IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN: 2278-487X, p-ISSN: 2319-7668 (April, 2017)
PP 01-06
www.iosrjournals.org

Privacy Preserving Horizontal partitioning of outsourced database for frequent pattern mining using paillier

Manasi Dhage[1], Dr. A.B.Banubakode[2]
[1] *(Department of Computer Engineering,*
RajarshiShahu College of Engineering, Pune
SavitribaiPhule, Pune University)

Abstract: *In recent years data sharing is an important task. This data publication is conducted by various organizations under some important rules and regulations. such data is useful in various researchers. Also With high demand for cloud services there are serious concerns about the privacy of individuals and also the outsourced database. In such cases there is high demand to check the utility of data and integrity of data. Data integration is another form of data sharing, where data owners sends there data to server for aggregation and then performed preprocessing on aggregated data before storing on third party server or cloud. There are several challenges while designing secure system like hiding the original and sensitive information of the individual and whole database from attacker. So in this system to tackle these challenges, the proposed system uses the technique of homomorphic encryption which works on encrypted data, which results in increased security of outsourced data and also increased system performance. We are adding the concept of horizontal partitioning to spilt the database in two different parts as well as apply the rule generation algorithm. By making use of horizontal partition we are able to save the time needed for rule generation.*

Keywords: *Collaborative data publishing, Differential privacy, Horizontal partitioning, Homomorphic Encryption, Utility verification***.**

I. Introduction

The demand for gathering and sharing information is increment strongly because of the quick development of data. An large amount of information is utilized for investigation, insights and calculation to discover general pattern or guideline which is advantageous to social improvement and human advancement. In the interim, dangers emerge when enormous information accessible for the general population. For instance, individuals can burrow protection data by getting together sheltered appearing information, subsequently; there is an extraordinary plausibility uncovering people security. As per the study, roughly 87 % of the number of inhabitants in the United States can be uniquely distinguished by given dataset distributed for people in general[8]. To stay away from this circumstance deteriorating, measures are taken by security division of numerous nations, for instance, declaring protection direction. The prerequisite for information distributer is that information to be published must fit for the predefined conditions. Distinguishing credit should prevent from distributed dataset to ensure that individual's security can't be construed from dataset straightforwardly. Expelling identifier trait is only the readiness work of information handling, a few cleansing operations should be done further. Though, after information preparing, it might be diminish information utility significantly, while, information security did not get completely protected.

In face of the challenging risk, some researchers have been proposed as a remedy of this uncomfortable circumstance, which focus at achieving the equalization of information utility and data security when publishing dataset. The continuous examination is called Privacy Preserving Data Publishing (PPDP). In the previous couple of years, specialists have responded to the call and undertaken a lot of examines. Numerous attainable methodologies are proposed for various security saving situation, which illuminate the issues in PPDP successfully. New strategies and hypothesis turn out consistently in experts' effort to complete privacy preserving.

In this paper we study about the related work done, in section II, the proposed approach modules description, mathematical modeling, algorithm and experimental setup in section III .and at final we provide a conclusion in section IV.

II. Literature Survey

In this section discuss the existing method developed for privacy preserving of data. Now we discuss different methods developed by the researchers, the different methods are as follows:

K-Anonymity

The various phenomena arise when analyzing and publishing the data in high-dimensional space. K-Anonymity was a technique to hold up the blight. Speculation on K-Anonymity was connected to cover the careful estimation of an attribute [2]. The annoyance strategy on K-Anonymity was reasonable for total circulation of an individual than the inter-attribute relation of an individual. 2-anonymity and Gaussian cluster strategies proposed on K-Anonymity strategy, guarantee protection by assessing likelihood and assigning its value to zero. As per the authors view, this method tried to understand the probability distribution which would have maximum likelihood of its attributes. As per the authors, there would be a loss for high-dimensional data.

□-Diversity

Data around an individual couldn't be distributed without uncovering delicate attribute [3].

K-Anonymization was insufficient to secure the information which incorporate homogeneity attack and background learning attack. ℓ-Diversity method portrayed that sensitive attribute would have at most same recurrence. For instance, with positive disclosure, if Alice needs to find Bob, Alice would decide Bob with high-likelihood appropriation. The negative exposure would happen when an adversary could accurately dispense with some conceivable estimation of the delicate traits. There could be a minimum distinction between the earlier conviction and back conviction.

T-Closeness

Anil Prakash, RavindarMogili found that K-Anonymity and ℓ - Diversity was not used to avert quality revelation [4]. ℓ-Diversity would have very much represented sensitive characteristic esteem that was allotted just with certain number of restrictions. t-closeness has been proposed to depict the appropriation of sensitive characteristic with comparability class. Earth Mover Distance was used to gauge the separation between the two probabilistic distributions. Conjunction has been proposed to consolidate machine learning and factual investigation. Closeness among the segments was diminished by aggregation.

Km Anonymity

Km Anonymity has been proposed for an anonymize value-based database [5]. Km Anonymity go for ensure the database against an enemy who knows about m items in the transaction. The speculation was utilized to keep up the set esteemed information. For any exchange on K-1 records, other indistinguishable transaction would likewise show up. Km secrecy has been presented by means of top down nearby speculation procedure to record the quantity of exchange records. The segment based methodology was utilized to gathering (parcel) the comparative items in a top-down way. The km secrecy model would keep protection breaks raised from an enemy who might discovered m things in an transaction database.

Distributed K-Anonymity framework (DKA)

The gathering of information from various locales can't be shared specifically. The key step was to anonymize the data in order to generalize a specific value [6]. A protected 2-party structure was intended for multiparty calculation that has been utilized to join the dataset from various sites. Appropriated K-Anonymity prevent recognizable proof of an individual by make utilization of worldwide Anonymization in the encrypted form. DKA give a protected structure between two parties. Two parties would agree on Global Anonymization algorithm that could produce local Anonymous dataset. Additionally, DKA give a protected dispersed protocol which would require that two parties could commonly semi-honest. Still the exchange off amongst utility and capability of information was misused in DKA.

K-Anonymity Clustering

Among various clustering methods, hierarchal clustering was mostly used to achieve KAnonymity. Weighted Feature C-means Clustering [WFC] utilized to diminish the data deformation. WFC segment all records into proportionality class and would combine the class utilizing class merging mechanism [7]. The numerical values of quasi identifier were used to evaluate the Weighted Feature C-means Clustering technique. The authors also try to provide the dissimilarity evaluated approach which would take different types of feature values for class merging mechanism.

Differentially private mechanism

Privacy-Preserving Utility Verification of the Data Published by Non-interactive Differentially Private Mechanisms [1] proposed privacy-preserving utility verification mechanism based upon cryptographic technique for DiffPart-a differentially private scheme designed for set-valued data. The mechanism Improve the security and efficiency of the system but Association rule mining over huge data may increase the execution time. The technique has side effect on utility of data or data publisher may even cheat in this process of anonymization.

III. Proposed Approach

We presented homomorphic Paillier encryption and secure apriori Association rule creation method on horizontal partitioned dataset for privacy preserving mining of frequent patterns from outsourced transaction database. The implementation details proposed system are shown below.

A. Proposed System Overview

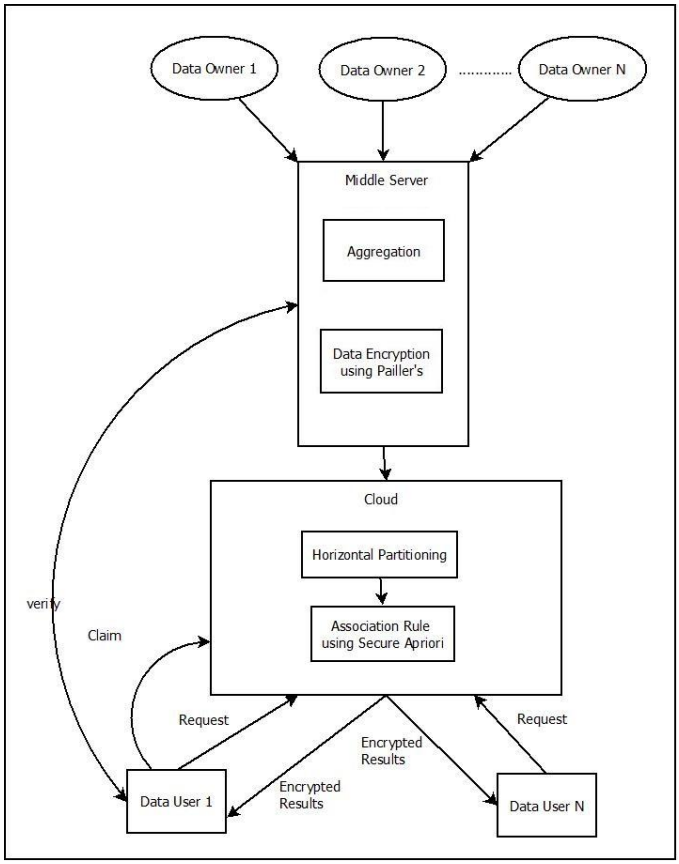

Figure 1. Proposed System Architecture

Techniques used to implement this system:
The system consisting of following modules:

1. Input Dataset and Aggregation
   Input:- Original data files
   Output:- Aggregated data.
   In this module the data owners provide the Set-valued Data as a dataset and send it to the middle client or server where all data is aggregated.

2. Encryption and Horizontal Partitioning
   Input:- Aggregated data.
   Output:- Encrypted Horizontally Partitioned data.
   After the data aggregation process, the aggregated data is encrypted using public key paillier encryption algorithm and send it to third party server or cloud where it is horizontally partitioned.
   Paillier cryptosystem is an additively homomorphic public key encryption scheme.

3. Rule Generation

Input:- Encrypted Horizontally Partitioned data.
Output:- Encrypted Rules
Procedure:-
After horizontally data partitioning the secure apriori algorithm is implemented for rule generation which is works on encrypted dataset. So Third party server or cloud does not get any original information related to dataset and secure transactions are carried out here.

4. Decryption

Input:- Encrypted Rules
Output:- Decrypted Rules
Procedure:-
In this module all encrypted rules decrypted using private key generated in paillier homomorphic cryptosystem algorithm.

5. Verification and Claiming

Input:- Decrypted Rules
Output:- Verified Rules
This model is implemented to verify that generated rules are valid or not, is server cheated with users or not is verify here.

B. Algorithm

Algorithm 1: Apriori Algorithm

Apriori (T, ϵ)
$L_1 \leftarrow$ {large 1 - itemsets}
$K \leftarrow 2$
while $L_{k-1} \neq \phi$
$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \dot{\subseteq} c \wedge |s| = k - 1\} \dot{\subseteq}! L_{k-1}\}$
for transaction $t \in T$
$C_t \leftarrow \{c \mid c \in C_k \wedge c \dot{\subseteq} t\}$
for candidates $c \in C_t$
count [c] $\leftarrow$ count [c] + 1
$L_k \leftarrow \{c \mid c \in C_k \wedge \text{count [c]} \geq \epsilon\}$
$k \leftarrow k + 1$
return $\bigcup_k L_k$

Algorithm 2: Paillier Algorithm

1) Key generation :

a) Select two large prime numbers a and b arbitrary and independent of each other such that gcd(n, Φ (n)) = 1, where Φ (n) is Euler Function and n=ab.
b) Calculate RSA modulus n = ab and Carmichael's function is given by λ = lcm(a-1, b-1).
c) Select g called generator where $g \in \mathbb{Z}^*_{n2}$ Select α and β randomly from a set $\mathbb{Z}_n^*$ then calculate g = $(\alpha n + 1)\ \beta^n m \bmod n^2$.
d) Compute the following modular multiplicative inverse $\mu = (L\ (g^{\lambda} \bmod n^2)^{-1} \bmod n$. Where the function L is defined as L(u) = $(u-1)/n$.

The public (encryption) key is (n and g).
The private (decryption) key is (λ and μ).

2) Encryption:

a. Let mess be a message to be encrypted where $mess \in \mathbb{Z}_n$.
b. Select random r where $r \in \mathbb{Z}^*_{n2}$.
c. The cipher text can be calculated as: cipher = $g^{mess} \cdot r^n \bmod n^2$.

3) **Decryption:**

a. Cipher text $c \in \mathbb{Z}^*_{n^2}$
b. Original message: mess = L (cipher$^{\lambda}$ mod n^2).μ mod n.

C. Mathematical Model

Set Theory
Let S be a system, such that, S= {Input, Process, Output}

Input:
I = {i1}
Where,
i1 = BMS-POS Dataset

Process:
P = {P_1, P_2, P_3, P_4, P_5,}
Where, P represent the total number of steps perform in system to get output.

P1 = Data Aggregation

P2 = Data Encryption
In this step, the aggregated data encrypted using Paillier encryption algorithm.

P3= Horizontal Partitioning
In horizontal partitioning all encrypted data is horizontally partitioned.

P54= Frequent Patterns Rules
P5 = {$P5_1$, $P5_2$,...,P5n}
Where, P5 represents the set of frequent patterns rules.

Output:
Verified Rules
R = {R1,R_2, ,...,R_n}
Where, R represents the set of verified rules.

Mathematical Equations
Key Generation:

$$n = 2pq + 1$$

Encryption:

$$E(m,r) = g^m h^r$$

Decryption:

$$E(m)^q = (g^m, h^r) = (g^q)^m$$

Homomorphic addition:
$E(m1 + m2) = g^{m1+m2} h^{r1+r2}$=E(m1)E(m2)

IV. Result And Discussion

A. Experimental Setup

The system is built using Java framework on Windows platform. The Net beans IDE are used as a development tool. The system doesn't require any specific hardware to run ,any standard machine is capable of running the application.

B. Dataset

The proposed system used Set-value data. In that we create a more than one number of client and they contain the client data.

C. Expected Result

In this section discussed the experimental result of the proposed system.

In table 1 shows the time requires for implementing the proposed and existing system according to mathematical model. Implementation time of the existing system is more than the proposed system.

Table 1: Time Comparison

System	Time in ms
Existing system without horizontal partitioning	10000
Proposed system with horizontal partitioning	5000

Following figure 1 shows the comparison of proposed system and existing system on the basis of their implementation time and mathematical model equations. From the graph it shows that time of the proposed system is less than the time of the existing system.

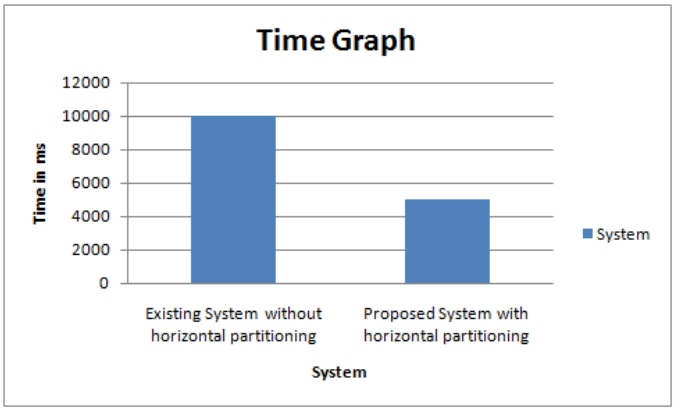

Fig. 2: Time Graph

V. Conclusion

In this system we proposed Homomorphic Encryption I e. paillier cryptosystem method that are suitable for outsoursed transactional association rule mining, which preserves security and free from man in middle attack. Also the issue of verifying utility of mined frequent patterns is checked. Finally, the concept of horizontal partitioning this proposed o spilt the database in two different parts and secure apriori rule generation algorithm is applied. By making use of horizontal partition we are able to save the time needed for rule generation and also increased system performance. Experimental results show that the proposed system has better performance in terms of time, security and rule generation than the existing system. We extend our work for relational database.

Acknowledgements

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule University, Pune and concern members of cPGCON2017 conference, organized by, for their constant guidelines and support. We are also thankful to the reviewer for their valuable suggestions.

References

[1] Jingyu Hua, An Tang, Yixin Fang, Zhenyu Shen, and Sheng Zhong, ``Privacy-Preserving Utility Verification of the Data Published by Non-interactive Differentially Private Mechanisms", IEEE Transactions on Information Forensic and Security.

[2] Charu C. Aggarwal, (2005), "On k-Anonymity and the Curse of Dimensionality", Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp.901-909

[3] AshwinMachanavajjhala, Daniel Kifer, Johannes Gehrke, MuthuramakrishnanVenkita Subramanian, (2006), "l-Diversity : Privacy Beyond K-Anonymity", Proc. International conference on Data Engineering.(ICDE),pp.24.

[4] Anil Prakash, RavindarMogili ,(2012), "Privacy Preservation Measure using t closeness with combined l-diversity and k-anonymity", International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCEE)Volume 1, Issue 8,pp:28-33

[5] Yeye He, Jeffery Naughton .F, (2009), "Anonymization of Set Valued Data via Top Down Local Generalization", Proc. International Conference on Very Large Databases (VLDB), pp.934-945.

[6] Wei Jiang, Chris Clifton, (2006), "A secure distributed framework for achieving k anonymity^a, the VLDB Journal, Vol.15, No.4, pp.316-333.

[7] Chuang-Cheng Chiu, Chieh Yuan Tsai, (2007), "A k Anonymity Clustering method or Effective Data Privacy Preservation", Springer journal on Verlag Berlin Heidelberg, pp.88-99..

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

[9] Xiaojian Zhang, Xiaofeng Meng, and Rui Chen. Differentially private set-valued data release against incremental updates. In Database Systems for Advanced Applications, pages 392–406, 2013.

[10] R.Natarajan,Dr.R.Sugumar,Mahendran,K. Anbazhagan , "Asurvey on Privacy Preserving Data Mining", InternationalJournal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 1,MARCH 2012.

[11] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N.Mamoulis, "Security in outsourcing of association rule mining,"in Proc. Int. Conf. Very Large Data Bases, 2007, pp. 111-122.

[12] Evfimievski A,Srikant R,Agrawal R, et al. , "Privacy preserving mining of association rules," In: Proc. of t he Eighth ACM SIGK2DD International Conference on Knowledge Discovery and Data Mining, ACM Press,2002, pp.217-a228. P. K. Prasad and C. P. Rangan, "Privacy preserving birchalgorithm for clustering over arbitrarily partitioned databases,"in Proc. Adv. Data Mining Appl., 2007, pp. 146-15