New Approach to Reduce the Data Loss in Privacy Preserving Data Analysis

Kirti B. Sable
PG Department
MBES's College of Engineering
Ambajogai, India, 431 517.

B. M. Patil
PG Department
MBES's College of Engineering
Ambajogai, India, 431 517..

ABSTRACT

Recent advances in information and communication technologies shows that privacy is one of the important concept regarding the data sharing, maintaining the data confidentiality and data loss issues. This paper mainly focus on data loss issues and overcome some of the problems exist in pervious system of privacy preserving such as role base accessibility, high time complexity and privacy breaching issue. This uses different techniques such as slicing which is combined with C constraint. This will reduce the data loss and provide more privacy to the data. Here considered a health care system as a framework. SQL injection and Aho-Chorasick algorithm is used for database security.

Keywords

Privacy, L-diversity, slicing, attacker, k-anonymity, C-constraint

1. INTRODUCTION

There is lots of information or data out there containing valuable information. Privacy and security are the essential things regarding the information sharing and confidentiality of the identity of individual users. Apart from this data loss is also a significant issue. Slicing combined with C constraint uses K-anonymity and L-diversity where L- diversity is always less than K-anonymity. The C- constraint is applied over different techniques such as anonimization, generalization; random permutation gives the privacy view. The records which don't follow the K-anonymity and L-diversity rule goes to bucket i.e temporary data. The final bucket data count is equal to the count of database which results in 0% data loss without generating the background knowledge of data. The scenario used here is health care system.

Privacy preserving in data mining contains the several issues such as:

- Privacy: In data mining huge data is analyzed and useful information is extracted hence it is also called as knowledge discovery. Releasing this information reveals information about individual. Here it used heath care system; this may lead to reveal information about the patient.
- Attacker: In privacy preserving there is no encryption and decryption techniques such as used in security. This leads to advantage for an attacker to attack on database.
- Data Loss: In privacy preserving techniques there is an issue of missing data which is not efficient which result in loss of data and inappropriate output.
- Time complexity: High time complexity when data work with vertical and horizontal format.

To overcome these problems the slicing technique with C-constraint is used. And verification is used to verify data privacy for every section of data against number of providers of data.

1.1 Updated Slicing

In data analysis, the slicing is the systematic reduction of a data into smaller views that will yield more knowledge about the information. Slicing is basically depends on Attribute and tuple partitioning. Attribute partitioning is vertical partition while tuple partition is horizontal partition.

For example, suppose a data table D is sliced into two tables as p and q. In table p, the attribute partition is {{Name}, {Age}, {Gender}, {zip}, {Provider}} and tuple partitioning is {d1; d2; d3; d4; d5}. In table q, the attribute partition is {{Name, Age}, {Gender, zip, Provider}} and tuple partitioning is {d1; d2; d3; d4; d5}.

- L DIVERSITY: L diversity is the concept of maintaining uniqueness within data. In this system this concept is used on sensitive attribute.
- C CONSTRAINT: C is a privacy constraint in which D* should fulfill slicing condition with L diversity.
- K ANONYMITY: K anonymity is the concept of maintaining the number of dataset records in single view. If at least k specific records are indistinguishable in its identifying information then that data record is called as k-anonymous.

2. LITERATURE SURVEY

The incentive compatible privacy preserving data analysis was given by Murat et al. [1]. This scheme tried to convince that by conducting privacy preserving data analysis task the participating parties will motivate to provide truthful input data. But the data provided by the participating parties cannot verify, it is truthful or not. It contains the important problem of data loss issue and privacy. It also has the problem of role base accessibility.

In Secure multi-party computation (SMC) [3], [4] the participating parties got to know only the final result and their own input for analysis. But this model does not guarantees that data provided was truthful that leads to inaccurate result. The Yao's millionaire problem is discussed here [5]. But the SMC also contain some inference problems it deals with the inequality. W. Jiang et al. [2] discussed the k-anonymity for this inference problem where data set having identifying information was maintained in such an indistinguishable k-specific records. He used one way to preserve privacy while enabling beneficial use of data is to utilize k-anonymity for publishing. But disadvantage of this is that it may not produce accurate data. A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols [2] can be used to guarantee lack of intermediate information disclosure during the anonimization. However, not only TTP but also SMC cannot protects against analyze information from the anonymized data [9]. The attacker can use this information. S. Goryczka et al. [6] used m- privacy technique to prevent the insider attack. But this does not considered other attacks such as SQL injection.

A. Machanavajjhala et al. [7] states that l-Diversity requires each QI (quasi-identifier equivalence group) group to contain at least “well-represented” sensitive values. He shows that privacy against attacker using back ground knowledge does not guaranteed by k-anonymity. But this does not give any information regarding the data loss. Tiancheng Li et al. [8] used slicing for privacy preserving data analysis.

In proposed system, the updated slicing with C constraint, k-anonymity and l- diversity is used. This will gives the 0% data loss with improving privacy of existing system. To protect the database from attacker or adversaries it implements the SQL injection and Aho-Chorasick algorithm. And it also reduces the high time complexity as shown in analysis.

3. PROPOSED SYSYTEM ARCHITECTURE

For the proposed system contribution a system is developed with the help of distributed as well databases. The proposed system architecture is shown in figure 1. This proposed system can be understood easily using the example of health care system.

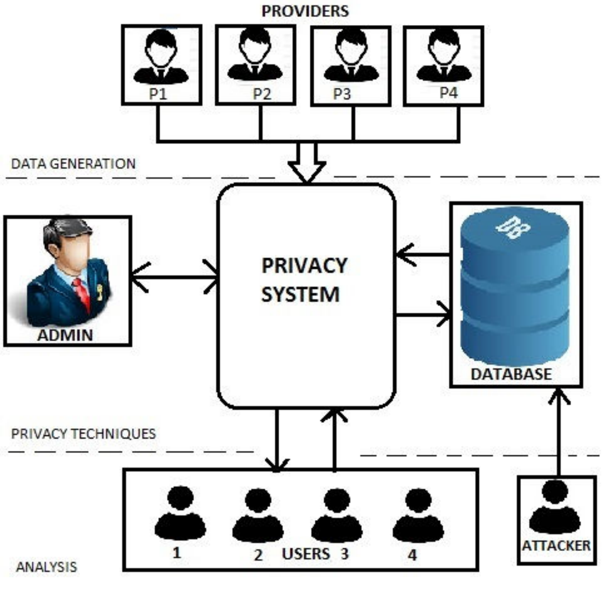

Figure1. Proposed System Architecture

The above figure can be divided into 3 layers as data generation, privacy techniques and analysis. The first layer generates the data and passes it to the 2 layer and analyzed by last layer.

1) Data generation: The first layer is data generation. It consists of data providers’ p1, p2, p3, p4. In health care system the providers are the one who provide the patient data as Name, Age, Zip, Diseases and treatment.

2) Privacy Techniques: Privacy Techniques consist of different components as Admin, Privacy system, Database. In this layer the data is sliced using C constraint to give the privacy view i.e anonimized view.

   - Admin: Admin can analyze the whole data, also set the role base accessibility rules for users and providers. Admin can view all doctor records, patient records and providers’ data. It also contains the attacker term. The attacks such as SQL injection, brute force attack, data collusion can be viewed by the admin as attack, pattern, date and time of the attack.
   - Privacy System: privacy system consist of different privacy techniques and algorithms such as updated sliced algorithm to reduce the time complexity and data loss issue using C constraint. This will give the privacy view. SQL injection and prevention algorithm implemented between the database and attacker to eliminate the data injection, data collusion, brute force base attacks. The request and response is carried out through the privacy system.
   - Database: The providers provide the data. After providing the data, it is stored in database. The database is used to store the data or information. This information can be accessed under privacy and role base accessibility with request and response.

3) Analysis: The users shown in above figure can analyze the data in privacy view. The data stored in database is in plain text so there are chances of attacks on the information to prevent this attack SQL injection and prevention algorithm implemented between the database and attacker.

4. ALGORITHM

Here the updated slicing algorithm is used to reduce time complexity and C-constraint to reduce data loss. SQL injection and Aho-Chorasick algorithm is used for database security.

4.1 Updated Slicing Algorithm:

```
Input: Data set with D, providers n, with C
Output: Slice view (T*) with provider
Step 1: read data from (D up to null)
Step 2: For each (attributes in table)
For each (tupels in tables)
Step 3: set quasi identifier (QIfr) and sensitive attributes (SA)
Step 4: Apply generalization technique it will classify the tupples in QIfr groups
Step 5: Apply anonimization on relative information attributes
Step 6: While (verify data-privacy(D, n, C) = 0) do
        if (Di → D) verified with QIfr then
add Di up to when K-anonymity
else stop
Bucket (i1) → D;
Step 7: permute the data with (I=(I( null-1)))
Step 8: Apply Pruning on (D)
Step 9: Apply step 1, 2, 3 on Becket(i1)
Step 10: if (C fails with (D) && (p#1))
    Bucket (i2) → Bucket(i1(j))
Step 11: Display all (Bucket (i2)!=null)
Step 12: end while
Step 13: end for
```

4.2 SQL Injection Algorithm

```
1: Procedure SPMA (Query, SPL[ ])
INPUT: Query=User Generated Query
SPL[ ]=Static Pattern List with m AnomalyPattern
2: For j = 0 to m do
3: If (AC (Query, String.Length(Query), SPL[j][0]) = =0 )then
4: Calc anomaly score
5: If ( ) Score Value Anomaly = Threshold
6: then
7: Return Alarm .Administrator
8: Else
9: Return Query. Accepted
10: End If
11: Else
12: Return Query. Rejected
13: End If
14: End For
End Procedure
```

4.3 AHO-Chorasick Algorithm

```
1: Procedure AC(y, n, q0)
INPUT: y= array of m bytes representing the text input (SQL Query Statement)
        n= integer representing the text length (SQL Query Length)
q0=initial state (first character in pattern)
2: State: q0
3: For i = 1 to n do
4: While g (State, y[i] = = fail) do
5: State ← f (State)
6: End While
7: State ← g(State, y[i])
8: If o (State) == NULL then
9: Output i
10: Else
11: Output
12: End If
13: End for
14: End Procedure
```

5. RESULT AND DISCUSSION

For the proposed system performance evaluation, here a system is deploying on java with INTEL 3.0 GHz i7 processor and 8 GB RAM and 1000 records. Here each graph shows the system performance with different experiments that has been classified in graphs.

Here the graph shown below is the graph for data insertion i.e. time required for execution to number of providers. Here the total execution time is given by (EndTime-StartTime). Initially the StartTime considered as zero.

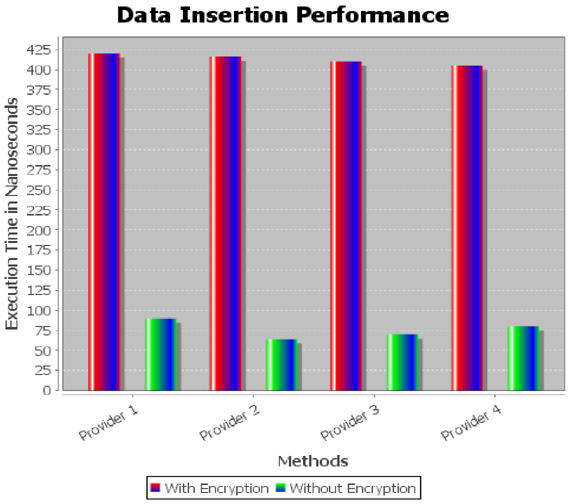

Figure 2.Data Insertion Time

The second graph shown below is the graph for data extraction i.e. the time required to extract the data to the data provided by the provider. Here the total execution time is given by (EndTime- StartTime).

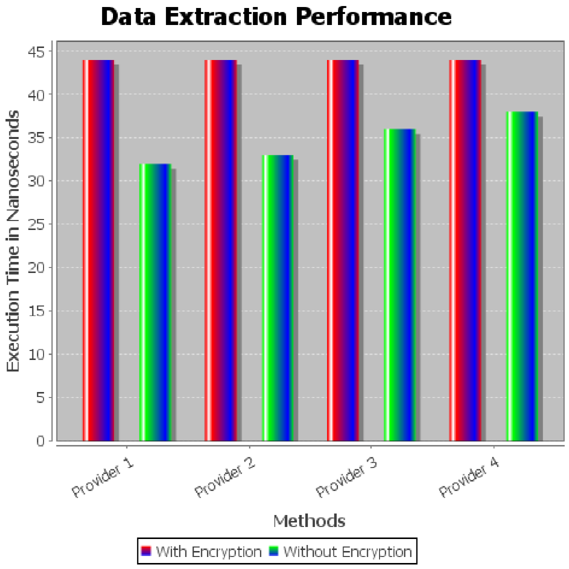

Figure 3. Data Extraction Time

Next graph shown below is the comparison graph for data slicing in existing and proposed system. The data slicing in proposed system gives the accurate result with better privacy view as compared to existing system. The proposed system gives 0% data loss using the C constraint in slicing i.e updated slicing.

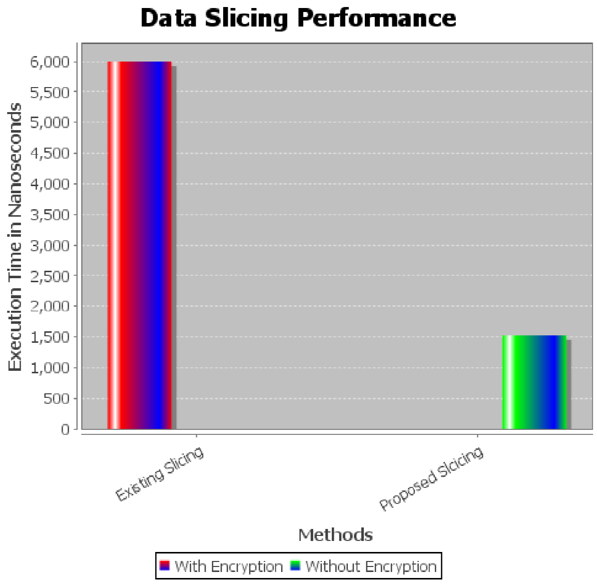

Figure 4. Comparison Graph for data slicing in existing and proposed system.

6. CONCLUSION

System discussed that there are some data loss as well some privacy issues in existing systems, with using different privacy techniques we can eliminate the drawbacks of system. Here we use slicing technique with C constraint that can be easily providing end user data privacy or it can improve the system accuracy. Proposed scheme is very useful in many practical applications, especially where privacy is required. This technique also reduces the time complexity. Proposed scheme also provides more security to the data available in the database using SQL injection and Aho-chorasick algorithm. System also considers a potential attack on collaborative data publishing. Here slicing algorithm is used for anonimization and L diversity and verify it for security and privacy by using binary algorithm of data privacy. Slicing algorithm is very useful when we are using high dimensional data. It divides data in both vertical and horizontal fashion. Due to encryption the security can be increased. But the limitation is there could be loss of data utility.

For the future enhancement system can work with big data in hadoop framework. The HDFS framework will provide the parallel processing with structure as well as semi structured data in distributed environment. Also need to work on database security from different attacks. The ad hoc network will also provide the drastic security to any type data security applications.

7. REFERENCES

[1] Murat Kantarcioglu and Wei Jiang, “Incentive Compatible Privacy-Preserving Data Analysis”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 6, JUNE 2013

[2] W. Jiang and C. Clifton, “Privacy-preserving distributed k-anonymity, “in DBSec, vol. 3654, 2005, pp. 924–924

[3] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game - a completeness theorem for protocols with honest majority. In 19th ACM Symposium on the Theory of Computing, pages 218–229, 1987.

[4] Andrew C. Yao. Protocols for secure computation. In Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science, pages 160–164. IEEE, 1982.

[5] Andrew C. Yao. How to generate and exchange secrets. In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pages 162–167. IEEE, 1986.

[6] S. Goryczka, L. Xiong, and B. C. M. Fung, “m-Privacy for joint data publish,” in Proc. of the 7th Intl. Conf. on joint compute: Networking, Applications and Work sharing, 2011.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,“l-Diversity: Privacy beyond k-anonymity,” in ICDE, 2006,p. 24

[8] Tiancheng Li, Ninghui Li, Jian Zhang,Ian Molloy,”Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE transactions on knowledge and data engineering, vol. 24, no. 3, March 2012.

[9] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, “Centralized and distributed anonymization for high dimensional healthcare data,” ACM Trans. on Knowledge detection from Data, vol. 4, no. 4, pp. 18:1–18:33, October 2010.