Collaborative Data Publishing Using Privacy Preserving Technique

[1]Varsha Gaikwad, [2] Nikita R. Khare, [3[Chaitanya N. Kalantri
[1] Asst. Professor, [2] [3] Student Information Technology,
[1][2][3] Government College of Engineering Aurangabad,
[1] varshagaikwad2006@gmail.com, [2] nikitar.khare@gmail.com, [3]chaitanyakalantri835@gmail.com

***Abstract*- In current years, for public advantage data need to be shared. Generally data is collected from distributed databases for e.g. in case of Health care and researches, data is collected from different providers and gathered in central network. In health care all information related to patient is present in central network which includes disease details, corresponding treatment and test details.**

In this paper, we consider the collaborative data publishing for anonymizing horizontally partitioned data at multiple data providers. Here, we are trying to come yet with some of the most basic yet unseen conclusions which will help both the government as well as the individual hospitals to identify the situation of their city people. By using anonymization technique the data is modified and then released to the public. This process is known as the privacy preservation data publishing.

With the help of trusted third party data insertion of data, we are even considering "insider attack" and trying to make sure that the patient's data is safe. This paper addresses this new thread, and makes several contributions.

First, in order to make the patient's information safe we are anonymizing the data using generalization and suppression algorithm.

Second, displaying the results in the tabular form and graphical user interface form and implementing jFreeChart algorithm to display the graphical user interface data in pie chart and bar graph.

Third inserting the data with the help of a third party who is trusted and no other person will be allowed to access the data. This is avoid "external" attacks.

***Index Terms*: Generalization, Integrity, Privacy, Protection, Security, Suppression**

I. INTRODUCTION

Most work has focused on a single data provider setting and considered the data recipient as an attacker. A large body of literature assumes limited background knowledge of the attacker, and defines privacy using relaxed adversarial notion by considering specific types of attacks. Representative principles include *k*-anonymity and *t*-closeness. A few recent works have modeled the instance level background knowledge as corruption, and studied perturbation techniques under these syntactic privacy notions.

A. Disadvantages of existing system

1. Collaborative data publishing can be considered as a multi-party computation problem, in which multiple providers wish to compute an anonymized view of their data without disclosing any private and sensitive information.

2. The problem of inferring information from anonym zed data has been widely studied in a single data provider setting. A data recipient that is an attacker, e.g., *P*0, attempts to infer additional information about data records using the published data, T^* , and background knowledge, *BK*.

II. PROPOSED SYSTEM

We consider the collaborative data publishing setting with horizontally partitioned data across multiple data providers, each contributing a subset of records Ti. As a special case, a data provider could be the data owner itself who is contributing its own records. This is a very common scenario in social networking and recommendation systems. Our goal is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties.

A. *Scope of work*

Proposed concept can be used in many applications like hospital management system, many industrial areas where we like to protect a sensitive data like salary of employee. Pharmaceutical company where sensitive data may be a combination of ingredients of medicines, in banking sector where sensitive data is account number of customer, this system can be used. It can be used in military area where data is gathered from different sources and need to secure that data from each other to maintain privacy.

B. *Advantages of existing system*

Compared to our preliminary version, our new contributions extend above results.

First, we adapt privacy verification and anonymization mechanisms to work for *n*-privacy with respect to any privacy constraint, including no monotonic ones. We list all necessary privacy checks and prove that no fewer checks are enough to confirm n-privacy.

Second, we propose to display very easy to understand conclusions in the tabular as well as graphical user interface form. For all protocols we prove their security, complexity and experimentally confirm their efficiency.

III. ARCHITECTURE

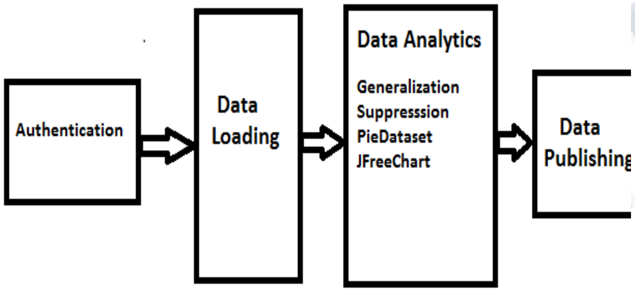

Figure 1. Flow chart of the project

IV. MODULES

a) ***Start of the Application***.
b) ***Third party login***: Add patient data to system by third party.
c) ***Patient's Registration***.
d) ***Admin Login***.
e) ***Tabular Form Display***
f) Display data in generalized form.
g) Display data in suppressed form.
h) **Graphical User Interface Form Display.**
i) Display data using Pie Chart.
j) Display data using Bar Graph.
k) Module Description
l) Start of the Application

Simply depending upon the requirement the respective module could be selected. There are basically two modules displayed in the start of the application i.e. Registration and Collaboration.

- ***Third party login***

In the Registration module, the third party have to initially log in to their account, so as to prevent the access of any other person. Only authorized user can access this application and then he or she can be able to insert patient details in a database.

- ***Patient's Registration***

So authorized user can now enter the patient's details such as Name, Disease, Email, DOB, Hospital, Pin code, Age, Area.

Furthermore, there is an "Import" named button which directly reads data from any excel file and adds the distinct records in the table.

Important steps for reading an excel file are:

- Add the respective jar file.
- Open the file to be read.
- Finally use the methods in order to retrieve the data from the file. Generally one record at a time and if the record is distinct add it.
- Continue this for all the records.
- Admin Login
  In the Collaboration module, initially the admin has to log in. Admin is the only person to have access to this module. Collaboration module contains all the conclusions and the details of the patient.
- Tabular Form Display

- Display data in generalized form.
  Generalization:
  In this module we have generalized data in such a way that it falls into some set of ranges.
  For ex. Age attribute.
  Age attribute can have any numerical value and user can reveal patients data based on his or her age so we have used a generalization algorithm so that particular age of patient
  Falls in range. For Ex. from 10-20 or 30-40 etc.
  Generalization Algorithm:
- Retrieve Age from database.
- Calculate unit place digit from age.
- Subtract above result from age this gives starting range of age
- Add 10 to starting range to get ending range
- Concatenate start age range and end age range to obtain generalization result.

Pseudo code

```
Accept age from user
unitplace=age % 10 start = age-unitplace  end = start + 10
Concatenate result of start and end  =>  start + "-" + end
```

- ***Display data in suppressed form.***
- ***Suppression:***

In this module we have suppressed data in such a way that it does not reveal patient details.

For ex. Name attribute

Name attribute can have any string of characters and user can reveal patients data based on his or her name so we have used a suppression algorithm so that particular name of patient is presented to administrator in the form of some characters from patients name and '*' character appended.

Ex. Name 'Abhilasha Joshi'

Can be presented to user as 'Ab******* J****'

Suppression Algorithm

- Retrive name from database.
- Calculate length of name.
- Take first two letters of name to display using substring function.
- Append '*' length-2 times to hide complete name.

Pseudo Code

```
Accept name from user
name=name.SubString(0,2)
Loop 2 to length of name
name=name+"*"
End loop
Accept last name and store it in variable lname
lname=lname.subString(0,1)
Loop 1 to length of lname
Lname=lname+"*"
    Endloop
```

Name	Age	Disease	Pincode
Pr*** B***	30-40	Asthma	43****
Ap**** V*****	60-70	Asthma	43****
Su****** J****	20-30	Cancer	51****
Ch****** P*****	30-40	Cancer	43****
Ra**** K***	50-60	Skin Disease	43****
Ru***** D******	50-60	Skin Disease	62****
Ta**** D*****	10-20	TB	43****

Figure 2. Generalization and Suppression table

Note that order by clause is applied on disease field while retrieving the data which helps the data to be easily differentiated and understood.

- Graphical User Interface Form Display

- There are few ways of representation of the conclusions in the GUI format. They are as follows:
- City wise conclusion.
- Area wise disease distribution.
- Hospital wise disease distribution.
- Age group wise distribution.

The first two types have been shown in detail in the following sub points.

V. CITY WISE CONCLUSION

Using the dummy data, the concept is shown. The above pie chart has Skin disease count equal to 3, similarly for Cancer, Asthma, and TB its 3, 1 and 4 respectively.

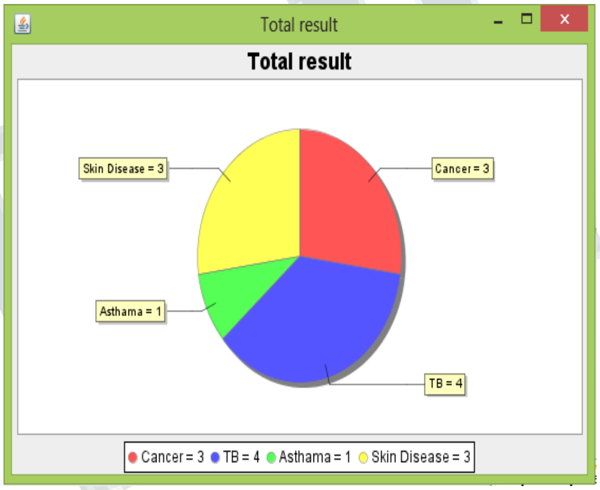

Figure 3. Output image of the total results using some dummy data in the form of pie chart

a) *Area wise disease distribution*

Using the dummy data, the concept is shown. The below bar graph has area wise disease distribution. For instance in Cidco area there are 1 unit of Cancer and Asthma patients. Similarly in Hudco area there are 2 units of Cancer, in Osmanpura there are 4 units of TB, and in Garkheda there are 3 units of Skin disease.

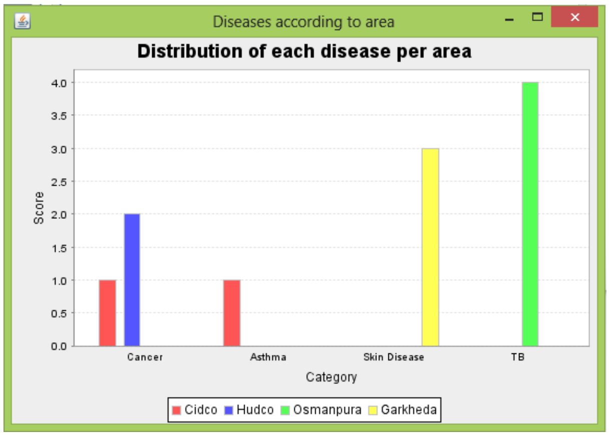

Figure 4. Output image of the distribution of each disease per area using some dummy data in the form of bar graph

b) Hospital wise disease distribution

Consider a particular hospital X, there are n1 number of TB patients, n2 number of Cancer patients and similarly n3 and n4 for Asthma and Skin disease respectively. The illustration of which is demonstrated in the below output figure.

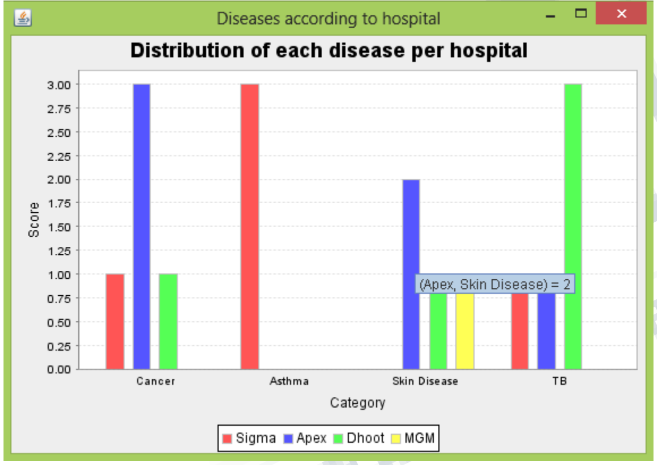

Figure 5. Output image of the distribution of each disease per hospital using some dummy data in the form of bar graph.

c) Age group wise distribution

Consider any age group X, there are n1 number of patients suffering from TB, n2 number of people suffering from Asthma and similarly n3 and n4 for Cancer and Skin disease respectively. The illustration of the above concept is demonstrated in the below output figure.

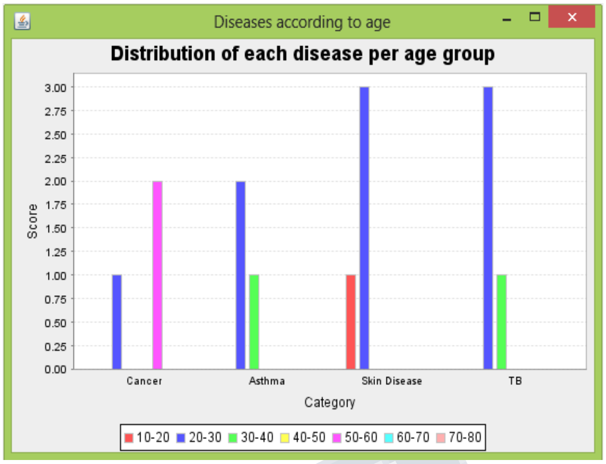

Figure 6. Output image of the distribution of each disease per age group using some dummy data in the form of bar graph.

VI. CONCLUSION

In this paper, we tried to anonymize the data of the user by using algorithms such as Generalization and Suppression. Generalization helps to specify the range of age instead of making it particular or fixed. Suppression helps to hide the name of the patient. Our experiments confirmed that our approach achieves better or comparable utility than existing algorithms while ensuring privacy efficiently. There are many remaining research questions.

Moreover, the conclusions are pretty easy to understand even for any nomad person. The conclusions are represented in the form of Pie Chart and Bar Graph. There are various conclusions that we are trying to come up with. For instance, area wise disease distribution and age wise disease distribution.

Furthermore, it would be also interesting to verify if our methods can be adapted to other kinds of data such as set-valued data.

FUTURE SCOPE

Above discussed approaches help to enhance the data privacy and security when data is collected from various resources and output should be in collaborative style. In future, this system can consider for data, which are distributed in ad hoc grid computing. Also the system can be considered for set valued data. The consumption of various protocols can address various data publishing paradigms. The consumption of these protocols can make collaborative data publishing more effective and enhanced the privacy.

REFERENCES

1) Mingxuan Yuan, Lei Chen, Member, IEEE, Philip S. Yu, Fellow, IEEE, and Ting Yu-"Protecting Sensitive Labels in Social Network Data Anonymization"-IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013.

2) B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput.Surv., vol. 42, pp. 14:1–14:53, June 2010

3) Karthikeyan.B,Manikandan. G,Vaithiyanathan. V," A FUZZY BASED APPROACH FOR PRIVACY PRESERVING CLUSTERING", Journal of Theoretical and Applied Information Technology, 2011, Vol. 32 No.2.

4) http://www.tutorialspoint.com/jfreechart/jfreechart_bar_chart.htm

5) http://www.tutorialspoint.com/jfreechart/jfreechart_pie_chart.htm