**INTRODUCTION**

The World Wide Web consists of plethora of information. There are about 2 billion web pages on web. The main source of the information on these web pages is from huge databases which are increasing tremendously day to day. This explosive increase in storage and processor capacity has made the data sharing very easy which possesses a greater threat to personal data. Data mining plays a key role in extraction of this information from the databases. These Data mining services require accurate variations, distributions, forecasting, inventory holdings etc. However, the customer wants the details about their buying patterns to be kept confidential. The reason for this may be the worry that the requested information may be misused by the service provider to harass the customer. For example, some companies conduct panel surveys in order to take decisions regarding advertisements, product variations, distributions, forecasting, inventory holdings etc. However, the customer wants the details about their personal purchases to be confidential even though the company may be acquiring the data solely for genuine data mining purposes that would eventually reflect itself in better service to the client. A number of privacy preservation techniques have been proposed for hiding sensitive details of the individuals. But all these results in over anonymity. As a result, the accuracy of data mining results is reduced. To avoid such problems a cluster based cryptographic technique is used. In this paper a patient database is considered and privacy is provided for patient's data by applying clustering algorithm, cryptographic technique on patient database. The further organization of the paper includes: considering a dataset, apply dbscan clustering algorithm on certain attributes and apply cryptographic technique on other attributes.

**RELATED WORK**

Privacy concerns are increasing as the World Wide Web makes it very easy for the new data to be automatically collected and added to databases. The individuals provide information based on any of the following scenarios [1].

1. **Individuals provide information when they are not identified.** The users provide information only if they are not personally identified. For example in case of medical domain research works are carried out to find new medicines based on the disease patterns of the patients. In this case the patients will provide information about their problems only if their identities are not revealed.

2. **Individuals do not provide sensitive details.** The respondents often feel very uncomfortable to provide personal identification information like credit card numbers, social security numbers, phone numbers, postal address etc because they are worried that this information can be used to link with other external data sources to re-identify person specific information.

3. **Information sharing.** When deciding whether to provide information to researchers or not, the respondents report that the most important factor is whether or not information will be shared with other companies and organizations.

4. **Purpose of gathering information from users.** Some of the users consider the purpose of collecting the information. If it is really for a valid research work then they will be ready to provide even the sensitive details.

5. **Benefits:** The individuals provide information also based on the benefits they obtain by revealing their private details. This may be in terms of money, property, gifts etc;
The respondents show two different kinds of behavior. One is shown when they are not asked for their personal identified information and the other kind of behavior is shown when asked for their personal identified information. This is shown below by considering two scenarios.

![Graph comparing individual behavior](image)

**Privacy preservation techniques**

In this paper a dataset containing different attributes is considered initially and the privacy preservation techniques are applied on the quasi attributes. Quasi attributes are the attributes which can be used to re-identify an individual by linking original dataset with some external datasets. Consider the following patient database table.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Birthdate</th>
<th>Zip code</th>
<th>Gender</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linda</td>
<td>31</td>
<td>12-01-1980</td>
<td>512343</td>
<td>Female</td>
<td>cancer</td>
</tr>
<tr>
<td>Jeremy</td>
<td>09</td>
<td>12-12-2001</td>
<td>512742</td>
<td>Male</td>
<td>Flue</td>
</tr>
<tr>
<td>Tada</td>
<td>14</td>
<td>22-10-1996</td>
<td>512598</td>
<td>Male</td>
<td>Flue</td>
</tr>
<tr>
<td>Justine</td>
<td>49</td>
<td>12-07-1951</td>
<td>512347</td>
<td>Female</td>
<td>Parkinson</td>
</tr>
<tr>
<td>Nick</td>
<td>35</td>
<td>04-05-1976</td>
<td>512221</td>
<td>Male</td>
<td>Cancer</td>
</tr>
<tr>
<td>Karen</td>
<td>54</td>
<td>06-09-1957</td>
<td>617623</td>
<td>Female</td>
<td>Parkinson</td>
</tr>
<tr>
<td>Nichol</td>
<td>59</td>
<td>07-08-1952</td>
<td>617612</td>
<td>Female</td>
<td>Alzheimer</td>
</tr>
<tr>
<td>Warren</td>
<td>11</td>
<td>21-06-2000</td>
<td>617624</td>
<td>Male</td>
<td>Flue</td>
</tr>
<tr>
<td>Mollie</td>
<td>27</td>
<td>14-04-1984</td>
<td>617656</td>
<td>Female</td>
<td>Cancer</td>
</tr>
<tr>
<td>Deand</td>
<td>57</td>
<td>17-03-1954</td>
<td>512711</td>
<td>Male</td>
<td>Parkinson</td>
</tr>
<tr>
<td>Myah</td>
<td>40</td>
<td>16-02-1972</td>
<td>512765</td>
<td>Female</td>
<td>Heart disease</td>
</tr>
<tr>
<td>Pedro</td>
<td>63</td>
<td>18-07-1948</td>
<td>612341</td>
<td>Male</td>
<td>Alzheimer</td>
</tr>
</tbody>
</table>

In the above table the quasi attributes are to be identified. The quasi attribute are name, age, date of birth, zip code, gender. In these attributes name, gender is anonymized using cryptographic technique and age, date of birth, zip codes are anonymized using clustering technique.

Since we have used two different techniques the over anonymity problem occurred due to application of single technique on all dimensions can be reduced.

**Applying clustering technique**

The DBSCAN clustering technique is a popular technique used to group the attributes based on the density values. The center based approach is commonly used approach for defining density. The density is estimated for a particular point in the dataset by counting the number of points within a specified radius. The points within the density are grouped to one cluster. These points are categorized to three types. 1. Core points. 2. Border points 3. Noise points.

**Core points**: A point is a core point if the number of points within a specified distance parameter and a user specified parameter exceeds a certain threshold.

**Border points**: Border points are the points that falls within the neighborhood of several core points.

**Noise points**: A noise point is one that is neither a core point nor a border point.

![DBSCAN Algorithm](image)

**DBSCAN**

1. Plot the points in a graph
2. Specify the distance parameter Eps and number of points to be present within Eps distance
3. Specify the user defined parameter that determines the number of points to be present within Eps distance
4. Now identify core noise and border points
5. Remove noise points
6. Put an edge between all core points that are within eps of each other
7. Combine each group of connected core points to one cluster
8. Assign border points to one of the clusters depending on its associated core points.

Now consider the patient database table. The attributes name, age, birthdates, zip code, gender has to be anonymized using the dbscan algorithm. This is shown below.

Now let the distance parameter Eps be 30 and number of points within Eps be 4.

Now consider the attribute age
Here the numbers represent the names of different people in table from top to bottom that is 1- Linda, 2- Jerome and so on. Two different clusters are formed one represent the age between 1-30 and other between 31-40. So we can replace the age with cluster names as 1-30 and 31-40. As a result the identification becomes difficult.

The next attribute is birth date the birth date can be given the same name as like age because we can determine the year of birth based on age. The next attribute gender is clustered as below

After applying the clustering techniques the values of the three attributes (age, gender, and birth date) are replaced as below

<table>
<thead>
<tr>
<th>Age</th>
<th>Birthdate</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster B</td>
</tr>
<tr>
<td>1-30</td>
<td>Cluster 1</td>
<td>Cluster A</td>
</tr>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster A</td>
</tr>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster A</td>
</tr>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster A</td>
</tr>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster A</td>
</tr>
<tr>
<td>1-30</td>
<td>Cluster 1</td>
<td>Cluster A</td>
</tr>
<tr>
<td>1-30</td>
<td>Cluster 1</td>
<td>Cluster A</td>
</tr>
<tr>
<td>1-30</td>
<td>Cluster 1</td>
<td>Cluster A</td>
</tr>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster A</td>
</tr>
<tr>
<td>30-60</td>
<td>Cluster 2</td>
<td>Cluster A</td>
</tr>
</tbody>
</table>

The other attributes of patient database remains the same.

**Applying cryptographic technique**

The next attribute is name. It is anonymized using the cryptographic technique. Hill cipher cryptographic technique is used for anonymization. In this technique the encryption algorithm takes m successive plain text letters and substitutes for them m cipher text letters. The substitution is done in m linear equations in which each character is assigned a numerical value. i.e; a -0, b-1......z=25. For m=4 the system is described as

\[
\begin{align*}
C1 &= (k11p1+k12p2+k13p3) \text{ mod } 26 \\
C2 &= (k21p1+k22p2+k23p3) \text{ mod } 26 \\
C3 &= (k31p1+k32p2+k33p3) \text{ mod } 26 \\
C4 &= (k41p1+k42p2+k43p3) \text{ mod } 26 \\
C &= PK \text{ mod } 26
\end{align*}
\]

For example let the key be

\[k = \begin{pmatrix}
17 & 17 & 5 \\
21 & 18 & 21 \\
2 & 2 & 19 \\
10 & 5 & 7
\end{pmatrix}\]

Now consider the word Linda from the patient database table above. It is encrypted using hill cipher technique as below

At first each alphabet is assigned its numeric equivalent

\[
\begin{align*}
L &= 11, I &= 8, N &= 13, D &= 3, A &= 0
\end{align*}
\]

Now let the key be

\[K = \begin{pmatrix}
17 & 15 & 7 \\
13 & 2 & 8
\end{pmatrix}\]

We have \[C=PT \text{ mod } 26\]. Therefore

\[
\begin{align*}
C &= \begin{pmatrix}
11 & 81 & 11 & 17 & 15 & 7 \\
2 & 0 & 5 & 13 & 2 & 8
\end{pmatrix} \text{ mod } 26 = \begin{pmatrix}
5 & 25 & 8 \\
0 & 19 & 10 & 14 & 0
\end{pmatrix}
\end{align*}
\]

Now replace these digits with corresponding letters. We get

\[
\begin{align*}
5 &= f, 25 &= z, 8 &= i, 0 &= a, 19 &= t, 24 &= y, 13 &= n, 10 &= k, 14 &= 0
\end{align*}
\]

Therefore we have \[C = fziatynk\]. Since we have added one more digit randomly to make the matrix multiplication to be valid, we can now remove the last digit so we get the correct encrypted word as fziatynk. Similarly the same process is followed for all other names. Different keys can be used for different names. As a result it becomes highly impossible to guess and decrypt the original names. This cryptographic technique provides high confidentiality. Even if the other attributes that are anonymized using clustering technique are identified then it becomes difficult to reidentify the original individual as the name has been encrypted.

Some other names are encrypted with different randomly generated keys as follows

<table>
<thead>
<tr>
<th>Name</th>
<th>Key</th>
<th>Encrypted Name</th>
</tr>
</thead>
</table>
| Jerome | \[k = \begin{pmatrix}
9 & 4 \\
17 & 12 \\
14 & 24 \\
11 & 16
\end{pmatrix}\] | klycnoyki |
| Jerome | \[k = \begin{pmatrix}
6 & 8 \\
10 & 5
\end{pmatrix}\] | kwsy |
| Tada | \[k = \begin{pmatrix}
10 & 22 \\
10 & 5
\end{pmatrix}\] | k|

The same procedure is applied for the field zip code. All the zip codes in the database table consist of 6digits. So the key used for this is a 2X3 matrix. A different key can be chosen randomly for every zip code to provide high privacy.

Since the zip code contains digits instead of characters. We can directly encrypt them without the need of identifying the numeric equivalents as in case of name. Now consider the zip code 512343

Let the key be

\[k = \begin{pmatrix}
10 & 5 \\
7 & 2
\end{pmatrix}\]

Therefore we get \[C = 511251520239161\]
It is highly impossible to determine the original zipcode from this encrypted one because infinite number of combinations has to be tried to determine the original zipcode. Even if the encryption method is known the zip code cannot be determined because it is difficult to try a vast number of combinations as the key cannot be guessed each time.

Therefore by using this cluster and cryptographic techniques privacy for information of users will be provided. The data will be displayed but it is the one that has been anonymized.

This process is done for all the attribute values of name and zip code. Finally the private information of the individuals that is the disease will be hidden by these clustering and cryptographic techniques.

**CONCLUSION**

Multidimensional anonymization techniques results in over anonymity problem if the same technique is applied over a number of dimensions. In this paper a cluster based cryptographic technique is used which can avoid over anonymity problem and provides better privacy for private details of the individuals.

**REFERENCES**


