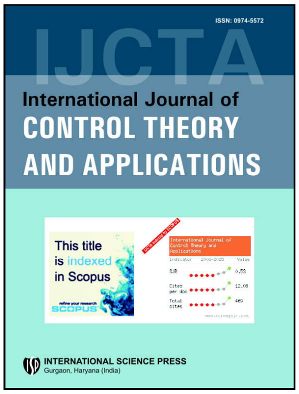


International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 6 • 2017

A Framework to Preserve the Privacy of Electronic Health Dynamic Data Streams Using Parallel Architecture

Ganesh D. Puri[1] and D. Haritha[2]

[1] Department of Computer Science and Engineering, KL University, Vaddeswaram, Guntur, Andhra Pradesh, India, Email: puriganeshengg@gmail.com
[2] Department of Computer Science and Engineering, KL University, Vaddeswaram, Guntur, Andhra Pradesh, India, Email: haritha_donavalli@kluniversity.in

Abstract: Existing privacy preserving methods for electronic health data are working in the sequential order. Such processing has disadvantages like time consuming operation, more memory requirement, and performance reduction. To avoid this, delay free anonymization method working in the parallel form on large volume data is proposed. In this method the preprocessing is applied in parallel form on the data before it is applied to privacy preserving method. After preprocessing, the data is grouped followed by the operations of counterfeit generation and l-diversity. To generate the counterfeit values analysis of past data is used. This reduces the time required to calculate counterfeit sensitive values. The late validation is applied to reduce the effect of counterfeit value added in the group. The data utility measures are applied to calculate the information loss and measure the processing time. This proposed method is advantageous for real time health data applications, health monitoring systems.

Keywords: Anonymization, Privacy, Parallel.

1. INTRODUCTION

Nowadays many privacy preserving methods are available working on static data, dynamic data and data streams. Data for electronic health records can be categorized as database, dataset and data stream. In the database, data is present in the rows and columns format. It is systematically available in the table format. In the dataset, the data is available in the file format. This data can be converted in the database format. In the data stream, the data is generated in real time. In the static form of electronic health data, it is taken in the batches to process. Batch data processing in sequential form has certain disadvantages. It suffers from time consuming operation, more memory requirement, and performance reduction problems. If new tuple of electronic health data is inserted in static data, it is required to process whole data again to apply privacy preserving method [1]. In the real time data in the form of stream insert, update and delete operations are possible to make the data dynamic [2]. The data can come from any distributed source [3]. The privacy preserving method is applied to the newly inserted tuple only and not on the whole dataset. This reduces the time to process the data. The general approach for privacy

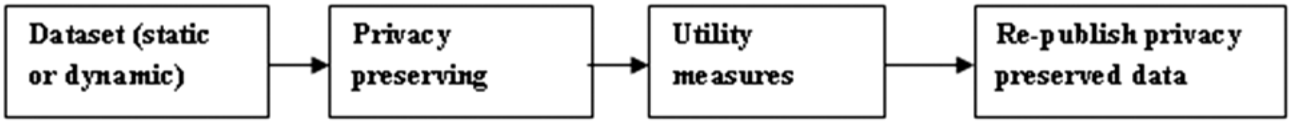

Dataset (static or dynamic)
Privacy preserving
Utility measures
Re-publish privacy preserved data

Figure 1: General steps in privacy preserving

preserving method is shown in Figure.1. In this utility measures are applied to the data after privacy preserving. There should be balance between the privacy preservation and information loss. While doing the privacy preservation, delay should be minimized for anonymization process.

The possible privacy attacks on the republished data after anonymization, are listed out in the Table 1. The attacker can insert or delete his/her own fake record and observe the changes made to the privacy preserved dataset to get the pattern or to identify the grouping of quasi and sensitive attributes. To avoid this possibility the counterfeit past data analysis is added. In this paper we introduce new dynamic data streams privacy preserving algorithm with reduced delay, works on parallel architecture and efficiently handles large amount of data.

The work in the paper is organized as follows. Section II is about existing methods and models and related work. Section III and IV describes research method and algorithm of the framework. Section V gives distributed execution flow. Section VI comparison and section VII gives conclusion of this work

2. EXISTING METHODS AND MODELS FOR PRIVACY PRESERVING

Lot of work is carried out in data privacy. According to big data working group the security and privacy challenges are divided in main four parts as 1) Infrastructure security 2) Data privacy 3) Data management 4)Integrity and reactive security. Focusing on data privacy different methods for privacy preserving are like cryptography, attribute based encryption, data anonymization [4], notice and consent and differential privacy.

2.1. Privacy preserving methods in demographics

In the Table 1 the privacy attacks and privacy models applicable to demographics [5] are listed. These methods can be considered for relational data and the electronic health data coming from individual hospital [6]. In this case the dataset is of patients of individual hospital and attacker can try to link the information in the dataset with another dataset publicly available. For example in the hospital dataset even though the direct attributes like name, SSN (social security numbers) are removed, it is possible to link the attributes with other dataset by using indirect attributes.

Table 1
Privacy attacks and privacy models on demographics

Sr. no	*Attack*	*Threat*	*Privacy model*
1	Record linkage	Identity disclosure	k-anonymity[7], (X-Y) anonymity[8], Multi R-anonymity[9]
2	Attribute linkage 1) Homogeneity 2) Background knowledge	Attribute disclosure	Bayes optimal privacy, entropy l-diversity, multiattribute l-diversity, recursive(c,l) diversity[10]
3	Attribute linkage 1) Skewness attack 2) Similarity attack	Attribute disclosure	t-closeness, (N,T) closeness[11] , personalized privacy
4	Table linkage	Membership disclosure	δ- Presence, C- confidence δ-presence

Table 2
Aggregation privacy preserving techniques

Type of data	Operation/ algorithm	Advantage	Drawback
Structured	Differential privacy[12]	Inferences of record are individual, independent of record present or not in dataset	Generate noisy summary statistics and does not guarantee all attacks
Unstructured	Homomorphic public key encryption[13]	Popular data collecting technique for event statistics.	Purpose specific, unsuitable for complex data types.

The attacks are mainly categorized in the Record linkage, attribute linkage, table linkage [18]. Identification of individual can be done from the linkages [5]. To avoid the attacks, the privacy models mainly used are k-anonymity, l-diversity, t-closeness, personalized privacy.

Table 3
Operations over encrypted data

Operation over encrypted data	*Technique of encryption*	*Advantage*	*Application*
Privacy preserved query[14]	Ciphertext–policy attribute based encryption	Confidentiality and user privacy in cloud	Suitable for multidimensional big data search in heterogeneous distributed system such as cloud.
Secure full text retrieval over encrypted data[15]	Hierarchical bloom filter index	Secure and efficient privacy preserved index	Cloud storage application
Privacy preserving data sharing scheme[16]	Ciphertext–policy attribute based encryptionKey-policy attribute based encryption	Fine grained access control [17], dynamic social attribute management, multiuser searchable scheme	Cloud for social application

2.2. Privacy preserving aggregation

In the Table 2 the methods for aggregation technique for privacy preservation are listed. These are mainly according to the structured and unstructured data.

2.3. Operations over encrypted data

Sensitive documents are kept private by using encryption technique. The Table 3 contains different techniques of encryption. The sensitive documents and related keywords are encrypted. The privacy preserved query and privacy preserved data sharing scheme can be used to access data from cloud server [19]. The applications for encrypted data are widely available in cloud computing.

2.4. Privacy preserving Cosine similarity measure

Inter organization big data processing of the sensitive documents can be done using cosine similarity measure. In this method the documents are represented as vector [20], [21].Rongxing Lu et.al in [22] used lightweight multi-party random masking and polynomial aggregation techniques to calculate cosine similarity in privacy preserved form.

3. RESEARCH METHOD

3.1. The need and importance of the problem

The preprocessing of the health data after collection on single machine, takes much time. While applying the anonymization process on preprocessed health data, sequential processing in terms of counterfeit generation, l-diversity and utility measures create delay. This delay should be reduced by parallelizing these tasks. Large amount of data is created for patient related health data. There is need to process this large amount data and republish the health data in privacy preserved form with minimum information loss. Our proposed framework gives solution for all these problems.

3.2. Architecture

Figure 2 shows the architecture of proposed method. First part in the architecture is input dataset. Different hospitals send the data. Such electronic health data is collected in one dataset. Collected data may be in the large volume [3]. In the data model we are converting data in the suitable form. This step is called as preprocessing and it is applied in the distributed fashion. This is done in parallel. Record source is kept to identify in case of privacy breach.

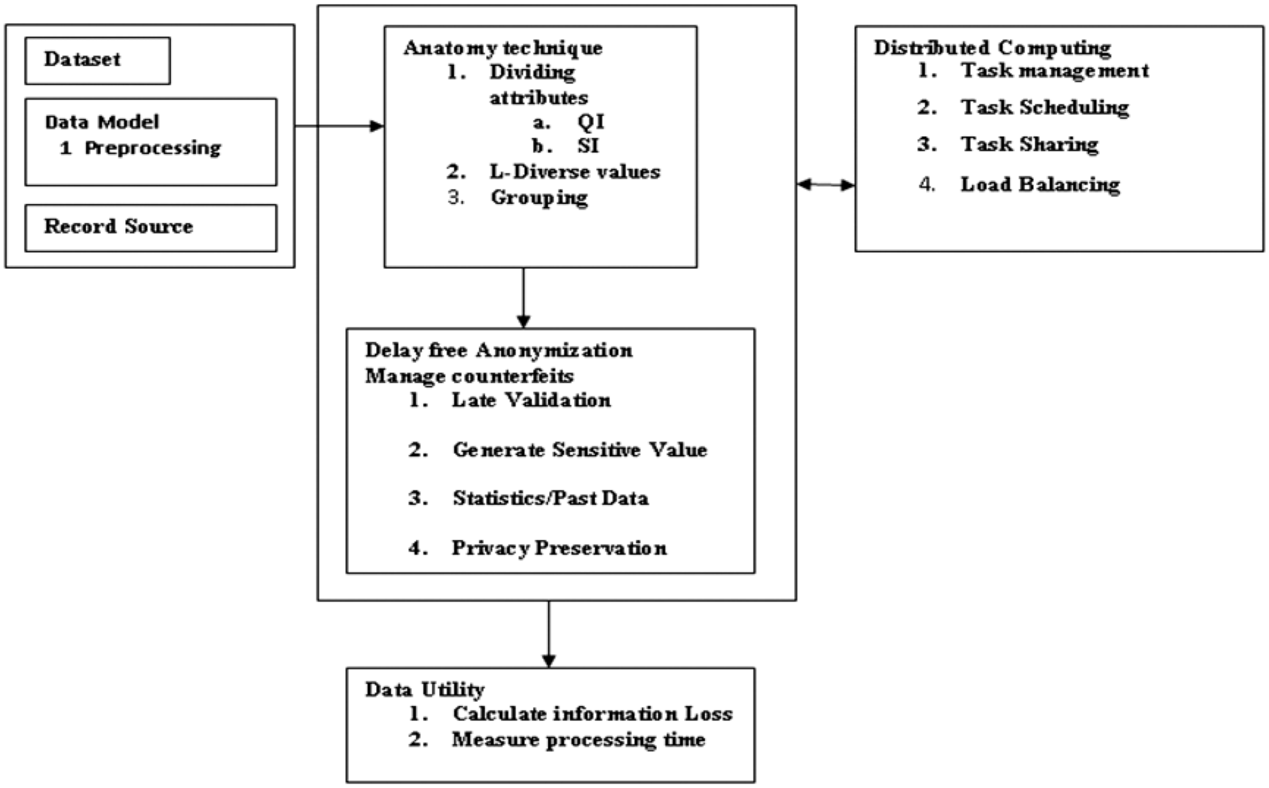


Figure 2: Architecture of proposed method

In the second stage the Anatomy technique [23] is applied. The grouping is done as per the quasi and sensitive attributes. Counterfeit generation, l-diversity and data utility calculations are dependant tasks and need to be performed in sequential order. But in the parallel architecture, counterfeit generation can be separated from other tasks. While one incoming tuple is applied for counterfeit generation, after finishing this task that tuple is applied to l-diversity method to make sure attacker cannot apply similarity or skewness attack [11]. Meanwhile new incoming tuple can be applied for counterfeit generation. After ensuring l-diversity the tuple is applied to data utility calculations. So counterfeit generated new tuple can be applied to l-diversity. In this way the task of counterfeit generation, l-diversity measure and utility measure is parallelized to reduce the delay in anonymization. To generate the counterfeit value of sensitive attribute, analysis of past data is used. Parallel processing can be applied in counterfeit generation, l-diversity and data utility calculations.

3.3. Past data analysis for counterfeit

The past data analysis can be used to make the counterfeit value management. The attacker can try to get the pattern of counterfeit value generation by inserting repeated quasi identifier and changing sensitive values [2]. To avoid this in our framework we are using past data analysis. It stores the records for user and the counterfeit value generated for those quasi identifiers. If record with same quasi identifiers is inserted again then counterfeit value is not generated. It may be attacker trying to get the pattern of counterfeit values. In Table 4 input data from the stream is collected. Different users inserted data with quasi identifiers age, sex and sensitive value as Disease. In Table 5 past data analysis of counterfeit value is generated. Alice entered the record twice with different disease. So first time the counterfeit is generated, but next time it is not generated. For Dis.B once counterfeit is generated so next time for same disease it is not generated.

Table 4
Input data

User	*Age*	*Sex*	*Disease*
Alice	24	Male	Dis.A
Bob	32	Male	Dis.B
Jon	28	Male	Dis.C
Lisa	27	Female	Dis.B
Alice	24	Male	Dis.D

Table 5
Analysis of counterfeit past data

User	*Age*	*Sex*	*Disease*	*Counterfeit (yes/no)*
Alice	24	Male	Dis.A	yes
Bob	32	Male	Dis.B	yes
Jon	28	Male	Dis.C	yes
Lisa	27	Female	Dis.B	no
Alice	24	Male	Dis.D	no

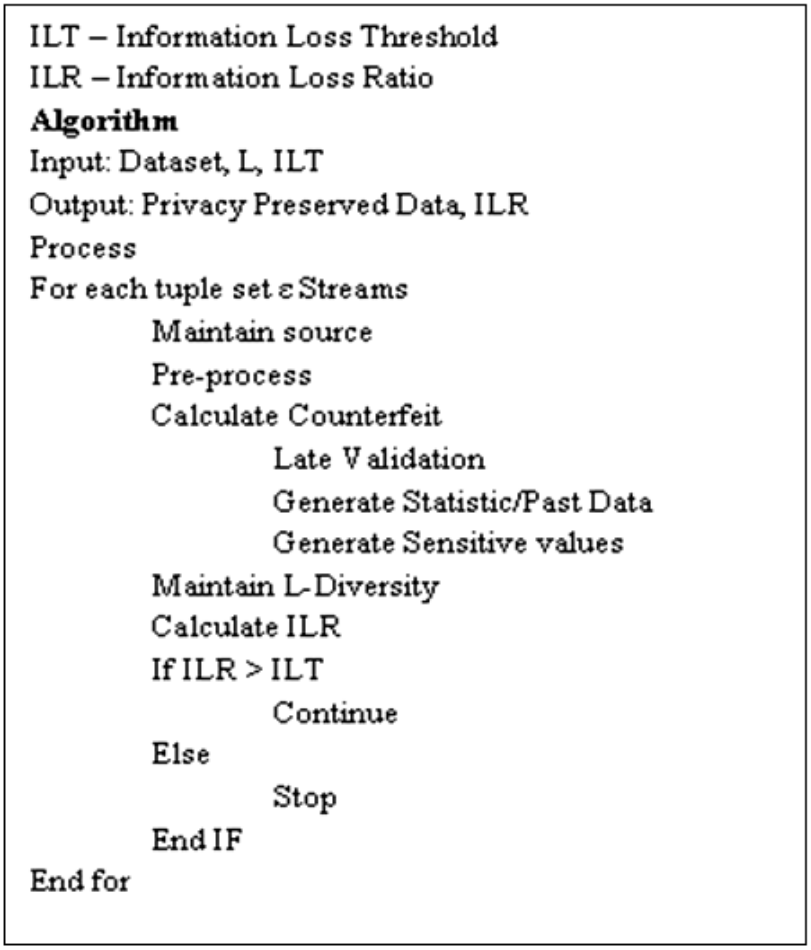
4. ALGORITHM

In the Figure 3 algorithm for our framework is given. In the algorithm information loss ratio (ILR) measure is taken. It is compared against the information loss threshold (ILT). For the incoming input tuple the source information is maintained. Preprocessing is performed in distributed fashion. Distributed processing flow is explained in Figure 4.

For calculating the counterfeit values statistical analysis or past data is used to generate the sensitive values. After generating the counterfeit values, l-diversity is maintained and information loss is calculated. If diversity of sensitive values is more information loss may increase. To keep the balance between information loss ratio and l-diversity, information loss threshold is maintained. If loss is more the counterfeit generation process is repeated to maintain it.

5. DISTRIBUTED EXECUTION FLOW

In the Figure 4 distributed task execution flow is shown. On the master node our algorithm explained in Figure 3 is running. In that independent tasks are processed in distributed manner. Distributed execution supports large



```
ILT – Information Loss Threshold
ILR – Information Loss Ratio
Algorithm
Input: Dataset, L, ILT
Output: Privacy Preserved Data, ILR
Process
For each tuple set ε Streams
        Maintain source
        Pre-process
        Calculate Counterfeit
                Late V alidation
                Generate Statistic/Past Data
                Generate Sensitive values
        Maintain L-Diversity
        Calculate ILR
        If ILR > ILT
                Continue
        Else
                Stop
        End IF
End for
```

Figure 3: Algorithm for framework

amount of data as input [24]. In preprocessing data model is formed by classifying the attributes in quasi and sensitive identifiers.

This step can be performed in parallel fashion as data stream is coming from different hospitals. The master node is running the algorithm. The input is taken from task queue. Efficiency of framework is increased due to running the main algorithm on master node and the rest of the work is distributed. It generates the delay in enqueue and dequeue operation but it is negligible over the advantages provided by this framework.

6. COMPARISON

6.1. Information loss

Comparison for the framework is done in two aspects. First is information loss and second is execution time. Information loss (IL) after anonymization in tuple based method can be calculated as

$$IL = IL(\text{quasi attributes}) + IL(\text{sensitive attributes})$$

Anonymization based on the generalization methods for privacy preservation affect with information loss [23]. It is due to generalization of quasi attributes. Information loss in generalization method depends upon number of quasi attributes taken for generalization. In continuous valued attributes, information loss depends upon range of the attributes whereas it depends upon level of taxonomy for categorical attributes [2].

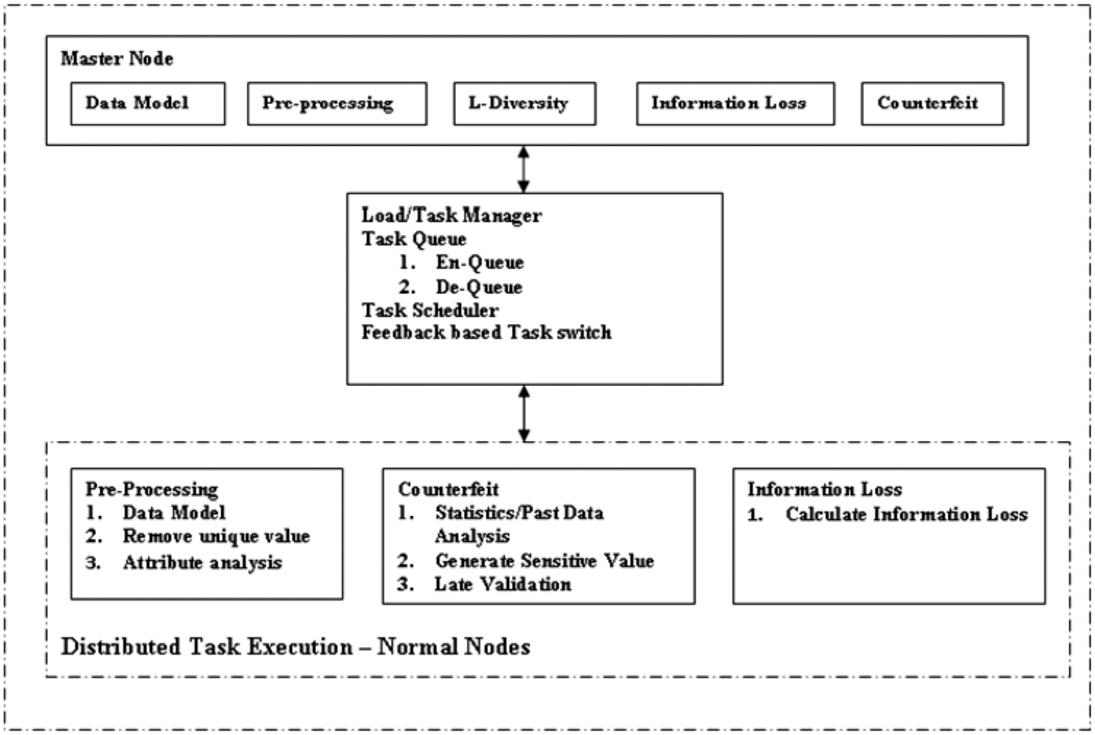


Figure 4: Distributed Execution flow for framework

In our framework information loss due to anonymization is less than generalization techniques. As quasi attributes are not going to generalize, information loss is comprised due to counterfeit value of sensitive attribute. Counterfeit values are validated in the next stage so information loss due to counterfeit value is reduced [2].

6.2. Execution time

In the existing sequential approach, for the anonymization of data the total time required for execution is

$$\text{Time} = \sum \text{preprocessing} + \sum \text{algorithm}$$

In our framework the preprocessing is performed in distributed form. The N number of nodes is added to work in parallel architecture. Algorithm will work on master node. In algorithm counterfeit generation, l-diversity and information loss calculation parts can be parallelized. So in our framework using parallel architecture execution time is

$$\text{Time} = \frac{\sum \text{preprocessing}}{\text{N}} + \frac{\sum \text{algoreithm}}{3}$$

In our framework past data analysis is used. It avoids executing algorithm for creation of the counterfeit values repeatedly. If user and quasi identifiers are matching with existing records then counterfeit algorithm will not executed. It will be useful to enhance performance of framework.

7. CONCLUSION

The framework proposed for the electronic health data stream, working in parallel fashion. It produces better results than sequential flow of anonymization. The delay free anonymization is improved by distributing the dependant task like counterfeit generation, l-diversity and utility measure in parallel fashion.

REFERENCES

[1] Lorenzo Bossi, Alberto Trombetta, Elisa Bertino,Wei Jiang,"Privacy-Preserving Updates to Anonymous and Confidential Databases",IEEE transactions on dependable and secure computing, vol. 8, no. 4, July/August 2011.

[2] Soohyung Kim, Min Kyoung Sung, Yon Dohn Chung, "A framework to preserve the privacy of electronic health data streams", Journal of Biomedical Informatics 50 (2014) 95–106.

[3] Ganesh D. Puri, D. Haritha, "Survey Big Data Analytics, Applications and Privacy Concerns", Indian Journal of Science and Technology, Vol 9(17), DOI: 10.17485/ijst/2016/v9i17/93028, May 2016.

[4] Rashid Hussain Khokhar, Rui Chen, Benjamin C.M. Fung, Siu Man Lui,"Quantifying the costs and benefits of privacy-preserving health data publishing", Journal of Biomedical Informatics 50 (2014) 107–121.

[5] Aris Gkoulalas-Divanis, Grigorios Loukides, Jimeng Sun,"Publishing data from electronic health records while preserving privacy: A survey of algorithms",Journal of Biomedical Informatics 50 (2014) 4–19.

[6] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH, "Automatic de-identification of textual documents in the electronic health record: a review of recent research" BMC Med Res Methodol 2010;10(70).

[7] L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, 2002.

[8] KeWang and Benjamin C. M. Fung, "Anonymizing sequential releases", In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.

[9] M.E. Nergiz, C. Clifton, and A.E. Nergiz, "Multi relational k-anonymity", Knowledge and Data Engineering, IEEE Transactions on, 21(8):1104 –1117, August. 2009.

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkita subramaniam, "L-diversity:Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):146, 2007.

[11] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, "Closeness: A new privacy measure for data publishin", IEEE Trans. Knowl. Data Eng., 22(7):943–956, 2010.

[12] Guha S, Rastogi R, Shim K, "Cure: an efficient clustering algorithm for large databases", In: SIGMOD; 1998. p. 73–84.

[13] P. Paillier, "Public-Key Cryptosystems based on Composite Degree Residuosity Classes," EUROCRYPT, 1999, pp. 223–38.

[14] Rong Jiang,Rongxing Lu,Kim-Kwang Raymond Choo,"Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data",Future Generation Computer Systems 2016.

[15] W. Song, B. Wang, Q. Wang, Z. Peng, W. Lou, Y. Cui, "A privacy-preserved full-text retrieval algorithm over encrypted data for cloud storage applications", J. Parallel Distrib. Comput. (2016), http://dx.doi.org/10.1016/j.jpdc.2016.05.017.

[16] Chen Lyu, Shi-Feng Sun, Yuanyuan Zhang, Amit Pande,Haining Lu and Dawu Gu, "Privacy-Preserving Data Sharing Scheme over Cloud for Social Applications", Journal of Network and Computer Applications, http://dx.doi.org/10.1016/j.jnca.2016.08.00

[17] Krishna Keerthi Chennam, M. Akka Lakshmi,"Cloud Security in Crypt Database Server Using Fine Grained Access Control",IJECE Vol 6, No 3 June 2016 PP. 915-924.

[18] Andrei Manta,"Literature Survey on Privacy Preserving Mechanisms for Data Publishing", Literature Survey submitted in partial fulfillment of the requirements for the degree of MS November 1, 2013

[19] Antorweep Chakravorty, Tomasz Wlodarczyk, Chunming Rong,".Privacy Preserving Data Analytics for Smart Homes",2013 IEEE Security and Privacy Workshops.

[20] Mr. Puri G. D., Prof. Gawali S. Z. "Web Text Extraction and Classification using Vector Space Model method."ISSN: 0974-3596, IJ-CA-ETS April 11-Sept 11, Volume 3: Issue 2, Page 153-157.

[21] Mr. Puri G. D., Prof. Y. C. Kulkarni "Realization of Framework for Web text Extraction and Classification" International Journal of Computer Applications (0975 – 8887) Volume 32 No.6, October 2011(22-26).

[22] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao,"Toward Efficient and Privacy-Preserving Computing in Big Data Era", IEEE Network,July/August 2014 46-51.

[23] Xiaokui Xiao Yufei Tao. "Anatomy: Simple and Effective Privacy Preservation", ACM.VLDB '06, September 1215,2006, Seoul, Korea.

[24] Mohammed Erritali, Abderrahim Beni-Hssane, Marouane Birjali, Youness Madani,"An Approach of Semantic Similarity Measure between Documents Based on Big Data", IJECE Vol. 6, No. 5, October 2016, pp. 2454-2461.