JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

A hybrid cryptography based decision tree classifier for privacy preserving databases

Aaluri Seenu[1], Dr. G Samba Siva Rao[2]

[1]Research Scholar, [2]Professor

[1,2] Department of CSE, Acharya Nagarjuna University (University College of Engineering and Technology) Nagarjuna Nagar-522510, Guntur, A.P, India

[1]Email: cnuaaluri@gmail.com

Abstract:

Individual decision patterns can be protected from unauthorised access using the privacy-preserving data privacy concept. As the data owner's decision-making processes are made public, they can be misused in distributed applications. It is necessary to encrypt sensitive information about businesses, industries, and individuals before it is shared or published in order to protect their privacy. Using huge dispersed datasets, this research develops and implements a unique chaotic privacy-preserving model. Traditional privacy preserving models are utilised to compare the suggested model to the traditional models in this article. In this work, a hybrid cryptographic chaotic based privacy preserving model is developed to improve the overall efficiency of different databases.

Keywords: privacy preserving , classification, support vector, privacy applications.

1.**Introduction**

In recent years, data has been piling up in practically every sector, including industrial organisations, scientific research, educational institutions, medical science, economic sectors, and government agencies. The likelihood of privacy protection [1] is high. In order for PPDM to function, Secure Multi-party Computation must be used (SMC). Private information can be protected using this way. PPDM is in charge of enhancing the security of data mining's most sensitive information. Businesses are worried that intruders will gain access to their confidential data if they don't protect it. As a result, PPDM must be applied in a wide range of data mining applications. PPDM research has been going on for decades, and numerous algorithms have been developed by a variety of academics. The development of a safe channel for the exchange of complicated data is now absolutely necessary. Each and every transaction is recorded in the database during secure sharing. Terabytes and petabytes of data are generated. The performance of mining algorithms is determined by their efficiency, scalability, and confidentiality. Different data mining methods implement numerous privacy analysis tools to mine vast amounts of data [1]. The preservation of individual privacy is critical in the data mining industry.

People's personal information and company-related data are collected and retained by organisations for various purposes, and hence cannot be shared with third-party authorities. Machine learning algorithms can also expose sensitive information by generating patterns in data. In order to ensure the privacy of data, owners of the data disclose it only when there is a guarantee of confidentiality. High computational time and memory requirements make typical privacy-preserving methods unsuitable for picking privacy patterns. It's because of these methods that the data can be used in any computation. Non-data mining uses of the public information are permitted. Transformation operations like scaling, rotation, and noise addition can be used to alter the original value of data. In the course of the randomization process, any mapping or correlation that might have existed between the various pieces of information is destroyed forever. In order to protect data from unauthorised access, cryptographic methods employ a variety of encryption techniques. However, after the data has been decrypted, it can be attacked by a malicious third party who is at least trying to be honest.

Cluster analysis can be used in two instances when data partitioning is required in a distributed manner. In the first case, there is a substantial amount of data to be analysed. A large amount of computational work has to be put into this, and it is sometimes impossible to finish this process. Rather than combining the scattered results, it is preferable to separate the data and cluster it in a distributed manner. In a centralised

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

database, all of the data is held in one location, whereas in a distributed database, data can be spread across multiple locations. Data is saved in one location when the database is centralised. The distributed database is not the same as this sort of database. Data bottlenecks can emerge at important stages in the release or assimilation process when using a centralised database since the data is housed in a single location. Therefore, the retrieval efficiency is not as high as it would be in a distributed database system while looking for data availability.

This work makes use of a classification approach to identify a variety of techniques that could be ideal in a specific situation. There have been a few ideas put forth. A single privacy-preserving strategy could perhaps surpass all others given the wide range of contexts in which it is employed. There are advantages and disadvantages to every method. A thorough review of all privacy-protection methods is therefore critical. First developed by Agrawal and Srikanth, the Randomization method is one of the most often used privacy-preserving data mining algorithms (2000).

2. Related works

In unsupervised self-organizing learning with support vector ranking [2], the unbalanced data problem is alleviated. To deal with this difficulty, support vector machines use a model that selects variables for this method. Classification of ranker features using ESOM, or Emergent Self-Organising Map, can be done without supervision. A decision tree approach (KS tree) [3]-based Kolomogrov-Smirnov statistic (KS tree) is the newest method for breaking down complicated problems into smaller, more manageable pieces. This technique is also employed for the removal of superfluous features from feature selection. Before logistic regression models can be built, the data is rebalanced using a two-way resampling method, which determines the appropriate sampling criteria. Considered in its entirety, simpleMIL performs exceptionally well. It emphasises the distribution of intensity and texture in the image. When scanning many domains, these classifiers are more appropriate. The above-mentioned classifiers are ineffective when used to cross-domain scenarios, such as numerous scanners. To get around this problem, a multi-class dataset is used to create the classification algorithm. Texture features are more effective than intensity features when it comes to analysing data. Database scans from various domains can be distinguished using an effective weighting strategy based on classifiers. This methodology has the potential to improve on standard categorization methods.

Security of communications and encryption are two of the most important concerns in a distributed data mining approach. As a result, privacy-preserving and encryption techniques are frequently used in distributed systems. Both the vertically partitioned and the horizontally partitioned data structures are available to all distributed applications. Data that is vertically partitioned can be defined as data that is spread across multiple locations and stored in such a way that no data overlaps. Data that is horizontally partitioned is defined as data that is located in a specific number of locations. Records from other sites aren't available on these servers. [6,7] Horizontally partitioned data generates several relevant and legitimate association rules. To address the problem of maintaining privacy, a significant body of work has been done in the discipline of cryptography. An example of Secure Multi-Party Computation may be found here (SMC). Each user in a multi-user network has the obligation of working collaboratively with other users to fulfil computational tasks while also ensuring privacy[8]. [9].

Various privacy-preservation mechanisms are being used in current research, which is then processed through data mining models to uncover patterns of decision-making. Developing a data mining model for sensitive information is the primary purpose of the privacy protection strategy. 1. The most typical concern is to protect sensitive information such name, ID number, address, and income level.

An efficient solution for tiny datasets has been provided in ([11]) to address privacy preserving data mining issues. They used the ID3 algorithm, which uses horizontally partitioned databases, to protect privacy. Privacy can be preserved by considering horizontally partitioned data when using the ID3 technique, according to [12]. As opposed to vertically partitioned data techniques in which only two parties are examined, they additionally evaluated several parties. For privacy-preserving ID3 algorithms, the Gini feature selection measure is used in place of entropy in order to generate decision tree patterns. The value of qualities can be evaluated by all parties. When the database is partitioned into two or more parties, this method works wonderfully.. As a general rule, entropy aids in the creation of balanced trees.

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

Various privacy-preserving data mining techniques were examined and analysed in [10]. Data mining may be hindered by the use of specific methodologies. Privacy Preserving Record Linkage (PPRL) is the name given to a novel way of preserving privacy (PPRL). Using this method, you can link separate databases to different organisations while maintaining the confidentiality of important information.

Traditional data mining methods can be used to enhance privacy. Data distribution, distortion, mining strategies, and data concealment are all key components of these methodologies. [10] Homomorphic encryption and the standard digital differential approach were combined to create a novel privacy-preserving model. Using a multi-party communication mechanism, this paradigm is implemented. In addition, this protocol relies on homomorphism encryption to ensure that the decision tree building remains private.

3. Chaotic Maps : Preliminaries

A chaotic system has a number of key features, among them pseudo-randomness and sensitivity to the beginning conditions.. Diffusion and confusion traits are critical for encryption, and these features have them in spades. The usage of chaotic hash functions improves the security and randomness of the data integrity and message authentication process. In recent years, key hash functions have been built using chaos. For a given set of initial conditions, a chaotic map's behaviour can take on a wide range of new and unexpected manifestations. Changing the initial circumstances of a chaotic function causes it to produce radically distinct patterns of output. It's a great feature for encrypting sensitive data! There have been a variety of chaotic maps used in encryption methods to date. In the last decade, chaos has seen a lot of use because of its intriguing properties, such as sensitivity to modest changes in initial conditions and parameters, mixing property, and so on. However, the majority of chaotic hash functions can only be used in a sequential fashion [1–5]. As a matter of fact, until the previous message unit is processed, the present message unit cannot begin processing. As a result, productivity suffers dramatically.

Chaotic Encryption Model

The skew tent map is defined as follows:

$$f_{\varphi}(m) = \begin{cases} \dfrac{m}{\varphi}, & 0 < m \leq \varphi \\ \dfrac{m-1}{\varphi - 1}, & \varphi < m \leq 1 \end{cases}$$

The inverse function of the skew tent map is given by

$$\mathrm{f}_{\varphi}^{-1}(\mathrm{m}) = \varphi \mathrm{x} \text{ or } \mathrm{f}_{\varphi}^{-1}(\mathrm{m}) = 1 + (\varphi - 1)\mathrm{m}$$

$$\text{Hqper} - \text{chaoti c Lü sqstem} = \begin{cases} \dot{\mathrm{p}} = \mathrm{a}(\mathrm{q} - \mathrm{p}) \\ \dot{\mathrm{q}} = -\mathrm{pr} + \mathrm{cq} \\ \dot{\mathrm{r}} = \mathrm{pq} - \mathrm{br} \end{cases}$$

$$\text{Hqper} - \text{chaoti c Chen sqstem} = \begin{cases} \dot{\mathrm{p}} = \mathrm{a}(\mathrm{q} - \mathrm{p}) \\ \dot{\mathrm{q}} = -\mathrm{pr} + \mathrm{dp} + \mathrm{cq} - \mathrm{q} \\ \dot{\mathrm{r}} = \mathrm{pq} - \mathrm{br} \\ \dot{\mathrm{q}} = \mathrm{p} + \mathrm{k} \end{cases}$$

$$\text{Hqper} - \text{chaoti c Rossler sqstem} = \begin{cases} \dot{\mathrm{p}} = -\mathrm{q} - \mathrm{r} \\ \dot{\mathrm{q}} = \mathrm{p} + \mathrm{aq} + \mathrm{w} \\ \dot{\mathrm{r}} = \mathrm{b} + \mathrm{pr} \\ \dot{\mathrm{w}} = -\mathrm{cr} + \mathrm{dw} \end{cases}$$

The chaotic tent map used here is a one-dimensional calculation and a piece-wise linear map.

For the iteration $\mathrm{r}_{n+1} = \mathrm{f}_{\mu}(\mathrm{r}_n) = \mathrm{f}(\mu, \mathrm{r}_n)$, any period-2 point $\bar{\mathrm{r}}$ satisfies $\bar{\mathrm{r}} = \mathrm{f}_{\mu}^{2}(\bar{\mathrm{r}}) = \mathrm{f}(\mu, \mathrm{f}(\mu, \bar{\mathrm{r}}))$

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

For the map $T(\mu, r) = \mu\sin(\pi r)$, we have

$$T(\mu, r) = rk(\mu, r) \text{ where}$$

$$k(\mu, r) = \mu\left[\pi - \frac{\pi(\pi r)^2}{3!} + \frac{\pi(\pi r)^4}{5!} - \cdots + (-1)^n \frac{\pi(\pi r)^{2n}}{(2n+1)!} \pm \cdots\right]$$

$$\text{At } (\mu, r) = \left(\frac{1}{\pi}, 0\right), \text{ we have}$$

$$\frac{\partial T(\mu, r)}{\partial r} = 1, \frac{\partial k(\mu, r)}{\partial \mu} = \pi \neq 0, \frac{\partial^2 T(\mu, r)}{\partial r^2} = 0, \frac{\partial^3 T(\mu, r)}{\partial r^3} = -\mu\pi^3 \neq 0$$

$$\frac{\partial^2 k(\mu, r)}{\partial r^2} = -\frac{1}{3}\pi^2 < 0$$

The high randomised hash value is computed using the data from each node in this process. The integrity algorithm uses the client ID and cloud node data M as input parameters.

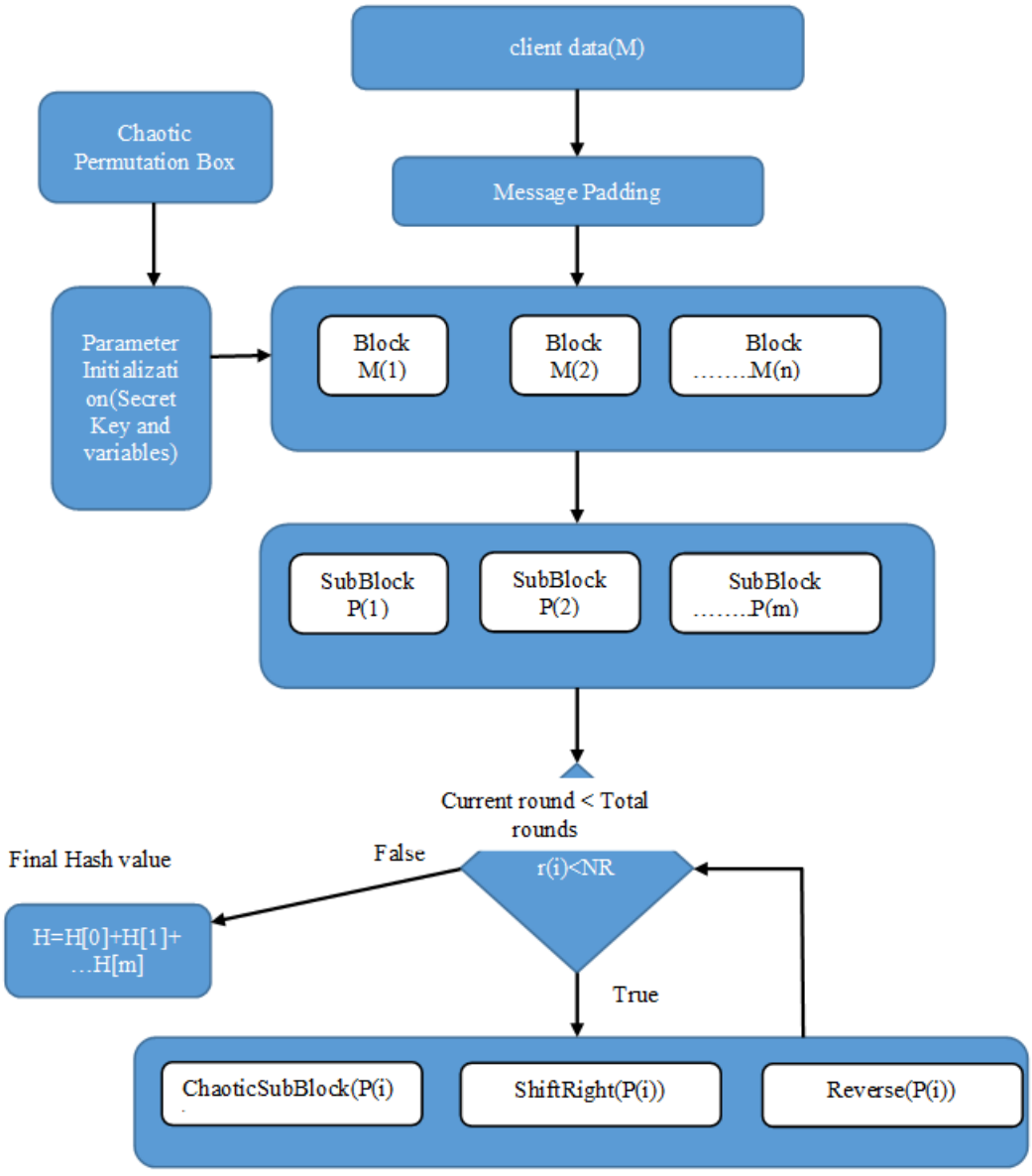

Figure 1: Proposed model

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

Ensemble Deep learning framework

Input : Pre-processed dataset PD-1,PD-2..PD-N, Fast Random Decision tree with Node N, Tree count |N|, Attributes list A.

Procedure:

- For each attribute $A\tau$
- Do
- Partition $A\tau$ using the different class labels using the following measure.
- $$\text{RandomForest Partitioning Measure=RPM[Att,m]}=\frac{-PD[i].\log(\sum_{i=1}^{|PD|} PD[i])*Prob(PD[i]/C_m)}{PD[i]^3.\sqrt{Chisquare(PD[])}}$$
- End for
- End for

Choose the node with class m as the best split attribute in partitioning list.

- Create root node N[0].
- Repeat the process until all the nodes in the Tree T.
- End for
- **Non-linear SVM**
- Apply SVM multi-class optimization models as

$$\min_{W_k,a_k} \frac{1}{2}||W_k||_1^2+\alpha\sum_{i=1}^{n} ker<x,y>+\tau_m$$

$$s.t$$

$$W_k^T D_i+b_k \geq 1-\tau_m, if\ y_i=k$$

$$W_k^T D_i+b_k \leq -1+\tau_m, if\ y_i \neq k$$

$$\tau_m>0; m=1...classes$$

- Here kernel function ker(x,y) defines the x input values that are mapped to y dimensional space as:
- Ker<x,y>= $e^{-||y||log\sum||x||^2}.max\{2exp||x||,log||y||\}$ if x==y

QR decomposition measure Q and R matrices are used to produce new keys. The computation and rank of the Cauchy polynomial may be found using these matrices. Until the number of rounds is reached, this procedure is repeated It is used to verify the cloud node's integrity against assaults by generating a generated integrity value. There is no predetermined size for the input hash in this algorithm. The suggested algorithm generates hash values of 512, 1024, 2048, and 4096. The suggested approach is more time and sensitivity sensitive than the current integrity algorithms. It is then utilised in step 2 to verify the cloud node's signature in the exchange of data.

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

4. Experimental Results

Table1:Performance analysis of different classification models in terms of computational runtime(ms) on credit card data

K-value	Naive_Bayes	RandomForest	SVM	J48
K-5	8571.32	7912.54	7229.48	6583.72
K-15	7961.29	8037.46	7597.72	7562.06
K-25	7963.88	8091.71	7079.92	7628.22
K-35	8859.8	7967.64	6868.45	7271.53
K-45	8475.12	8007.62	7132.04	6978.83
K-55	8663.92	8058.39	7566.6	6882.88
K-65	8535.01	7989.65	7389.15	7443.58
K-75	8050.95	8158.75	7000.84	6894.32
K-85	8153.9	7924.23	7264.11	7476.07

Table.1, describes the average computational runtime(ms) of different classification models on credit card dataset. Here, the average computational runtime(ms) of different k values {5,15,25,35,45,55,65,75,85} are used in experimental results. To each k value, the average runtime computational of original input credit card data, perturbation credit data, anonymization credit data and perturbation anonymization credit card data are used in experimental results.

Table 2: Performance analysis of different classification models in terms of computational runtime(ms) on bank data

K-value	Naive_Bayes	RandomForest	SVM	J48
K-5	5883.61	5987.71	5989.86	5610.85
K-15	5869.61	5677.5	5978.22	5604.58

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

K-25	5599	5658.62	5635.24	6032.53
K-35	6094.93	6126.69	5844.52	5779.1
K-45	6094.82	5942.49	5794.66	5830.87
K-55	6171.86	5645.32	6036.28	5764.6
K-65	5605.26	5807.02	6064.73	5768.94
K-75	6001.9	5999.63	5611.58	5716.78
K-85	5599.22	5765.5	6018.92	5942.3

Table 2, describes the average computational runtime(ms) of different classification models on bank dataset. Here, the average computational runtime(ms) of different k values {5,15,25,35,45,55,65,75,85} are used in experimental results. To each k value, the average runtime computational of original input bank data, perturbation bank data, anonymization bank data and perturbation anonymization bank card data are used in experimental results. The re-identification risks for original anonymized bank dataset are shown in fig.5.9. The reidentification risks for bank dataset after applying proposed method are shown in fig.2. It is observed from the results shown in Fig. 2 and 3 that re-identification risks have reduced by using the proposed approach.

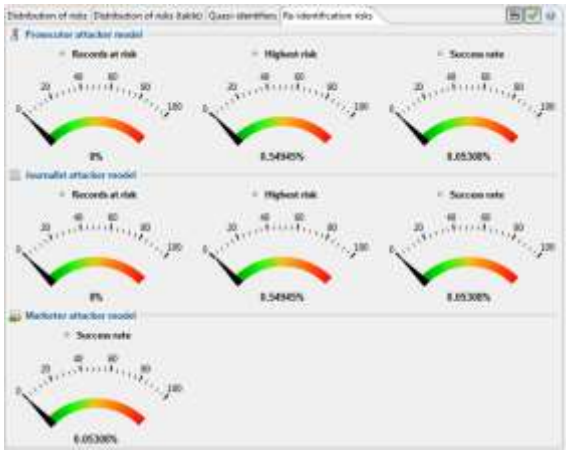

Fig.2 Re-identification Risks for bank dataset

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

K- Anonymization is performed on pre-processed bank dataset with k=55 and utility metric set as kldivergence. After anonymization, the probability of various attacker models is shown in Fig 3.

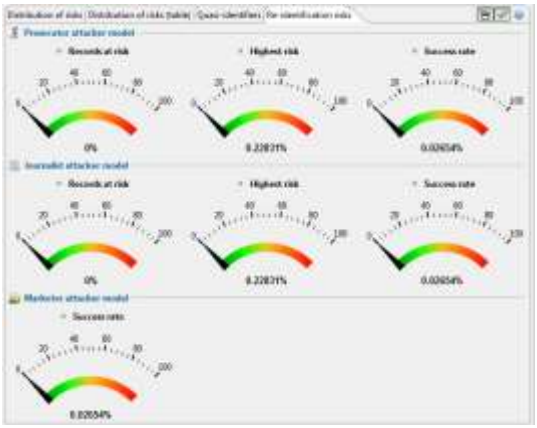

Fig.3 Re-identification risks for bank dataset after applying proposed method

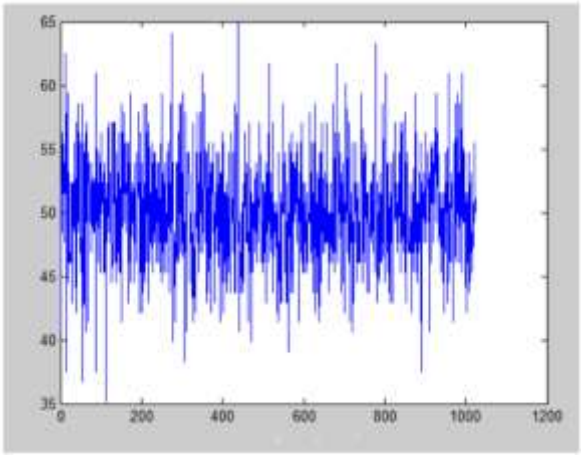

Figure 4: Proposed Chaotic hash

References

[1] Mahesh, R., & Meyyappan, T. (2013, February). Anonymization technique through record elimination to preserve privacy of published data. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on (pp. 328-332). IEEE. [2] Usha, P., Shriram, R., & Sathishkumar, S. (2014, February). Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining. In Information Communication and Embedded Systems (ICICES), 2014 International Conference on (pp. 1-5). IEEE.

[3] Prakash, M., & Singaravel, G. (2015). An approach for prevention of privacy breach and information leakage in sensitive data mining. Computers & Electrical Engineering, 45, 134-140.

[4] J. Le Ny and G. Pappas, "Differentially private filtering," Automatic Control, IEEE Transactions on, vol. 59, no. 2, pp. 341–354, Feb 2014.

[5] Z. Huang, S. Mitra, and G. Dullerud, "Differentially private iterative synchronous consensus," in Proceedings of the 2012 ACM .

[6] Lior Rokach and Oded Maimon "TopDown Induction of Decision Trees Classifiers – A Survey", IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11, NOVEMBER 2002

[7]. Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao-Kui. Privacy Preserva- tion in Database Applications: A Survey.C hinese journerl of computer,2009

[8]. Yan Zhao1 Ming Du2 Jiajin, Le1 Yongcheng Luo1, A Survey on Privacy Preserving Approaches in Data Publishing. First International Workshop on Database Technology and Applications, 2009

ISSN- 2394-5125 VOL 06, ISSUE 01, 2019

[9] Agrawal, Shashank, et al. ”Function Private Functional Encryption and Property Preserving Encryption: New Definitions and Positive Results.” IACR Cryptology ePrint Archive 2013 (2013): 744

[10]. Attrapadung, Nuttapong, and Benot Libert. ”Functional encryption for public-attribute inner products: Achieving constant-size ciphertexts with adaptive security or support for negation.” J. Mathematical Cryptology 5.2 (2012): 115-158.

[11] Stefano Guarino,"Provable Storage Medium for Data Storage Outsourcing",IEEE TRANSACTIONS ON SERVICES COMPUTING,2014.

[12] Jinguang Han,"Improving Privacy and Security in Decentralized Ciphertext-Policy Attribute-Based Encryption",IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 3, MARCH 2015.

[13] Jianting Ning,"White-Box Traceable Ciphertext-Policy Attribute-Based Encryption Supporting Flexible Attributes",IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 6, JUNE 2015.

[14] Changji Wang,"An Efficient Key-Policy Attribute-Based Encryption Scheme with Constant Ciphertext Length",Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 810969