Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 7, July 2014, pg.700 – 705

RESEARCH ARTICLE

Clustering and Hiding Sensitive Data for Social Network Dataset

N.Revathi[1], Dr. A.Padmapriya[2]

[1]Department of Computer Science & Engineering, Alagappa University Karaikudi, India

[2]Department of Computer Science & Engineering, Alagappa University Karaikudi, India

[1] arharsha2012@gmail.com, [2] mailtopadhu@yahoo.co.in

Abstract: Data Mining is one of the main processing step in KDD (Knowledge Discovery in Database), which provides the potentially useful but unknown information to the user. The mining of the databases includes the preprocessing, removing duplication and irrelevant data from the database. For the mining process several techniques were evolves such as association, classification, clustering, feature selection etc. Each of the technique is specifically designed and in use. This proposed work concentrates on clustering technique. The clustering is made in sequential manner, in particular to provide privacy to the data. Hence, the proposed work with the sequential clustering technique, hide very sensitive information from the third party authorization. The application considered here is the Social Networks, in which the sequential clustering methodology is applied. The sensitive data that needs the privacy is computed here with the symbols. The results show that the proposed method remains better at its performance in authentication.

Keywords: Data Mining, Sequential Clustering, Data Preprocessing, Sensitive Data, Privacy Preservation

I. INTRODUCTION

Data Mining is used for analyzing data from various perspectives that converts the previously unknown information into potentially useful data from huge databases. It is the major step in KDD (Knowledge Discovery in Database). The data Mining process is used to provide novel and hidden patterns in the data. Data Mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in the database. They are used in areas such as decision support, predictions, forecasting. Data mining technique includes, Association rules, Clustering, Classification and Prediction.

Clustering technique, discussed in this paper is an important methodology used in various fields such as finance, medicine, and social network databases etc. The clustering is defined as the grouping of similar datasets under single sequential fold. The entity that performs the clustering operation has access to the whole database, while in some other cases; databases from different resources are merged to improve the performance of the clustering algorithms [9]. The clustering is a class of data mining task in which algorithms are applied to discover interesting data distributions in the underlying data space. The formation of clusters is based on the principle of maximizing similarity between patterns belonging to distinct clusters.

Cluster Analysis categorizes the data into groups that are useful and meaningful. Sometimes they are used for the summarization of the data. Clustering is mainly focus in the field of unsupervised learning methods. Clustering is the grouping of similar instances under, some sort of measure that can determine whether two objects are similar or dissimilar. The two main types of measures used to estimate this relation is, distance measures and similarity measures. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects.

A valid distance measure should be symmetric and obtains its minimum value in identical vectors. The distance measure is called a metric distance measure [4]. In this paper, the privacy preservation in the social networks based on the sequential clustering methods is provided. It is necessary in the social network to hide the sensitive information, from the knowledge of unauthorized user. The paper is summarized as follows. Section 2 consists of related works to the clustering technique and privacy preservation. Section 3 shows the proposed methodology with the procedure and their utility. Section 4 consists of the experimental results and section 5 provides conclusion for the proposed work.

II. BACKGROUND STUDY

Seema Kedar et.al [6] proposed a new survey is made in which a technique for privacy preservation in data mining is explained. The primary goal of this survey paper is to understand the existing privacy preserving data mining techniques and to achieve efficiency.

R. Zaiane [7] revisited a family of geometric data transformation methods (GDTMs) that distort numerical attributes by scaling, rotations, translations functions or by the combination of all above transformations. This method was designed to specify privacy-preserving clustering, as well as guarantee valid clustering results. Authors also provided a particularized, broad and advanced picture of methods for privacy-preserving clustering by data transformation.

Tamir Tassa and Dror J. Cohen,[8] explains in details the problem of privacy-preservation in social networks. Authors consider the distributed setting in which the network data is split between several data holders. To that end, authors start with the centralized setting and offer two variants of an anonymization algorithm which is based on sequential clustering. The algorithms significantly outperform other algorithms which are the leading algorithm for achieving anonymity in networks by means of clustering.

Zekeriya Erkin et.al [9] provides an efficient privacy-preserving variant of K-means clustering. The scenario authors consider involves a server and multiple users where users need to be grouped into K clusters. In devised protocol the server is not allowed to learn the individual user data and users are not allowed to learn the cluster centers. The experiments on the analyzed dataset show that deployment of the system for real use is reasonable as its efficiency even on conventional hardware is promising.

Elena Zheleva and Lise Getoor,[2] discuss Preserving the Privacy of Sensitive Relationships in Graph Data, focus on the problem of preserving the privacy of sensitive relationships in graph data. The problem of inferring sensitive relationships from anonymized graph data as link reidentification is provided. Authors propose five different privacy preservation strategies, which vary in terms of the amount of data removed) and the amount of privacy preserved. Authors show experimentally the success of different link re-identification strategies under varying structural characteristics of the data.

III. PROPOSED METHODOLOGY

The sensitive data hiding by sequential clustering approach is explained in this section. The clustering technique is chosen so as to group the similar datasets together. The database considered here is the social network with the records of 8000 in size. The sensitive information to hide in the networking is the gender, which is hidden based on the sequential splitting and clustering groups.

The algorithmic steps are as follow. The dataset is extracted with 8000 records. Now, the records are split into 3 different tables each of size carrying 4000, 2000, and 2000 records. It is now stored at 3 various cluster tables. After this, the clustering step is followed. In clustering, the data with each of cluster records are loaded preprocessed. The individual dataset with the records is now split into 3 clusters based on the attribute age. If the age is between 17 – 25, it is clustered and stored at next sequential level table.

The next cluster group is age limit under 26 – 45, clustered and separately stored, and final cluster group is age limit under 46- 90. Now, the sequential split and cluster is done here. After the cluster groups obtained, now the privacy preservation to the data is taken place. A new zip code is created for each cluster tables for 3 different age groups. The gender information for every code is now computed. For the cluster table 1, with 4000 records, divide it to two divisions based on gender, if the gender key is male add "*" to the gender key and hide the data Male.

If the computed key is female then add "**" instead of indicating the data female to the cluster table. Hence, all the cluster groups show the gender in wither of the two symbols instead of the data. Iteratively the process is done for every cluster table. Hence a sequential order is maintained in the case of large databases. A random number, values between 10000 to 90000 is generated for the zip code and the random number is taken into account for the cluster identification. Based on the random number the male is represented using single * and female is represented using ** symbols. The working strategy is explained in flowchart as follows

Procedure

Step 1: Read the database to cluster.

Step 2: Verify the total number of records in database. The records in the provided set were 8000.

Step 3: Based on the clustering technique, divide the database into 3 groups with each of 4000, 2000 and 2000 records.

Step 4: Develop separate tables to view the attributes of the records. Cluster groups are based on particular condition.

Step 5: The clustering process is taken place. Within each of the clustered records, the cluster group is created, based on the age attribute.

Step 6: If the age limit is between 17 to 25, place it under cluster 1, else if between 26-45 load it into cluster 2 else otherwise if between 46 – 90 places under cluster group 3.

Step 7: The sensitive data to be hide is gender, hence create secret code to each age group.

Step 8: Iteratively for each cluster group check the identity whether gender is male or female.

Step 9: Generate a random number values from 10000 to 90000, followed by the zip code, based on the gender value.

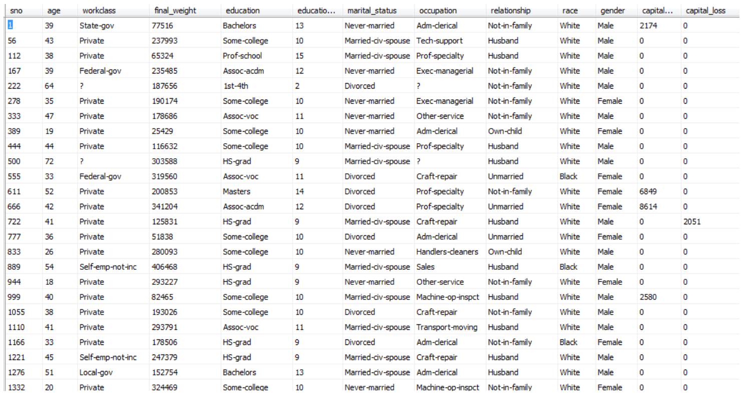
Step 10: If the key found is = Male, then indicate it through the symbol "*"

Step 11: If the key found is = Female, then indicate it through the symbol "**"

Hence, with the help of sequential clustering technique, the cluster groups are divided and for each group the sensitive data is preserved from the knowledge of the unauthorized user.

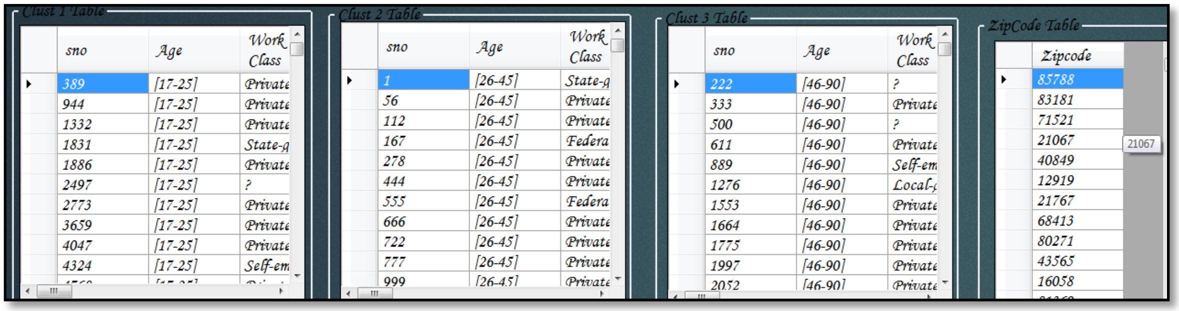
IV. EXPERIMENTAL RESULTS

The experimental analysis is done in the real time data of about 8000 records with 15 of attributes. The results show that, when cluster analysis taken place and after the cluster groups were formed, the computation and the privacy preservation time for the data is considerably reduced. The figures from 4.1 to 4.6 show the whole process of zip code generation and hiding technique of sensitive data. The following graph indicates the time variation between the individual dataset and the cluster groups.



sno	age	workclass	final_weight	education	educatio...	marital_status	occupation	relationship	race	gender	capital...	capital_loss
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0
56	43	Private	237993	Some-college	10	Married-civ-spouse	Tech-support	Husband	White	Male	0	0
112	38	Private	65324	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0
167	39	Federal-gov	235485	Assoc-acdm	12	Never-married	Exec-managerial	Not-in-family	White	Male	0	0
222	64	?	187656	1st-4th	2	Divorced	?	Not-in-family	White	Male	0	0
278	35	Private	190174	Some-college	10	Never-married	Exec-managerial	Not-in-family	White	Female	0	0
333	47	Private	178686	Assoc-voc	11	Never-married	Other-service	Not-in-family	White	Male	0	0
389	19	Private	25429	Some-college	10	Never-married	Adm-clerical	Own-child	White	Female	0	0
444	44	Private	116632	Some-college	10	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0
500	72	?	303588	HS-grad	9	Married-civ-spouse	?	Husband	White	Male	0	0
555	33	Federal-gov	319560	Assoc-voc	11	Divorced	Craft-repair	Unmarried	Black	Female	0	0
611	52	Private	200853	Masters	14	Divorced	Prof-specialty	Not-in-family	White	Female	6849	0
666	42	Private	341204	Assoc-acdm	12	Divorced	Prof-specialty	Unmarried	White	Female	8614	0
722	41	Private	125831	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	2051
777	36	Private	51838	Some-college	10	Divorced	Adm-clerical	Unmarried	White	Female	0	0
833	26	Private	280093	Some-college	10	Never-married	Handlers-cleaners	Own-child	White	Male	0	0
889	54	Self-emp-not-inc	406468	HS-grad	9	Married-civ-spouse	Sales	Husband	Black	Male	0	0
944	18	Private	293227	HS-grad	9	Never-married	Other-service	Not-in-family	White	Female	0	0
999	40	Private	82465	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	2580	0
1055	38	Private	193026	Some-college	10	Divorced	Craft-repair	Not-in-family	White	Male	0	0
1110	41	Private	293791	Assoc-voc	11	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0
1166	33	Private	178506	HS-grad	9	Divorced	Adm-clerical	Not-in-family	Black	Female	0	0
1221	45	Self-emp-not-inc	247379	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0
1276	51	Local-gov	152754	Bachelors	13	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0
1332	20	Private	324469	Some-college	10	Never-married	Machine-op-inspct	Not-in-family	White	Female	0	0

Fig 4.1 Data set with 15 attributes to cluster



Clust 1 Table

sno	Age	Work Class
389	[17-25]	Private
944	[17-25]	Private
1332	[17-25]	Private
1831	[17-25]	State-g
1886	[17-25]	Private
2497	[17-25]	?
2773	[17-25]	Private
3659	[17-25]	Private
4047	[17-25]	Private
4324	[17-25]	Self-em

Clust 2 Table

sno	Age	Work Class
1	[26-45]	State-g
56	[26-45]	Private
112	[26-45]	Private
167	[26-45]	Federa
278	[26-45]	Private
444	[26-45]	Private
555	[26-45]	Federa
666	[26-45]	Private
722	[26-45]	Private
777	[26-45]	Private
999	[26-45]	Private

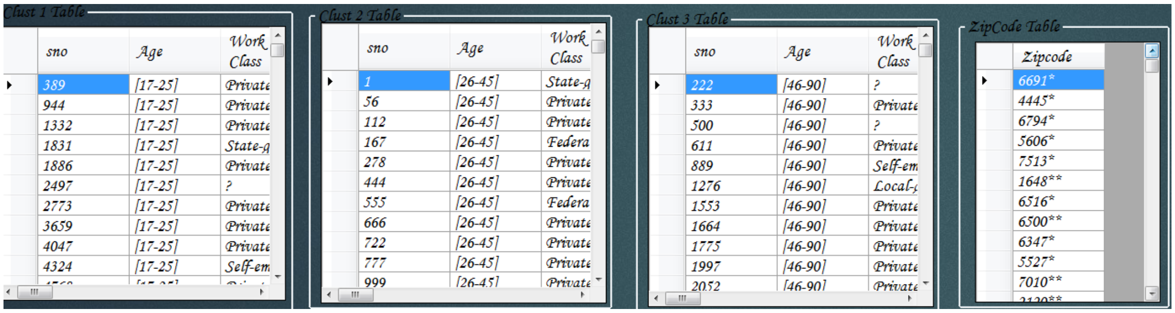
Clust 3 Table

sno	Age	Work Class
222	[46-90]	?
333	[46-90]	Private
500	[46-90]	?
611	[46-90]	Private
889	[46-90]	Self-em
1276	[46-90]	Local-g
1553	[46-90]	Private
1664	[46-90]	Private
1775	[46-90]	Private
1997	[46-90]	Private
2052	[46-90]	Private

ZipCode Table

Zipcode
85788
83181
71521
21067
40849
12919
21767
68413
80271
43565
16058

Fig 4.5 Random Number generation followed by Zip Code



Clust 1 Table

sno	Age	Work Class
389	[17-25]	Private
944	[17-25]	Private
1332	[17-25]	Private
1831	[17-25]	State-g
1886	[17-25]	Private
2497	[17-25]	?
2773	[17-25]	Private
3659	[17-25]	Private
4047	[17-25]	Private
4324	[17-25]	Self-em

Clust 2 Table

sno	Age	Work Class
1	[26-45]	State-g
56	[26-45]	Private
112	[26-45]	Private
167	[26-45]	Federa
278	[26-45]	Private
444	[26-45]	Private
555	[26-45]	Federa
666	[26-45]	Private
722	[26-45]	Private
777	[26-45]	Private
999	[26-45]	Private

Clust 3 Table

sno	Age	Work Class
222	[46-90]	?
333	[46-90]	Private
500	[46-90]	?
611	[46-90]	Private
889	[46-90]	Self-em
1276	[46-90]	Local-g
1553	[46-90]	Private
1664	[46-90]	Private
1775	[46-90]	Private
1997	[46-90]	Private
2052	[46-90]	Private

ZipCode Table

Zipcode
6691*
4445*
6794*
5606*
7513*
1648**
6516*
6500**
6347*
5527*
7010**

Fig 4.6 Representation of gender clustering by * symbol

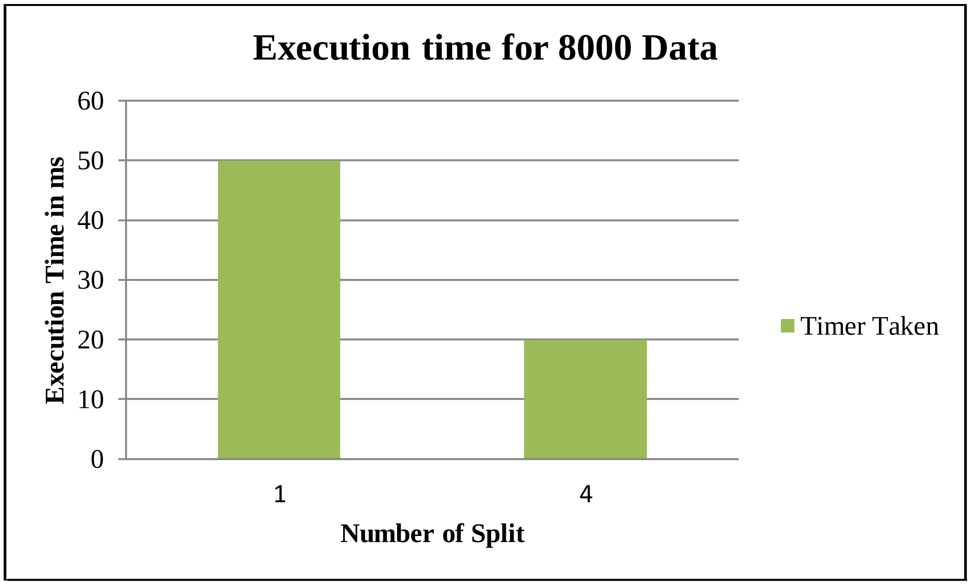


Fig 4.7 Graph comparison between the splits in database

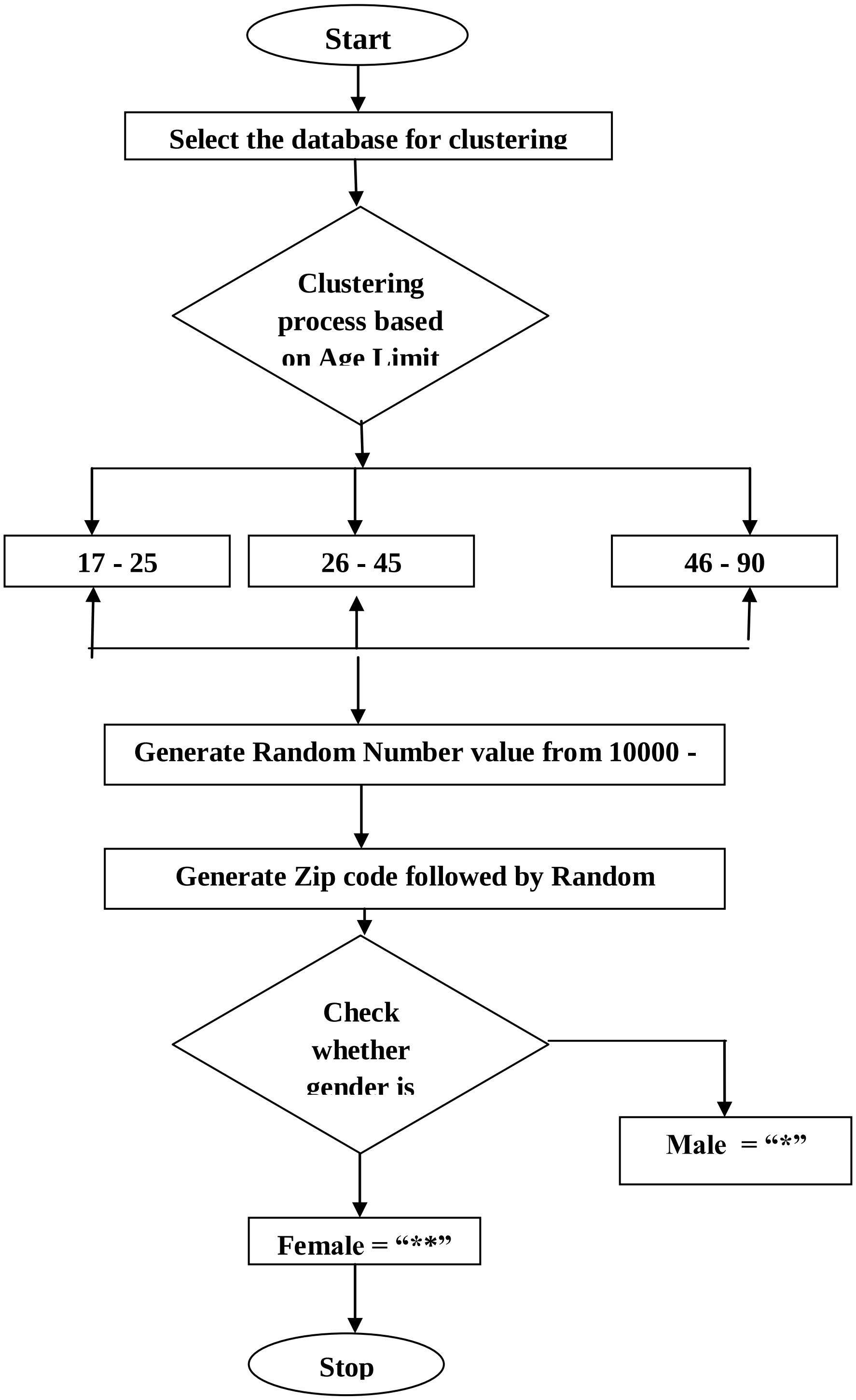

Start
Select the database for clustering
Clustering process based on Age Limit
17 - 25
26 - 45
46 - 90
Generate Random Number value from 10000 -
Generate Zip code followed by Random
Check whether gender is
Male = "*"
Female = "**"
Stop

Figure 4.8: Block Design of the proposed Architecture

V. CONCLUSION

In this paper we present an efficient, privacy-preserving Sequential clustering algorithm in a social network setting. It is simple and strong algorithm to hide the gender information in large databases, by dividing it into multiple cluster groups. The secret code used here splits the dataset into multiple cluster groups based on the age limit of the user. The proposed work achieves more privacy by hiding all sensitive user data from the server and the cluster locations from the users. The algorithm produces secrecy by means of clustering with better utility than achieved by existing algorithms. The random number generation is the crucial step that is further followed by zip code process. This zip code is used in the representation of the sensitive data. The time graph chart shows while creating the clusters the transactions in the datasets becomes easier to access and the time taken for the process is minimally reduced. For this, purpose the clustering technique between multiple datasets remains better.

REFERENCES

[1] Agrawal, R. and R. Srikant, 1998. Fast Algorithms for Mining Association Rules. In: Readings in Database Systems, Stonebraker, M. and J.Hellerstein (Eds.). Morgan Kaufmann, Massachusetts, ISBN: 1558605231, pp: 580-592.

[2]Elena Zheleva and Lise Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data", pp: 1-20.

[3] Ila Chandrakar, Yelipe Usha Rani, Mortha Manasa andKondabala Renuka, "Hybrid Algorithm for Privacy Preserving Association Rule Mining ", in Journal of Computer Science, Vol: 6, No .12, pp: 1494-1498, 2010.

[4] Lior Rokach and Oded Maimon," Data Mining And Knowledge Discovery Handbook"

[5] A.S. Naveenkumar and Punithavalli, "New Framework in Sensitive Rule Hiding", in International .Journal of Computer Techology & Applications,Vol 3, No.1,pp:182-186, Feb 2012.

[6] Seema Kedar, Sneha Dhawale, , Wankhade Vaibhav, "Privacy Preserving Data Mining", in International Journal of Advanced Research in Computer and Communication Engineering,Vol. 2, Issue 4, April 2013, pp:1677- 1680002E

[7] Stanley R. M. Oliveira, and Osmar R. Zaiane, "Revisiting Privacy Preserving Clustering by Data Transformation," Journal of Information and Data Management, vol. 1, no. 1, 2010.

[8] Tamir Tassa and Dror J. Cohen, "Anonymization of Centralized and Distributed Social Networks by Sequential Clustering", In Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 2, February 2013 pp:311-324.

[9] Zekeriya Erkin,Thijs Veugen, Tomas Toft, and Reginald L Lagendijk, "Privacy-preserving distributed clustering ", in EURASIP Journal on Information Security2013, pp:1 -15.