PRIVATIZATION OF SENSITIVE INFORMATION IN DATA PUBLISHING USING CLOSEENESS

S.SARANYA, Assistant professor, SNS College of Engineering, Coimbatore, Tamilnadu. .

ABSTRACT-

Privacy is an important issue in data mining while publishing the data in dataset. Many organizations distribute non-aggregate personal data for research, and they must take steps to ensure that an adversary cannot predict sensitive information pertaining to individuals with high confidence. Show that when the data contains a large number of attributes which may be considered quasi-identifiers; it becomes difficult to anonymizing the data without an unacceptably high amount of information loss. This is because an exponential number of combinations of dimensions can be used to make precise inference attacks, even when individual attributes are partially specified within a range and provide an analysis of the effect of dimensionality on k-anonymity methods. And conclude that when a data set contains a large number of attributes which are open to inference attacks, faced with a choice of either completely suppressing most of the data or losing the desired level of anonymity. To overcome these limitations, to implement a concept called “closeness”. The base model of this concept is t-closeness which requires the distribution of a sensitive attribute in the overall table. In the slicing process we can perform the better data utility and membership disclosure protection.

Index Terms—Privacy preservation, dataanonymization, data publishing, data security.

I. INTRODUCTION

Data mining is the process of extracting useful, interesting, and previously unknown information from large data sets. The success of data mining relies on the availability of high quality data and elective information sharing. The collection of digital information by governments, corporations, and individuals has created an environment that facilitates large-scale data mining and data analysis. Moreover, driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for sharing data among various parties.Each record has a number of attributes, which can be divided into the following three categories:1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date,and Gender. 3) Attributes that are considered sensitive, such as Disease and Salary.When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified ,identity disclosure and attribute disclosure. 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.

II. PROBLEM DEFINITION

The privacy preserving data mining problem has gained considerable importance in recent years because of the vast amounts of personal data about individuals stored at different commercial vendors and organizations. In many cases, users are willing to divulge information about themselves only if the privacy of the data is guaranteed. This creates the natural challenge of mining the data in an effective way with a limited data representation. Government agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature.

- Identity disclosure
- Attribute disclosure

III. CLOSENESS: A NEW PRIVACY MEASURE

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 than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief.

The novelty of our approach is that we separate the information gain into two parts: that about the population in the released data and about specific individuals.

Table 1

Original Patient Table

	Zipcode	Age	Disease	Count
1	47632	23	Flu	500
2	479203	34	Dia	200
3	637234	22	Cancer	900
A4	563843	25	Flu	100
l5	234355	34	Cancer	300
l6	643656	36	Fever	900
7	476567	39	Flu	100
t8	346535	42	Flu	400
u9	344656	54	Dia	500
p10	256565	59	Flu	900
l11	637422	51	Cancer	100
e12	237565	55	Fever	300

s

in Table 1 have to be generalized into a single equivalence class. This results in substantial information loss. If we examine the original data in Table 1, we can discover that the probability of cancer among people living in zipcode 476__ is as high as 500 1;000 ¼ 0:5, while the probability of cancer among people living in zip code 479__ is only 2002;000 ¼ 0:1. The important fact that people living in zip code 476__ have a much higher rate of cancer will be hidden if 0.1-closeness is enforced.

IV TUPLE PARTITIONING ALGORITHM

Introduce a new basic partitioning algorithm and propose POP(postponed partitioning algorithm)improvedfrom BAP algorithm.These algorithm are different only in partitioning procedure.

COMMON PARTS OF ALGORITHM BAP AND POP

The basic idea of algorithm BAP and POP is to partition a relation logically to find candidates.These algorithm consist of two phase. Phase partitioning a relation and selecting candidates and the phase computing exact average value of candidates.

ALGORITHM PARTITIONING

```
input:relation R,Thrshold T
Begin
// Phase 1: partitioning and selecting candidates
        Second_scan:=false
bucket_num:=0;
        for all tuple d in R do
        if there is the bucket for d target t in counter C
        then
        update (C,d);
        else

        if bucket_num<|c|then

        insert (C,d);
        bucket_num++;
                else
Gen_partitioning(C,T,bucket_num);
second_scan:=true;
end if
        end if
end for
//phase 2: Computing the extract value of candidates

if second_scan==false then
        print_result(C,T);
        else
        repeat
read as many candidates as Counter buckets to C;
        scan R and update C;
print_Result(C,T);
        until there no more candidates
        end if
        end
```

In phase I tuple is read from relation R. The treatment of d varies according to whether or not the counter bucket for d. target exists and whether or not there are any empty bucket in counter.If the bucket for d,target exists,sum_val and cnt_val in the bucket will be updated. If the bucket for d,target doesn't exists,but there some empty bucket s,the bucket for d, target will be generated.If neither the bucket for d target nor any empty bucket exists, Gen partition() will be called.

This procedure performs partitioning which means selecting candidates writing them and resetting counter. However Gen-partition() is different between IN BAP and IN POP.

The phase 2 is to compute the exact average value of candidates. If the number of partitions is one, them all results can be returned from counter immediately . Otherwise, candidates are loaded into memory and R is scanned again to compute sum_value and cnt_val of candidates. If there are two many candidates for buckets,these steps will be over and over.Each scan will return some part of result.

DATA FEED MODULE

Data is stored at different security classifications and users having different security clearances. Data on individuals and entities are being collected widely. These data can contain information that explicitly identifies the individual (e.g., social security number). Data can also contain other kinds of personal information (e.g., date of birth, zip code, gender) that are potentially identifying when linked with other available data sets. Data are often shared for business or legal reasons. Unprecedented amounts of data are being collected on individuals and entities. This is being fuelled by progress in various technologies like storage, networking and automation in various business processes. Of particular interest are data containing structured information on individuals.

MICRO DATA IDENTIFICATION

In the database, each record has a number of attributes, which can be divided into the following three categories 1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code,Birthdate,Gender. 3) Attributes that considered sensitive, such as Disease and Salary.

PRIVACY MEASUREMENT

First, an observer has some prior belief B0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed.The observer's belief is influenced by Q, the distribution of the sensitive attribute values in the whole table, and changes to belief B1. Finally, the observer is given the released table.

By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learns the distribution P of sensitive attrinobute values in this class.

V. CONCLUSION

The privacy measure is one of most important for datamining . In datamining informzation losses, security and unauthorised access are one of the problem. So avoiding this problem use the concept of slicing and closeness can be used. In the closeness concept the recently released data can be publish in different way then only the unknown cannot analysis the information. Slicing is used for partitioning the table into two horizontally and vertically and the overlapping can be reduced. In general the feeding the data from database or some of the update, search and save operation can be performed. And then the attributes can be classified as explicit identifier, Quasi identifier and sensitive identifier can be done. Then how the data can sliced and the overlapping of the slicing can be reduced. Finally the data can be published and to provide the privacy measure for all the data in the database.

REFERENCES

[1] Aggarwal C. (2005) "On k-anonymity and the curse of dimensionality " In VLDB, pages 901–909.

[2] Blum A, Dwork C, McSherry F, and Nissim K (2005)"Practical privacy: the sulq framework" In PODS,pages 128–138.

[3] Brickell.J and Shmatikov.V (2008) " The cost of privacy: destruction of data mining utility in anonymized data publishing" InKDD, pages 70-78.

[4] Chen B.C,Ramakrishnan R, and K. LeFevre. ,(2007)"Privacy skyline: Privacy with multidimensional adversarial knowledge". In VLDB, pages 770–781.

[5] Dinur I and Nissim K. (2003) " Revealing Information while Preserving Privacy", Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210.

[6] Koudas N, Srivastava D, Yu T and Zhang Q. (2007) "Aggregate query answering on anonymized tables". In ICDE,pages 116-125.

[7] Li J, Tao Y, and Xiao X.(2008) "Preservation of proximity Privacy in Publishing numerical sensitive data ". In SIGMOD, Pages 473-486.

[8] Machanavajjhala A,Gehrke J, D. Kifer,and M.Venkitasubramaniam.ℓ diversity:"Privacybeyond anonymity". In ICDE, page 24, 2006.

[9] Nergiz M. E, Atzori , and Clifton C. (2007) "Hiding the presence of individuals from shared databases". In SIGMOD, pages 665–676.

[10] Rastogi V, Suciu D, and Hong S.(2007)" The boundary between privacy and utility in data publishing". In VLDB, pages 531–542.

Ms S.SARANYA currently working as a ASSISTANT PROFESSOR in SNS COLLEG OF ENGINEERING Coimbatore. She did my Master of Engineering in COMPUTER SCIENCE and ENGINEERING at Vivekanandha College of Engineering for Women, affiliated to Anna University Chennai, Tamilnadu, India and received my B.E degree from Selvam College of Technology Namakkal, affiliated to Anna University Coimbatore. Her research interests include data mining and networking

.