ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Survey on Security Methods for Personal Credentials in Big Data

Varsha Bhasin [1], Nitin Bhil [2]

P.G. Student, Department of Computer Engineering, AEC Engineering College, Chikhli, MS, India[1]

Asst. Professor, Department of Computer Engineering, AEC Engineering College, Chikhli, MS, India[2]

ABSTRACT: The large amount of data comes from pubs to public agencies, mom and pops to multi nationals; nobody is unrelated to large amount of data. Large organization takes the influential decisions based on analysis of their data. To handle large amount of data, often third party enrolment is present as part of their client projects. The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy preserving data mining (PPDM), has been extensively studied in recent years. There are several methods of PPDM with the check of privacy. The methods are classified on three different approaches. The aim of this paper is to discuss the PPDM method, which is based on Data Modification framework. It provides a wider perspective and investigates various approaches that can help to protect sensitive information.

KEYWORDS: Privacy preserving data mining (PPDM), Data Modification, Perturbation, Anonymization, t-closeness

I.INTRODUCTION

With the rapid use of technology in almost every field, had emerged large amount of data to be effectively handled by the organization. The organization need to release it for various processes, this large amount of data is "Big data". The glossary meaning of big data is "extremely large data sets that may be analyzed computationally to reveal patterns, trends and associations". In simple terms it's a term for data sets that are so large or complex. Traditional data processing applications are inadequate to deal with them. The advantages of big data is that its capability of handling large amount of information with minimum processing time. On the reverse side the processing algorithm put the sensitive information on the edge of exposure. Generating an anonymized data set that is suitable for public release is essentially a matter of finding a good equilibrium between disclosure risk and information loss. Releasing the original data set provides the highest utility to data users but incurs the greatest disclosure risk for the subjects in the data set. On the contrary, releasing random data incurs no risk of disclosure but provides no utility [1]. So the term privacy preserving data mining is used for voluminous data in such manner so that mining algorithms can be applied effectively without compromising security of sensitive and confidential data.

II.LITERATURE SURVEY

Data modification technique can be categorized in two groups, perturbation approach and anonymization approach. In simple words data modification approaches are hiding the important data with aim that private data can't be reached directly or indirectly. In similar terms anonymization technique can be defined as preventing from identifying the important data, thus preserving one's privacy and in perturbation approach a part or whole dataset is modified for creating meaningful data mining models.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

1. Anonymization Techniques

In this technique the sensitive information about the one's record is to be hidden. This technique even assume that to retain the sensitive information for further analaysis.By the above statements, it's obvious that explicit identifiers will be removed, but quasi identifier can be linked with the available data. For example: consider the attributes such as Gender, Pincode and date of birth available in UID database. Such records can also be available in some other dataset where linking can be done with much possibility. Thus in simple terms the quasi identifiers can be easily linked with possible dataset. This approach is classified as follows:

1.1 k-anonymity:

In this technique each record within an anonymized table must be indistinguishable with at least k-1 other record within the dataset, with respect to a set of QI attributes. In particular, a table is K-anonymous if the QI attributes values of each record are identical to those of at least k-1 other records. To achieve the K-anonymity requirement, generalization or suppression could be used [2, 3]. This method simply removes some information in table such as explicit information such as name, telephone number which help to identify the identity of row. After removing the information the rows are divided in certain groups, each row in each group had same values in quasi code attributes. Thus the new release data could stand to attacks like linkage attack, i.e the new release data reduces the risk of identification of individual with the publically available data. The limitation of this technique is, it's very difficult for a DB designer to identify which the attributes common in both dataset and other external table. There are various algorithms available in recent areas to implement k-anonymity. This method ensures that the data transformed is true, but there is information is loss.

1.2 L-diversity Technique

The above technique, k-anonymity protects against identity disclosure but it cannot protect against attribute disclosure, to solve this limitation a new notion of PPDM known as l-diversity. In this technique, l-diversity requires the distribution of sensitive information in each equivalences class having well represented values.

An equivalence class is said to have l-diversity if there are at least "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. [2]In simple terms, well represented is to ensure that there are at least l-distinct values for the sensitive information of the row in the dataset. The advantage of this technique is that it overcomes the shortcoming in the k-anonymity model i.e. if there is homogeneity of sensitive data within the dataset, k-anonymity cannot protect identities to k-individuals level. To solve this concept a new concept, of intra group diversity of sensitive vales is introduced. L-diversity technique solves the homogeneity attack of k-anonymity that emphasizes not only on saving the minimum size of K group but also considers saving the variety of the sensitive attributes of each group. The l-diversity technique had shortcomings like; this technique is insufficient to solve attribute disclosure attack, in simple form similarity attack. If the sensitive information in anonymized group is distinct but semantically same, the challenger can find the important information about the attributes within the dataset. Another shortcoming with this technique is that it effectively assumes all attributes to be definite, the challenger either does or does not learn something sensitive. This is specific with the numerical values in the dataset being close to the value is often good enough.

1.3 T-Closeness Technique

L-diversity disadvantage is that it treats all values in the given tuple in the same regardless of its distribution in the datset.In particular it is neither necessary nor sufficient to prevent attribute disclosure. To overcome, a new technique known as t-closeness is introduced. In this technique a threshold value't' is used, i.e. the distribution of the sensitive data should not be more than the threshold value. T-closeness formalizes the idea of global background knowledge by distributing the sensitive information in any class to the distribution of the attribute in the given table. To calculate the information between sensitive values, Earth mover distance metric is used. Thus we state the t-closeness Principle: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. [3]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Thus by above discussion author can state that t-closeness protects against attribute disclosure of sensitive information, but can't deal with identity disclosure. So for effective privacy features use both t-closeness and k-anonymity techniques at the same time. T-closeness deals with the homogeneity and background knowledge attacks on *k*-anonymity not by guaranteeing that they can never occur, but by guaranteeing that if such attacks can occur, then similar attacks can occur even with a fully-generalized table. Thus t-closeness had advantage of anonymization technique other than generalization of secondary identifier. It simple state that hide the sensitive information, thus removing help to achieve smooth distribution and bring it closer the overall distribution. It also enhance the t-closeness on the concept of l-diversity i.e. treating all values of attributes in similar way, irrespective of the distribution of the attributes in the given dataset. The *t*-closeness approach tends to be more effective than many other privacy-preserving data mining methods for the case of numeric attributes. But the main challenge with this technique is to show the distance criteria reflect between attributes.

III.FEATURES OF DIFFERENT ANONYMIZATION TECHNIQUE

Anonymization based PPDM identify the sensitive information about the owner to be hidden. It shows the advantages of overcome of linking attack and heavy loss of information. The following table shows the comparison of different anonymization technique:

Technique	Features
k-anonymity	• able to preserve privacy • achieve the anonymization • protects against identity disclosure, while don't protect against attribute disclosure. • possible attacks on it are background attack and homogeneity attack.
l-diversity	• better than k-anonymity in preserving in data mining. • maintain the diversity of the sensitive attributes • takes in account the diversity of sensitive values in the group. • possible attack on it is similarity attack.
t-closeness	• allows us to take advantage of anonymization techniques other than generalization of quasi-identifier and suppression of records. • provides a way closer to the overall distribution of the information in dataset • the basic requirement of it is that the sharing of the sensitive information within each dataset be similar to the distribution of the confidential information values in the entire data set. • limits the amount of information that any impostor can obtain about the sensitive information on any specific subject. • the distribution of the sensitive information should be equally distributed within the dataset; this is requirement of t-closeness.

IV.CONCLUSION

Thus data mining is the major part of the today's technology era. To handle large amount of data, different application are available which come under 'Big Data' concept. Under big data concept the important term is privacy, to protect sensitive data. Privacy is important term because people are unease about their sensitive information to be shared. There are various PPDM techniques which can be broadly classified as anonymization and perturbation approach. Thus this paper list different anonmization techniques used in PPDM. From the above discussion it is clear that there is no single technique that is reliable in all aspects of technology. The type of data and the type of working domain affect the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

performing factors of the different techniques. The main aspect of the techniques is to create balance between privacy maintaining and accuracy maintenance, to improve the accuracy feature of PPDM algorithm.

REFERENCES

[1] Jordi Soria-Comas, Josep Domingo-Ferrer," T-Closeness Through Microaggregation: Strict Privacy With Enhanced Utility Preservation", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 11, NOVEMBER 2015
[2] Mohammadreza Keyvanpour, Somayyeh Seifi Moradi," Classification And Evaluation The Privacy Preserving Data Mining Techniques By Using A Data Modification–Based Framework", International Journal On Computer Science And Engineering (IJCSE)
[3] Ninghui Li ,Tiancheng Li, Suresh Venkatasubramanian," T-Closeness: Privacy Beyond K-Anonymity And L-Diversity".
[4] Yong Xu,Xiaolin Qin, Zhongxue Yang,Yitao Yang,Kun Li," A Personalized K-Anonymity Privacy Preserving Method ", Journal Of Information & Computational Science 10: 1 (2013) 139–155
[5] K. Sathiyapriya,Dr. G. Sudha Sadasivam,," A SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING", International Journal Of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, March 2013.
[6] K.Deepika1, P.Andrew2, R.Santhya3, Prof.S.Balamurugan4, S.Charanyaa5," Investigations On Methods Evolved For Protecting Sensitive Data", International Advanced Research Journal In Science, Engineering And Technology Vol. 1, Issue 3, November 2014
[7] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita," A Review On Privacy Preserving Data Mining: Techniques And Research Challenges", International Journal Of Computer Science And Information Technologies, Vol. 5 (2) , 2014, 2310-2315.
[8] Alexandre Evfimievski,Ramakrishnan Srikant, Rakesh Agrawal,Johannes Gehrke," Privacy Preserving Mining Of Association Rules"
[9] Ashwin Machanavajjhala,Johannes Gehrke,Daniel Kifer," ℓ-Diversity: Privacy Beyond K-Anonymity"
[10] Hina Vaghashia,Amit Ganatra,"A Survey: Privacy Preservation Techniques In Data Mining", International Journal Of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015
[11] Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh,Mohammad Abdur Razzaque," A Comprehensive Review On Privacy Preserving Data Mining", Aldeen Et Al. Springerplus (2015) 4:694 DOI 10.1186/S40064-015-1481-X
[12] Mr. Pranav O. Chitnis , Prof. Satish R. Todmal ," INFORMATION PRIVACY & SECURITY IN DATA MINING", International Journal OF Engineering Sciences & Management Research