Implementing Privacy of Data and Classification Approach using Data Anonymization Techniques

Mavooru Jyothsna[1], Mula Sudhakar [2]

Final M.Tech Student[1], Asst.professor[2]

[1,2]Dept of CSE, Sarada Institute of Science, Technology and Management (SISTAM), Srikakulam, Andhra Pradesh

Abstract:

Now a day's to provide privacy of personal information is a challenging to data management communities. So that the communities are allow to processing of personal information without loss of data. To provide privacy of personal information the communities are using so many data anonymization techniques. One of the data anonymization technique is tree structured data oriented for provide privacy of personal information. In the tree structured data anonymization technique will face the problem of time complexity for generalized data. So that to generalize the personal information it will build the data oriented tree structure and perform the search operation. To overcome this problem we are proposed Posteriori Probability of generalizations approach for performing classification of data. Before performing this process we can store data into database within format cipher format. So that by converting plain format data into cipher format we are using extended tiny encryption algorithm. By implementing encryption process we can provide privacy of personal information and also improve efficiency of stored data. So that by using those operations we can provide privacy of personal information and also get required treatment for particular diseases. In this paper we are take personal information related to the medical data and perform those operations on that data. By implementing those concepts we can overcome time complexity and also provide more security of data.

Keywords: *Cryptography, Anonymization, Generalization, classification, Privacy.*

I. INTRODUCTION

To provide privacy of personal information is increasingly organized by the companies and also concern significant challenges to data management organization. So that the data management organization or companies are trying to implement the data anonymization techniques for anonym zing personal information. By implementing data anonymization technique in order to allow the processing personal information without compromising the data privacy. So that by provide privacy of data the data anonymization techniques are face the problem of records for maintaining dependences of each record. In order to overcome that problem it will proposed tree structured data anonymization technique. In the tree structured data anonymization the data can be generalized in form of tree structure. In the tree structure data anonymization process is used to generalize data and get required classified data. By implementing tree structured data anonymization technique it will also face the problem classifying the data by performing searching each node in the tree. By performing generalizing data using tree structured data anonymization technique will take more time for performing searching.

In this paper we are proposed one of data anonymization technique for generalizing the personal information. Before performing classifying the data using the generalization process the personal data will be stored into database. Each record in the data base will contain with in format of cipher. Because storing data into database the administrator will perform encryption process for convert each record into cipher format. So that to provide privacy of data the administrator will convert data into cipher format. By converting plain format record into cipher format we can use the cryptography technique. By implementing this cryptography technique we can provide privacy of data in the database. After converting cipher format the administrator will stored into data into database. The analyst receives the cipher format training data set and applies generalization technique it will get the classified data. Before apply generalization technique the analyst will retrieve cipher format data and decrypt that data. After decrypting data we can implement generalization process for classifying data. By implementing those concepts we can take medical records of each patient.

In this paper we can take the medical information for performing generalization process. Before performing generalization process the analyst will retrieve cipher format medical data from the data base and decrypt that medical information. By performing decryption process the cipher format data will be converting into plain and apply the

generalization concepts. By implementing generalization concept's we can get classified data and get required output. Before performing generalization process the analyst also take testing dataset for each patient medical information and apply those generalization concepts on that data. So that by performing those operations we can get better generalization result and also reduce time complexity for performing generalization process. The remaining this paper contains related work of our paper, implementation of our proposed system and last one conclusion our paper.

II. RELATED WORK

The problem of protecting privacy in the publication of set-valued data is defined in Local and global recoding methods for anonymizing set-valued data[3],Consider a collection of supermarket transactions that contains detailed information about items bought together by individuals. Even after removing all personal characteristics of the buyer, which can serve as links to his identity, the publication of such data is still subject to privacy attacks from adversaries who have partial knowledge about the set. Consider a database D, which stores information about items purchased at a supermarket by various customers. A subset of items in a transaction could play the role of the quasi-identifier for the remaining (sensitive) ones and vice-versa. Another fundamental difference is that transactions have variable length and high dimensionality, as opposed to a fixed set of relatively few attributes in relational tuples. All items can act as quasi-identifiers, an attacker who knows them all and can link them to a specific person has nothing to learn from the original database. Her background knowledge already contains the original data. There are three classes of algorithm they are the optimal anonymization (OA) algorithm, which explores in a bottom-up fashion the lattice of all possible combinations of item generalizations, and finds the most detailed such sets of combinations that satisfy k_m-anonymity. The best combination is then picked, according to an information loss metric. Direct anonymization (DA) heuristic operates directly on m-sized itemsets found to violate k anonymity.

There need to share person-specific records in such a way that the identities of the individuals who are the subjects of the data cannot determined, it is determined in Achieving k- Anonimity privacy protection using generalization and suppression[4].Generalization involves replacing(or recoding)a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all. A value is replaced by a less specific, more general value that is faithful to the original. These techniques can provide results with guarantees of anonymity that are minimally distorted. Any attempt to provide anonymity protection, no matter how minor, involves modifying the data and thereby distorting its contents, so the goal is to distort minimally. Anonymizing Classification Data For Privacy Preserving [5] Data sharing in today's globally networked systems poses a threat to individual privacy and organizational confidentiality. First of all, knowing that the data is used for classification does not imply that the data provider knows exactly how the recipient may analyse the data. The recipient often has application-specific bias towards building the classifier. The objective of most anonymizing algorithms is to find an optimal recoding of the data that satisfies a given privacy guarantee and preserves as much data utility as possible. The latter is accomplished by minimizing a function which estimates the information loss. [32] Proved that optimal anonymity for multidimensional QI is NP-hard, under both the generalization and suppression models. For the latter, they proposed an approximate algorithm that minimizes the number of suppressed values with the approximation bound O (k _ log). [8] Improved this bound to O(k), while [38] further reduced it to O(logk).

III. PROPOSED SYSTEM

The paper propose concept of Posteriori Probability of generalizations approach for performing generalization process of classifying personal information. In this paper we are take medical information of each patient and apply generalization concept on that data. Before apply the generalization technique we can encrypt patient information and stored into data base. By performing encryption process we are using extended tiny encryption algorithm and also to provide privacy of stored data. By using both concepts we can improve privacy of data and also provide more efficiency for performing generalization technique. The implementation procedure of both concepts is as follows.

Store Data into Database:

In this module the administrator will store medical information of each patient into database. Before storing data into database the administrator will enter required filed of table and that attribute data can be encrypted. After performing encryption process that cipher format data will be stored into database. The encryption process can be done by using the tiny encryption algorithm. The implementation procedure of extended tiny encryption algorithm is follows.

i) Encryption process:

```
int delta= 0x9E3779B9;
```

```
            void encrypt (int [] buf)
            {
    buf.length % 2 == 1;
    int i, v0, v1, sum, n;
    i = 1;
    while (i<buf.length)
    {
     n = CUPS;
     v0 = buf[i];
     v1 = buf[i+1];
     sum = 0;
     while (n-->0)
     {

 v0 += (((v1 << 4) ^ (v1>>5)) + v1) ^ (sum +key
[sum & 3]);
 sum += delta;
v1 += (((v0 << 4) ^ (v0 >> 5)) + v0) ^ (sum +key
[(sum>>11) & 3]);

}
 buf[i] = v0;
 buf[i+1] = v1;
 i+=2;

}
}
```

After completion of encryption process the administrator will stored cipher format data into data. That cipher format data will be treated as training data sets and those data sets are used for performing generalization process.

Retrieve Data from Database:

In this module user will retrieve cipher training formatted data from the database. After retrieving cipher training format data it will decrypt by using extend tiny decryption process. The implementation procedure of decryption process is as follows.

I) Decryption Process:

```
  int delta= 0x9E3779B9;
  void decrypt (int [] buf)
   {
    buf.length % 2 == 1;
    int i, v0, v1, sum, n;
    i = 1;
   while (i<buf.length)
     {
      n = CUPS;
      v0 = buf[i];
      v1 = buf [i+1];
      sum = delta;
 while (n--> 0)
      {
 v1 -= (((v0 << 4) ^ (v0 >> 5)) + v0) ^ (sum +key
[(sum>>11) & 3]);
                sum -= delta;
          v0 -= (((v1 << 4) ^ (v1>>5)) + v1) ^ (sum
+key [sum & 3]);
                }
          buf[i] = v0;
          buf[i+1] = v1;
          i+=2;
    }
          }
           }
```

The completion of decryption process user will get plain format training data and using that data set we can perform the data generalization technique. Before apply the data generalization user will retrieve testing data for performing classification approach. By performing classification process we are using Posteriori Probability of generalizations approach. The implementation procedure of Posteriori Probability of generalizations approach is as follows.

Posteriori Probability of generalizations approach:

By performing Posteriori Probability of generalizations approach will contain two types data format i.e. training data set and testing data. The training data set is full-fledged data containing information related medical details each patient. The training data set contains attributes are patient id, patient name, age, hospital, disease and treatment. By using this training type of data we can apply generalization technique to get classified data. Before get the classified the user also gives testing data for getting classified data. By performing generalization we can implement Posteriori Probability of generalizations approach.

1. Read training and testing data set of medical information.

2. Take each record from the testing data and calculate probability of each attribute by using training data set.

3. In this paper we are considering age, hospital and disease attribute for finding related treatment.

4. By using those attribute we can calculate individual probability of each attribute.

5. P (age) =total occurrence of age\total number of records.

6. P (hospital) = total occurrence of hospital\total number of records.

7. P (disease) = total occurrence of disease\total number of records.

8. After calculation of probabilities we can calculate total probability.
P (total) = p (age) + p (hospital) + p (disease)

9. Calculate the total record probability i.e. P (record) = occurrence of total record\total number of records.

10. If(p(total)>p(records))
Retrieve treatment from the training data and put into testing data.

11. Repeat step 4 to 10 over the completion total records in testing data set.

After completion of this process we can get classified data and that data will be displayed. By implementing above concepts we can improve privacy of data and also improve efficiency for generalizing data.

IV. CONCLUSIONS

In this paper we are mainly focus on privacy of stored data and also perform generalization over data by getting classified data. By satisfying those concepts we can implement the two concepts i.e. extended tiny encryption algorithm and Posteriori Probability of generalizations approach. Using extend tiny encryption algorithm we can convert data into cipher format and stored that cipher format data into database. So that by implementing this process we can satisfy privacy of stored data. Another concept is Posteriori Probability of generalizations approach used to classify testing data and get required treatment. By implementing this process we can satisfy data generalization. So that by implementing both concepts we improve efficiency for getting classified and also provide more security of data.

REFERENCES

[1].M. Ercan Nergiz Chris Clifton, \MultiRelational k- Anonymity": , 2007 IEEE.

[2] L. Sweeney, \k-anonymity: a model for protecting privacy ":, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[3] Pierangela Samarati, Member, IEEE Computer Society, "Protecting Respondents Identities in Microdata Release ": , IEEE Transaction on Knowledge and Data Engineering, VOL. 13, NO. 6, Nov/Dec 2001.

[4]. A. Meyerson and R. Williams. On the Complexity of Optimal Kanonymity. In PODS, pages 223–228, 2004.

[5]. Olga Gkountouna, Student Member, IEEE and Manolis Terrovitis, "Anonymizing Collections of Tree-Structured Data,":IEEE Transaction on Knowledge and Data Engineering, VOL. 27, NO. 8, Aug 2015.

[6]. G. Cormode , Personal privacy vs population privacy: learning to attack anonymization.

[7].R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k-Anonymization. In ICDE, pages 217–228, 2005.

[8]. P. Samarati and L. Sweeney. Generalizing Data to Provide Anonymity when Disclosing Information (abstract). In PODS (see also Technical Report SRI-CSL-98-04), 1998.

[9]. G. Ghinita, P. Kalnis, and Y. Tao. Anonymous publication of sensitive transactional data. TKDE, 23(2):161–174, 2011.

[10]. P. Samarati and L. Sweeney. Generalizing Data to Provide Anonymity when Disclosing Information (abstract). In PODS (see also Technical Report SRI-CSL-98-04), 1998.

BIOGRAPHIES:

Mavooru Jyothsna is student in m.tech(CSE) in sarada institute of science technology and management, srikakulam. She has received her M.C.A from Sarada institute of science technology and management, srikakulam. Her interesting areas are data mining, network security and cloud computing.

Mula Sudhakar is working as an Assistant Professor in Sarada Institute of Science, Technology and Management, Srikakulam, Andhra Pradesh. He received his M.Tech (SE) from Sarada Institute of Science, Technology and Management, Srikakulam. Andhra Pradesh. His research areas include Computer Networks, Data Mining, and Distributed Systems.