IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 5 (Mar. - Apr. 2013), PP 36-43
www.iosrjournals.org

Privacy Preservation for Knowledge Discovery: A Survey

Jalpa Shah[1], Mr. Vinit kumar Gupta[2]
[1]*Department of Computer Engineering, Hasmukh Goswami College of Engineering, Vehlal, Gujarat*
[2]*Department of Computer Engineering, Hasmukh Goswami College of Engineering, Vehlal, Gujarat*

Abstract: *Today's globally networked society places great demand on the dissemination and sharing of information. Privacy Preservation is an important issue in the release of data for mining purposes. How to efficiently protect individual privacy in data publishing is especially critical. With releasing of microdata such as social security number disease by some organization should contain privacy in data publishing. Data holders can remove explicit identifiers to gain privacy but other attributes which are in published data can lead to reveal privacy to adversary. So several methods such as K-anonymity, L-diversity, T-closeness, (n,t) closeness, (α,k)-anonymization, p-sensitive k-anonymity and others method come into existence to maintain privacy in data publishing.*

Keywords *– Data anonymization, Generalization, Data suppression*

I. INTRODUCTION

There are so many organizations who publish their data in various forms. These forms contain various information. Information can be helpful for someone and at the same time can be useless for another one. Some information may be important for business point of view, industrial point of view that depends on person to person. So which information is sensitive i.e. we do not want to disclose it for general people and which information can be published. So caring of these issues, organization needs to publish their information. As for example in a hospital system a lot of patient comes for their treatment in respective departments. Hospital need to maintain their records and make a file for that which contains patient information. They want to publish reports such that information remains practically useful and the important thing is that identity of an individual can not be determined. So publishing of data is main concern here. Organization needs to publish microdata. Microdata e.g. Medical data, voter registration and census data for research and other purposes. These data are stored in a table. Each record corresponds to one individual. Microdata is a valuable source of information for the allocation of public funds, medical research, and trend analysis. However, if individuals can be uniquely identified in the microdata then their private information (such as their medical condition) would be disclosed, and this is unacceptable. Each record has number of attributes, which can be divided into three categories. (1) Explicit identifiers attributes that clearly identify an individual. E.g. - social security number. (2) Quasi-identifiers attributes whose value when taken together can identify an identify an individual.e.g. Zip-code, birth date and gender. (3) Attributes those are sensitive such as disease and salary. It is necessary to protect sensitive information of individuals from being disclosed. There are two types of information disclosure identity disclosure and attribute disclosure [9]. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. If there is only one female black dentist is in area and sequence queries reveal that she is in database then identification occurs. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, to limit the disclosure risk to an acceptable level while maximizing the benefit. This can be done by anonymizing the data before release. By knowing the quasi identifiers can lead to know the sensitive information. This can be done by knowing the individual personally or other publicly available database.

1.1 Motivation

Huge databases exist in the society example medical data, census data, data gathered by government agencies. The rapid evolution of storage, processing and computational technologies is changing the traditional information system architecture adopted both private companies and public organization. This change is necessary for two reasons First the amount of information held by organization is increasing very quickly thanks to large storage capacity and computational power of modern devices. Second the data collected by organization contains sensitive information (e.g. identifying information, financial data, and health diagnosis) whose confidentiality must be preserved. More and more healthcare system collect sensitive information about

historical and present hospitalization, and more in general health condition of patients. Since these data are associated with identity of patients. As a consequence, any healthcare system should adopt an adequate privacy protection system, which guarantees the protection of sensitive attribute. Companies and agencies that collect such data often need to publish and share the data for research and other purposes. However, such data usually contains personal sensitive information, the disclosure of which may violate the individual's privacy.

II. Theoretical Background And Review

2.1 Theoretical background

Some important terms regarding data preservation for data publishing

2.1.1 Information Disclosure Risk

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Three types of information disclosure have been identified in the literature [7–9]: membership disclosure, identity disclosure, and attribute disclosure. **Membership Disclosure** When the data to be published is selected from a larger population and the selection criteria are sensitive (e.g., when publishing datasets about diabetes patients for research purposes), it is important to prevent an adversary from learning whether an individual's record is in the data or not. **Identity Disclosure** Identity disclosure (also called re-identification**)** occurs when an individual is linked to a particular record in the released data. Identity disclosure is what the society views as the clearest form of privacy violation. If one is able to correctly identify one individual's record from supposedly anonymized data, then people agree that privacy is violated. In fact, most publicized privacy attacks are due to identity disclosure. In the case of GIC medical database [4], Sweeney reidentified the medical record of the state governor of Massachusetts. In the case of AOL search data [5], the journalist from New York Times linked AOL searcher NO. 4417749 to Thelma Arnold, a 62-year-old widow living in Lilburn, GA. And in the case of Netflix prize data, researchers demonstrated that an adversary with a little bit of knowledge about an individual subscriber can easily identify this subscriber's record in the data. When identity disclosure occurs, it has been said that "anonymity" is broken. **Attribute Disclosure.** Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm [8]. An observer of the released data may incorrectly perceive that an individual's sensitive attribute takes a particular value and behave accordingly based on the perception. This can harm the individual, even if the perception is incorrect. In some scenarios, the adversary is assumed to know who is and who is not in the data, i.e., the membership information of individuals in the data. The adversary tries to learn additional sensitive information about the individuals. In these scenarios, our main focus is to provide identity disclosure protection and attribute disclosure protection. In other scenarios where membership information is assumed to be unknown to the adversary, membership disclosure should be prevented. Protection against membership disclosure also helps protect against identity disclosure and attribute disclosure: it is in general hard to learn sensitive information about an individual if you don't even know whether this individual's record is in the data or not.

2.1.2 Data Anonymization

While the released data gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the data. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. Privacy attacks that use quasi-identifi ers to re-identify an individual's record from the data is also called *re-identification attacks* .To prevent re-identification attacks, further anonymization is required. A common approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. For example, age **24** can be generalized to an age interval **[20 − 29]**. As a result, more records will have the same set of quasi-identifier values. We define a *QI group* to be a set of records that have the same values for the quasi- identifiers. In other words, a QI group consists of a set of records that are indistinguishable from each other from their quasi-identifiers.

2.2 Review

Many researchers have found several approaches for data preservation for data publishing. So several methods such K-anonymity, L-diversity, T-closeness and others are come into existence to maintain privacy in data publishing. In this paper we discussed pros and cons of all these techniques.

2.2.1 K-anonymity

Samarati and Sweeney [10], [11] proposed a definition of privacy called k-anonymity. Each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In other words, k-anonymity requires that each equivalence class contains at least k records. So there is need to hide the information. So to remove these anomalies k-anonymity is an approach in this direction. The attacker can join the non-sensitive data to identify the sensitive attribute. Generalization and suppression are the techniques of anoymization. In generalization replace the original value by a semantically consistent but less specific value. In suppression data not released at all it is suppressed on cell level or tuple level. Figure 1(a) shows medical records from a hospital located in New York. This table does not contain uniquely identifying attributes like name, social security number, etc. In this example, division of the attributes into two groups: the sensitive attributes (consisting only of medical condition) and the non-sensitive attributes (zip code, age, and nationality). An attribute is marked sensitive if an adversary must not be allowed to discover the value of that attribute for any individual in the dataset.

TABLE-1 Inpatient Microdata

	Non sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	13053	28	Nepali	Heart disease
2	13068	29	Japanese	Viral
3	13068	21	Nepali	Viral
4	14853	50	Chinese	Cancer
5	14853	55	Indian	Viral
6	14850	47	Chinese	Viral
7	13053	31	Nepali	Cancer
8	13053	37	Nepali	Cancer
9	13068	36	Indian	Cancer

Here the collection of attributes {zip code, age, nationality} is the quasi-identifier for this dataset. Figure 3 represents a 3-anonymous table derived from the table in Figure 2(a) Here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range [14850−14859] and "age=3*"means the age is in the range [30-39]. Suppression is done by cell level or tuple level. Here there are 9 records in this table. So 3-anonymous table would be having three equivalence classes. In this 3-anonymous table, each tuple has the same values for the quasi-identifier as at least three other tuples in the table. So k-anonymity protects against identity disclosure.

TABLE-2 3-anonymous Inpatient Microdata

	Non Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	< 30	*	Heart disease
2	130**	< 30	*	Viral
3	130**	< 30	*	Viral
4	140**	> 40	*	Cancer
5	140**	> 40	*	Viral
6	140**	> 40	*	Viral
7	130**	3*	*	Cancer
8	130**	3*	*	Cancer
9	130**	3*	*	Cancer

2.2.2 Homogeneity attack

K-anonymity suffers from homogeneity attack [4] can create groups that leak information due to lack of diversity in the sensitive attribute. From the previous example Alice and Bob are neighbours. One day Bob falls ill and is taken by ambulance to the hospital. After seeing the ambulance, Alice wants to know what kind of

disease Bob is suffering from. Alice can see the 3-anonymous table of current inpatient records published by the hospital which is shown in the previous figure, and so she can know that one of the records in this table contains Bob's data. Since Alice is Bob's neighbour, she knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob's record number is 7, 8 or 9 Now, all of those patients have the same medical condition (cancer), and so Alice concludes that Bob has cancer. This is not uncommon consider a dataset containing 60,000 distinct tuples where the sensitive attribute can take 3 distinct values and is not correlated with the non sensitive attributes. 5-anonymization of this table will have around 12,000 groups and, on average, 1 out of every 81 groups will have no diversity (the values for the sensitive attribute will all be the same). Thus we should expect about 148 groups with no diversity. Therefore, information about 740 people would be compromised by a homogeneity attack. This suggests that in addition to k-anonymity, the sanitized table should also ensure "diversity" all tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes.

2.2.3 Background knowledge attack

K-anonymity does not protect against attacks based on background knowledge. Alice has a closed friend named Umeko who is admitted to the same hospital as Bob, and whose patient records also appear in the table shown in Figure 2. Alice knows that Umeko is a 21 year old Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko's information is contained in record number 1, 2 or 3 without additional information; Alice is not sure whether Umeko caught a virus or has heart disease. However, it is well known that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection. After knowing the background knowledge an intruder can identify the sensitive attributes. So k-anonymity does not protect against homogeneity attack and background knowledge attack. It protects against identity disclosure very well but failed to protect against attribute disclosure.

2.2.4 L-Diversity

To address the limitation of K-anonymity Machanavajjhala [4] an equivalence class is said to have l - diversity if there are at least "well represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. The meaning of well represented would be to ensure there are at least l-distinct values for the sensitive attribute in each equivalence class. Distinct l-diversity does not prevent probabilistic inference attacks. An equivalence class may have one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value. This leads to the development of stronger notions of l-diversity.

2.2.5 Similarity attack

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. For example consider this example [7], this is the original table and next figure [7] shows anonymized version which satisfy distinct and entropy 3-diversity. Here there are two sensitive attributes salary and disease. Suppose sum one knows that Bob's record corresponds to one of the first three records, then one knows that Bob's Salary is in the range [3K–5K] and can infer that Bob's salary is relatively low. This attack applies not only to numeric attributes like salary, but also to categorical attributes like disease. Knowing that Bob's record belongs to the first equivalence class enables one to conclude that Bob has some stomach-related problems, because all three diseases in the class are stomach-related. So here sensitive information is revealed. So l-diversity also has certain limitation. It does not well protect attribute disclosure very well.

TABLE-3 Original Salary/Disease table [7]

	Non sensitive			Sensitive
	Zip code	Age	Salary	Disease
1	47677	29	3k	Gastric ulcer
2	47602	22	4k	Gastritis
3	47678	27	5k	Stomach cancer
4	47905	43	6k	Gastritis
5	47909	52	11k	Flu
6	47606	47	8k	Bronchitis
7	47675	30	7k	Bronchitis

8	47603	36	9k	Pneumonia
9	47607	32	10k	Stomach cancer

TABLE-4 3-Diverse Version of Original Table [7]

	Non sensitive			Sensitive
	Zip code	Age	Salary	Disease
1	476**	2*	3k	Gastric ulcer
2	476**	2*	4k	Gastric
3	476**	2*	5k	Stomach cancer
4	4790*	>40	6k	Gastritis
5	4790*	>40	11k	Flu
6	4790*	>40	8k	Bronchitis
7	476**	3*	7k	Bronchitis
8	476**	3*	9k	Pneumonia
9	476**	3*	10k	Stomach cancer

2.2.5 T-Closeness [3]

L-diversity suffers from similarity attack. L-diversity requires that each equivalence class has at least l well represented values for each sensitive attribute. Ninghui li, Tiachang li and S. Venkatasubramanian proposed t-closeness to deal with Similarity attack. T-closeness says that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more then a threshold t. A table is said to have t-closeness if all equivalence class have t-closeness. T-closeness with EMD handles the difficulties of l-diversity. EMD has been choosing instead of other distance metric. For numerical attributes ordered distance has been used and for categorical attributes hierarchical distance formula has been used.

Say Q = { 3k,4k,5k,6k,7k,8k,9k,10k,11k},P1 ={3k,4k,5k} and P2 = {6k,8k,11k}.from the given values of Q, P1,P2 we can calculate D[P1,Q] and D[P2,Q] using EMD. We have D [P1, Q] = 0.375 and D [P2, Q] = 0.167.

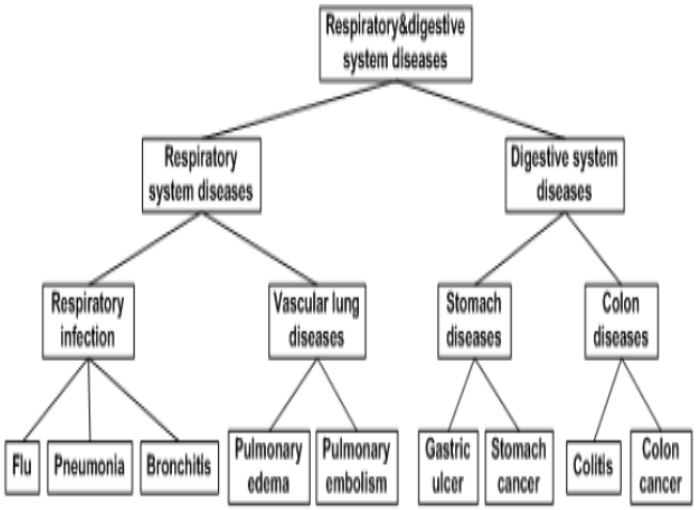

Fig.1. (a) Hierarchical for categorical attribute Disease [4]

For the disease attribute, use of hierarchy as shown in Figure to define the ground distances. For example, the distance between "Flu" and "Bronchitis" is 1⁄3, the distance between "Flu" and "Pulmonary embolism" is 2⁄3, and the distance between "Flu" and "Stomach cancer" is 3⁄3 = 1. Then the distance between the distribution {gastric ulcer, gastritis, stomach cancer} and the overall distribution is 0.5 while the distance between the distribution {gastric ulcer, stomach cancer, pneumonia} is 0.278. So anonymized version of previous table can be generated. It has 0.167 closeness w.r.t Salary. The similarity attack is prevented in the following table.

TABLE-60.167 closeness w.r.t salary [3]

	Non sensitive			Sensitive
	Zip code	Age	Salary	Disease
1	4767*	< 40	3k	Gastric ulcer
2	4767*	< 40	4k	Stomach cancer
3	4767*	< 40	9k	Pneumonia
4	4790*	> 40	6k	Gastritis
5	4790*	> 40	11k	Flu
6	4790*	> 40	8k	Bronchitis
7	4760*	< 40	4k	Gastritis
8	4760*	< 40	7k	Pneumonia
9	4760*	< 40	10k	Stomach cancer

For example, Alice can not infer that Bob has a stomach related disease based on this table**.**

2.2.5 Motivation for (n,t) Closeness [3]

T- Closeness limits the release of useful information through the following example table shown below that contains 3000 individuals and next table shows an anonymized version of it. The sensitive attribute is disease. Count is a column that indicates number of individuals. The probability of cancer among the population in the dataset is 700/3000 = 0.23. While the probability of cancer among individuals among first equivalence class is 300/600 = 0.5. Since (0.5-0.23 > 0.1) the anonymized table does not satisfy 0.1-closeness.

TABLE-7 Original Patient Table [3]

	Zip code	Age	Disease	Count
1	47673	29	Cancer	100
2	47674	21	Flu	100
3	47605	25	Cancer	200
4	47602	23	Flu	200
5	47905	43	Cancer	100
6	47904	48	Flu	900
7	47906	47	Cancer	100
8	47907	41	Flu	900
9	47603	34	Cancer	100
10	47605	30	Flu	100
11	47602	36	Cancer	100
12	47607	32	Flu	100

To achieve 0.1-closeness, all tuples in Table have to be generalized into a single equivalence class. This results in substantial information loss. If we consider the original patient data in Table, we can see that the probability of cancer among people living in zipcode 476** is as high as 500/1000 = 0.5. While the probability of cancer among people living in zipcode 479** is only 200/2000 = 0.1. We can see here people living in zipcode 476** have a much higher rate of cancer will be hidden if 0.1-closeness is enforced.

TABLE-8 Violating 0.1 Closeness [3]

	Zip code	Age	Disease	Count
1	476**	2*	Cancer	300
2	476**	2*	Flu	300
3	479**	4*	Cancer	200
4	479**	4*	Flu	1800
5	476**	3*	Cancer	200
6	476**	3*	Flu	200

2.2.6 (n, t) Closeness [3]

The (n, t) closeness model requires that every equivalence class of a table must contains at least n records and the distance between the two distributions of the sensitive attribute in the equivalence class is not more than the threshold t. An equivalence class E1 is said to have (n, t)-closeness if there exists a set E2 of records that is a natural superset of E1 such that E2 contains at least n records, and the distance between the two distributions of the sensitive attribute in E1 and E2 is no more than a threshold t. A table is said to have (n, t)-closeness if all equivalence classes have (n, t)-closeness. In the definition of the (n, t)-closeness, the value of n defines the size of the viewer's background knowledge. A smaller n means that the viewer knows the sensitive information about a smaller group of records. The value of t limits the amount of sensitive information that the viewer can get from the released table. A smaller t implies a stronger privacy requirement. Choosing the parameters n and t would affect the level of privacy and utility. The larger n is and the smaller t is, one achieves more privacy, and less utility. It does not deal with multiple sensitive attributes. Suppose we have two sensitive attributes like Salary and Disease. One can consider the two attributes separately, i.e., an equivalence class E has (n, t)-closeness if E has (n, t)-closeness with respect to both Salary and Disease.

2.2.7 (α, k)-anonymization

WONG R C et al. propose an (α, *k*)-anonymity model to protect both identifications and relationships to sensitive information in data and to limit the confidence of the implications from the quasi-identifier to a sensitive value (attribute) to within α. The model avoids the sensitive information is inferred by strong implications.

(α, *k*)-ANONYMIZATION: A view of a table is said to be an (α, *k*)-anonymization of the table if the view modifies the table such that the view satisfies both *k* anonymity and α- deassociation properties with respect to the quasi-identifier [16].

2.2.8 p-sensitive k-anonymity

Traian Marius Truta and Bindu Vinay introduce *p* sensitive *k*-anonymity that and protects against both identity and attribute disclosure on the base of extending *k* anonymity model.

***p*-sensitive *k*-anonymity property:** The masked microdata satisfies p-sensitive *k* anonymity property if it satisfies *k*-anonymity, and for each group of tuples with the identical combination of key attribute values that exists in masked microdata, the number of distinct values for each confidential attribute occurs at least p times within the same group [17]. *P*-sensitive *k*-anonymity protects against attribute disclosure. On this aspect, it is the same with *t*-Closeness.

2.3 The Approaches Based on Clustering

Clustering is the problem of partitioning a set of objects into groups, such that objects in the same group are more similar to each other than objects in other groups based on some defined similarity criteria. Various approaches based on clustering have been proposed. In [13], the idea of clustering to minimize information loss and ensure good data quality and formulate a specific clustering problem is proposed. The key idea is that data records are naturally similar to each other should be part of the same equivalence class. [14] Achieves anonymity via constraining each cluster must contain no fewer than a pre-specified number of data records.[15] Introduces a family of geometric data transformation methods (GDTMs) that distort confidential numerical attributes. The advantage of these proposes based on clustering are high-accurate and available result.

III. Conclusion And Prospect

Privacy here means logical security of data not the traditional security of data. Here adversary uses legitimate method. K-anonymity is an approach to protect microdata. K-anonymity is done by generalization and suppression techniques to publish the useful information. L-diversity is one step ahead to k-anonymity. But it does not protect attribute disclosure very well. T-closeness comes into picture to remove similarity attack. Some pitfalls in T-closeness leads to (n,t) closeness. We conclude five research directions of privacy preserving approaches for knowledge discovery by analyzing the existing work in future.

1) The research of finding K-anonymity solution with suppressing fewest cells by reducing complexity.
2) The research of personalized privacy preservation will become an issue.
3) How to improve the efficiency of implementation and ensure available of the result in order to meet the various requirements.
4) The research about how to combine the advantage of above approaches.
5) The research about improving the algorithm, generalized for both categorical and numerical values.

Acknowledgement

I would be thankful to my guide assistant professor Vinitkumar Gupta here for fervent help when I have some troubles in paper writing. I will also thank my class mates in laboratory for their concern and support both in study and life.

References

[1] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," proc. Of the int'l conf. on very large data base (VLDB), pp. 901909, 2005.
[2] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k- Anonymization," Proc. Int'l Conf. Data Engineering (ICDE), pp. 217-228, 2005.
[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 106115, 2007.
[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "'l-Diversity: Privacy Beyond k-Anonymity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 24, 2006.
[5] T. M. Truta and B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property," Proc. Int'l Workshop on Privacy Data Management (ICDE Workshops), 2006.
[6] X. xiao and Y. Tao, "Personalized Privacy Preservation," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 229- 240, 2006.
[7] Ninghui Li,tiancheng li and suresh venkatasubramanian. "Closeness: a new privacy measure for data publishing".IEEE, july 2010.
[8] G. T. Duncan and D. Lambert, "Disclosure-Limited Data Dissemination," Journal of The American Statistical Association, vol. 81, pp. 10-28, 1986.
[9] D. Lambert, "Measures of Disclosure Risk and Harm," Journal of Official Statistics, vol. 9, pp. 313-331, 1993.
[10] P. Samarati, "Protecting Respondent's Privacy in Microdata Release," IEEE Trans. on Knowledge and Data Engineering (TKDE) vol. 13, no.6, pp. 1010-1027, 2001.
[11] L.Sweeney,"k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertain.Fuzz.,vol.10,no.5,pp.557-570,2002.
[12] M. Barbaro and T. Zeller, "A face is exposed for AOL searcher no. 4417749," New York Times, 200
[13] Ji-Won Byun, Ashish Kamra, et al, "Efficient k-Anonymization Using Clustering Techniques", In Internal Conference on Database Systems for Advanced Applications (DASFAA), Berli: Spring-Verlag, April. 2007, pp. 188-200.
[14] G. Aggarwal, T. Feder, et al, "Achieving Anonymity via Clustering", Twenty Fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS. Chicago, Illinois. 2006.
[15] Stanley R. M. Oliveira, Osmar R. Zaiane, "Privacy Preserving Clustering By Data Transformation", In Proc. of the 18th Brazilian Symposium on Databases (SBBD), Manaus, Brazil, October 6-10, 2003, pp. 304-318.
[16] WONG R C, LI J, FU A W, et a1, "(α, k) Anonymity：an enhaned kanonymity model for privacy-preserving data publishing", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, New York, 2006, pp. 754-759.
[17] TRUTA T M， VINAY B, "Privacy protectio: p-Sensitive k-anonymity property," Proceedings of the 22nd on Data Engineering Workshops, IEEE Computer Society, Washington Dc, 2006.