Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Privacy Preserving updates to Personalized Anonymity Based Anonymous and Confidential Database

Rajeshwari Suryawanshi*
Department of Computer Science& Engineering
Rashtrasant Tukdoji Maharaj Nagpur University
Abha Gaikwad Patil College of
Engineering,Nagpur, India

Sulabha Patil
Department of Computer Science& Engineering
Rashtrasant Tukdoji Maharaj Nagpur University
Tulsiramji Gaikwad Patil College of
Engineering,Nagpur, India

Abstract— Privacy of individual's information in datasets is main concern in the present technological phase. Thus it is becoming an increasingly important issue in many data mining applications in various fields like medical research, hospital records maintenance, intelligence agencies etc. Many previous works has focused on generalization and suppression based anonymity which provides same amount of privacy preservation to all individuals. The Paper focuses on devising private update techniques to database systems that supports notions of anonymity different than k-anonymity. Therefore the concept of personalized anonymity is used which performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata. In this paper, Personalized Privacy is achieved by using Greedy Framework and optimal SA(sensitive Attribute)-generalization to protect privacy of individual . On the personalized anonymous database updates are performed and an comparative analysis of information loss for k-anonymity and personalized anonymity is carried out.

Keywords: Anonymous database, Generalization, K-Anonymity, Personalized Anonymity, SA-generalization

I. INTRODUCTION

The database becomes an important asset for many applications and thus their security is crucial. Today there is an increased concern for privacy. The availability of huge numbers of databases recording a large variety of information about individuals makes it possible to discover information about specific individuals by simply correlating all the available databases. Data confidentiality and privacy are important concept for achieving Security of databases, but they are different concepts: data confidentiality is about the difficulty (or impossibility) by an unauthorized user to learn anything about data stored in the database. Usually, confidentiality is achieved by enforcing an access policy, or possibly by using some cryptographic tools. Privacy relates to what data can be safely disclosed without leaking sensitive information regarding the legitimate owner.

Anonymized or anonymization means remove personal identifier to protect private information. There are many ways of anonymization. Data anonymization enables transferring information between two organizations, by converting text data into non human readable form using encryption method. There have been lots of techniques developed. K-anonymization is one of the approaches. This technique protects privacy of original data by modification. So problem arises at this point where database needs to be updated. So when tuple is to be inserted in the database problems occurs relating to privacy and confidentiality that is database owner decide that whether database preserve privacy without knowing what new tuple to be inserted. The existing methods focus on a universal approach that exerts the same amount of preservation for all persons, without catering for their concrete needs. The consequence is that we may be offering insufficient protection to a subset of people, while applying excessive privacy control to another subset. Motivated by this, we present a new generalization framework based on the concept of personalized anonymity. Our technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata. It is often necessary to publish personal information for research purposes. For example, a hospital may release patients' diagnosis records so that researchers can study the characteristics of various diseases. The raw data, also called microdata, contains the identities (e.g. names) of individuals, which are not released to protect their privacy. However, there may be other attributes that can be used, in combination with an external database, to recover the personal identities.

II. LITERATURE SURVEY

Privacy preserving techniques use some form of transformation on the data in order to perform the privacy preservation. Some of the techniques to achieve privacy on the database before publishing it for research purpose are discussed below. Along with these privacy preserving updates using K-Anonymity technique is also discussed. The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records [10]. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. The paper in [1] author proposed a formal protection model named k-anonymity for privacy

de-identification. It prevents the attack by suppressing and generalizing the Quasi-identifier attributes which can combine with public records and uniquely identify the records. A microdata release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals in microdata. This paper also verifies re-identification attacks that can be realized on releases that adhere to k-anonymity. Since k-anonymity does not provide sufficient protection against attribute disclosure. The paper in [2] proposes the model of l-diversity. As k-anonymity protects the microdata released table against identity disclosure, it is insufficient to provide attribute disclosure. L-diversity requires that each equivalence class of dataset should have at least l represented values for sensitive attribute. Its limitation is that it is possible for an adversary to gain information about the sensitive attribute if the attacker has knowledge about global distribution of the attribute. In the paper [3] author proposes novel privacy called t-closeness and showed that l-diversity has number of limitations as it is difficult to achieve and insufficient to prevent attribute disclosure. If the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the whole table is less than or equal to threshold t then the equivalence class is having t-closeness. These highly limit the amount of individual specific knowledge an attacker can learn**.** In the paper [4] the author proposed technique that performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the micro data. It illustrates how the k-anonymity requirement can be translated, through the concept of quasi-identifiers, in terms of a property on the released table.

The personalized anonymity specifies degree of privacy for his/her sensitive values. K-anonymity has several drawbacks as discussed in the paper [5]. A k-anonymous table may lose considerable information from the microdata and may allow an adversary to derive the sensitive information of an individual with 100% confidence. To achieve the concept of personalized anonymity is proposed and develop new generalization framework that takes into account customized privacy requirements. These technique prevents fully privacy intrusion and results in generalized tables that permit accurate aggregate analysis In the paper [6] the author suggested paper deals with problems concerning privacy and confidentiality such that updates can be performed without revealing the contents of tuples and DB to the user or data provider. It exerts the same amount of preservation for all persons, resulting in more information loss in microdata release. The first protocol is aimed at suppression-based anonymous databases which allow the database owner to anonymized the tuple without gaining any information about the individual specific data and without sending new tuples owner newly generated data. The second protocol is aimed at generalization-based anonymous databases, and it works mainly on a secure set intersection protocol, to provide privacy-preserving updates on a generalization-based k-anonymous database. In the paper [7] the author proposed the techniques which address the problems of efficiently and privately computing set intersection database oriented operations. It formalize the notion of minimal information sharing across In these paper the author proposed protocols for three operations Intersection, Intersection size and Equijoin and proved that these protocols disclose minimal information apart from query results. It then gives cost analysis for these protocols and estimation of execution times of the application examples. It has two limitations. It do not address the problem of what the parties might learn by combining the results of multiple queries and how to find which database contains which tables and what are the attributes names.

In the paper [8] they discuss the relationship between privacy preserving and SMC and problems involved. It reviews definitions and constructions for secure multiparty computation and discusses the issue of efficiency and demonstrates the difficulties involved in constructing highly efficient protocols. In this paper [9] the two protocols are proposed to perform private updates on anonymous database. But these protocols have limitations, of not supporting to generalization-based updates, which is the main strategy adopted for database anonymity. Therefore, if the database is not anonymous with respect to a tuple to be inserted, the insertion cannot be performed. One of the protocols proposed in the paper is not efficient.

III. PROPOSED WORK

The Existing system [6] has K- anonymized Database DB by generalizing and suppressing the tuples before performing private updates. In addition, k-anonymity fails to guarantee safe publication, even in the scenario with no personal preferences. Assume that the information related to a patient is stored in a tuple t of Database is kept confidentially at the server. The insertion of information about new patient in the anonymous database DB can be performed if the updated database DB U t is still anonymous. Since Database contains privacy sensitive data, main concern is to protect the identity of patient. So the database is K-anonymized by perform.ing Generalization and suppression. While inserting a tuple in anonymous database, the main concern is to protect the identity of patient. Therefore before inserting the tuple it is anonymized and then it is inserted in Anonymous database. But the existing method provides same amount privacy to all person which leads to unnecessary information loss.

As k-anonymity has several drawbacks, the concept of personalized anonymity is used. The proposed system is a new generalization framework based on the concept of personalized anonymity, as k-anonymity has several drawbacks. A simple taxonomy on attribute Disease is accessible by the public It organizes all diseases as leaves of a tree as shown in Fig 3.1. An intermediate node carries a name summarizing the diseases in its sub tree. A personal preference can be taken when she/he is supplying their data. In our approach, a preference is formulated through a node in the taxonomy. To achieve personalized anonymity greedy Framework algorithm is used. It works in two steps. In the first steps a generalization function for every QI attribute is chosen and the generalized value is obtained for all tuple t Є T. The Generalized tuple are divided into QI-Group. In the second step SA-generalization uses a different function for each group. This strategy achieves less Information loss, by allowing each group to decide the amount of necessary generalization. SA-generalization results in less precise values on sensitive attribute, it retains more information on the QI attributes

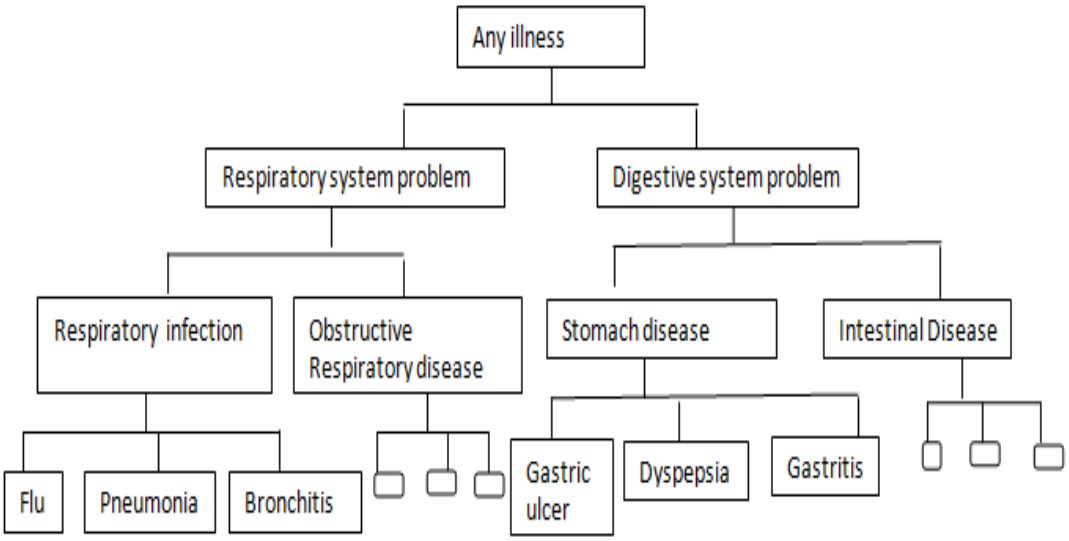

Fig 3.1: Taxonomy for Disease

A. Design View of system

The Modules will have different views to ensure the privacy of patient as shown in Fig 3.2. As the requirements is to achieve the personalized privacy by having access to different attributes as per their needs. In the project there are different views for Administrator, Patient, Doctor and Researcher.

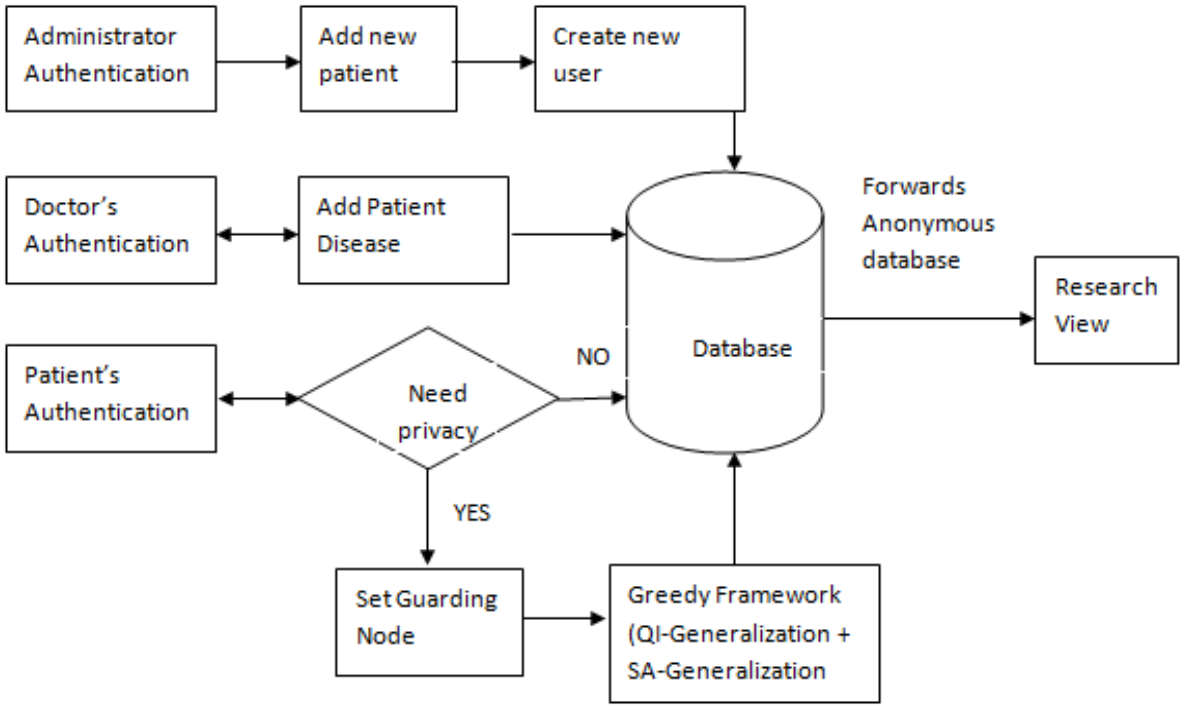

Fig 3.2: Design view of system

IV. **Personalized Anonymity**

The steps to carry out to perform personalized Anonymity needs a Careful Study. Firstly, we study the concept that underlies a greedy framework for computing privacy-conscious information taking into account individual preferences. As opposed to k-anonymity, the approach applies direct protection against the association between individuals and their sensitive values. In the Second step analysis of theory behind the methodology personalized anonymity to get the formulas for quantifying privacy breach and information loss. The mathematical equations prove that K-anonymity can /cannot ensure safe data publication. In particular, k-anonymity (even its improved version "l-diversity" [2]) cannot guarantee privacy protection if an individual may correspond to multiple tuples in the microdata.

A. Formulas to calculate Probability breach and Information Loss

1) Probability Breach: As discussed in previous section ,to infer the sensitive value A_s for an target individual O_{tar}

an adversary can reconstruct ℰG(o_{tar}) then n_{recon} represents the number of reconstruction and n_{breach} represents the number of reconstructions from which adversary ca.n associate the sensitive value with the individual. It means the reconstructions that violates the privacy enforced by $t_{tar.}$GN equation to calculate the probability breach for tuples are given as

$$P_{breach}(t_{tar}) = \begin{cases} b/n & \text{if } t^*_{tar}.A^s \text{ is in SUBTR}(t_{tar}.GN) \\ b.c/n & \text{otherwise} \end{cases} \quad (1)$$

2) Information loss: Let v be a value in the domain of attribute A. The $IL_{value}(v^*)$ to calculate the amount of information loss in generalizing v to v^*, which is a partition in the corresponding general domain . The equation to calculate the information loss :

$IL_{value}(v^*)$ =((the number of values in v*) – 1)/ (the number of values in the domain of A) (2)

For instance, if the domain of Age is [1, 60], generalizing age 5 to [1, 10] has information loss $IL_{value}(v^*)$ ([1, 10]) = (10 − 1) / 60. Similarly, since the taxonomy of Disease has 12 leaves, generalizing flu to respiratory-infection results in $IL_{value}(v^*)$ (respiratory-infection) = (3 − 1)/12, where 3 is the number of leaves under respiratory infection. if v is not generalized (i.e., v = v^*), $IL_{value}(v^*)$ (v^*) equals 0, i.e., no information is lost. The information loss of a generalized tuple t^* is given by:

$$\mathbf{w^s.IL_{value}(t^*.A^s) + \sum_{i=1}^{d} w_i^{qi}.IL_{value}(t^*.A_i^{qi})} \quad (3)$$

V. Implementation

In this section, the performance of Personalized Anonymity algorithm is evaluated. We compare information loss of personalized anonymity algorithm with the K-anonymity algorithms. This section experimentally evaluates the effectiveness of our project. The dataset contains a relation with 10 tuples, each contain information of an individual. The relation has 6 columns: Name, Age, sex, Zip code, Disease, treatment. The second and fourth columns are numerical, whereas Gender are categorical; these 3 columns are the QI attributes. As described in earlier sections Disease is the sensitive attribute. The domain for the disease includes disease names for respiratory system problem and digestive system problem which constitutes the leaves of the taxonomy. The disease belonging to the same group is grouped as child of level-2 node. Recursively the every two level nodes are grouped under three level nodes. These results in two level three nodes which are the children of the root. A unique ID is added as Patient_ID field to each tuple to obtain a primary relation.

The values to be used for age and zip code are fixed. The age for the patient should be inserted in between 1 to 60. The zip code values should be inserted in between 1-40000. It should be of five digits. The maximum permissible breach probability P_{breach} is fixed to 0.25. As mentioned in Section 4, our generalization algorithm requires penalty factors $w_1^{qi} \dots \dots \dots \dots w_5^{qi}$ (for the 5 QI attributes) and w_s. In all cases,$w_1^{qi} \dots \dots \dots \dots w_5^{qi}$equal 1. The value of ws will be varied in different experiments. For each tuple in the first (or second) group, its guarding node is the parent of its sensitive value. The guarding nodes of the tuples in the last group are their sensitive values. The QI generalization is performed on the attributes age, sex, and zip code. After generalization QI-groups are formed and each QI group is given as input to the SA-generalization where probability breach is calculated. In the greedy framework information loss for generalized Table is calculated.

Probability breach for each new updated tuple will be displayed in the demo view. The original view of patient table contains the guarding node attribute. These attribute contains the value set by the patient as their guarding node if they select that he/she needs the privacy. If patient does not need the privacy then the all information will be given to the researcher as seen in the research view. If the need privacy value from patient view is yes, then greedy framework executes to get generalized tuple for that patient information to be displayed in the research view.

The value of k for k-anonymity equals 2. The value of w_s is fixed to 1. In the following experiments, each breach probability is computed from formula given in section 4(B) equation (1). K-anonymity cannot achieve the required level of protection, because the breach probabilities of some tuples are significantly higher than p_{breach} = 0.25. As mentioned in Section 1, k- anonymity prevents accurate association between individuals and tuples, but does not provide direct protection against association between individuals and sensitive values. Both l-diversity and personalized guarantee adequate preservation. SA-generalization is beneficial compared to pure quasi-identifier generalization; it results in generalized tables that permit more accurate data analysis.

A. Results for personalized Anonymity Information Loss

In the experiment, the information loss of generalized table is calculated for k-anonymity and personalized anonymity. The information loss is also calculated for each generalized tuple in personalized anonymity. The analysis of information loss is carried out after update is performed on the patient table as shown in Table5.1. The formula used to calculate the information loss is given in section 4. The information loss for table is calculated by considering information loss for each tuple by using equation (3). To get the information loss of each tuple information loss for one attribute value is calculated. The personalized anonymity information loss for patient table is calculated for quasi Identifiers along with sensitive attribute. But for k anonymity information loss is only calculated for quasi identifier. The Final research view to be published is as shown in Fig 5.1

Greedy Framework (QI-generalization + Optimal SA generalization) RESULTS

After Actual Algorithm Implementation (Module 2)

System Breach Probability = 0.25
Probability Breach For The recent Tuple = 0
Informatio Loss = 13.766433333333

Sr. No.	Name.	Age	Sex	Zipcode	Disease
1	Andy	[1-30]	Male	[1-30000]	Digestive System Problem
2	Bill	[1-30]	Male	[1-30000]	Dyspepsia
3	Ken	[1-30]	Male	[1-30000]	Respiratory System Problem
4	Nash	[1-30]	Male	[1-30000]	Bronchitis
5	Joe	[1-30]	Male	[1-30000]	Pneumonia
6	Sam	[1-30]	Male	[1-30000]	Pneumonia
7	anu	[1-30]	Male	[1-30000]	Digestive System Problem
8	Linda	21	Female	58000	Respiratory Infection
9	jame	[1-30]	Female	[30001-60000]	Stomach disease
10	sarah	[1-30]	Female	[30001-60000]	Respiratory System Problem
11	ritu	[1-30]	Female	[30001-60000]	Respiratory Infection
12	ritu	25	Female	34567	

Figure 5.1 : Research view to be published

TABLE I PERFORMANCE OF PERSONALIZED ANONYMITY in TERMS of INFORMATION LOSS

Sr.No.	**Need privacy**	**Probability Breach**	**Level of guarding node from taxonomy tree**	**Information loss for Personalized Anonymity**	**Information Loss for K-anonymity**
1	YES	0.33	1	8.87	11.8
2	YES	0.33	2	9.77	12.78
3	YES	0.67	3	10.67	13.77
4	NO	0	0	10.67	14.75
5	YES	0	0.	11.32	15.73

The Fig 5.2 shows the comparison for information loss of both personalized and K-anonymity techniques. It is clear from the figure that information loss is always less in personalized anonymity. The research view has a link Results as shown in Fig 5.1. The Results which will display Bar graph of information loss for personalized anonymity and K-anonymity as shown in Fig 5.2.

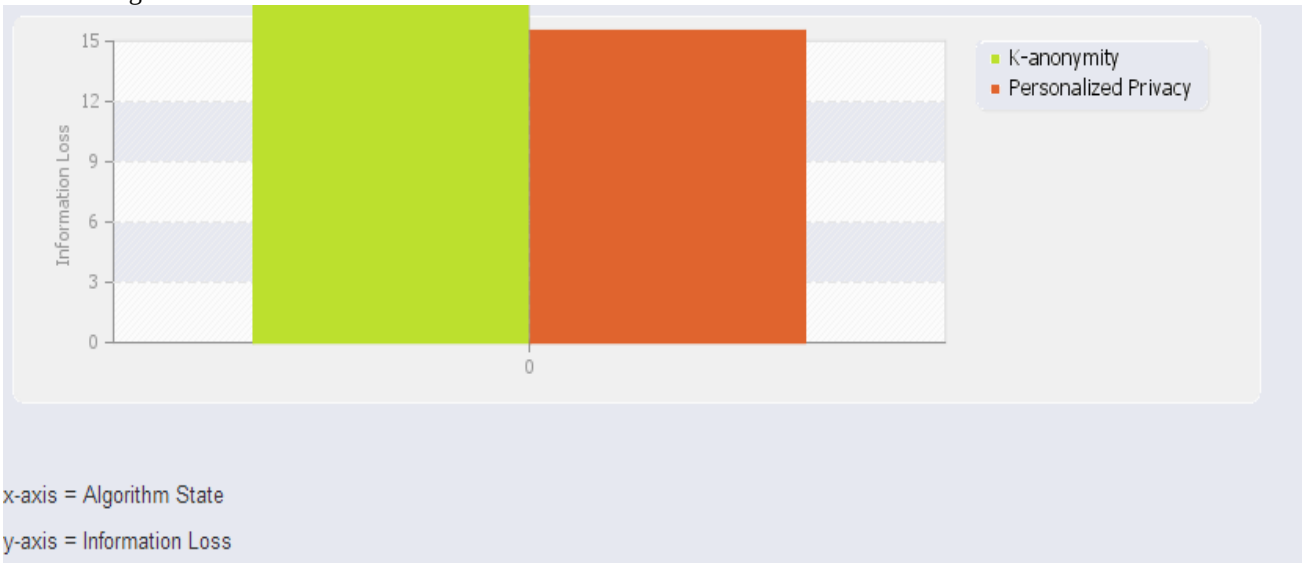

Fig 5.2: Information Loss Comparison Graph.

VI. Conclusion And Future Scope

Generalization using k-anonymity is inadequate because they cannot guarantee privacy protection in all cases, and often incur unnecessary information loss by performing excessive generalization. So the concept of Personalized Anonymity is becoming more important .In this paper, we work with the concept of personalized anonymity, and updates will be performed on these personally anonymized databases by using Greedy Framework .Whenever a new tuple is inserted the individual will decide the level of privacy from taxonomy tree for sensitive attributes . These works serve as a foundation to develop generalization strategies. The greedy framework implemented is not optimal as it does not necessarily achieve lowest information loss. In future another anonymity technique that further reduces the information loss. The researcher might need the data for performing different analytical task. The information may be utilized to release a table that will be used for these analytical tasks.

References

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[2] A.Machanavajjhala, J.Gehrke, et al., *ℓ-diversity: Privacy beyond k-anonymity*, In Proc. of ICDE, Apr.2006.B

[3] N. Li, T. Li, and S. Venkatasubramanian, *t-Closeness: Privacy Beyond k-anonymity and l-Diversity*, In Proc. Of ICDE, 2007, pp. 106-115.

[4] P.Samarati, "Protecting Respondent's Privacy in Microdata Release", IEEE Trans.Knowledge and Data Eng.,vol.13,no.6,pp. 1010-1027, Nov./Dec. 2001. W. and Marchionini, G. 1997.

[5] Xiaokui Xiao, Yufei Tao "Personalized Privacy Preservation", SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA. Copyright 2006 ACM 1595932569/ 06/0006

[6] Alberto Trombetta ,Wei Jaing,Elisa Bertino and Lorenzo Bossi, "Privacy Preserving Updates to anonymous and Confidential database" IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 8, NO. 4, JULY/AUGUST 2011.

[7] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing across Private Databases," Proc. ACM SIGMOD Int'l Conf.Management of Data, 2003.

[8] Yehuda Lindell and Benny Pinkasy,"Secure Multiparty Computation for Privacy-Preserving Data Mining" 2005

[9] A.Trombetta and E. Bertino, "Private Updates to Anonymous Databases," Proc. Int'l Conf. Data Eng. (ICDE), 2006.

[10] Aggarwal C. C., *On Randomization, Public Information and the Curse of Dimensionality*, ICDE Conference, 2007.

