Privacy Preserving Mining of Transaction Databases

Sunil R[1] Dr. N P Kavya[2]
[1]M.Tech Scholar [2]Prof. & Head, MCA Dept.
[1,2] Department of Computer Science and Engineering
[1, 2] RNS Institution of Technology, Bangalore VTU, Belgaum, Karnataka, India

Abstract— In this paper, A company (data owner) lacking in expertise or computational resources can outsource its mining needs to a third party service provider (server) taking the advantage of cloud computing. However, a business can have many suppliers and/or customers and may have a set of transactions associated with each one and also the items in the outsourced database and the patterns of items that can be mined from the database are considered as the private property of the corporation (data owner). To protect the corporate privacy, the data owner transforms its data and ships it to the server. The server sends extracted patterns to the owner in response to the mining queries. The owner recovers the true patterns from the extracted patterns received. Here the problem of outsourcing the association rule mining task within a corporate privacy-preserving framework is studied. Proposed model based on background knowledge of Fast Encryption Algorithm (FEAL) Cryptosystem an encryption/decryption strategy for providing security and hence enabling privacy-preserving of outsourced mining task, and protection against the privacy violation attack.

Keywords: - FEAL 4, k-anonymity, l-diversity, Cryptography, Randomization Method, Privacy – preserving outsourcing.

I. Introduction

Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. In recent years, privacy-preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. Data mining has been viewed as a threat to privacy because of the widespread proliferation of electronic data maintained by corporations. This has led to increased concerns about the privacy of the underlying data. In recent years, a number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy.

Most traditional data mining techniques analyze and model the data set statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure individual data records. This domain separation points to the technical feasibility of PPDM. Historically, issues related to PPDM were first studied by the national statistical agencies interested in collecting private social and economical data, such as census and tax records, and making it available for analysis by public servants, companies, and researchers. Building accurate socio economical models is vital for business planning and public policy. Yet, there is no way of knowing in advance what models may be needed, nor is it feasible for the statistical agency to perform all data processing for everyone, playing the role of a trusted third party. Instead, the agency provides the data in a sanitized form that allows statistical processing and protects the privacy of individual records, solving a problem known as privacy-preserving data publishing.

With the advent of cloud computing and its model for IT services based on the internet and big data centre, the outsourcing of data and computing services is acquiring a novel relevance, which is expected to skyrocket in the near future. Business intelligence and knowledge discovery services, such as advanced analytics based on data mining technologies, are expected to be among the services amenable to be externalized on the cloud, due to their data intensive nature, as well as the complexity of data mining algorithms. Thus, the paradigm of mining and management of data as service will presumably grow as popularity of cloud computing grows [6]. This is the data mining-as-a-service paradigm, aimed at enabling organizations with limited computational resources and/or data mining expertise to outsource their data mining needs to a third party service provider. Although it is advantageous to achieve sophisticated analysis on tremendous volumes of data in a cost-effective way, there exist several serious security issues of the data-mining as-a-service paradigm. One of the main security issues is that the server has access to valuable data of the owner and may learn sensitive information from it. For example, by looking at the transactions, the server (or an intruder who gains access to the server) can learn which items are always co-purchased. However, both the transactions and the mined patterns are the property of the data owner and should remain safe from the server. This problem of protecting important private information of organizations/ companies is referred to as corporate privacy [5]. Unlike personal privacy, which only considered the protection of the personal information recorded about individuals, corporate privacy requires that both the individual items and the patterns of the collection of data items are regarded as corporate assets and thus must be protected. Here studied the problem of outsourcing the association rule mining task within a corporate privacy preserving framework.

A substantial body of work has been done on privacy-preserving data mining (PPDM) in a variety of contexts. A common characteristic of most of the previously studied frameworks is that the patterns mined from the data (which may be distorted, encrypted, anonymized, or otherwise transformed) are intended to be shared with parties other than the data owner. The key distinction between such bodies of work and this problem is that, in the latter, both the underlying data and the mined results are not intended for sharing and must remain private to the data owner.

II. Related Work

The research of PPDM has caught much attention recently. The main model here is that private data is collected from a number of sources by a collector for the purpose of consolidating the data and conducting mining. The collector is not trusted with protecting the privacy, so data are subjected to a random perturbation as it is collected. Techniques have been developed for perturbing the data so as to preserve privacy while ensuring the mined patterns or other analytical properties are sufficiently close to the patterns mined from original data. This body of work was pioneered by [9] and has been followed up by several papers since [10]. This approach is not suited for corporate privacy, in that some analytical properties are disclosed.

Another related issue is secure multiparty mining over distributed datasets. Data on which mining is to be performed is partitioned, horizontally or vertically, and distributed among several parties. The partitioned data cannot be shared and must remain private but the results of mining on the union of the data are shared among the participants, by means of multiparty secure protocols [11]–[13]. They do not consider third parties. This approach partially implements corporate privacy, as local databases are kept private, but it is too weak for our outsourcing problem, as the resulting patterns are disclosed to multiple parties.

III. Basic Concepts Of Cryptographic Approach To PPDM

The cryptographic approach to PPDM assumes that the data are stored at several private parties who agree to disclose the result of a certain data mining computation performed jointly over their data. The parties engage in a cryptographic protocol; that is, they exchange messages encrypted to make some operations efficient while others computationally intractable. In effect, they blindly run their data mining algorithm. Classical works in secure multiparty computation such as Yao (1986) and Goldreich, Micali, and Wigderson (1987) show that any function F(x1, x2, …, xn) computable in polynomial time is also securely computable in polynomial time by n-parties, each holding one argument, under quite broad assumptions regarding how much the parties trust each other. However, this generic methodology can only be scaled to database-sized arguments with significant additional research effort. The first adaptation of cryptographic techniques to data mining is done by Lindell and Pinkas (2000) for the problem of decision tree construction over horizontally partitioned data; it was followed by many papers covering different data mining techniques and assumptions. The assumptions include restrictions on the input data and permitted disclosure, the computational hardness of certain mathematical operations such as factoring a large integer, and the adversarial potential of the parties involved: The parties may be passive (honest but curious, running the protocol correctly but taking advantage of all incoming messages) or malicious (running a different protocol), some parties may be allowed to collude (represent a single adversary), and so forth. In addition to the generic methodology such as oblivious transfer and secure Boolean circuit evaluation, the key cryptographic constructs often used in PPDM include homomorphic and commutative encryption functions, secure multiparty scalar product, and polynomial computation. The use of randomness is essential for all protocols. The privacy guarantee used in this approach is based on the notion of computational indistinguishability between random variables. Let X_k and Y_k be two random variables that output Boolean vectors of length polynomial in k; they are called computationally indistinguishable if for all polynomial algorithms A_k(alternatively, for any sequence of circuits of size polynomial in k), for all c> 0, and for all sufficiently large integers k,

$$Prob\ [A_k\ (X_k) = 1] - Prob\ [A_k\ (Y_k) = 1]\ |\ < 1 / kc.$$

The above essentially says that no polynomial algorithm can tell apart X_k from Y_k. To prove that a cryptographic protocol is secure, we show that each party’s view of the protocol (all its incoming messages and random choices) is computationally indistinguishable from a simulation of this view by this party alone. When simulating the view of the protocol, the party is given everything it is allowed to learn, including the final data mining output. The exact formulation of the privacy guarantee depends on the adversarial assumptions. Goldreich (2004) and Stinson (2006) provide a thorough introduction into the cryptographic framework.

Scalability is the main stumbling block for the cryptographic PPDM; the approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

IV. Proposed Framework

First developed an encryption scheme, called FEAL 4, that the encrypt/decrypt(E/D) module can employ to transform client data before it is shipped to the server.

Second, to allow the E/D module to recover the true patterns and their correct support up on receiving mining query, here proposed that it creates and keeps a compact structure, called synopsis.

Third, then conducted a formal analysis based on attack model and proved that the probability that an individual item, a transaction, or a pattern can be broken by the server can always be controlled by FEAL 4 encryption/decryption scheme.

A. System Architecture

Here client can be represented as a company (data owner) lacking in expertise or computational resources can outsource its mining needs server, this server is the third party service provider which conducts data mining and sends the (encrypted) patterns to the owner or client. Encrypt/ Decrypt(E/D) module which is responsible for transforming the input data into an encrypted database from the client or data owner then according to encryption scheme has the property that the returned supports are not true supports. The E/D module recovers the true identity of the returned patterns [7].

Modules used are Client/ Data owner, Encrypt/ Decrypt, Centralized Server/ Service provider. These are explained below and shown in Fig. 1.

1) Client/ Data owner

In this module client inputs data a raw file, ie., transaction database. Here the data owner sends data and mining query to the application that is privacy preserving process, and this application will perform all the operation and sends the result to data owner. Transaction databases are in .csv format.

2) Encrypt/ Decrypt

The module is about the encryption and decryption, where up on receiving the data owner data it will encrypt it and stores in server and based on mining query of the client it performs the decryption on returned pattern from server to get true patterns.

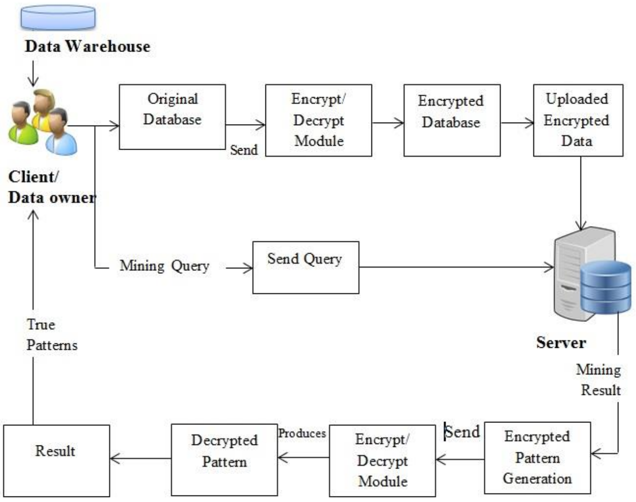

Fig. 1. The System Architecture

3) Centralized Server/ Service Provider

With the advantage of cloud computing the Centralized Server/ Service Provider is a third party service provider which provides the Specific human resources or Technological resources like Storage, processor etc., When a Company(data owner) does not have In-house expertise for doing data mining or Computing infrastructure.

B. Data Owner Side Flow Chart for Sending Datasets/ Mining queries

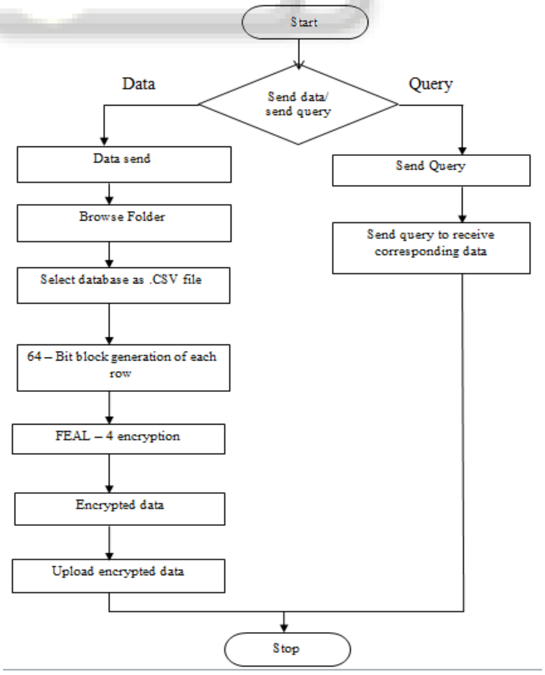

Fig. 2: Data owner side flow chart for outsourcing needs

Fig. 2. shows the data owner side flow chart for sending datasets/mining queries, above algorithm checks whether if the data owner wants to outsource its data to server or the mining queries to the server, if it was the data means then it allows the data owner to select a data file by browsing a data file with in the paradigm of data mining as a service data owner selects a .csv file and then forwards to encrypt/ decrypt module. Since this encrypt/ decrypt module that is the FEAL – 4 algorithm use 64-bit plain text and 64-bit key to encrypt and decrypt the given plain text, so the algorithm then performs a 64 –bit block generation of each row, after this FEAL – 4 encrypt module will perform encryption on the each generated 64-bit block and then encrypted data is uploaded to server.

C. Server Side Flow Chart

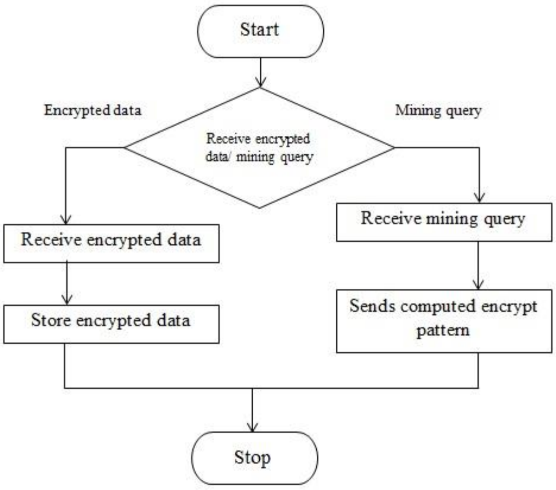

Fig. 3: Server side flow chart for storing data or receiving mining query

Fig. 3. shows the server side flow chart for storing data or receiving mining query, if received data is encrypted data means then algorithm stores the encrypted data, when server receives mining query it performs required computation based on the mining query and sends computed encrypted pattern to data owner.

D. Encryption/ Decryption Scheme

In this section, we introduce the encryption scheme, called FEAL 4, which transforms a TDB D into its encrypted version D∗.

FEAL is a Fast Encryption Algorithm (FEAL) is a symmetric encryption algorithm, also called as Japanese Encryption algorithm. FEAL works almost similar to Data Encryption Standard algorithm (DES), but it is faster than DES. FEAL works in different standards like FEAL-4, FEAL-6 and so on up to FEAL-n. Here, “n” indicates the number of Feistel permutation rounds. The function in FEAL-4 uses two S-box functions in general these S-box are used based on inputs such as two bytes x,y and delta[8] .

In general S-box is

$S(x,y,delta) = ((x+y+delta)mod256)ROT2$

Where if delta = 0 then,

$S0\ (x,y) = ((x+y)mod\ 256)<<2$

If delta = 1 then,

$S1\ (x,y) = ((x+y+1)mod\ 256)<<2$

Example: (delta = 1)

$S(X,\ Y,\ delta) = ROT2((X + Y + delta)\ mod\ 256\ )$

$X = 00010011\ and\ Y = 10110011$

$$\begin{array}{rr} & 00010011 \\ + & 10110011 \\ + & 1 \end{array}$$

= 11000111 Rot2 of result is 00011111

Function f(a, b) as shown in Fig. 4. is meant for performing the linear functionality in FEAL encryption algorithm according to the Fig. 5. Key generation function f_k as shown in Fig. 7. uses both the S-boxes. The 64-bit key is divided into two equal parts of 32-bit each represented by a and b respectively. The keys a and b are further divided into 8-bits of the form a1, a2, a3, a4, b1, b2, b3 and b4. The key generation function f_k is shown in Fig. 6.

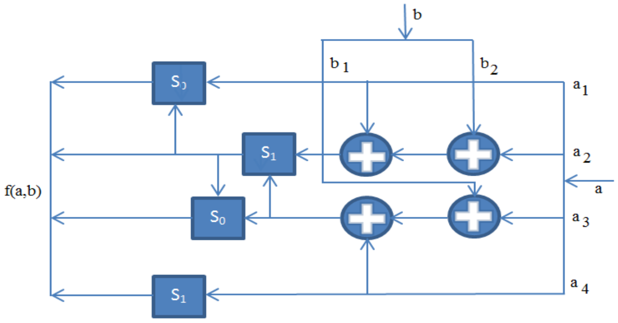

Fig. 4: Function f(a,b)

The iterations of linear encryption functions in basic FEAL algorithm use 64-bit plain text and 64-bit key to encrypt and decrypt the given plain text.

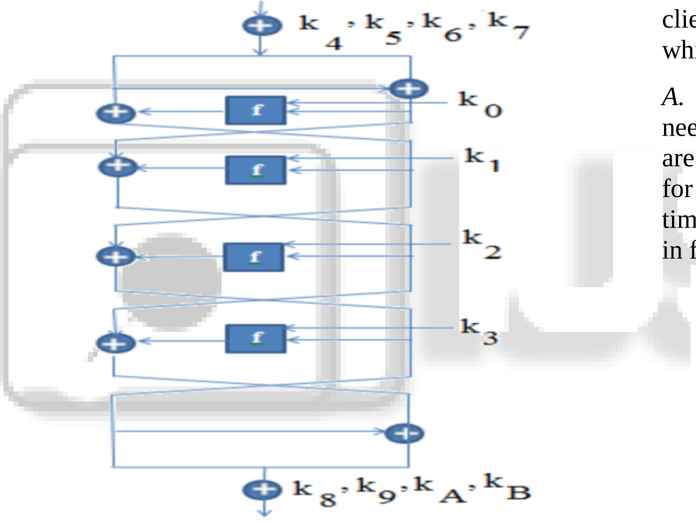

Fig. 5: Encryption and Decryption process using FEAL

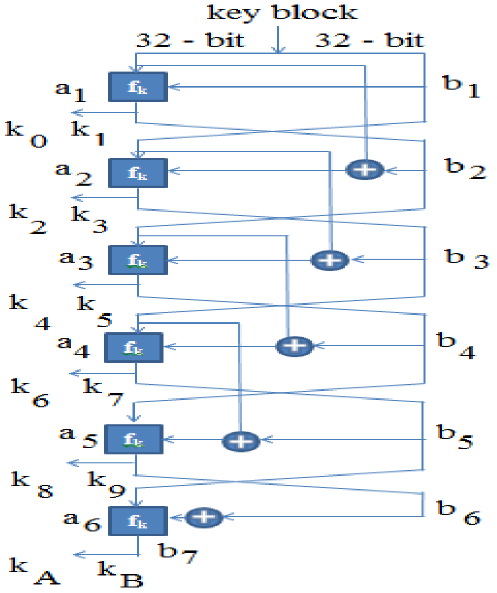

Fig. 6: Key Generation

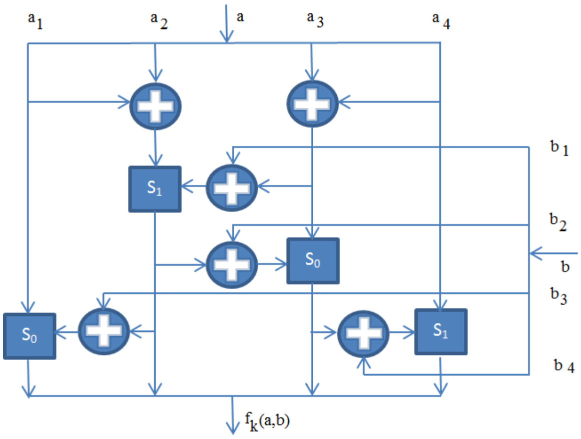

Fig. 7: Key generation function f_k

V. Experiments

Experiments were conducted on a real-world database. The implemented FEAL 4 encryption scheme, as well as the decryption scheme performs encryption and decryption on this example database. Here we ignore the time for transmitting TDBs between the client and server as we assume that the TDB streams into the ED module and the client can send the data that has been encrypted to the server while encrypting the remaining data.

A. Encryption Overhead: First, we assessed the total time needed by the ED module to encrypt the database, timings are reported in Figure 8 different number of transactions and for different database size the results shows that encryption time is always small when compared with AES is as shown in figure 10.

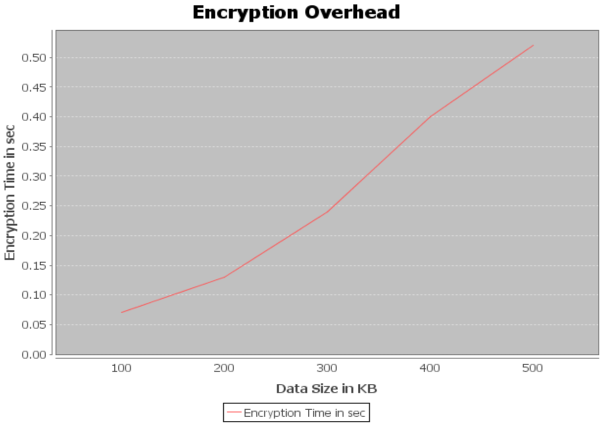

Fig. 8: Encryption Overhead

B. Decryption Overhead: We now consider the feasibility of the proposed outsourcing model. The ED module encrypts the TDB once which is sent to the server. Mining is conducted repeatedly at the server side and decrypted every time by the ED module ED and timings are reported in Figure 9 decryption time is always small when compared with AES is as shown in figure 11.

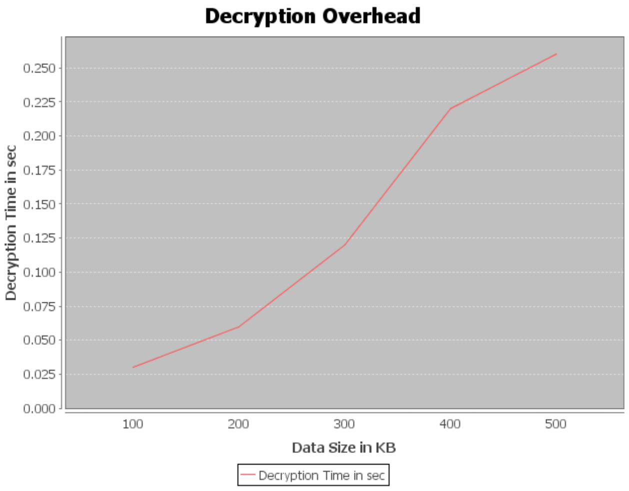

Fig 9: Decryption Overhead

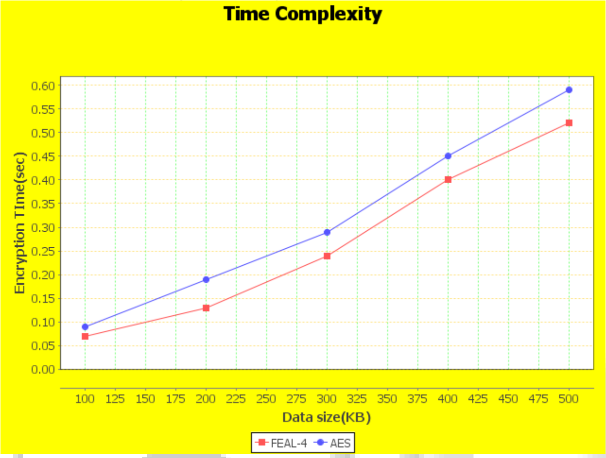

Fig 10: Encryption time of FEAL-4 and AES

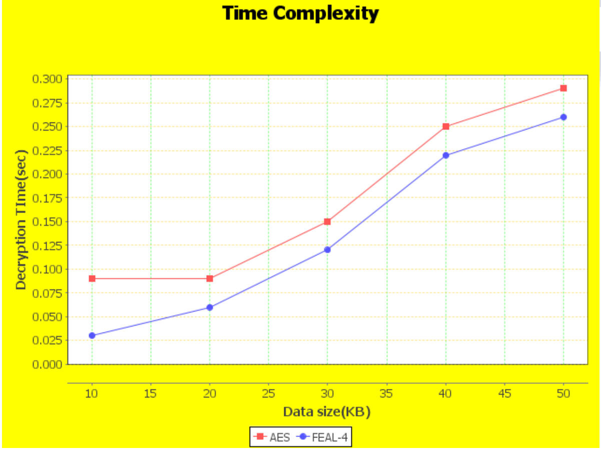

Fig 11: Decryption time of FEAL-4 and AES

VI. CONCLUSION

We studied the problem of (corporate) privacy-preserving mining of frequent patterns (from which association rules can easily be computed) on an encrypted outsourced TDB, and also studied problem of mining task on an encrypted outsourced TDB. Here proposed an encryption scheme called FEAL – 4 for providing corporate privacy for outsourced data. Here since attacker has access only to the encrypted items this method is robust against an adversarial attack based on the original items.

FEAL works almost similar to Data Encryption Standard (DES) algorithm, but it is faster than DES. To overcome the problem of security issues associated with AES algorithm a modification is done by adjusting the Shift Row Transformation.

It would be interesting to enhance the framework and the analysis by appealing to cryptographic notions such as perfect secrecy that is it could be interesting to consider other attack models where the attacker knows some pairs of items and their cipher values.

ACKNOWLEDGEMENT

We would like to thank Director Dr. H N Shivashankar, Principal Dr. M K Venkatesha and Dr. G.T Raju, professor and Head, Dept of Computer Science and engineering, RNSIT, for constant encouragement in for implementing this project and pursuing this paper.

REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[2] Hegland, M., Algorithms for Association Rules, Lecture Notes in Computer Science, Volume 2600, Jan 2003, Pages 226 – 234

[3] Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, 487-499.

[4] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms.ACM PODS Conference, 2002.

[5] C. Clifton, M. Kantarcioglu, and J. Vaidya, “Defining privacy for data mining,” inProc. Nat. Sci. Found. Workshop Next Generation Data Mining, 2002, pp. 126–133.

[6] R. Buyya, C. S. Yeo, and S. Venugopal, “Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities,” inProc. IEEE Conf. High Performance Comput. Commun., Sep. 2008, pp. 5–13.

[7] F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, “Privacy-preserving outsourcing of association rule mining,” ISTI-CNR, Pisa, Italy, Tech Rep. 2013.

[8] Nithin N, Anupkumar M Bongale, G. P. Hegde, “Image Encryption based on FEAL algorithm”, International Journal of Advances in Computer Science and Technology, Volume 2, No.3, March 2013.

[9] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 439–450.

[10] S. J. Rizvi and J. R. Haritsa, “Maintaining data privacy in associationrule mining,” in Proc. Int. Conf. Very Large Data Bases, 2002, pp. 682–693.

[11] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,” IEEE Trans.Knowledge Data Eng., vol. 16, no. 9, pp. 1026–1037, Sep. 2004.

[12] B. Gilburd, A. Schuster, and R. Wolff, “k-ttp: A new privacy model for large scale distributed environments,” in Proc. Int. Conf. Very Large Data Bases, 2005, pp. 563–568.

[13] P. K. Prasad and C. P. Rangan, “Privacy preserving birch algorithm for clustering over arbitrarily partitioned databases,” in Proc. Adv. Data Mining Appl., 2007, pp. 146–157.

IJSRD