

Privacy-Preserving in Data Mining using Anonymity Algorithm for Relational Data

Karan Dave¹, Chetna Chand²

¹ Gujarat Technological University, Kalol Institute of Engineering, Kalol, Gujarat, India

² Assistant Professor, Gujarat Technological University, Kalol Institute of Engineering, Kalol, Gujarat, India

ABSTRACT

Data mining is the process of analyzing data from different perspectives. To summarize it into useful information, we can consider several algorithms. To protect data from unauthorized user in this case is a problem to solve. Access control mechanisms protect sensitive information from unauthorized users. But if the privacy protected information is not in proper format, again the user will compromise the privacy and quality of data. A privacy protection mechanism can use suppression and generalization of relational data to anonymized and satisfy privacy requirements, e.g., k -anonymity and l -diversity, against identity and attribute disclosure. However, privacy is achieved at the cost of precision of authorized information. In this paper, we propose an accuracy-constrained privacy-preserving access control framework. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k -anonymity or l -diversity. An additional constraint that needs to be satisfied by the PPM is the imprecision bound for each selection predicate. The techniques for workload-aware anonymization for selection predicates have been discussed in the literature. However, to the best of our knowledge, the problem of satisfying the accuracy constraints for multiple roles has not been studied before. In our formulation of the aforementioned problem, we propose heuristics for anonymization algorithms and show empirically that the proposed approach satisfies imprecision bounds for more permissions and has lower total imprecision than the current state of the art.

Keyword : - Data mining, Data Integrity, Data privacy, Anonymization, K anonymity, L diversity.

1. Introduction

The issue of information protection is getting progressively critical for our general public. This can be demonstrated by the very truth that the responsible administration of touchy learning is explicitly being ordered through laws. The difficulties of protection mindful access control are like the issue of workload-mindful anonymization. In our investigation of the related work, we concentrate on inquiry mindful anonymization. They additionally present the issue of exactness compelled anonymization for a given bound of adequate data misfortune for every identicalness class [9]. Databases inside the globe zone unit are regularly gigantic and modern. The test of questioning such implant in a convenient manner has been concentrated on by the database, information handling and learning recovery groups, however rarely examined inside the security and protection area.

1.1 Privacy Preservation

Data Protection or safeguarding information has risen to address the protection issues in information mining. Implanting protection into information mining has been a dynamic and productive exploration zone. Late research in the zone of PPDM has dedicated much push to decide an exchange off amongst protection and the requirement for learning disclosure, which is essential with a specific end goal to enhance basic leadership process and other human exercises. Speculation comprises of substituting quality qualities with semantically

reliable however less exact values. Concealment alludes to evacuating a specific trait esteem and supplanting events of the quality with an exceptional esteem "?", showing that any quality can be put information bother (Applying Rotation, interpretation, and turmoil.)

1.2 Motivation

Most of the Privacy preserving technique includes data mining concepts of clustering and classification. Preservation of large data for different purposes at a different time will be useful to many researchers and analyst to handover secure data. Data Privacy should be preserved in case of data handover for further analysis.

2. Literature Review and Motivation

The formulation of the accuracy and privacy constraints as the problem of k-anonymous partitioning with Imprecision Bounds (k-PIB) and give hardness results. Second, they introduced the concept of accuracy-constrained privacy-preserving access control for relational data. Then, heuristics approach to approximate the solution of the k-PIB problem and conduct empirical evaluation.

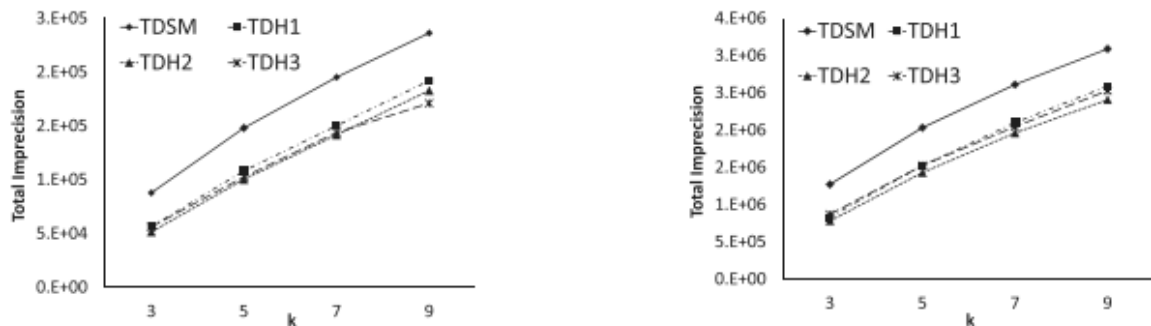


Fig -1: No of queries violating bounds for k-anonymity for given data set

2.1 A Survey on Security and Accuracy Constrained Privacy Preserving Task Based Access Control Mechanism for Relational Data

Since it's a review paper, they have studied and proposed distinctive techniques for protection safeguarding system. Here they have proposed it on the base of errand based control where work process administration frameworks come into center where there are different undertaking to be apportioned. While apportioning errand, the director or dependable individual ought to remember protection for security of information and data.

2.2 Privacy Preserving Suppression Algorithm for Anonymous Databases

Confidentiality means only authorized users can read the data. Usually confidentiality can be achieved by using some cryptographic tools. Not only confidentiality, but anonymization [1] is still required to provide privacy. Suppose a medical facility connected with a research institution and the researchers can use the medical details of a patient without knowing the personal details. Thus the research data base used by the researchers must be anonymized (Sanitized).

2.3 Algorithms

In suppression based anonymization, mask the Quasi-Identifiers value using a special symbol like * and in Generalization based anonymization method, replace a specific value with a more general one using Value Generalization Hierarchies (VGH).

2.3.1 Steps

STEP 1: X encrypt the tuple T, and send it to Y.

STEP 2: Y can decrypt tuple T and then suppress the personal identifiers in the tuple.

STEP 3: After the suppression check the nonsuppressed attributes in the tuple T and loaded tuples.

STEP 4: If any match found, insertion can be performed and send a status message "INSERTED".

STEP 5: If no match found, discard the tuple and send the status message "IGNORE".

3. Implementation strategies

The heuristics proposed in this paper for accuracy-constrained privacy-preserving access control are also relevant in the context of workload-aware anonymization. The anonymization for continuous data publishing has been studied in literature. In this paper the focus is on a static relational table that is anonymized only once. To exemplify our approach, role-based access control is assumed. However, the concept of accuracy constraints for permissions can be applied to any privacy-preserving security policy, e.g., discretionary access control.

Advantages:

1. Accuracy constrained privacy preserving access.
2. It maintains data in secure manner.

	QI ₁	QI ₂	S ₁
ID	Age	Zip	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	Diarrhea
4	25	15	Fever
5	22	28	Flu
6	32	35	Fever
7	38	32	Flu
8	35	25	Diarrhea

(a) Sensitive table

	QI ₁	QI ₂	S ₁
ID	Age	Zip	Disease
1	0-20	10-30	Flu
2	0-20	10-30	Fever
3	20-30	10-30	Diarrhea
4	20-30	10-30	Fever
5	20-30	10-30	Flu
6	30-40	20-40	Fever
7	30-40	20-40	Flu
8	30-40	20-40	Diarrhea

(b) 2-anonymous Table

Fig -2: Display of sensitive and anonymous tables

3.1 ANONYMIZATION WITH IMPRECISION BOUNDS:

The problem of k-anonymous Partitioning with Imprecision Bounds and present an accuracy-constrained privacy-preserving access control framework. Imprecise data means that some data are known only to the extent that the true values lie within prescribed bounds while other data are known only in terms of ordinal relations. Imprecise data envelopment analysis (IDEA) has been developed to measure the relative efficiency of decision-making units (DMUs) whose input and/or output data are imprecise. In this paper, we show two distinct strategies to arrive at an upper and lower bound of efficiency that the evaluated DMU can have within the given imprecise data. The optimistic strategy pursues the best score among various possible scores of efficiency and the conservative strategy seeks the worst score. In doing so, we do not limit our attention to the treatment of special forms of imprecise data only, as done in some of the studies associated with IDEA. We target how to deal with imprecise data in a more general form and, under this circumstance, we make it possible to grasp an upper and lower bound of efficiency.

3.2 Accuracy-Constrained Privacy-Preserving Access Control:

An accuracy-constrained privacy-preserving access control mechanism. (Arrows represent the direction of information flow), is proposed. The privacy protection mechanism ensures that the privacy and accuracy goals are met before the sensitive data is available to the access control mechanism. The permissions in the access control policy are based on selection predicates on the QI attributes. The privacy protection mechanism is required to meet the privacy requirement along with the imprecision bound for each permission.

3.2.1 Algorithms:

CALCULATION: L-DIVERSITY CALCULATION IN LIGHT OF PROTECTION MEASURES

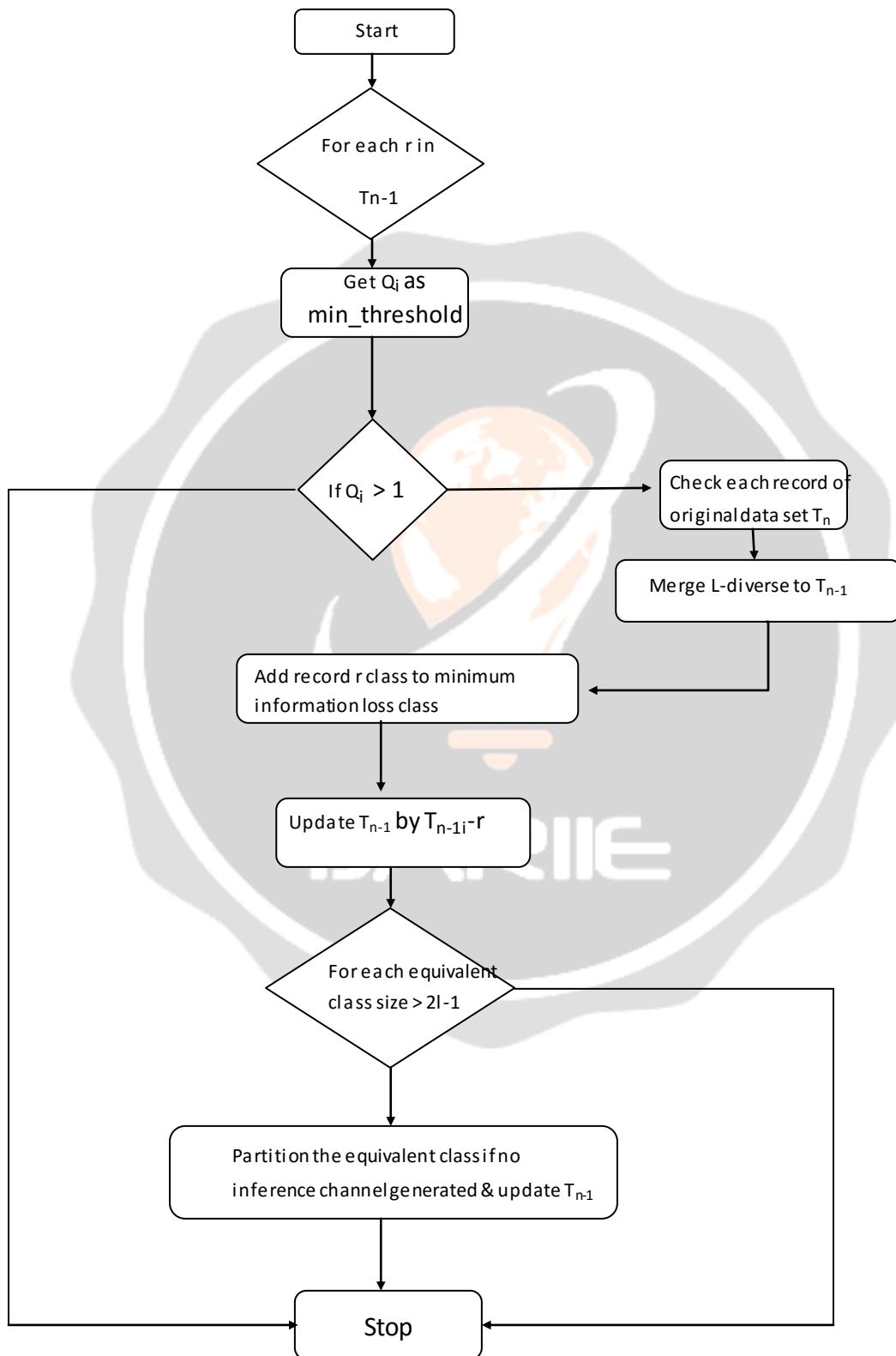
INFORMATION: A RELEASABLE DATASET $TN-1$, AN INCREMENTAL DATASET $\Delta TN-1$, AND AN ASSORTED QUALITIES LIMIT ESTEEM L .

Δ : COMMUNICATING THE RELATIONSHIP BETWEEN A SECTION AND AN ENTIRETY.

YIELD: A RELEASABLE DATASET TN , WHICH GUARANTEES THAT EVERY EQUALITY CLASS HAS THE SAME TOUCHY CHARACTERISTIC QUALITIES SET PRIOR AND THEN AFTERWARD UPGRADE AND HAS NEGLIGIBLE DATA MISFORTUNE.

1. GO TO STEP 2 IF THE QUANTITY OF TOUCHY TRAIT VALUES IN $\Delta TN-1$ IS NOT AS MUCH AS L .
- A. $TN = TN-1$.
- B. MERGE THE AUTONOMOUS L-DIFFERENT COMPARABILITY CLASSES PRODUCED FROM $\Delta TN-1$ WITH TN .
- C. REMOVE THE COMPARING RECORDS FROM $\Delta TN-1$.
- D. FOR EVERY RECORD R IN $\Delta TN-1$
 2. GENERATE THE HOPEFUL PROPORTIONALITY CLASSES C_r IN TN AS PER ITS DELICATE QUALITY WORTH;
 3. INSERT THE RECORD R INTO A CHOSE COMPETIT OR IDENTICALNESS CLASS, WHICH COMES ABOUT THE NEGLIGIBLE DATA MISFORTUNE;
 4. $\Delta TN-1 = \Delta TN-1 - R$.
 5. FOR EVERY PROPORTIONALITY CLASS WHOSE SIZE IS MORE THAN $2L-1$ AND EVERY DELICATE PROPERTY ESTIMATION EXISTS NO LESS THAN TWO TIMES
 6. PARTITION THE IDENTICALNESS CLASS IF NO DERIVATION CHANNELS ARE CREATED.
 7. RETURN.

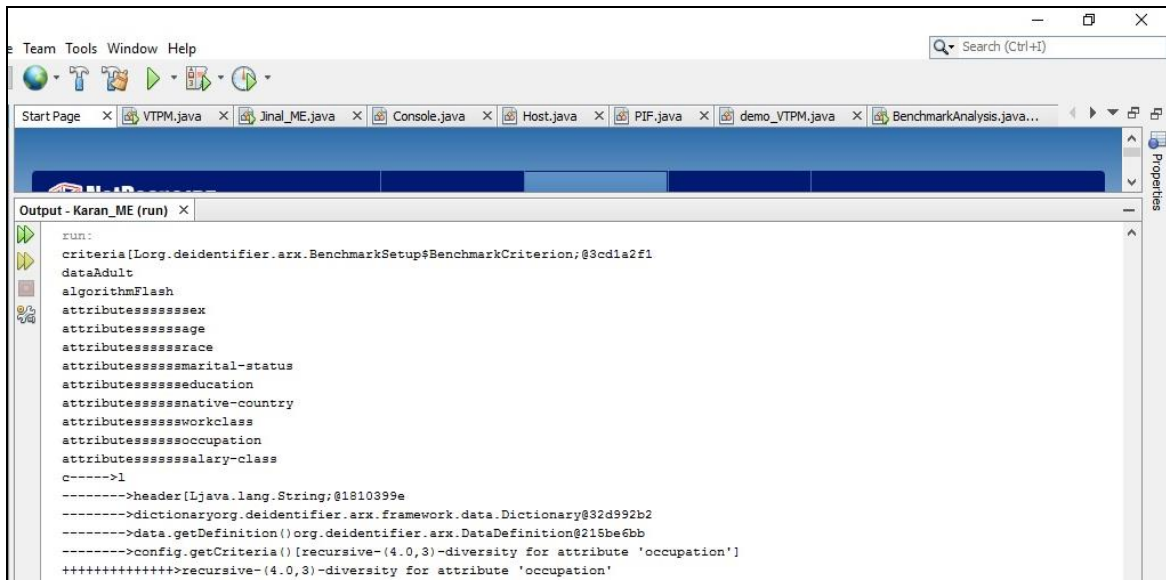
3.2.2 FLOWCHART OF PROPOSED WORK



4. Experiment Result

We ran our trials on the Adult Database from the UCI Machine Learning Repository and the Germen Credit Database. The Adult Database contains 45,222 tuples from Germen Credit information and the Germen Credit Database contains 1000 occasions of credit data. We evacuated tuples with missing values and received the same space speculations.

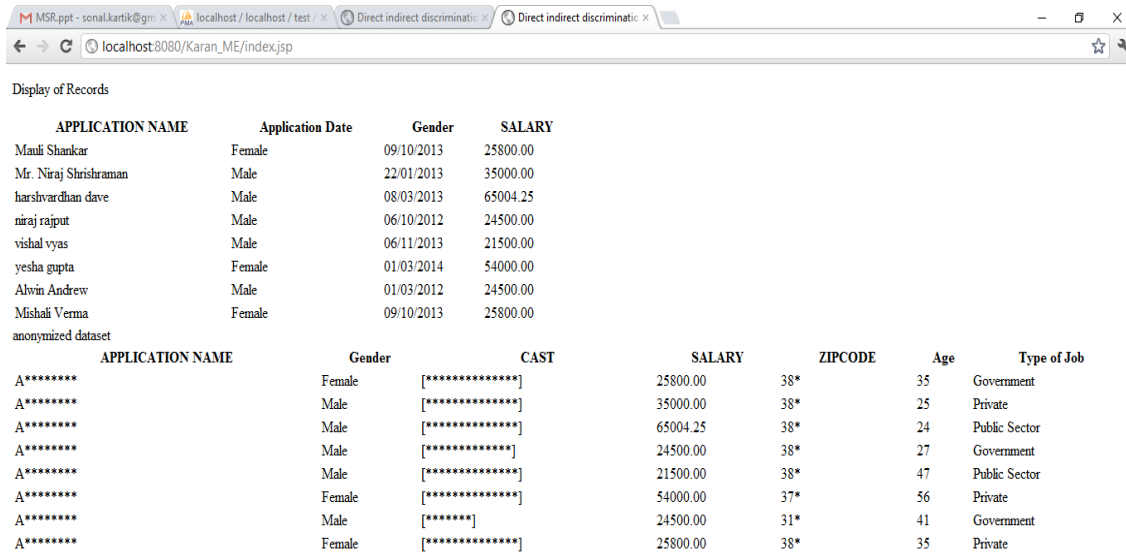
4.1 ANONYMIZATION OVER ADULTDATA SET



4.2 SAMPLE DATA SET

	Appl_No	Applicant_Name	Gender	cast	Comp_Name	Appl_Date	Salary	ZipCode	Age	Employer_Name	Type_Of_Job	Email_Id
<input type="checkbox"/>	10351001	Mauli Shankar	Female	hindu, brahmin	M.H.Logistics	2013-10-09	25800.00	380065	35		Government	ms01299ak@mhlogistics.co
<input type="checkbox"/>	10351002	Mr. Niraj Shrishraman	Male	hindu, brahmin	nisant pharma	2013-01-22	35000.00	380061	25	Keyur Patel	Private	nishant012322k@gmail.com
<input type="checkbox"/>	10351003	harshvardhan dave	Male	hindu, brahmin	tisan consturctions	2013-03-08	65004.25	380061	24	mr. n.k.charurvedi	Public Sector	harsh345@gmail.com
<input type="checkbox"/>	10351004	niraj rajput	Male	hindu, rajput	G.E.P.C	2012-10-06	24500.00	380055	27		Government	abc_niraj@abcpharma.com
<input type="checkbox"/>	10351005	vishal vyas	Male	hindu, brahmin	Kiran Motors	2013-11-06	21500.00	380061	47	Mr. Narendra Patel	Public Sector	v_vishal@gmail.com
<input type="checkbox"/>	10351006	yesha gupta	Female	hindu, solanki	V.K.Engineers pvt ltd	2014-03-01	54000.00	370025	56	Mr. Mahesh Sharma	Private	yesha1112@gmail.com
<input type="checkbox"/>	10351007	Alwin Andrew	Male	English	Guj Agro	2012-03-01	24500.00	310025	41		Government	andrew_anenn@gmail.com
<input type="checkbox"/>	10351008	Mishali Verma	Female	hindu, brahmin	M.H.Logistics	2013-10-09	25800.00	380065	35		Private	ms01299ak@mhlogistics.co

4.3 ANONYMIZATION OVER SAMPLED DATA SET OVER APPLICANT'S NAME



Display of Records

APPLICATION NAME	Application Date	Gender	SALARY
Mauli Shankar	Female	09/10/2013	25800.00
Mr. Niraj Shrishraman	Male	22/01/2013	35000.00
harshvardhan dave	Male	08/03/2013	65004.25
niraj rajput	Male	06/10/2012	24500.00
vishal vyas	Male	06/11/2013	21500.00
yeshu gupta	Female	01/03/2014	54000.00
Alwin Andrew	Male	01/03/2012	24500.00
Mishali Verma	Female	09/10/2013	25800.00

anonymized dataset

APPLICATION NAME	Gender	CAST	SALARY	ZIPCODE	Age	Type of Job
A*****	Female	[*****]	25800.00	38*	35	Government
A*****	Male	[*****]	35000.00	38*	25	Private
A*****	Male	[*****]	65004.25	38*	24	Public Sector
A*****	Male	[*****]	24500.00	38*	27	Government
A*****	Male	[*****]	21500.00	38*	47	Public Sector
A*****	Female	[*****]	54000.00	37*	56	Private
A*****	Male	[*****]	24500.00	31*	41	Government
A*****	Female	[*****]	25800.00	38*	35	Private

5. REFERENCES

- [1]. I-Diversity: Privacy Beyond k-Anonymity Ashwin , Machanavajhala Johannes Gehrke Daniel Kifer Muthuramakrishnan Venkitasubramaniam, Department of Computer Science, Cornell University {mvnak, johannes, dkifer, vmuthu} @cs.cornell.edu – release -2012
- [2]. Privacy Preserving Suppression Algorithm for Anonymous Databases Ebin P.M 1, Brilley Batley. C 2 1,2 AMIE, Assistant Professor Department of Computer Science & Engineering, Hindustan University, Chennai, India pmebin74@gmail.com .(IJSR), India Online ,ISSN: 2319- 7064. Volume 2 Issue 1, January 2013
- [3]. [14]A Survey on Security and Accuracy Constrained Privacy Preserving Task Based Access Control Mechanism for Relational Data Pratik Bhingardev 1, D. H. Kulkarni21, 2 Pune University, Smt. Kashibai Navale College of Engineering, Vadgaon (BK), Pune-411041, India –IJSR-Feb-2013.
- [4]. Aggarwal, G., Feder, G., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D. and Zhu, A.: Achieving Anonymity via Clustering, In Proc. of ACM PODS, (2006), pp.153-162.
- [5] Atzori, M., Bonchi, F., Giannotti, F., and Pedreschi, D.: Anonymity Preserving Pattern Discovery, VLDB Journal, accepted for publication, (2008).
- [6] Bayardo, R. J. and Agrawal, R.: Data Privacy through Optimal k-Anonymization, In Proc. of ICDE, (2005), pp.217-228.

[7] Ghinita, G., Karras, F P., Kalnis, P., and Mamoulis, N.:Fast Data Anonymization with Low Information Loss, InVLDB, (2007), pp.758-769.

[8] Ghinita, G., Tao, Y., and Kalnis, P.: On the Anonymization of Sparse High-Dimensional Data, In Proceedings of ICDE, (2008).

[9] Privacy Preserving Data Mining* Yehuda Lindell , Department of Computer Science Weizmann Institute of Science Rehovot, Israel. lindell@wisdom.weizmann.ac.il

[10]Privacy Preserving Data Mining Cynthia Dwork and Frank McSherry. 2012.

[11]IJRITCC ISSN: 2321-8169 Volume: 3 Issue: 4 Security Management Methods in Relational Data Suhasini Gurappa .Metri PG Student, CSE Dept Cambridge institute of technology ,Bangalore ,India.

[12]Zahid Pervaiz, Walid G.Aref, Arif Gafoor, “Accuracy constrained privacy preserving access control mechanism for relational databases” IEEE Transaction on Knowledge Engineering, vol.26, No.4, April 2014, pp.795-807 .

