IJCSN International Journal of Computer Science and Network, Volume 5, Issue 2, April 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Anonymous Publication of Sensitive Transactional Data with Low Information Loss

[1] N. Nandhini, [2] V. Bindu

[1] M.E., Computer Science and Engineering, VELS University, Chennai, India

[2] Assistant Professor, Computer Science and Engineering, VELS University, Chennai, India

Abstract - A collection of real world data causes very complex data structure. In large scale companies, there will be serious issues in Data storage, Data Management, Data Retrieving. It's really a hard thing to do the data anonymization in live environment. Data Anonymization is a type of information disinfect whose intention is to protect the data for information loss and it will provide data security. It is also a process of removing the personal identifiable information from data set, so that the people whom the data can be described remains anonymous. In this paper "We have expressed the association between the different values and different structures in different databases using syntax, e.g. XML values. Its concentrates on the privacy guarantee and the data with very simple data structure. In this paper, we focus on the tree structured data from various applications, even when the structure is not directly applied into the syntax. This paper defines the k^{m-n} anonymity which provides the complete data privacy protection against unique and it's proposes the greedy cut search GCS algorithm, which is able to disinfect the high level datasets.

Keywords - *Tree Structured Data, Data Privacy, Anonymity, Data Sanitation, Structural Disassociation, Data Generalization, Synopsis Tree.*

1. Introduction

In large scale organizations like hospitals, MNC, textile industries etc., Data storage is really a high level issues in real time environment. So, organizations should maintain the records for each and every person, they might be patients or employers etc. Because the diseases and the medicine for the treatment of each patient will vary and also a patient will be having a different medical records. As per the hospitals and textile industries, each person's data may exist for 'n' number of times. By tracing the number of transactions and repetition of data attributes for each data generation, the performance complexity and time consumption can be analyzed.

All the user's information is stored into the same databases, there might be an information loss. To avoid such cases, data anonymization technique is used. This technique is used, in order to allow processing, the personal data of the single person without information loss. Data backup can be done using this algorithm formation. Certain information can be collected from the different tables by the foreign key relationship. Even the information can be derived from different databases and it is kept in the more flexible representation as as XML record. Such tree structured data cannot be get anonymized effectively with the table based anonymization methods.

This paper proposes about the k^{m-n} anonymity, which gives us guarantees that an attacker who knows up the 'm' element of a record and 'n' is a structural relations between the 'm' of the elements will not able to match her background knowledge to less anonymization procedure could not generalize the values, which participate in rare item combinations.

2. Methodology

Methodology is the theoretical analysis of data anonymity with very low loss of information. The following are the four modules involved in the data sanitization on tree structured data efficiently.

1) Synopsis Tree

2) Candidate Solution Check

3) Applying Anonymization algorithm. All Cut Search (ACS)

4) Greedy Cut Search Algorithm(GCS)

3. Information Loss

In this section, the value generalization and the structural disassociation destroy the original data transformation and it introduces the data loss in the k^{m-n} anonymity. When we are evaluating the effect of anonymization technique, we need the measurement of both generalized values and structural disassociations. We also need to measure of loss of anonymized data for the tree structured dataset D. To measure the information loss, we can use a simple metrics formulae called Reverse Path Domain (RPD), which calculates the reduction in the value generalization and structurally disassociated paths. The Dataset function d () gives the depth of a node. The distance between the node and the root can be also evaluated by using the same function d (). The Instinctive behind the dataset function d () is very nearer to their root, which is more important. Instinctive was verified experimentally by using the RPD paths.

3.1 Architecture

In this paper, data anonymization may be carried out in two ways. 1. Authenticated User, and 2. Unauthenticated User. The Admin accessed person can add the new person and also can modify the features of existing users. Authenticated user access the data using different functionality called Value Check and Candidate Check Solution. This user can also simplify and modify the sensitive data from large datasets.

In System operation, the user can access the data from different databases and also the different trees structured data. When the user accesses the data from the global data storage, the data security methodology is applied.

The transactional data can be accessed only after the data generalization and data anonymity. When the user takes the data from the transactional data from the global database, where will be the data checking takes place. Check Priority called Candidate check solution checks the authentication for the user.

If the user is authenticated user, he can directly perform the Greedy Cut Search algorithm (GCS) and Anonymization algorithm called All Cut Search (ACS) to the sensitive data from the large scale organization.

Algorithm can be used for the classification of data from the dataset. Sensitive data can be accessed only the admin, who has the entire login rights for adding or removing the data from the dataset in different databases.

User, who doesn't have the admin rights will get failed in check priority. They cannot add the new user for the data manipulating functionalities. Unauthenticated user can access only the data from the Dataset without applying any functionalities and algorithms. At last the transactional data can be derived by the authenticated user and not by the unauthorized user, because all the admin rights can be given to the all users who have works on the transactional data.

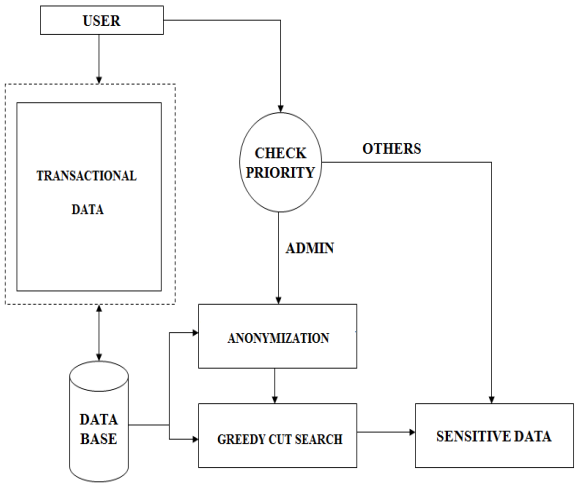

Fig 1. Architecture Diagram

4. Synopsis Tree

This module generates dataset. This module, it generates two types of dataset. A tree structure, which is created by super positioning the records in the Dataset D. Every record in the dataset is mapped to the single node called the root rs of the synopsis tree. Each node n has two elements:

1. Label which represents the records mapped to it.
2. ID's of all the records are sorted by the exact path from the root node to current node.

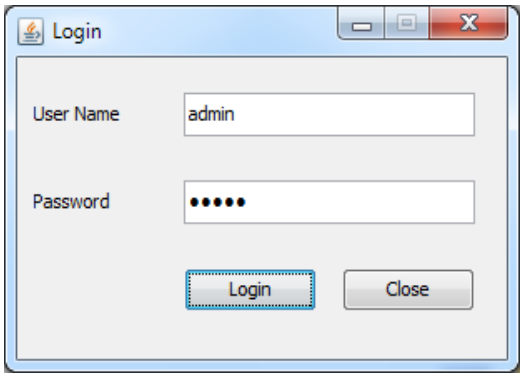

Fig 1. User Login

Here the user login allows, checking whether the user is authenticated user or not. Still the only the authorized user only create a new user and can give enough rights to them. Even the user can close the screen, if there are not interested to data anonymization at that particular time period.

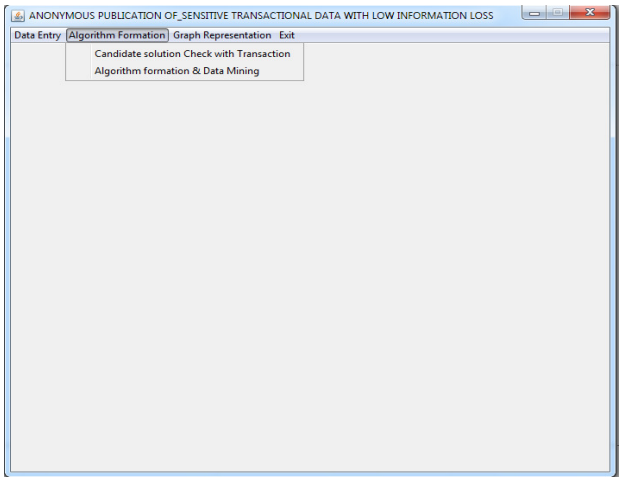

Fig 2.Transactional Data Set

The Data from the database are transactional data which can be accessed by the users, who has authenticated rights.

5. Candidate Solution Check

Candidate solution check is the process that can be used to quickly verify, if the solution are sufficient for providing the k^{m-n} anonymity which can be applied to the dataset D. This process can be performed into two phases:

1. Generalization Check: The first process generalization check is used to check whether the item sets i of the same size m is contained in the Dataset Dc appear at least k times.

2. Structural Relation Disassociation: This Structural Relation Disassociation will examine the Dataset D, whether they occur at least k records that contains combinations between them, when 'n' times structural relations occurs.

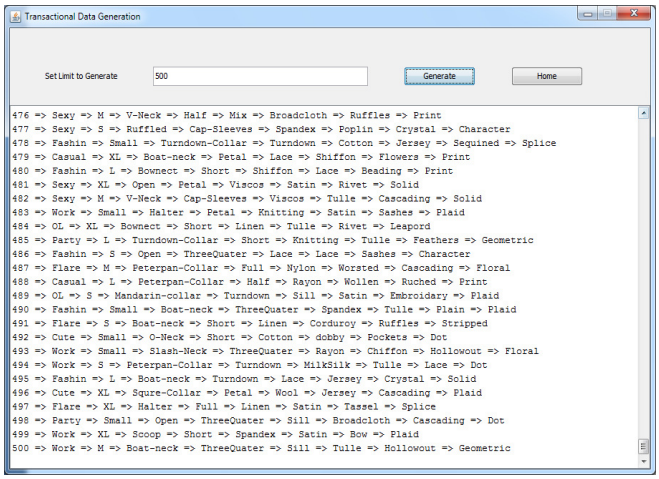

Fig 3. Transactional Data Generation

Candidate check solution uses pseudo codes for the two functionality.

1. The ValueCheck() calculates the repeated date items with different combinations in n number of transactions, that does not appear at least k times in dataset combinations Dc.
2. The StructureCheck() is to examine whether the Dc supports more than the k times of combinational items. It returns true if and only if the items with the relation between them is true.

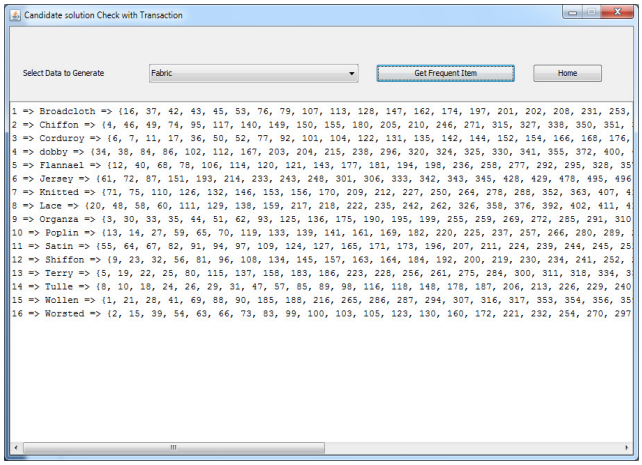

Fig 4: Candidate Solution Check

6. Anonymizaton Algorithm

All Cut Search (ACS)

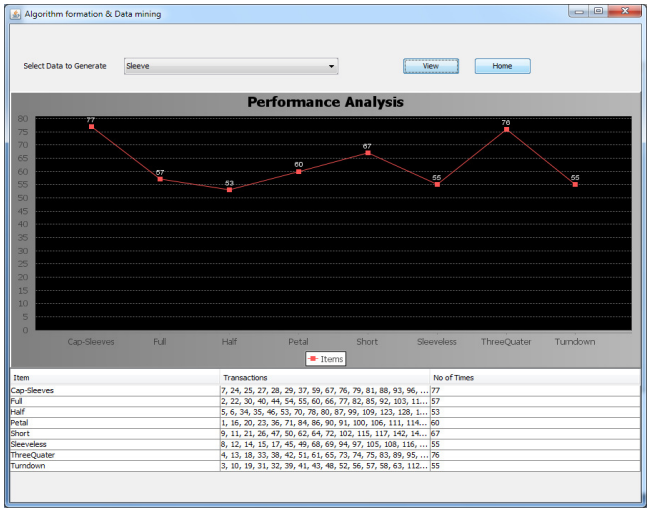

Fig 5. Algorithm Formation and Data mining

The characteristics of the anonymization algorithm is not symmetric. Both the ValueCheck and the StructureCheck functionality are completely symmetric. There are asymmetric when they will examine with different combinations of Data in the Dataset D. The complete solution for the problem comprises for all the possible cuts and all the possible disassociations rules for them, This process uses all algorithm called All Cut Search ACS Algorithm. Using this algorithm, we can define the performance analysis of the data transformation.

It explains us about the k times (no of times), in which the items can be repeated throughout the entire classifications. It shows us the items that can be transacted clearly, that help us to classify the sensitive data easily. Each and every records of the dataset have their identity. Using the unique identification ACS algorithm can be performed easily.

7. Greedy Cut Search Algorithm

The All Cut Search ACS Algorithm is not completing satisfies the law. ACS algorithm avoids exploring the whole solution space, because it doesn't satisfy the entire data domain or the dataset is large.

Greedy Cut Search GCS algorithm performs the entire assumption of the time analysis of the dataset. This algorithm can be entirely used for the time consumption analysis and that is also used to determine the entire dataset of the difference databases.

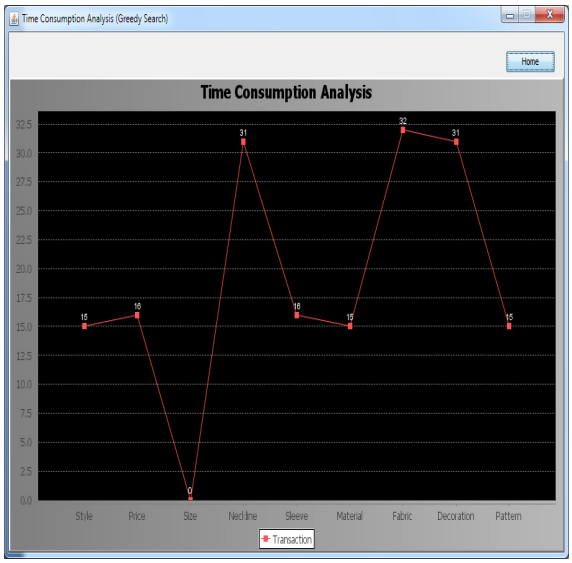

Fig 6. Time Analysis Performance

8. Conclusion

In this paper, the proposed method is based on k(m-n) anonymity. Addressing the problem of anonymizing the tree structured data in the presence of structural knowledge. We propose k^{m-n} anonymity privacy guarantee which addressing the background knowledge of both structure and value. This anonymization algorithm is used to create the k (m-n) anonymous datasets, by examine the value generalization and a data transformation of novels, which we determined about the structural disassociation.

References

[1] Olga Gkountona, Anonymizing Collections of Tree Structured Data, 1041-4347© 2015 IEEE

[2] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L.Xiong. Publishing set valued data via differential privacy. PVLDB, 4(11):1087–1098, 2011.

[3] J.Cheng, A.W.-c.Fu, and J. Liu. K-isomorphism: privacy preserving network publication against structural attacks. In SIGMOD, 2010.

[4] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. In PRIVDB, 2013.

[5] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In SIGKDD, pages 1253–1261, 2011.

[6] G. Cormode, C. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Empirical privacy and empirical utility of anonymized data. In PRIVDB, pages 77–82, 2013.