ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Implementation of Partial Generalization for Data Publishing with Minimal Information Loss

Urvashi Hiranwar[1], Amit Pimpalkar[2]

M.Tech Student, Dept. of CSE, RTMNU University, G.H.Raisoni Academy of Engineering and Technology, Nagpur, India. [1]

Assistant Professor, Dept. of CSE,RTMNU University, G.H.Raisoni Academy of Engineering and Technology, Nagpur, India[2]

ABSTRACT: Presently, there is a demand for trade and distribution of data among specific gatherings driving these bodies to be spurred by common intrigue that includes certain data to be distributed for research and examination reason; individuals who are subjects to this data have different necessities, benefits and commitments, they require ensure from the gathering that their data is not adjusted, corrupted or uncovered in some other frame. Most past research on privacy-preserving data publishing, in view of the k-secrecy display, has taken after the shortsighted approach of homogeneously giving the same generalized an incentive in every single quasi-identifier inside a parcel. We watch that the anonymization blunder can be decreased on the off chance that we take after a non-homogeneous generalization approach for gatherings of size bigger than k. Such an approach would permit tuples inside a segment to take distinctive generalized quasi-identifier values. Anonymization taking after this model is not unimportant, as its immediate application can without much of a stretch disregard k-obscurity. At that point, we propose a method that produces a non-homogeneous generalization for a segment and demonstrate that its outcome fulfills k-anonymity, however by straight forwardly applying it; privacy can be bargained if the aggressor knows the anonymization calculation. In light of this, we propose a randomization technique that keeps this sort of assault and demonstrate that k-anonymity is not bargained by it. The paper we show here starts the idea of Limited Generalization with ostensible data misfortune which makes the data loose in this manner data can't be re-distinguished while data stays helpful. Non-homogeneous generalization can be utilized on top of any current partitioning way to deal with enhance its utility. Likewise, we demonstrate that another partitioning system customized for non-homogeneous generalization can additionally enhance quality.

KEYWORDS: Limited Generalization; nominal information loss; information privacy

I. INTRODUCTION

The issue of privacy-preserving data publishing has been broadly contemplated since it was first presented in [20]. Consider an extensive table which must be discharged to people in general for research purposes. Privacy is normally traded off via reckless publishing of the table [3], since touchy data might be spilled. In this way, the objective of data publishing is to change the table, to such an extent that people may not be connected to particular tuples with high conviction. In the meantime, the distributed data ought to at present be helpful, so a streamlining issue emerges: anonymizes the data with the end goal that a specific level of privacy is protected while data utility is amplified.

In the table to be distributed, aside from the keys that are smothered before production, there is an arrangement of characteristics called the quasi-identifier (QID). The QID of each tuple is known to the assailant and might be utilized to distinguish a person. A run of the mill case of QID is {ZIP code, sex, date of birth}, which can interestingly distinguish 63% of the populace in 2000 US Census data [8]. The famous k-anonymity rule requires that the likelihood of an enemy having the capacity to discover the character of an anonymized tuple is at most 1/k. The most well-known

DOI: 10.15680/IJIRCCE.2017. 0505076

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

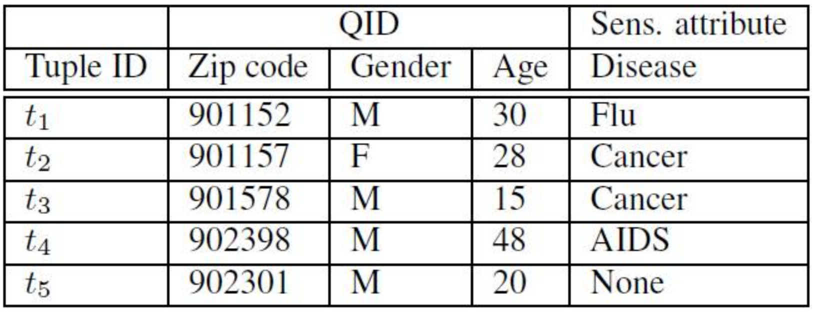
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

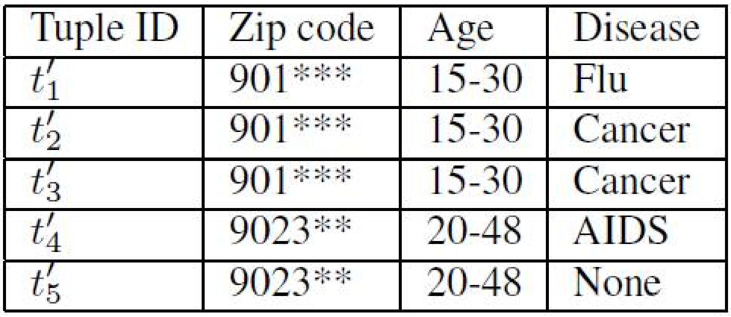
Vol. 5, Issue 5, May 2017

strategy for accomplishing k-anonymity is generalization [13, 14, 10, and 6]. The table is isolated into gatherings having k tuples or increasingly and the QID values in each gathering are generalized to a range containing every unique esteem. Table 2 demonstrates a praiseworthy 2-anonymized table utilizing generalization.



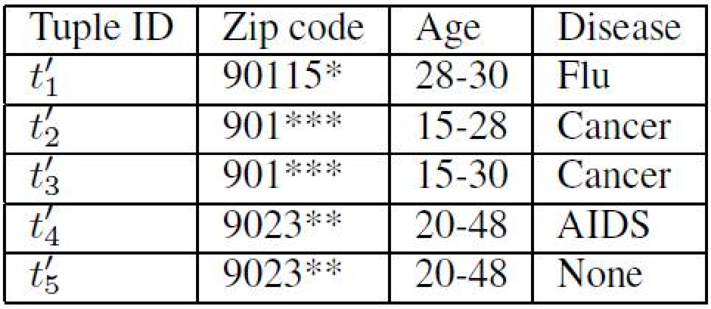
	QID			Sens. attribute
Tuple ID	Zip code	Gender	Age	Disease
t_1	901152	M	30	Flu
t_2	901157	F	28	Cancer
t_3	901578	M	15	Cancer
t_4	902398	M	48	AIDS
t_5	902301	M	20	None

Table 1 Original table



Tuple ID	Zip code	Age	Disease
t'_1	901***	15-30	Flu
t'_2	901***	15-30	Cancer
t'_3	901***	15-30	Cancer
t'_4	9023**	20-48	AIDS
t'_5	9023**	20-48	None

Table 2 2-anonymity using Limited generalization



Tuple ID	Zip code	Age	Disease
t'_1	90115*	28-30	Flu
t'_2	901***	15-28	Cancer
t'_3	901***	15-30	Cancer
t'_4	9023**	20-48	AIDS
t'_5	9023**	20-48	None

Table 3 2-anonymity using Limited generalization and Suppression

The first data are appeared in Table 1. (t'_i in Table 2 is the generalized form of ti in Table 1 for simple reference.) For instance, the period of t3 is initially 15 and after generalization, it is supplanted by the range 15-30. Aside from microdata production, k-anonymity has been generally embraced in applications like area based administrations [19, 11], to secure the personality of question backers. An extensive variety of calculations utilizing generalization are proposed for tending to k-anonymity [10, 14, and 6].

They share a typical structure: first segment the tuples into gatherings, and then appoint the same generalized QID to tuples in a similar gathering. The gathering of tuples with the same QID is called proportionality class. Such an approach, to which we allude as homogeneous generalization, brings up an imperative issue: does generalization need to be homogeneous? For instance, consider the conceivable distribution of Table 1, as appeared in Table 3. t'1, t'2 and t'3 have an alternate generalized QID. This generalization is non-homogeneous. Accepting the enemy knows the QIDs of all people contained in Table 1, he can discover the character of any anonymized tuple in Table 3 with likelihood at most 1/2. Consequently, 2-anonymity is fulfilled, as this is likewise the case for Table 2. Then again, on the off chance that we think about the utility of the two tables, we can watch that Table 3 is superior to Table 2, paying little mind to the utility measure utilized; for each tuple and QID trait of Table 3, the generalized range is littler than or equivalent to the comparing range in the relating tuple and characteristic in Table 2. This illustration demonstrates that it is conceivable to accomplish higher utility utilizing non-homogeneous generalization. The possibility of non-homogeneous generalization was first presented in [7], which contemplates strategies with a certification that a foe can't relate a generalized tuple to not as much as k people. Be that as it may, the proposed arrangements don't offer limits for the likelihood of every affiliation. Henceforth, a few people may have higher likelihood to be related to an anonymized tuple than others and this may prompt privacy ruptures.

In this paper, we deliberately concentrate the utilization of non-homogeneous generalization in anonymizing tables. We give a philosophy to checking whether a non-homogeneous generalization damages k-anonymity. At that point, we propose a method that produces a non-homogeneous generalization and demonstrate that its outcome fulfill K-anonymity, however by direct applying it, privacy can be traded off if the aggressor knows moreover the anonymization calculation. In view of this, we propose a randomization strategy that keeps this sort of assault and demonstrate that k-anonymity is not bargained by it. Despite the fact that non-homogeneous generalization can be utilized on top of any current partitioning way to deal with enhance its utility; we demonstrate that another partitioning strategy custom fitted for non-homogeneous generalization can additionally enhance quality. Our fundamental concentration all through the paper is k-anonymity; be that as it may, we likewise talk about how our approach can be stretched out to enhance utility for other privacy standards. An intensive trial assessment shows that our system extraordinarily enhances the nature of anonymized data by and by.

II. RELATED WORK

V. Ciriani, et al. [21] has examined how k-anonymity can be combined with data mining for shielding the uniqueness of the receiver to whom the data being mined refer also exemplify the two main ways to combine k-anonymity in data mining.

Sara Hajian, et al. [22] studied the influence of full domain generalization on discrimination prevention also making original data privacy-protected by combining k-anonymity and α- protection which can take as parameters several legally grounded trials of discrimination and generate privacy and discrimination-protected full domain generalization.

The notion of k-anonymity was proposed by Samrati and Sweeny [23] based on full domain generalization, replacing quasi-identifier attribute values with a generalized version of them and explored the concepts of domain generalization hierarchy and corresponding value generalization hierarchy.

L-diversity model introduced by A.Machanavajjhala, et al. [24] was premeditated to grip some Achilles' heel in the k-anonymity model since shielding identities to the level of k-individuals is not the same as protecting the equivalent sensitive values particularly when there is homogeneity of sensitive values of group.

Benjamin C. M. Fung, et al. [25] presented views on the differentiation between privacy-preserving data publishing (PPDP) and PPDM, and presented a number of advantageous properties of a PPDP method which are compared and reviewed against other methods in terms of anonymization algorithms, privacy models, anonymization operations and information metrics. Most of these approaches implicitly uses a single release from a single publisher, and thus only protected the data up to the first release or the first recipient.

K-optimize optimal algorithm presented by Bayardo, et al. [26] employ sub-tree generalization and record suppression schemes; it is the only competent optimal algorithm that uses the flexible sub-tree generalization and effectively removes non-optimal anonymous tables by molding the search space by means of set enumeration tree.

Iyengar, et al. [27] proposed a flexible generalization method and applies a genetic algorithm to achieve k–anonymization on the better search space that resulted from it; although the method can maintain a good result quality, it has been argued for being a time-consuming iterative process.

In this perspective, Lunacek et al. [28] introduces a novel crossover operator that can be applied with a genetic algorithm for guarded attribute generalization, and successfully shows that Iyengar's approach can be made quicker.

Incognito algorithm by Lefevre et al.[29] implements a dynamic programming approach which assures subset property which states that a relation T cannot be k-anonymous if it's subset of quasi-identifiers does not satisfy k-anonymity.

III.PROPOSED METHODOLOGY

In this section we propose the technique to achieve k-anonymity by limited generalization with nominal information loss.

A. *Algorithm:*

```
  Input: Table T including set of tuples t1, t2, t3….
Precondition: Tuple should include quasi identifier and sensitive attribute.
  Step 1: Identify Quasi-Identifier
  Step 2:  Calculate Domain
  Step 3:  For each tuple
             T→anonymized (attribute)→(suppression/generalization)
              T→T'
  Step 4:  Calculate Information Loss and Privacy Gain.
  Step 5: T'→compare (External data source)
  Step 6:  T→identifiable
  Step 7: Consider next Quasi-Identifier
  Step 8: Repeat step 2-6
  Step 9: End
```

B. *Framework Description*

Step 1: Data Collection Phase: "Adult" dataset in multiple csv formats are uploaded and saved in user_files_master table.

Step 2: Process Phase: In process phase all the data sets uploaded in csv format are clubbed and uploaded in mysql database table imported _data; there are now total 300 records.

The Adult data set has total 10 attributes with 300 records. Cleaning is performed based on the attribute City containing zip codes values and Age having continues values that must not have any incomplete data, missing data, ambiguous data.

Step 3: Generalization phase 1:We first select the minimum and maximum zipcode values from the CITY attribute.We next generalize the zipcode and arrange it in the ascending order.For each value of generalized zipcode we suppress the last digit by *.We assign the group numbers by labelling them as $C_{groupnumber.}$

While calculating zip code information loss for each value of generalized zip code and then summed it to get total zip code information loss.

The information loss for numeric attribute is slightly different from categorical attribute .Usually, the value of numeric attribute with domain [U, L] is generalized to a specific range, like [U_g, Lg].So, the information loss of this generalization can be calculated by formulae.

$IL_{zip\ code}$= (Current zip code value- minimum zipcode)/maximum zipcode.eq. (1)

Step 4:Calculate sum of zip code information loss is 127.29439219291.The privacy gain is always 1/ sum of zipcode information loss.

$PG_{zipcode}$=1/sum of $IL_{zipcode}$ eq. (2)

Therefore,1/127.29439219291=0.0078558056075597

Step 5:Generalization Phase 2:

Generalization being one of the most used technique to create shielded microdata that satisfies not only k-anonymity but also any of the improved anonymization models.

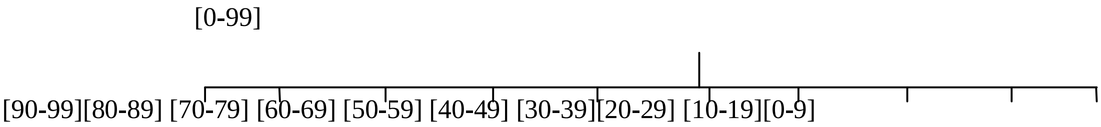


Figure 1: Structure of numeric quasi-identifier Age

The AGE attribute is generalized to specific ranges as shown above.

Step 6:The information loss for numeric attribute is slightly different from categorical attribute .Usually, the value of numeric attribute with domain [U,L] is generalized to a specific range, like[U_g,L_g].So, the information loss of this generalization can be calculated by formulae

IL_{AGE}= (Currrent age value- minimum age)/maximum age….………eq. (3)

Total IL_{AGE}=127.26262626263.

PG_{AGE} =1/sum of IL_{zAGE}….………eq. (4)

Total Privacy Gain=0.0078577664894039.

Step 7: Published Data:Our final released data set is K-anonymized dataset with nominal information loss.

IV. Analysis

A. *Scalability* :

Our analysis is based on the comparison of scalability with Bottom Up generalization.In the base paper of Bottom Up Generalization "Adult " data set have total 45222 records ,ρ(rho) is the novelty factor =3,σ is the distinct non –leaf values of VID =30 where the data has 45,222 * 30=1,356,660 records.For 1,356,660 records, the algorithm took 730 seconds. So, execution time to complete one record is 730/1,356,660 and for 255 records the algorithms is taking (730/1356660)* 255=0.1372 seconds.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Partial generalization:

In our Adult dataset total records are 255 after cleaning,ρ (rho) is the novelty factor=3,σ is the distinct non –leaf values of VID =12 where the data has 255*12=3060 records. For, 3060records, our algorithm take 0.0917 seconds.

Algorithm	Scalability
Bottom Up generalization	0.1372 seconds
Partial Generalization	0.0917 seconds

Table4: Algorithm and their Scalability

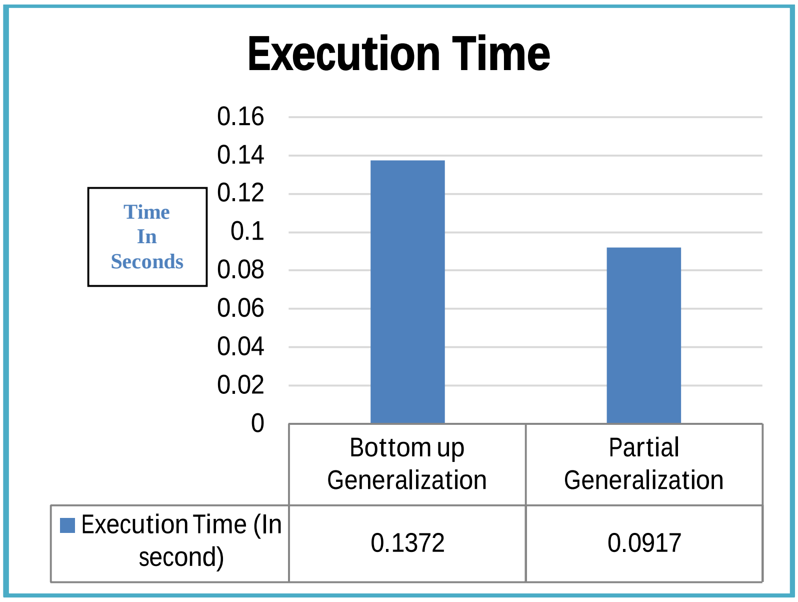


Figure2:Execution Time of Bottom UP generalization and Partial Generalization

Figure 2 represents the time comparison for existing system i.e. Bottom Up Generalization and Proposed mechanism Partial Generalization. The graph shows that the execution time for the corpus of 255 records comes out to be better in Partial Generalization as compared to Bottom Up Generalization, which proves that the time efficiency of our system is better as compared to previous one.

B. *Data Quality* :

To represent the data quality we are considering the data parameters for the calculation as follows:

Training set	100
Testing set	155
Corpus for final Result	255

Following are the results obtained after the implementation

Data Corpus	255	100	155
Information Loss	127.2943	40.4314	86.8629
Published Data	127.2626	40.4208	86.8418
Total	254.5569	80.8522	173.7047

DOI: 10.15680/IJIRCCE.2017. 0505076

From the above results if we calculate the data quality, it comes out to 86.23%.

C. Time Complexity :

The time complexity of our algorithm is O (log k) where k is quasi identifiers generalized present in final published data.

D. Experimental Setup:

The experiment has been implemented using frontend as PHP & MySQL as Backend. As mentioned earlier the dataset for the experiment is Adult dataset obtained from UCI Machine Learning Repository [33].

V. Conclusion and Future Work

In today's information age preserving privacy has become significant.In this paper, our objective is to implement partial generalization a data mining privacy based technique .Our idea is to examine data generalization for masking detailed information in order to achieve k-anonymity along with minimal information loss but also to prevent excess use of generalization that can impact data utility. The released data set we get, after applying partial generalization is inferable to external information sources and also does not have linkage with external data set. It also aims to prevent, excess generalization to gain k-anonymity thus reducing the amount of generalization that give rise to more information loss making data set much more competent for research purpose while preserving privacy.We have focussed on the issue of scalability as compared to Bottom up Generalization our algorithm has higher scalability.We believe our framework is open tofurther extension we can incorporate different metrics thus we plan to investigate in future.

References

1. R. Agrawal and R. Srikant. Privacy-preserving data mining. In SIGMOD, 2000.
2. A. Asuncion and D. Newman. UCI Machine Learning Repository, 2007.
3. M. Barbaro and T. Zeller. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times, 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html.
4. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In ICDE, 2005.
5. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In KDD, 2002.
6. G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis. Fast data anonymization with low information loss. In VLDB, 2007.
7. A. Gionis, A. Mazza, and T. Tassa. K-anonymization revisited. In ICDE, 2008.
8. P. Golle. Revisiting the uniqueness of simple demographics in the US population. In WPES, 2006.
9. Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. In SIGMOD, 2005.
10. T. Iwuchukwu and J. F. Naughton. K-anonymization as spatial indexing: toward scalable and incremental anonymization. In VLDB, 2007.
11. P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias. Preventing location-based identity inference in anonymous spatial queries. IEEE Trans. Knowl. Data Eng., 19(12), 2007.
12. D. Kifer. Attacks on privacy and definetti's theorem. In SIGMOD, 2009.
13. K. Lefevre, D. J. Dewitt, and R. Ramakrishnan. Incognito: efficient full-domain k-anonymity. In SIGMOD, 2005.
14. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In ICDE, 2006.
15. N. Li, T. Li, and S. Venkatasubramanian. T-closeness: Privacy beyond k-anonymity and l-diversity. In ICDE, 2007.
16. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In ICDE, 2006.
17. D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge in privacy. In ICDE, 2007.
18. A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In PODS, 2004.
19. M. F. Mokbel, C. Y. Chow, and W. G. Aref. The new casper: Query processing for location services without compromising privacy. In VLDB, 2006.
20. P. Samarati. Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng., 13(6):1010–1027, 2001.
21. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati ,"K-Anonymous Data Mining: A Survey", Springer US, Advances in Database System (2008).
22. Sara Hajian, Josep Domingo-Ferrer, OriolFarr`as," Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining", Data Mining and Knowledge Discovery, volume 28, Sep 2014.
23. P. Samarati and L. Sweeney "Generalizing data to provide anonymity when disclosing information". In Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS 98), Seattle, WA, June 1998, p. 188.
24. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam , " l-Diversity: Privacy beyond k-anonymity", In the Proceedings of the IEEE ICDE 2006 .

25. Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu,” Privacy-Preserving Data Publishing: A Survey of Recent Developments”, ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010.
26. R. J. Bayardo and R. Agrawal, “Data Privacy through Optimal k Anonymization,” in ICDE 2005: Proceedings of the 21st International Conference on Data Engineering, Tokyo, Japan, 2005, pp. 217–228.
27. V. S. Iyengar, “Transforming Data to Satisfy Privacy Constraints,” in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, 2002, pp. 279– 288.
28. M. Lunacek, D. Whitley, and I. Ray, “A Crossover Operator for the kAnonymity Problem,” in GECCO 2006: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, Seattle, Washington, USA, 2006, pp. 1713–1720.
29. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan,”Incognito: Efficient full-domain k-anonymity," in Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 49{60, ACM, 2005.
30. G. Loukides, A. Gkoulalas-Divanis, “Utility-preserving transaction data anonymization with low information loss”, Expert Systems with Applications, Elsevier 2012.
31. Pingshui WANG, “Survey on Privacy Preserving Data Mining”, International Journal of Digital Content Technology and its Applications, December 2010.
32. CharuAggarwal , Philip Yu,“Models and Algorithms : Privacy-Preserving Data Mining”, Springer 2008
33. U.C.Irvine Machine Learning Repository, http://www.ics.uci.edu/mlearn/repository.html.
34. X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In VLDB ’06: Proceedings of the 32nd international conference on Very large data bases, pages 139–150. VLDB Endowment, 2006.
35. X. Xiao and Y. Tao. Personalized privacy preservation. In Proceedings of ACM Conference on Management of Data (SIGMOD’06), pages 229–240, June 2006.