ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A SURVEY ON TO PRESERVE PRIVACY FOR COLLABORATIVE DATA PUBLISHING

Payal Patel, Sheetal Mehta
Master of computer engineering, Parul Instituteof Engineering and Technology, India

ABSTRACT

The collaborative data publishing issue for anonymizing horizontally partitioned data at various data providers is considered. Another kind of "insider attack" by colluding data providers who may utilize their own particular data records (a subset of the general data) notwithstanding the external background knowledge to construe the data records helped by other data providers. This new risk makes a few commitments. The idea of m-privacy, which ensure that the anonymized data fulfils a given protection requrements against any group of up to m colluding data providers. Two algorithms for collaborative data are: one is a heuristic algorithm which is checking m-privacy anonymized data from the provider and second algorithm is provider aware anonymization which ensures m-privacy methodology and it is highest rated anonymized data providing efficiency. Secure multi-party computation (SMC) protocol and trusted third party (TTP) protocol can be used to guarantee that there is no disclosure of intermediate information during the anonymization. This protocol is used in the system at server side for keeping the data secure. Experiments on real-life datasets recommend that this methodology accomplishes better or similar utility and efficiency than existing and baseline algorithms while giving m-privacy ensure.

KEYWORDS: Data Anonymization, Privacy Preservation, Data Mining, Slicing.

INTRODUCTION

Privacy means Individual's desire and ability to keep certain information about themselves hidden from others. Privacy is the quality or condition of being secluded from the presence or view of others. On relating privacy with data mining, privacy implies keep information about individual from being available to others. Privacy is a matter of concern because it may have adverse effects on someone's life. Privacy is not violated till one feels his personal information is being used negatively. Once personal information is revealed, one cannot prevent it from being misused. Let us take an example, date of birth, mother's maiden name, or sex etc. may not become a threat for an individual, but if one more attribute like the unique identification number or voter ID are also known then it may cause a serious effect like identity theft.

There are various security issues arising in day to day various data i.e. medical data healthcare data marketing and sales data.it is necessary to provide privacy for such data. This will lead to security (protection) of data.

There are many issues in privacy preserving. One of the major problems of privacy preserving data mining is the abundant availability of personal data. Second is Intimidation of data sharing is majorly caused by recent trends in data mining

The main aim of privacy is to get the global result without affecting on security. Security and privacy (confidentiality) is of utmost importance in any kind of large scale data-mining.

Whenever we are concerning with data mining, Security is measure issue while extracting data. Privacy Preserving Data Mining concerns with the security of data and provide the data on demand as well as amount of data that is required. Sometimes it may happen that we get the information but not complete. The Privacy Preserving algorithm is concerns on the basis of its performance, data utility. There are various techniques and tools for security are used.

Privacy preserving data mining deals with hiding an individual's sensitive identity without sacrificing the usability of data. It has become a very important area of concern but still this branch of research is in its infancy. People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information .

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

DETAILS OF METHODS

Generalization[1]

Generalization works by first removing identifiers from the data and then partitioning tuples into buckets and then transforming the QI values in each bucket into less specific but semantically consistent values such that the tuples in the same bucket cannot be distinguished by their QI values. this technique fails for high-dimensional data and forces a large amount of generalization which greatly reduces the utility of the published dataset. Also, since the specific value of a generalized interval cannot be determined, the data analyst has to assume a uniform distribution for each value in the interval. This further reduces the utility of the anonymized dataset.

Table 2.1 generalized database[1]

Age	Sex	Zip	Disease
[20-52]	*	4790*	Dyspepsia
[20-52]	*	4790*	Flu
[20-52]	*	4790*	Flu
[20-52]	*	4790*	Bronchitis
[54-64]	*	4730*	Flu
[54-64]	*	4730*	Dyspepsia
[54-64]	*	4730*	Dyspepsia
[54-64]	*	4730*	Gastritis

Bucketization[1]

Bucketization too works by first removing identifiers from the data and then partitioning tuples into buckets but then it separates the SAs from the QIs by randomly permuting the SA values in each bucket. The anonymized dataset then consists of a set of buckets with randomly permuted sensitive attribute values. This technique does not provide protection against membership disclosure and an adversary can find out whether an individual has a record in the published dataset or not because the QI values are published in their original forms.

Slicing[1]

In slicing, values are randomly permuted in each column. So the break linking between different column. There is Important advantage of slicing its ability to handle high-dimensional data. slicing preserves better data utility than generalization and is more effective than bucketization.

Table 2.2 Sample Database [2]

Age	Sex	Zip	Occupation	Education	Disease
20	F	12578	Student	12th	Flu
41	M	12589	Government	Post-Graduate	Dyspepsia
26	M	12160	Sales	10th	Dyspepsia
23	F	12216	Student	Graduate	Flu
29	M	12903	Agriculture	12th	Gastritis
32	M	12093	Army	Graduate	Bronchitis

Table 2.3 Sliced database[2]

Sex	Occupation	Zip	Education	Age	Diseases
M	Sales	12460	10th	32	Bronchitis
M	Army	12578	12th	26	Dyspepsia
F	Student	12093	Graduate	20	Flu
M	Agriculture	12216	Graduate	29	Gastritis
F	Student	12589	Post-Graduate	23	Flu
M	Government	12903	12th	41	Dyspepsia

Improved Slicing[2]

Improved slicing works by first finding the correlations between each pair of attributes and then clustering these attributes into columns by overlapped clustering on the basis of their correlation coefficients. The columns within each bucket are then randomly permuted with respect to one another to give an improved sliced dataset.

Overlapped Clustering

Tables 2.3 and 2.4 show the anonymized tables after applying slicing and improved slicing techniques respectively. In Table 2.3, *Disease* is grouped with *Age* and *Sex* is grouped with *Occupation*. Even if *Occupation* also had a reasonably high correlation with *Disease* but *Sex* did not, they could not be combined into a bigger group and thus the data utility due to the correlation between *Disease* and *Occupation* is lost. In Table 2.4, the attributes *Occupation* and *Disease* are present in more than one column i.e. they are overlapping. This allows highly correlated attributes to group together. This also solves the problem of lone columns by merging

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

correlated attributes into a new column instead of just leaving out an attribute with a low correlation.

Table 2.4 Improved sliced database[2]

Sex	Occupation	Zip	Education	Age	Disease	Disease	Occupation
M	Sales	12460	10th	32	Bronchitis	Dyspepsia	Sales
M	Army	12578	12th	20	Dyspepsia	Flu	Student
F	Student	12093	Graduate	20	Flu	Bronchitis	Army
M	Agriculture	12216	Graduate	29	Gastritis	Gastritis	Agriculture
F	Student	12589	PG	23	Flu	Dyspepsia	Government
M	Government	12903	12th	41	Dyspepsia	flu	Student

Tuple partitioning

$Q = \{T\}$;
$SB = \emptyset$;
while Q *is not empty* **do**
  Remove the topmost bucket B from Q;
  Split B into m buckets;
  for *each of the* m *buckets* **do**
    Randomly allot half the tuples to B_1;
    Allot rest of the tuples in bucket to B_2;
  end
  if ***diversityCheck***$(T, Q \cup \{B_1, B_2\} \cup SB, l)$ **then**
    $Q = Q \cup \{B_1, B_2\}$;
  else
    $SB = SB \cup \{B\}$;
  end
end
return SB;

Algorithm 1:tuple partitioning[2]

Here, B is the bucket containing the set of tuples, SB is a null variable, Q is the set having tuples. While Q is not empty, topmost tuple is to be removed; suppose it's B. Then split operation is performed; where B is split into m number of buckets. Then randomly, half the tuples of the split tuples are allotted to another bucket B1 and rest of the tuples to B2. And if Q is empty and the condition is satisfied, then all the tuples are present in Q otherwise SB contains the tuples of B.

k-anonymity

It is appropriately controlled by certain anonymized information.K namelessness does not ensure for protection.characteristics are stifled or summed up until each one line is indistinguishable with in any event k-1 different columns then that strategy is called as a k-obscurity. It avert database linkages furthermore ensures that the information discharged is exact.k-privacy can't be connected to high-dimensional information without complete loss of utility.

l-diversity

l-diversity is a form of group based anonymization that is used to preserve privacy in database.Method overcomes the drawbacks of k-anonymity but fails to preserve the privacy against skewness and similarity attacks.

t-closeness

Method is called as t-closeness when the distance between the distributions of a sensitive attribute in same class.Distribution of the attribute in the whole table is no more than a threshold t.It preserves the privacy against homogeneity and background knowledge attacks.

Limitations of the Existing System

- Bucketization does not prevent membership disclosure.
- Overlapping of Data in Improved Slicing.
- In slicing, randomly permutated values in column so the lost of data utility.

CONCLUSION

During my literature survey the work I have concluded provide privacy on sensitive data's. That is medical data. The patient's data are store on the database by applying various comparison techniques so that the data cannot be retrieved or view by third party or other users. A new type of potential attackers in collaborative data publishing – a coalition of data providers, called m-adversary is considered. To prevent privacy disclosure by any m-adversary we showed that guaranteeing m-privacy is enough.

When the user searches for sensitive query, the results were derived with k- anonymous property. The results shown were name, age, disease and pin code. The age value is generalized and pin code value is suppressed to preserve the anonymous property. The user can search any disease available in database.

ACKNOWLEDGEMENTS

I m thankful to mrs.sheetal Mehta for guiding me with this paper including its content and survey regarding the same.

REFERENCES

[1] Tiancheng Li, ninghui Li, jian zhang, "Slicing: A new approach for Privacy Preserving Data Publishing" IEEE Volume 24– No.3, March 2012.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

[2] Ajinkya A.Dhaigude, Preetham Kumar "Improved Slicing Algorithm For Utility In Privacy Preserving Data Publishing" International journal of data Engineering(IJDE) Volume 5 Aug 2014
[3] Francesco Bonchi ,Aristides Gionis, Antti Ukkonen "Overlapping correlation clustering" 11th IEEE International Conference on Data Mining, 2011
[4] Amar paul Singh ,ms.Dhanshri Paritar, "A Review Of Privacy Preserving Data Publishing Technique" International journal Of Emerging Research in Management & Technology volume 2 june 2013
[5] V.V.Nagendra Kumar, C. Lavanya, "Preserve Privacy For Collaborative Data Publishing" International Journal Of Computer Science and Technology volume 5 May 2014
[6] M V R Narasimha rao, J.S.VenuGopalKrishna, R.N.V.Vishnu Murthy, Ch. Raja Ramesh, "Closeness: Privacy Measure For Data Publishing Using Multiple Sensitive Attributes" International Journal Of Engineering science & Advanced Technology Volume 2 Mar-Apr 2012
[7] Slawomir Goryczka, Li Xiong, Benjamin C.M.Fung "m-privacy for Collaborative data publishing" IEEE Jan 2013
[8] Abou-ela Abdou Hussien, Nermin Hamza, Hesham A. Hefny, "Attacks on Anonymization Based Privacy Preserving: A Survey for data mining and data publishing " Journal of Information Security April 2013
[9] Sathish,Silambarashi.g,Saranya.p,Santosh kumar.B, "A New Approach For Collaborative Data Publishing Using Slicing and M-Privacy" International journal of Innovative Research in Computer and Communication Engineering Volume 2 march 2014
[10] Mohamed Nabeel and Elisa Bertino, "Privacy Preserving Delegated Access Control in Publish Clouds" IEEE Volume 26 September 2014
[11] Kishori Pawar, Y. B. Gurav "Overview of Privacy in Horizontally Distributed Databases" International Journal of Innovative Research in Advanced Engineering, may 2014