing the protocol overhead to large extent. Apart from the reduction in data size, there is also an issue related to security during transmission of data. Data should be encrypted into another form without losing accuracy to ensure protection from attackers.

2. Literature Review

There are various techniques and mechanisms developed to encrypt and decrypt the textual data into other formats.

Pinkas[1] describes privacy preserving in distributed computing where data is stored on distributed sites. Privacy preservation is applied in order to hide sensitive association rules. The author also mentions the extent of privacy required in any data and explains the role of various Cryptographic techniques for secure distributed computation, and their applications in Data Mining.

Maram[2] discusses the use of cryptography in communication of data in secret manner which means privacy of database. The author discussed about a new technique which used UNICODE for color encryption using private key cryptography for sharing the colors.

Aphtesi [3] discussed Visual Cryptography and Pixel Shuffling techniques. His proposed research work was based on medical images digital encryption which encrypts the image before transmission or storage. According to the author, medical images may be sensitive so encryption of images was done using visual cryptography in which the image was broken up into several parts before encryption.

Panchami et.al. [4] used RGB color model and Unicode for encryption and decryption. Text was decrypted using colors. The color chart was prepared for basic colors and colors were used from the mixture of these colors. RSA algorithm was used for secure transfer of data. The major drawback was that a limited numbers of colors were used that makes the encryption weak. To overcome this drawback RSA algorithm was used for secure data transfer.

Palanisamy and Shanthi [5] proposed that a plain text be encrypted to hex codes by using AES Algorithm and then the output cipher block is divided into 3 characters per block. After that 3 characters are converted to their respective RGB values.

Lee and Tsai [6] proposed a new and secure image transfer technique used to convert a given secret image into a mosaic image of the same size. Mosaic image was equal to randomly selected target image and it can be used as a concealment of the secret image. Techniques were used to keep such experiment lossless at the time of recovery at the receiver end to maintain the data accurate.

Rajesh et.al. [7] proposed two step random color matrix passwords for authentication. Traditional passwords are textual passwords. Such kinds of passwords are vulnerable for an attacker. In their paper, color matrix passwords were used which was stronger and less vulnerable for any kind of algorithm. Color matrix passwords were more robust as compare to ordinary passwords.

Sabeena and Haroon[8] focused on the main issue with image encryption in which images are distorted when we decrypt them at receiver's end. Some of the data is lost due to this distortion. The authors discussed lossless technique for image encryption by using 'Reversing Room before Encryption Method'.

Yuanyuan [9] proposed a new compression and encryption technique based on the fractal dictionary and Julia set technique. According to this technique image compression using fractal dictionary is beneficial because it saves time and also provides receiver a good quality of image reconstruction at receiver end.

3. The Model

3.1 The Model at Data Publisher's Side

Data Publisher receives the data from several distributed data providers. These are the sites that have agreements with the data publisher for sending their data. Data preprocessing is required at data publisher's side if senders at various sites are not aware about the homogeneity of the database. It becomes difficult to convert database from heterogeneous to homogeneous form. If it happens then it will always increase the burden for data publisher. To reduce the burden of data publisher horizontal partitioning must be used by every site, for which, the data must be send in the same assortment by every sender to the publisher. Here the main focus is market basket data. The technique discussed below is most suitable for market basket analysis. Market basket analysis is done to check associations between purchased items by customers in retail store. It is also used to see purchasing behavior of various customers. Each retail store is the site in our scenario. These sites are sending transactions to Data Publisher. Fig.1. shows the model for data publishing.

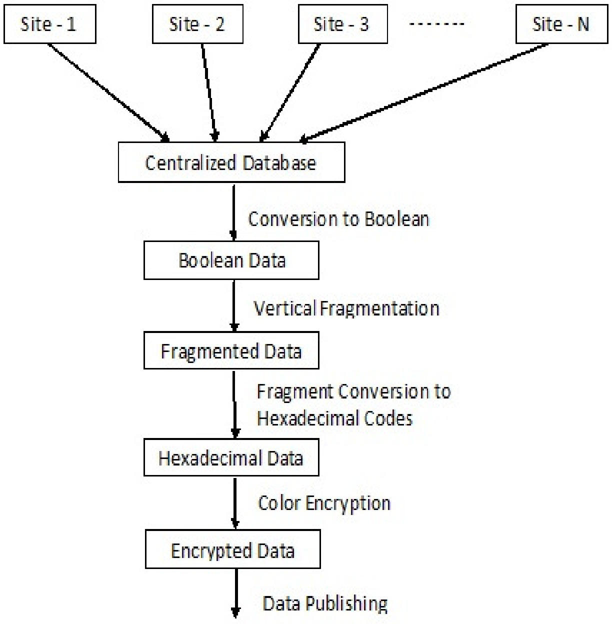


Fig 1 The Data Publishing Model

At the publisher's end, the centralized database is converted to Boolean form where a 0 represent the presence of an item and a 1 represent the absence of an in each transaction. Thereafter, the vertical fragments consisting of N attributes are extracted and converted to hexadecimal codes.

The dimensions are reduced to 1/N. of the original size of the database. The compressed transactions in hexadecimal form are then color encrypted with a predefined color system. The compressed and encrypted data will then be published as a sequence of colors. The data received at the other end will be secure as the color does not have any information about the transactions in the database.

3.2 An Example

Table 1. shows transactional database where T_Id represents the set of items purchased.

Table 1: Transactional Database

T_ Id	Items
T1	I1, I4, I6, I7, I9, I10, I11, I14, I15, I16
T2	I1, I2, I3, I5, I8, I10, I12, I13, I15
T3	I1, I5, I6, I7, I9, I11, I12, I15, I16
T4	I2, I3, I4, I7, I11, I12, I13, I14

The above centralized transactional database is converted to Boolean and compressed as Hex Code Matrix. Advantage of this matrix is that it can summarize the purchasing of four items into a single code. To create hex code matrix information of items I1-I4, I5-I8, I 9 -I12 and I13- I16 is represented into consecutive blocks of hex code matrix respectively. Overall, the size of the database is decreased by a factor of ¼.

Table 2. shows the intermediate transformation of Boolean to hexadecimal codes. The code 96E700, E95A00, 8EB300, 723C00 are converted into respective color codes, which contains the information about 24 items. Each pixel can contains the information of 24 items. Here as the number of items were 16 only, the transactions were padded with zeros to make it 24. Fig 2 shows the sequence of colors finally published using RGB as the color system.

Table 2: Boolean converted to Hexadecimal codes

9	6	E	7	0	0
E	9	5	A	0	0
8	E	B	3	0	0
7	2	3	C	0	0

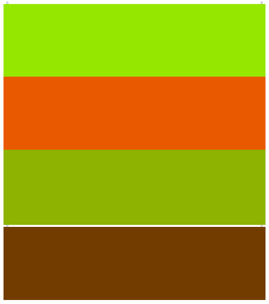


Fig 2 The Published Data

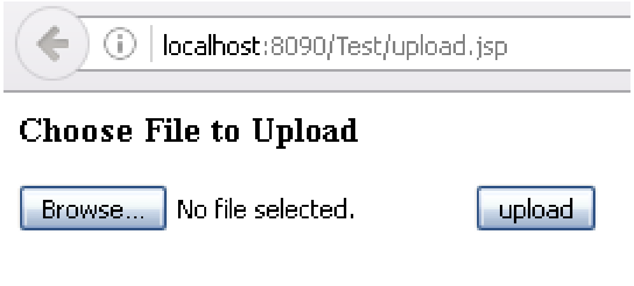
For a resolution of 1024X 768, we can store information about 1024 by 768 by 24 items.

3.3 The Receiver's Side

At the receiver side reverse procedure is followed to get transactions in Boolean form colors. First Color codes are converted to Hex Codes and after that Hex Codes are converted to Boolean form. If number of transactions are less than or equal to size of 24 bit output comes into four colors. If numbers of bits are more than 24 but less than 48, then 8 colors are produced. We can summarize each color into a single pixel. So we can say that 24 bit information of a single transaction can be summarized into one pixel. Next 24 bit information can be summarized to another pixel. So an array of colors in one row of pixels can be created to store the details of one transaction. A second array can be created in same manner to store the detail of second transaction. In such a manner an image can be created. This image is send to miner for retrieval of actual transactions.

4. ComEncColor: The Tool

An algorithm was designed based on the proposed model. A tool has also been developed to implement the algorithm. The tool is named as ComEncColor (CEC). CEC is developed in JAVA which can encrypt transactional information into colors for secure data publishing. CEC takes files with .xls extension as input and generated output as colors. This method is very efficient when the size of transactions is very large. Its interface is very simple. It will simply ask to upload the file for encryption. Select the excel file and run. The interface of the tool is shown as Fig. 3. given below. Initially, it prompts for uploading the file to be encrypted .



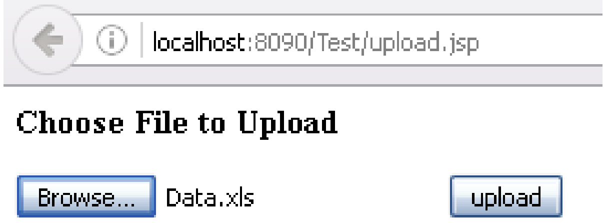
localhost:8090/Test/upload.jsp

Choose File to Upload

No file selected.

Fig 3 Interface of CEC prompting to upload file.

After the file is uploaded, the interface looks like the one given in Fig. 4. below.



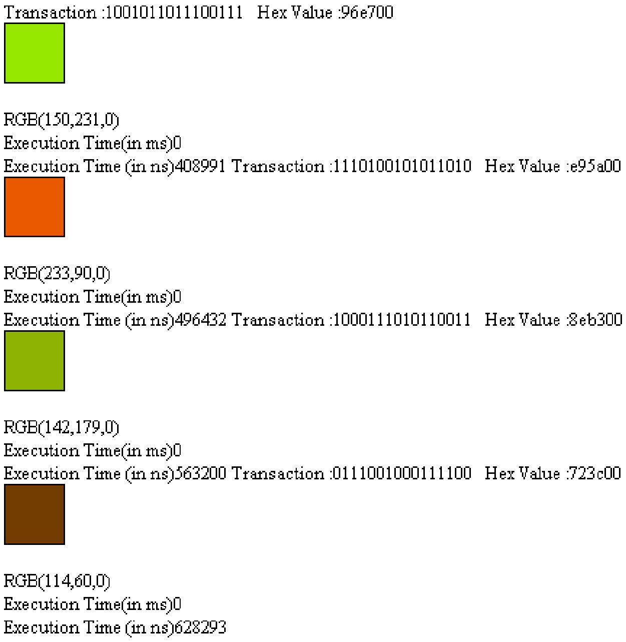
localhost:8090/Test/upload.jsp

Choose File to Upload

Data.xls

Fig 4 Interface of CEC after uploading the file.

In order to understand how it works, let us consider a sample file as shown in Table 1. as the input to CEC. The Transaction T1 is encoded as 96E700, T2 is encoded as E95A00, T3 is encoded as 8EB300 and T4 is encoded as 723C00. The RGB system was chosen as the preferred color system. Finally the output is received as shown in Fig 5, when the number of bits are less than or equal to 24.

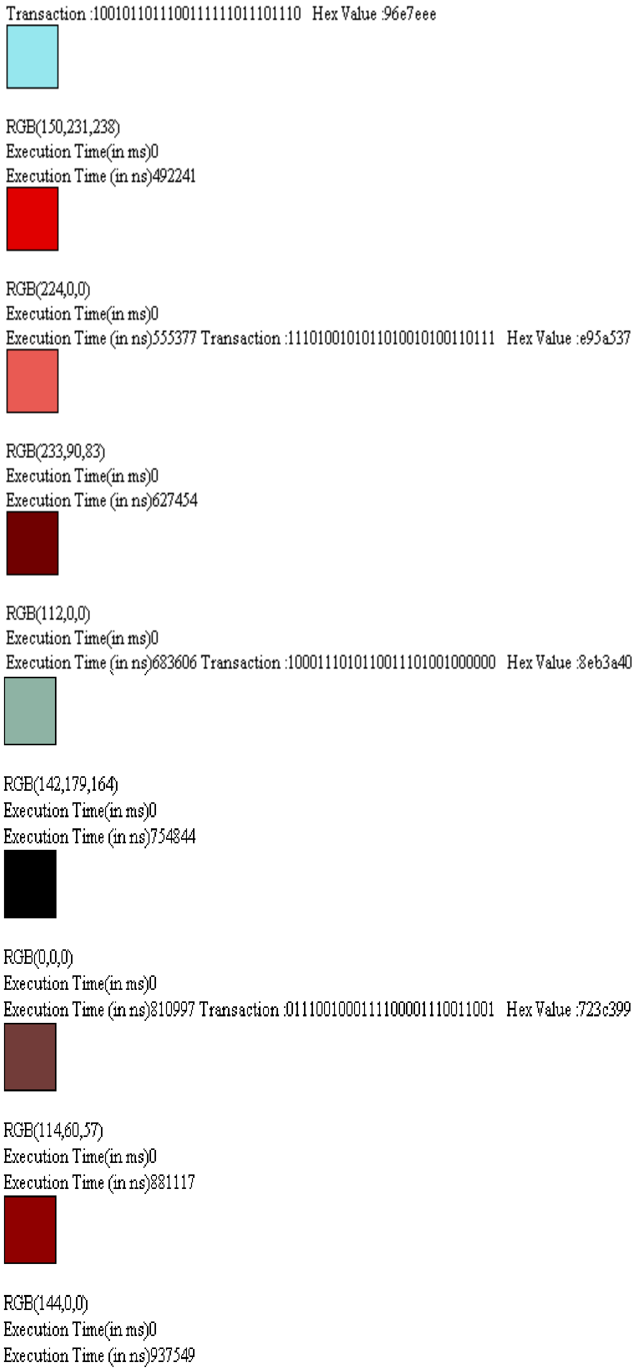


```
Transaction :1001011011100111   Hex Value :96e700

RGB(150,231,0)
Execution Time(in ms)0
Execution Time (in ns)408991 Transaction :1110100101011010   Hex Value :e95a00

RGB(233,90,0)
Execution Time(in ms)0
Execution Time (in ns)496432 Transaction :1000111010110011   Hex Value :8eb300

RGB(142,179,0)
Execution Time(in ms)0
Execution Time (in ns)563200 Transaction :0111001000111100   Hex Value :723c00

RGB(114,60,0)
Execution Time(in ms)0
Execution Time (in ns)628293
```

Fig. 5 Output generated for the sample data when the number of bits are less than equal to 24

The output screen also shows the Boolean form, the RGB code along with the execution time for complete transformation of a database transaction into colors. Fig. 6. shows the output for some sample data where the number of bits were greater than 24.

The Tool CEC encrypts the transactional databases reducing the size of centralized Boolean database and encrypting it efficiently for secure data publishing. A 24 bit information of a single transaction can be summarized into a single pixel without losing accuracy.



```
Transaction :10010110111001111110111011101110   Hex Value :96e7eee

RGB(150,231,238)
Execution Time(in ms)0
Execution Time (in ns)492241

RGB(224,0,0)
Execution Time(in ms)0
Execution Time (in ns)555377 Transaction :11101001010110100101001101111   Hex Value :e95a537

RGB(233,90,83)
Execution Time(in ms)0
Execution Time (in ns)627454

RGB(112,0,0)
Execution Time(in ms)0
Execution Time (in ns)683606 Transaction :1000111010110011101001000000   Hex Value :8eb3a40

RGB(142,179,164)
Execution Time(in ms)0
Execution Time (in ns)754844

RGB(0,0,0)
Execution Time(in ms)0
Execution Time (in ns)810997 Transaction :0111001000111100001110011001   Hex Value :723c399

RGB(114,60,57)
Execution Time(in ms)0
Execution Time (in ns)881117

RGB(144,0,0)
Execution Time(in ms)0
Execution Time (in ns)937549
```

Fig. 6 Encryption by CEC Tool when number of bits greater than 24.

5. Conclusion

In this paper a Tool was developed to encrypt the transactional databases. Technique used is efficient because the database is compressed to another encrypted form with compression but without losing the accuracy of database. Accuracy, security and size of database are major concerns while publishing the data. Secure data publishing is achieved without any loss of accuracy. The information of Hex Code matrix is compressed in such a manner that it becomes more compressed and secure. Such a technique is beneficial when the size of database is very large. In future, few more security levels can be added. Overheads and related challenges are to be discussed in detail as an extension to the proposed work.

Acknowledgments

We express our most sincere thanks to all researchers and authors of various technical papers for sharing their work without which it was very difficult to achieve successful completion. We also thank the department of CSE, Sir Padampat Singhania University, Udaipur for providing us with all necessary laboratory facilities required.

References

[1] Pinkas B. "Cryptographic Journal of Information and Education Technology, Vol. 1, 2011, 137-141.

[2] B. Maram, "Unicode and Colors Integration Tool for Encryption and Decryption", International Journal of Computer Science and Engineering, Vol. 3, 2011, 1197-1202.

[3] K. Q. Aphtesi, "A Visual Cryptographic Encryption Technique for Securing Medical Images", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, 2013, 496-500.

[4] Panchami V, Paul V. and Wahi A, "A New Color Oriented Cryptographic Algorithm based Unicode and RGB Color model". International Journal of Research in Engineering and Technology, Vol. 3, 2014, 82-87.

[5] Palanisamy V. and Shanthi, "A Novel Text to Image Encryption Technique by AES Rijndael Algorithm with Color Code Conversion", International Journal of Engineering Trends and Technology, Vol. 13, 2014.

[6] Y. L. Lee and W.H. Tsai, "A New Secure Image Transmission Technique via Secret- Fragment-Mosaic Images by Nearly Reversible Color Transformations", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 24, 2014, 695-703.

[7] Rajesh N, S Sushmashree, V Varshini, NB Bhavani and D Pradeep, "Color Code Based Authentication and Encryption", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4 , 2015, 403-405.

[8] O.M. Sabeena and R.P. Haroon, " Reversible Data Hiding in Encrypted Color Images by Reversing Room before Encryption with LSB Method", International Journal of Computational Engineering Research, Vol. 4, 2014, 68-72.

[9] S. Yuanyuan, X. Rudan, L. Chen and X. Hu," Image Compression and Encryption Scheme using fractal dictionary and Julia set", IET Digital Library, Vol. 9, 2015, 173-183.

[10] M.M. Laham, "Encryption-Decryption RGB Color Image Using Matrix Multiplication", International Journal of Computer Science & Information Technology, Vol. 7, 2015, 109-119.

[11] R.D. Yalene, N.N. Mhala and B.J. Chilke, "Security Approach by Using Visual Cryptography Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, 2015, 192-195.

[12] M. Karolin and T. Mayappan, "RGB Based Secret Sharing Scheme in Color Visual Cryptography", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, 2015, 151-155.

[13] A.M. Sharadqah, "RGB Color Image Encryption-Decryption Using Gray Image", International Journal Techniques for Privacy Preserving Data Mining "SIGKDD Explorations 2002, Vol. No. 4, 2002, 12-19.

[14] K Sakthidasan and B.V.S. Krishna, "A New Chaotic Algorithm for Image Encryption and Decryption of Digital Color Images", International Journal of Computer Science Issues", Vol. 12, 2015, 137-139.

[15] A.G. Gokul and N. Kumaratharan, "A Secure and Verifiable Visual Cryptography for Color Images", Universal Journal of Electrical and Electronic Engineering, Vol. 3, 2015, 165-172.

Gaurav Kumar Ameta is a Research Scholar, Department of Computer Science and Engineering at Sir Padampat Singhania University, Udaipur, India. His area of specialization is Data Mining, Computer Graphics, Cryptography, and Privacy Preservation.

Divya Bhatnagar is working as Professor in the department of Computer Science and Engineering in School of Engineering, Sir Padampat Singhania University, Udaipur, India. She holds 18 years of teaching and research experience. Her specialization areas include data mining and Neural Networks.