Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/1721
DOI URL: http://dx.doi.org/10.21474/IJAR01/1721
ISSN NO. 2320-5407

RESEARCH ARTICLE

Efficient Personalized Privacy Preservation Using Anonymization.

Ashwini N. Patil[1] and Prof. R. N. Phursule[2].
1. Student, Dept. of Information Technology, ICOER, Pune.
2. Professor, Dept. of Information Technology, ICOER, Pune.

Manuscript Info

Manuscript History

Received: 12 July 2016
Final Accepted: 19 August 2016
Published: September 2016

Key words:-
Data Publishing, Privacy Preserving, Anonymity Algorithms, Information Metric, Generalization, Suppression.

Abstract

The k-anonymity privacy for publishing micro data requires that each equivalence class contains at least k records. Many authors have studied that k-anonymity cannot prevent attribute disclosure. The technique of l-diversity has been introduced to address this; l-diversity requires that each equivalence class must have at least well-represented values for every sensitive attribute. In this paper, we show that l-diversity has many limitations. In particular, it is not necessary or sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new method to detect privacy which is called as closeness. We first present the base model t-closeness, which includes the distribution of sensitive attributes in any of the equivalence classes is near to the distribution of the attribute in the overall table (i.e., the difference between the two given distributions should be no more than threshold value t). tcloseness that gives higher utility. We present our method for designing a distance measure between given two probability distributions and give two distance measures. Here we discuss the method for implementing closeness as a privacy concern and illustrate its advantages through examples and experiments.

Introduction:-

PRIVACY is very important issue when onewants to make use of data that includes sensitive information. Studies on protecting the privacy of individuals and the confidentiality of data is contributed from many fields, including computer science, statistics, economics. This is an field that attempts to answer the problem of how an organization, such as a hospital, government agency or any organisation, can release data to the people without harming the confidentiality of personal information. We focus on privacy measures that provide legal safety, present algorithms that protect data to make it safe for accessing while preserving useful information, and discuss methodes for analyzing the sensitive data. Many challenges still remain. It provides a summary of the current state, based on which we expect to see advances in years to come. As personal information is collected in increasingly detailed level by various organizations, privacy related concerns are introducing significant challenges to the data management organisations. Data anonymizationmethodes have been proposed in order to allow processing of personal data without compromising users privacy. Nevertheless, practical problems like dependencies between values in personal records do not obtain a satisfying solution. Here, we focus on the anonymization of tree-structured personal records links. Personal information do not comprise just a single tuple in modern information systems. The information concerning a single person usually spans over several tables or it is kept in a more flexible representation as an XML

Corresponding Author:- Ashwini N. Patil.
Address:- Student, Dept. of Information Technology, ICOER, Pune.

record. Such tree structured data could not be anonymized effectively with table based anonymization techniques since the structural relation between different fields substantially differentiates the problem. The difficulty in anonymizing tree structured data has been considered in existing research literature, in the technique of multirelationalkanonymity. In our methode we consider general case for tree structured data and we propose an anonymization method that is not dependent
solely on the generalization of values, but also on the simplification of the data tree.

Literature Survey:-
To introduce the concept of Efficient PersonalizedPrivacy Preservation Using Anonymization. This paper analyzes many concepts of different authors as mentioned below:

In the paper Anonymizing Collections of Tree-Structured Data, Olga Gkountouna and ManolisTerrovitis [1] introduces real-world data which have implicit or explicit structural relations. Privacy preservation has focused on data with a very simple structure, e.g. data with very complex structure such as network graphs, but has ignored intermediate cases. Here we focus on tree structured data. Such data is required from various applications, e.g. XML documents. A example is a database where information about a person is scattered amongst tables that are associated through foreign keys.k(m;n) anonymity, which provides protection and proposes a greedy anonymization technique that sanitizes large datasets.

Q Wang, C. Wang [21] introduces Enabling Public Verifiability and Data Dynamics for Storage Security in Computing, Computing has been thought as the next generation architecture of IT Enterprise. It moves the application software and databases to large data repositories, where managing data and services may not be fully trustworthy. This brings about many new challenges, which are not understood. This work studies the problem of ensuring the integrity of data storage in Computing. We consider the job of allowing a third party auditor (TPA), as a client.TPA removes the involvement ofthe client through the checking if the data stored in the is indeed intact. The support for data by the most general forms of operations performed ondata, such as insertion and deletion, is also a important step toward practicality, since services in Computing are not limited backup data only.

Ateniese [3] developed a dynamic provable data possession protocol based on cryptographic hash function and symmetric key encryption. The main thing is to pre compute a certain number of metadata during the setup period, so that the number of challenges is prevented and fixed beforehand. The author construct a highly efficient and secure PDP technique based largely on symmetric key cryptography. This technique allows outsourcing of dynamic data, that is, it efficiently supports operations, such as block modification, deletion and append.

Juels and B. S. Kaliski [4],introducesHLA Based Solution. It supports public auditing without retrieving data block. It requires constant bandwidth. It is possible to compute an HLA which authenticates a linear combination of the individual data blocks.

N. Cao, S. Yu, S. Yang [5],tells us about Using Virtual Machine. They proposed Virtual machines that use RSA algorithm, for client data encryption and decryptions. Also SHA 512 algorithm is used which makes message digest and check the data integrity. Digital signature is used as a identity measure for client. It solves the problem of unauthorized access, integrity, privacy and consistency.

C.Erway, A.Kupcu [6] ,introduces Non Linear Authentication in which they suggested Homomorphicnon linear authenticator with randomized masking techniques to obtain security. K. Gonvinda proposed digital signature method to protect the privacy and integrity of data. RSA algorithm is used for encryption and decryption which uses the process of digital signatures for message authentication.

S. Marium [7] introduced Extensible authentication protocol through hand shake with RSA. They proposed identity based signature for class conscious architecture. They provide an authentication protocol for computing (APCC) . APCC is more easy and efficient as compared to SSL authentication protocol. Here, Challenge handshake authentication protocol (CHAP) is used. When make request for any data or any service on the . The Service provider authenticator (SPA) orders the first request for
client identity. Following are the steps:
1) When Client request for any service to service provider, SPA sends CHAP request challenge to the client.
2) The Client sends CHAP response or challenges which is calculated by using a hash function to SPA.

3) SPA compares the challenge value and its own calculated value. If they are similar then SPA sends CHAP success message to the client.

Proposed System:-

We have proposed a novel method of privacycalled closeness. We introduce two instantiations: a base model called t-closeness and a more flexible privacy method called (n, t) - closeness. We explain the rationale of the (n, t)-closeness model and show that it gives a better balance between privacy and utility. The (n, t)-closeness model better protects the data while improving the utility of the released data. The t-closeness model was introduced to overcome attacks which were possible onldiversity(like similarity attack). l-diversitymodel uses all values of a given attribute in a similar way(as distinct) even if they are semantically related. All values of an sattribute are not equally sensitive. The algorithm to check (n,t) closeness could be given as follows.

input: P is partitioned into r partitions $\{P_1, P_2, ..., P_r\}$
output: true if (n,t)-closeness is satisfied, false otherwise

```
1. for every P_i
2.     if P_i contains less than n records
3.         find=false
4.         for every Q ∈ Parent(P) and |Q| ≥ n
5.             if D[P_i, Q] ≤ t, find=true
6.         if find==false, return false
7. return true
```

Figure 1.:- Algorithm used

The algorithm consists of followingthree subsections:

1) Choosing a dimension on which we have to partition : Find Number of rows in patient-enq

2) selecting a value to split and start Suppression : Here we suppress using a zipcode. This zipcode is having 5 digits like 46982. The variable inc is the value to split if we set inc= 4. The zipcodeis displayed as first 4 digit numbers like 4698**. And we set threshold value t=0.5F and n is the second highest value of table age-count according to patients age in the table. For example we contain this data in our patient table,

Age Count
2* 6
3* 3
4* 10

3) Checking weather partitioning violates the privacy requirement : After that we check this following calculation.

t =(rowcount)/n (1)

If Each row of our table satisfies the condition, our privacy requirement is satisfied . Else we decrement our incvalue and again we test this condition satisfied by each row or not till this condition will satisfied.

Flow of algorithm :

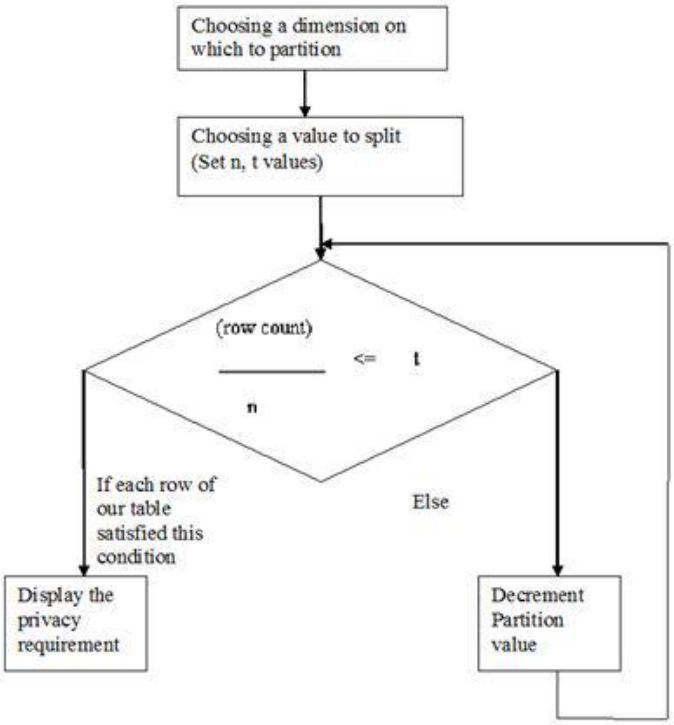
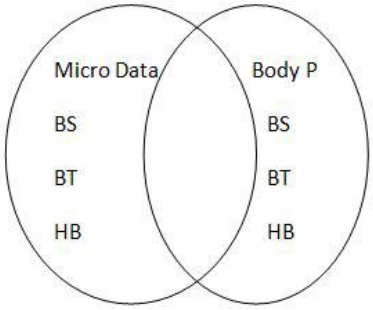


Figure 2.: Flow of Algorithm

Mathematical Model:-

Let ItemSet be a category table are sensitive attribute
model define to protect against attribute disclosure.



Privacy and sensitivity (Δf) of a function $f : \mathcal{D} \rightarrow \mathbb{R}^d$ is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

First, sort out the values of ItemSetaccording to their sensitivity, forming an ordered value domain D, and then partition the attribute domain into m-categories (S1,S2,. . . ,Sm),
such that

ItemSet = Um i=1ItemSet i, ItemSet i ∩ ItemSet j = Ø (for i 6= j) and Sl is more sensitive than the ItemSet k (for 1 ≤ l ≤ k).

For example, Consider the Body Condition ItemSet = (BP, BS, BT, HB, HT) Model has been partitioned into four categories according to the sensitivity of the Body Parameter. whereItemSet 1 (Prior Secret) is the most sensitive and ItemSet 4 (No Secret) is the least one. In order to measure the distance between two categories (attributes) and

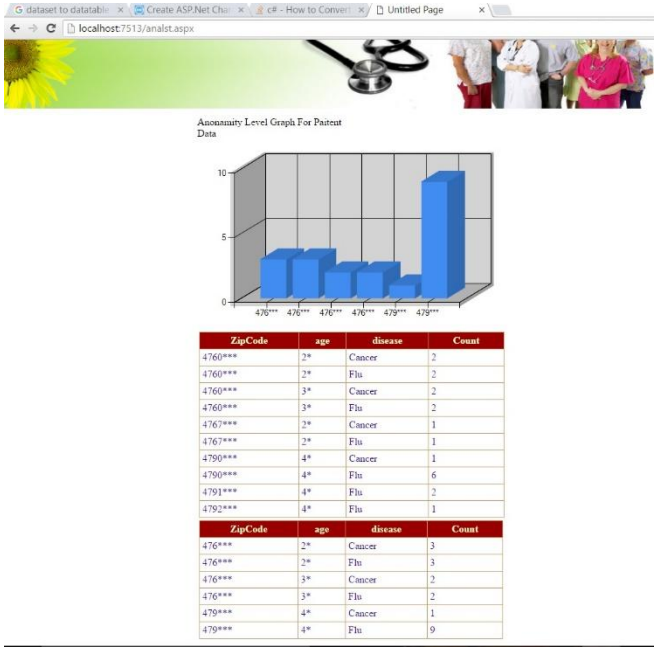
the degree that sensitive attribute values contribute to one QI-group, we introduce the following ordinal metric system. Let D(S) denote a categorical domain of an attribute ItemSet and D(S) be the total number of categories in domain D(S). The normalized distance between two categories Si and Sj of the attribute ItemSet with Si is greater than or equal toSj is:

The distance between two sensitiveattribute values is set same to the distance between the categories that they fall into. Moreover, we put an ordinal measured weight to each category to represent the degree that each specic sensitive attribute value in ItemSet contributes to ItemSet. Let D(ItemSet) = ItemSet1, ItemSet 2,. . . , ItemSet k denote a partition of categorical domain of an attribute ItemSetand let measured weight(ItemSeti) denote the measured weight of category ItemSeti. Then,

$$\text{measured weight (ItemSet 1)} = 0,$$
$$\text{measured weight(ItemSet i)} = i-1\ k-1;\ 1 < i < k$$
$$\text{measured weight(ItemSet k)} = 1,$$

the measured weight of the specic sensitive value is set same to the measured weight of the category that the specic value belongs to. The measured weight of the Quasi Identifier is the total measured weight of each specicsensitive values that the Quasi matrics contains four corresponding values set A= (BS, BT, HB, HT). The distance between BS (ItemSet 1) and BT(ItemSet4) is 3/3=1, while the distance between HB (ItemSet 2) and HT (ItemSet 3) is 1/3. According to (1), measured weight(V1) = 0, measured weight(V2) = 1/3 and measured weight(BS) = 2/3, measured weight(BT) = 1, the total measured weight of A is 0+1/3+2/3+1=2.

Results And Discussion:-



ZipCode	age	disease	Count
4760***	2*	Cancer	2
4760***	2*	Flu	2
4760***	3*	Cancer	2
4760***	3*	Flu	2
4767***	2*	Cancer	1
4767***	2*	Flu	1
4790***	4*	Cancer	1
4790***	4*	Flu	6
4791***	4*	Flu	2
4792***	4*	Flu	1

ZipCode	age	disease	Count
476***	2*	Cancer	3
476***	2*	Flu	3
476***	3*	Cancer	2
476***	3*	Flu	2
479***	4*	Cancer	1
479***	4*	Flu	9

Figure 3.:- Graphical representation of result

Evaluation consists in substituting the valuesof a given attribute with more general values. For this reason, the set of domain (i.e., the set of values that an attribute can assume) is extended to capture the Evaluation process by assuming that a set of generalized domains exist. The set of original domains together with their Evaluations is referred to as Domain. Each generalized domain contains generalized values and a mapping between each domain and its Evaluations exist. For example, ZIP codes can be generalized by dropping, at each Evaluation step, the least significant digit; home addresses can begeneralized to the street (dropping the number), then to the city, to county, state, and so on. This mapping is stated by means of aEvaluation relationship sigma of D. Given two domains Di and

Dj 2 Dom, Di Sigma of DDj states that values in domain. Dj is Evaluations of values in Di. The Evaluation relationship Sigma of D denotes a partial order on the Dom set of domains, and is required to satisfy the conditions.
Observation Table:

Observation 1 k-Anonymity with map reduce can create groups that leak information due to lack of diversity in the sensitive attribute	such a situation is common. As a back of the envelope calculation, suppose we have a dataset containing 60,000 different tuples where the sensitive attribute can take 3 distinct values and is not correlated with the attributes that are non-sensitive. A 5-anonymization of this table will have around 12,000 groups and, on average, 1 out of every 81 groups will not have diversity (the values for the sensitive attribute will all be the same). Thus we should expect about 148 groups will not have diversity. Therefore, information about 740 people would be compromised by a homogeneity attack. This indicates that in addition to k-anonymity, the sanitized table should also ensure "diversity" all tuples that share the similar values of their quasi-identifiers should have diverse values for their sensitive attributes
Observation 2 k-Anonymity with map reduce does not give us protection against attacks based on background knowledge	K-anonymous table may Disclose sensitive information. As both of these attacks are plausible in real life, we need a more capable definition of privacy that takes into account diversity and background knowledge.
Solution	**This project addresses this very issue i.e. noise adaption with map reduce (Positive Threshold)** Publishing the table T* that was derived from T results in a positive disclosure if the system can correctly identify the value of a sensitive attribute with high probability; i.e., when given a value $\delta > 0$, there is a positive disclosure if $\beta(q,s,T^*) > 1-\delta$ and there exists $t \in T$ such that $t[Q] = q$ and $t[S] = s$.

Conclusion And Future Work:-

Multiple sensitive attributes present additional challenges. Suppose if there are two sensitive attributes U and V. 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 U and V. Another approach is to consider the joint distribution of the two attributes. To use this approach, one has to choose the ground distance between pairs of sensitive attribute values. A simple formula for calculating EMD may be difficult to derive, and the relationship between (*n, t*) and the level of privacy become more complicated.

As seen above as k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The technique of l-diversity attempts to solve this problem. We have shown that –diversity has a number of limitations and especially discussed two attacks on l-diversity. Motivated by these limitations, we have proposed a novel privacy methode called closeness. We propose two techniques: a base model called t-closeness and a more flexible privacy technique called (n, t) closeness. We explain the logic of the (n, t)-closeness model and show that it achievesa better balance between privacy and utility. Finally, through experiments on real data, we show that similarity attacks are a real problem and the (n, t)- closeness model better protects the data while improving the utility of the

released data. (n, t)-closeness allows us to take advantage ofanonymization techniques other than generalization of quasiidentifier and suppression of records.

Acknowledgment:-

I feel great pleasure to submitting this project paper on EFFICIENT PERSONALIZED PRIVACY PRESERVATION USING ANONYMIZATION. I wish to express true sense of gratitude towards my project guide, Prof. R. N. Phursule who at very discrete step in study of this project, contributed his valuable guidance and helped to solve every problem that arose. My great obligation would remain due towards Prof. S.R.Todmal(Head of Department), who was a constant inspiration during my project. He provided with an opportunity to undertake the

project at JSPMs Imperial College Of Engineering and Research, Wagholi, Pune. I feel highly indebted to them who provided me with all my project requirements, and done much beyond my expectations to bring out the best in me. I sincere thanks to our respected Principal Dr.S.V.Admane proved to be a constant motivation for the knowledge acquisition andmoral support during our course curriculum.

References:-

1. Anonymizing Collections of Tree-Structured Data Olga Gkountouna,Student Member, IEEE, and ManolisTerrovitis, IEEE Transactions on Data and Knowledge Engineering Vol No 27 Year 2015
2. G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, Scalable and Efficient Provable Data Possession, Proc. Fourth Intl Conf. Security and Privacy in Comm. Networks (SecureComm08), 2008.
3. G. Ateniese, R. D. Pietro, dynamic provable data possession protocol based on cryptographic hash function and symmetric key encryption 2007
4. Juels and B. S. Kaliski HLA Based Solution International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1526-1532
5. N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou, LT CodesBasedSecure and Reliable Cloud Storage Service, Proc. IEEE INFOCOM, pp. 693-701, 2012
6. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia,
7. Dynamic Provable Data Possession, Proc. 16th ACM Conf.
8. Computer and Comm. Security (CCS09), pp. 213-222, 2009
9. S. Marium, Q. Nazir, A. Ahmed, S. Ahthasham and AamirM. Mirza, Implementation of EAP with RSA for Enhancing The Security of Cloud Computig, International Journal of Basic and Applied Science, vol 1, no. 3, pp. 177- 183, 2012
10. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D.Thomas, and A. Zhu, Achieving Anonymity via Clustering, Proc. ACM Symp. Principles of Database Systems (PODS), pp. 153- 162, 2006.
11. R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Inc., 1993
12. R.J. Bayardo and R. Agrawal, Data Privacy through Optimal k- Anonymization, Proc. Intl Conf. Data Eng. (ICDE), pp. 217-228, 2005.
13. F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller, From
14. Statistics to Beliefs, Proc. Natl Conf. Artificial Intelligence
15. (AAAI), pp. 602- 608, 1992.
16. J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, Secure Anonymization for Incremental Datasets, Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63, 2006.
17. B.-C. Chen, K. LeFevre, and R. Ramakrishnan, Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge, Proc. Intl Conf. Very Large Data Bases (VLDB), pp. 770- 781, 2007.
18. B.C.M. Fung, K. Wang, and P.S. Yu, Top-Down Specialization for Information and Privacy Preservation, Proc. Intl Conf. Data Eng. (ICDE), pp. 205-216, 2005
19. C.R. Givens and R.M. Shortt, A Class of Wasserstein
20. Metrics for Probability Distributions, Michigan Math J., vol. 31, pp. 231-240, 1984.
21. V.S. Iyengar, Transforming Data to Satisfy Privacy Constraints, Proc. ACM SIGKDD, pp. 279-288, 2002.
22. Kifer and J. Gehrke, Injecting Utility into AnonymizedDatasets, Proc. ACM SIGMOD, pp. 217-228, 2006.
23. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, Aggregate Query Answering on Anonymized Tables, Proc. Intl Conf. Data Eng.(ICDE), pp. 116- 125, 2007.
24. S.L. Kullback and R.A. Leibler, On Information and Sufficiency, Annals of Math. Statistics, vol. 22, pp. 79-86, 1951
25. Lambert, Measures of Disclosure Risk and Harm, J.
26. Official Statistics, vol. 9, pp. 313-331, 1993
27. Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing QianWang , Cong Wang , Jin Li , KuiRen , and Wenjing Lou Illinois Institute of Technology, Chicago IL 60616
28. C. Wang, Q. Wang, K. Ren, and W. Lou, Privacy- Preserving Public Auditing for Data Storage Security in Cloud Computing, Proc. IEEE INFOCOM, pp. 525-533, 2010