American Journal of Intelligent Systems 2017, 7(3): 50-53
DOI: 10.5923/j.ajis.20170703.02

Hiding Personal Detail using Overlapping Slicing

Rakshatha V.*, Supriya Salian

Department of Computer Science and Engineering, St Joseph Engineering College, Mangalore, India

Abstract Preserving the privacy while publishing the medical dataset is one of the techniques that can be implemented to preserve the privacy on the collected large scale of medical dataset. Medical data set contains the information that will include the personal identity of an individual therefore reproducing the same data to third party may gain privacy threats, which will include the personal detail of an individual. This paper proposes a data hiding technique called overlapping slicing for the better privacy preservation of the medical dataset that gets published.

Keywords Privacy, Overlapping Slicing, Privacy Preservation Data Publication

1. Introduction

Preservation on the privacy of the published medical dataset has been most significant issue. The modern era hangs on to the rules and regulations to limit to the different types of information and a bond to store and utilize the sensitive information. Contracts and agreements don't offer assurance that personal detail will not be revealed and end with wrong hands. Particularly when medical dataset gets published. Therefore publishing such kind of medical related datasets we have to make sure that certain techniques have been applied so the privacy is maintained while publishing the medical data. So when the medical dataset is been getting published to outside world proper measures should be taken especially while dealing with the sensitive information about any individual. This activity is called privacy preserving data publishing (PPDP). PPDP provides methods and tools for publishing useful information while preserving data privacy. PPDP offers methods and tools for publishing useful information while preserving the privacy of the medical dataset. In the most basic form of privacy-preserving data publishing (PPDP), there are different forms of identifiers namely:

- Explicit Identifier: These attributes contains set of attributes such as name and social security number.
- Quasi Identifier: These attributes contains set of attributes such as Birth date, zip code and sex.
- Sensitive Attributes: These attributes contains set of attributes such as disease and salary.

Published data becomes more useful if and only if the person's identity is preserved. In this paper we are making use of technique called overlapping slicing as it conserves the usefulness of data against privacy threats. In the information gathering stage, the data publisher will gather the required information from the record owners. Once the required medical datasets are collected from the record owners, and then that medical datasets will be released to the public or data miner called data receipt. Data miner plays an important role of performing the data mining operation on the collected medical dataset. In the current scenario shown in the figure in the information gathering stage, the data publisher collects the required medical dataset from the record owners i.e. Alice and Cathy. Once the information is gathered from record owners, data publisher will release the data to the public called the data recipient. In the figure shown below data publisher is hospital where it gathers information from patients and patient medical record history and then publish those data to the data recipient who refers to the medical center. We have to preserve the personal detail of an each individual.

1.1. Privacy-Preserving Data Publishing

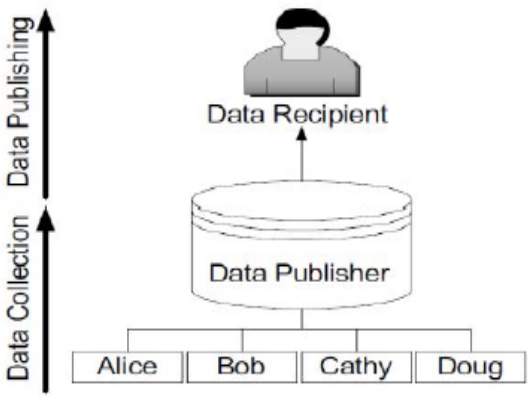


Figure 1. Data collection and publication

* Corresponding author:
rakshav700@gmail.com (Rakshatha V.)
Published online at http://journal.sapub.org/ajis

2. Literature Survey

Privacy preserving in the field of data mining, is the area where data mining techniques are applied to protect the sensitive information from illegal user. The problem of preserving the privacy of the data that gets published had become very important in current years due to the increasing ability for storing the personal data about each individual. Numbers of techniques have been introduced like bucketization, generalization, randomization etc. In order to preserve the privacy in data mining and handling the high dimension data, generalization has drawback of preserving the information that gets lost according to the recent researches. Even bucketization method doesn't prevent membership disclosure and at same time it is not applicable for the data that doesn't have a clear division between the sensitive attribute and quasi attribute [1].

A new method called slicing technique that partition data. Seems to be the most effective algorithm for figuring out the sliced data which follows l-diversity constraint, where we can prove that slicing prevents membership disclosure better than generalization and bucketization techniques. Compared to generalization and bucketization better utility is given by data slicing which doesn't require clear separation between sensitive attribute and quasi attribute. Data slicing conserves better data utility than generalization and much more effective than bucketization [2].

In overlapping slicing technique partitioning the data takes place vertically where highly co related attributes are grouped together so that it not reveals the information about any individuals and risk factor of identification is also reduced [3].

3. Problem Statement

Nowadays preserving the privacy in the data publication becomes difficulty. There is always risk involved in exposing the sensitive information especially when medical datasets get published to third part or agencies. In attribute partioning the attributes which are named sensitive are grouped together in one column where the violation of the privacy takes place in generalization method and in bucketization technique there exist less co-relation between the attributes. So in order to overcome this problem a technique called overlapping slicing is used. In overlapping slicing duplication of attributes takes place un more the one column which will lead to increase the correlation among the attributes.

4. Implementation

1) Generalization

Generalization is one of the most commonly used anonymization technique, where it replaces the quasi identifier with less specific values which are more reliable. Due to the high dimension of the medical data exists, it is likely that generalization technique will cause information loss. Generalization will be effective if and only if the records in the same bucket are very close to each other so there is always less chance of not losing the information in very large amount.

2) Bucketization

Bucketization technique preserves the better data utility than generalization, it has some limitation. Mainly because bucketization does not prevent from revealing the personal details.as bucketization publishes the QI values in their real form, so any type adversary can find out whether an individual has any record in the published medical dataset or not. An individual can be uniquely identified by means of three attributes namely birth data, sex and zip code. This means that most of the personal detail of the individual will be revealed from the bucketized table.

3) Slicing

Slicing technique is used to partition the data both in vertical and horizontal manner. Vertical partition is done by grouping attributes into columns and it is purely based on the concept of correlation that exists among the attributes. The idea behind the slicing is to reduce the dimensionality of the data and preserving the data usefulness in better manner compared to generalization and bucketization. Slicing handles high dimensional data and preserves the privacy by hiding the personal detail of an individual.

4) Overlapping slicing

In overlapping slicing, attributes are duplicated in more than one column. Attribute correlation exit in overlapping slicing. For example, in Table, one could choose to include the Disease attribute also in the first column. That is, the two columns are age, Sex, Disease and Zip code and Disease. This could give better data utility.

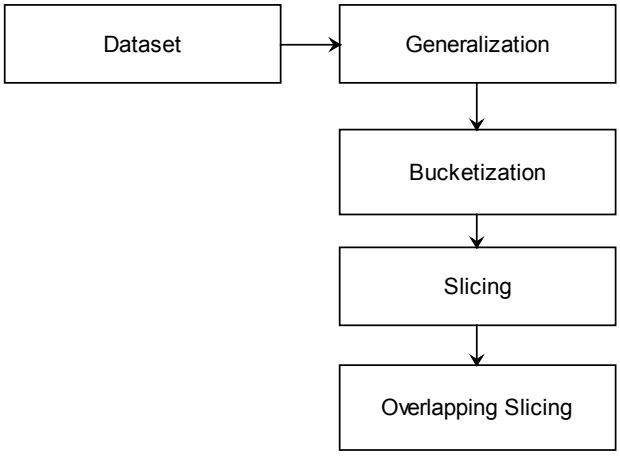
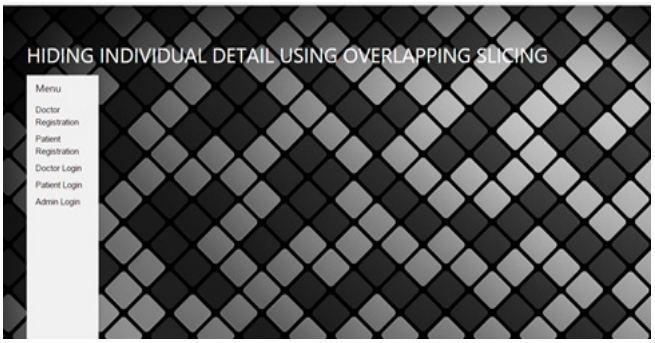


Figure 2. Overlapping slicing architecture

5. Results

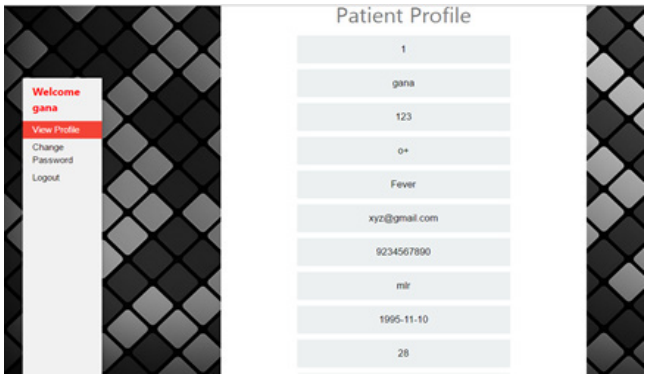
1) Homepage

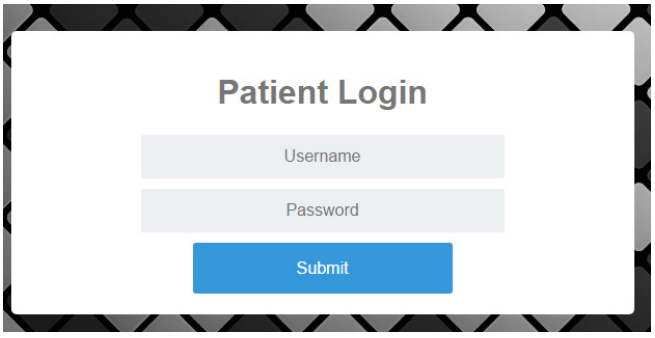
The concept of hiding individuals detail has been implemented and results are shown below. The proposed paper is implemented in Java technology on a Pentium-IV PC with minimum 20 GB hard-disk and 1GB RAM. The propose paper's concepts shows efficient results and has been efficiently tested on different Datasets. The home page describe the how the individuals privacy get disclosed and patient login, admin login is there.



2) Patient Profile and login

Here this page provides the details of patient and options to change password and logout.





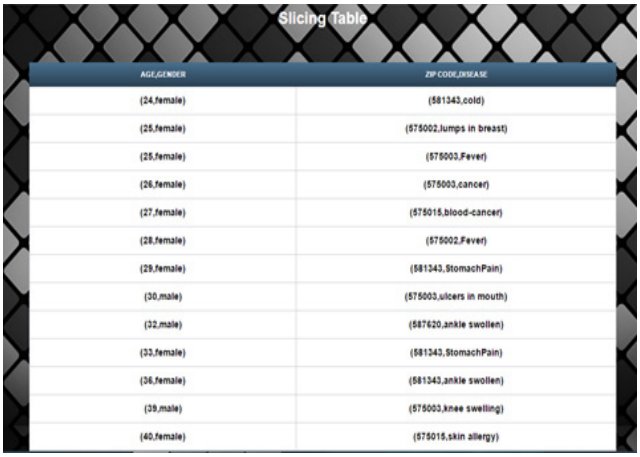
This is patient login page with option to enter his username and passsword.

3) Original data

This contains all the detailed information of patient. It includes name, zip code, sex, age, disease etc. publishing such data violates the privacy of patient.

4) Slicing

In slicing, two highly correlated attribute are grouped together in one column here age and gender are in one column so that privacy of data is preserved.



5) Overlapping Slicing

In overlapping slicing the attributes are duplicated in both columns so the privacy of the patients is preserved in better manner and if adversary trying to gain any personal information about any individual will lead to dilemma by knowing which is the original value and duplicate value.

6. Conclusions

The drawbacks of generalization and bucketization can be reduced in Slicing and overlapping slicing technique. Overlapping slicing has the potential to hold large amount of information. In overlap-slicing, the size of data is reduced by partitioning attribute into column. It conserves the data usefulness while protecting against privacy threats. In table, each column can be referred as sub-table with a lower dimensionality. Overlapping slicing is also different from the approach of publishing multiple independent sub tables in that these sub-tables are linked by the buckets in overlap slicing. Overlap-slicing can be used without such a separation of Quasi Identifiers attribute and sensitive attribute and preserves the identity of the individual.

REFERENCES

[1] D. Mohanapriya and Dr. T. Meyyappan "Slicing Technique For Privacy Preserving Data Publishing" International Journal of Computer Trends and Technology (IJCTT), Volume 4, Issue 2013.

[2] Alphonsa Vedangi and V. Anandam "Data Slicing Technique To Privacy Preserving And Data Publishing" IJRET: International Journal of Research in Engineering and Technology ISSN: 2321-7308.

[3] K. Vani and B. Srinivas "Enhanced Slicing For Privacy Preserving Data Publishing" The International Journal Of Engineering And Science (IJES), Volume 2, Issue 10, 2013 ISSN(e): 2319 – 1813.

[4] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao, "Anonymous Publication of Sensitive Transactional Data" in Proc. of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[5] G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[6] Tiancheng Li, Ninghui Li and Jian Zhang, Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, 2012.

[7] S. Giri and Mr. Nilav Mukhopadhyay "Overlapping Slicing With A New Privacy Model" International Journal of Scientific and Research Publications, Volume 4, Issue 6, 2014, ISSN 2250-3153.