

A SURVEY ON PRIVACY PRESERVING DATA PUBLISHING

S.Gokila¹, Dr.P.Venkateswari²

¹Computer Science and Engineering, Erode Sengunthar Engineering College,
Anna University Chennai, Tamilnadu

²Computer Science and Engineering, Erode Sengunthar Engineering College,
Anna University Chennai, Tamilnadu

ABSTRACT

Data mining is a computational process of analysing and extracting the data from large useful datasets. In recent years, exchanging and publishing data has been common for their wealth of opportunities. Security, Privacy and data integrity are considered as challenging problems in data mining. Privacy is necessary to protect people's interest in competitive situations. Privacy is an ability to create and maintain different sort of social relationships with people. Privacy Preservation is one of the most important factor for an individual since he should not be embarrassed by an adversary. The Privacy Preservation is an important aspect of data mining to ensure the privacy by various methods. Privacy Preservation is necessary to protect sensitive information associated with individual. This paper provides a survey of key to success and an approach where individual's privacy would be non-distracted.

KEYWORDS

Data mining, Privacy, Privacy Preserving

1. INTRODUCTION

Data mining is a multidisciplinary field, would drawn from areas including database technology, machine learning, information retrieval, knowledge-based system, high performance computing and data visualization. Data mining has attracted a great deal of attention in the industry to maintain information. Due to the wide availability of data and imminent need, data are turned into useful information and knowledge. The information and knowledge gained could be used for various applications ranging from market analysis, fraud detection and customer retention.

Due to the widespread of data, data mining has been viewed as a threat to privacy that are maintained by the industry. Data mining incorporate privacy as a functional component for the gained information and knowledge. Preservation of individual's information is an essential for the data owners to ensure his privacy. Privacy plays an important role in data publishing. Data mining process allows a company to use large amount of data to develop correlations and relationships among the data to improve the business efficiency. Therefore Privacy preserving data mining has become important field of research.

This paper is organised as follows. Section 2, introduce the need for privacy preservation. Section 3, summarize the importance of data collection. Section 4, discuss about various methods for partitioning of data collection for publishing microdata. Section 5, summarize various privacy

preserving approaches for microdata publishing. Section 6, deals with examples of several applications that ensure privacy. Section 7, discuss the possible privacy threats on microdata publishing. Section 8, present the study on various privacy preserving techniques. Section 9, introduce the comparative study on privacy preserving techniques.

2. NEED FOR PRIVACY PRESERVATION

Privacy preservation is considered as an important factor for effective utilization of the massive volume of data. This data have been stored in electronic format, without disturbing an individual. It preserves privacy during data collection and mining.

3. DATA COLLECTION AND MINING: Why and What:

In this information era, collected data and its transactions are recorded somewhere. Many data security-enhanced techniques have been developed to secure the privacy of the individual while collecting and mining the data. There is a demand to exchange and publish data among various parties after collecting it. The data would be collected from different sources and they cannot be shared directly. The owner might collect data and process it through various anonymization techniques in order to mine the data and release it to the recipient. The collected data would be prepared well before publishing in order to ensure confidentiality and privacy.

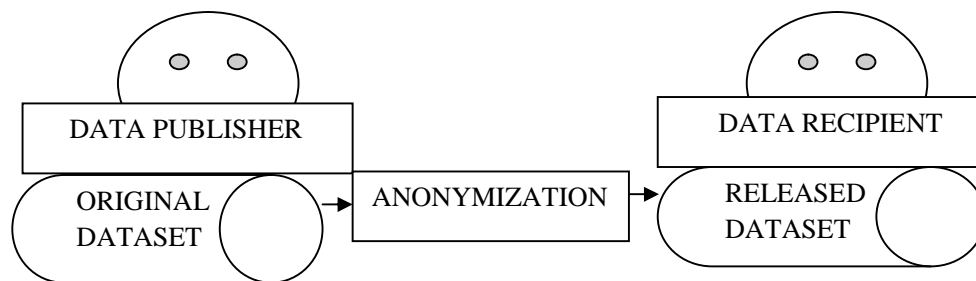


Fig 1: A Simple privacy model

4. MICRODATA PUBLISHING

Privacy preserving data publishing contains microdata. It provides information about an individual. The record holds information in the form of:

Identifier (I), Quasi Identifier (QI), Sensitive Attribute (SA). The Identifiers are explicitly known only to the data holder, Quasi Identifiers which are non sensitive attributes and sensitive Attribute contains individual's specific information that was unknown to an adversary.

5. PRIVACY PRESERVING APPROACHES

According to Sweeny [9], the specific information about a person are stored in the table (or a relation) of columns (or attributes) and rows (or records). Several privacy preserving approaches in data publishing such as randomization, sampling, cell suppression, data swapping and perturbation have been designed for microdata publishing [10]. Privacy preserving crossed over

various stages in its development. Due to the level of complexity in the existing technique, a thought to preserve privacy considered as a new area of research.

5.1 Randomization

Randomization was an ability to anonymize the entire dataset, to preserve certain semantics. In the existing privacy preserving data mining techniques, the randomization is considered as one of the most important technique. This provides knowledge discovery and a balance between privacy and utility [11]. The balance between privacy and utility was achieved by adding noise to the data. The randomized data after balanced would transmit to the recipient. The recipient would receive the data using distribution reconstruction algorithm.

5.2 Suppression

Suppression have not involved in releasing the actual value of the data. Suppression would replace the value of specific attribute description, typically the Quasi Identifiers as the attributes, with less specific description. Like generalization, Suppression would hide some details of Quasi Identifiers [13]. A specific value could be replaced by a generic value at the time of suppression. Suppression indicates that replaced value was not disclosed [13].

6. PRIVACY PRESERVING APPLICATIONS

6.1 Medical Database: Scrub System

As per the author perceptions [16], clinical information would be in the form of text, which contains information of patients like his family members, address, blood group and phone number. Traditional technique have been used only for global search and replace procedure in order to maintain privacy [16]. As per author Sweeny L findings, Scrub system used numerous detection algorithms in order to support the privacy.

6.2 Bioterrorism Application

It is essential to analyse the medical data for privacy preservation in the bioterrorism application. For example, Biological agents are widely found in the natural environment such as anthrax. It is important to find the anthrax attack from the normal attack [17]. It is necessary to track incidences of the common diseases. The corresponding data would be reported to the public health agencies. The respiratory diseases were not reportable-diseases. This provides a solution for more identifiable information in accordance with public health law [17].

7. PRIVACY THREATS

Releasing the result of data mining could cause privacy threats. Several privacy disclosure threats were possible in microdata publishing like identity disclosure, membership disclosure and an attribute disclosure. Privacy threats results in more disclosure risk. Anonymizing the data and preserving the data through various disclosure protections would result in better utility.

7.1 Identity Disclosure

Usually an individual was linked to a record in the published table. If his identity was disclosed, then the corresponding sensitive value of an individual would be revealed

7.2 Attribute Disclosure

Attribute disclosure was possible when information about individual record would be revealed. Before releasing the data, it is the must to infer attributes of an individual with high confidence. As per the authors view [19], matching multiple bucket was important to protect attribute disclosure.

7.3 Membership Disclosure

Membership information in the released table would infer an identity of an individual through various attacks. If the selection criteria were not a sensitive attribute value, then it would lead to have a membership disclosure [18].

8. RELATED WORK

8.1 K-Anonymity

The various phenomena arise when analysing and publishing the data in high-dimensional space. K-Anonymity was a method to support the curse [1]. Generalization on K-Anonymity was applied to mask the exact value of an attribute [1]. The perturbation technique on K-Anonymity was suitable for aggregate distribution of an individual than the inter-attribute relation of an individual. 2-anonymity and Gaussian cluster methods proposed on K-Anonymity technique, ensure privacy by evaluating probability and assigning its value to zero [1]. As per the authors view [1], this method tried to understand the probability distribution which would have maximum likelihood of its attributes. As per the authors [1], there would be a loss for high-dimensional data.

8.2 -Diversity

Information about an individual could not be published without revealing sensitive attribute [2]. K-Anonymization was not enough to protect the data which include homogeneity attack and background knowledge attack [2]. -Diversity technique described that sensitive attributes would have at most same frequency. For example, with positive disclosure, if Alice wants to discover Bob, Alice would determine Bob with high-probability distribution. The negative disclosure would happen when an adversary could correctly eliminate some possible value of the sensitive attributes. There could be a minimum difference between the prior belief and posterior belief [2].

8.3 t-closeness

Anil Prakash, Ravindar Mogili found that K-Anonymity and -Diversity was not used to prevent attribute disclosure [3]. -Diversity would have well represented sensitive attribute value that was assigned only with certain number of limitations [2]. t-closeness has been proposed to describe the distribution of sensitive attribute with equivalence class. Earth Mover Distance was used to measure the distance between the two probabilistic distributions [3]. conjunction has been proposed to combine machine learning and statistical analysis. Closeness among the columns was reduced by aggregation [3].

8.4 K^m Anonymity

K^m Anonymity has been proposed for an anonymize transactional database [4]. K^m Anonymity aim at protect the database against an adversary who has knowledge about almost m items in the transaction [18]. The generalization was used to maintain the set valued data. For any transaction on $K-1$ records, other identical transaction would also appear. K^m anonymity has been introduced via top down local generalization process to record the number of transaction records [4]. The partition based approach was used to group (partition) the similar items in a top-down manner [4]. The k^m anonymity model would help to prevent privacy breaches raised from an adversary who would discovered m items in a transaction database.

8.5 Distributed K-Anonymity framework (DKA)

The collection of data from different sites cannot be shared directly. The key step was to anonymize the data in order to generalise a specific value [5]. A secure 2-party framework was designed for multiparty computation that has been used to join the dataset from various sites [5]. Distributed K-Anonymity (DKA) prevent identification of an individual by make use of global Anonymization in the encrypted form. DKA provide a secure framework between two parties [5]. Two parties would agree on Global Anonymization algorithm that could produce local K-Anonymous dataset. In addition, DKA provide a secure distributed protocol which would require that two parties could mutually semi-honest [5]. Still the trade-off between utility and potential of data was misused in DKA [5].

8.6 K-Anonymity Clustering

Among various clustering methods, hierarichal clustering was mostly used to achieve K-Anonymity [14]. Weighted Feature C-means Clustering [WFC] used to reduce the information distortion. WFC partition all records into equivalence class and would merge the class using class merging mechanism [6]. The numerical values of quasi identifier were used to evaluate the Weighted Feature C-means Clustering technique. The authors also try to provide the dissimilarity evaluated approach which would take different types of feature values for class merging mechanism.

8.7 R-U Confidentiality Map

Normally, Anonymization would cause loss in potential utility gain. The most important factor that should be balanced are Risk(R) and Utility (U). The fundamental characteristic of generalization and bucketization was compared [8]. Privacy trade-off utility have fixed the privacy requirement with various privacy parameters and produced a n anonymized dataset. Generalization would work on three methods such as Apriori Anonymization (AA), the Constrained Based Anonymization Transaction (COAT) and Privacy-Constrained cluster Based Transaction Algorithm (PCTA). These three methods were used to maintain the association between Risk(R)-Utility (U). The trade-off was gained by combining the above three methods as (AA & COAT) and (COAT & PCTA) [8]. It was essential to maintain the histogram H for the occurrence of each sensitive item in the transaction database [8].

As per the authors [8], protection of the data requires:

- (i) Security, to ensure that there was no loss of the stored information.
- (ii) Confidentiality, to ensure that no one can drop data while the data were transmitted between authorized users.

(iii) Anonymity, to ensure private and sensitive information about an individual was not disclosed when a record was released.

8.8 Slicing

K-Anonymity could not guarantee the background knowledge attack of an adversary [1]. The data was partitioned in both horizontal and vertical direction to preserve the privacy by permuting the sensitive attribute [8]. Slicing was a popular data Anonymization technique. Slicing could be formalized by comparing with generalization and bucketization [18]. Generalization and Bucketization were used to replace a specific sensitive attribute value of transaction records [19]. Generalization changed the low conceptual level of data to high conceptual level. Bucketization used l-diversity checking technique for multiple matched buckets. Mondrian algorithm [18] support membership disclosure protection for different privacy threats. Privacy threats leads to identity disclosure, attribute disclosure and membership disclosure. Mondrian algorithm would support map between logical model and a physical model. Slicing support prevention on membership disclosure but still risk and utility was not achieved. Slicing preserve correlation on attributes. A tuple could potentially match with multiple buckets to avoid privacy threats with slicing [18]. Slicing supported data utility than the traditional techniques.

8.9 Overlapped Slicing

The problem was broken down into sub problems, which were reused several times. Overlapped slicing would duplicate an attribute with more than one column. Each column results in more attribute correlations [19]. Horizontal and vertical partition was done by duplicating the attributes in more than one columns. Tuples were grouped together in horizontal partitioning. Attributes were correlated in vertical partitioning. Sensitive attribute would be placed in each column. Random permutation would take place on Sensitive Attribute [SA] column [19]. The authors tried to provide protection against membership disclosure by distinguishing the original tuples from the fake tuples where the disclosure risk has been occurred. The matching strategy in the overlapped slicing table is low. Overlapped slicing enhanced with slicing provided the membership disclosure protection. But, it leads to more attribute correlations and there would be a secrecy loss of privacy in some extent.

9. COMPARATIVE STUDY

Techniques	Dataset	Parameter used	Advantages	Disadvantages
K-Anonymity	Market Basket Dataset	Number of data points, Dimensionality of data space	High correlation among the tuples	More Number of dimensions would be violated
-Diversity	Adult Database	Identifiers, Quasi-identifiers, Sensitive attribute	Sensitive attribute would have at most same frequency	Homogeneity and background knowledge attack has lacked
t-closeness	Pension scheme dataset	Identifiers, Quasi-identifiers, Sensitive attribute	Measure the distance between two probabilistic distribution that were indistinguishable from one another	Information gain was unclear
K ^m Anonymity	Market Basket Dataset	Distinct items, Maximum transaction size and Average transaction size on distinct items	Similar evaluated approach on k items	Loss of utility
WFC	Iris, Wine , Zoo Datasets	Single, Complete and Average link	Partition the records into equivalence classes	Utility was still not achieved.
Distributed K-Anonymity framework (DKA)	Employee Dataset	Public-key, Secret-key, Encryption	Global Anonymization to ensure privacy	Utility and potential were misused
R-U Confidentiality Map	Click Stream data	Maximum transaction size, Average transaction size	Maintain trade-off between privacy and utility	Vulnerable to homogeneity attack
Slicing	Health care Dataset	Identifier, Quasi-Identifier, Sensitive Attribute	Randomization on sensitive attribute	Utility and risk measures not matched
Overlapped Slicing	Health care Dataset	Identifier, Quasi-Identifier, Sensitive Attribute	Duplicate an attribute in more than one columns	Utility was not achieved

10. CONCLUSION

This paper, discussed about various anonymization techniques to preserve the privacy in data mining. Due to the large collection of information, it is important to maintain the Privacy. Number of anonymization techniques has been designed, still it remains an open problem on how to use the anonymization techniques effectively. The extra rows can be introduced to increase the redundancy in order to maintain the privacy. The more effective anonymization technique will be found to preserve the privacy and also ensure the privacy by cross-disciplinary fields.

REFERENCES

- [1] Charu C. Aggarwal, (2005), "On k-Anonymity and the Curse of Dimensionality", Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp.901-909
- [2] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkita Subramanian, (2006), "L-Diversity: Privacy Beyond K-Anonymity", Proc. International conference on Data Engineering (ICDE), pp.24.
- [3] Anil Prakash, Ravindar Mogili, (2012), "Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity", International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARC SEE) Volume 1, Issue 8, pp:28-33
- [4] Yeye He, Jeffery Naughton, F. (2009), "Anonymization of Set Valued Data via Top Down Local Generalization", Proc. International Conference on Very Large Databases (VLDB), pp.934-945.
- [5] Wei Jiang, Chris Clifton, (2006) "A secure distributed framework for achieving k-anonymity", the VLDB Journal, Vol.15, No.4, pp.316-333
- [6] Chuang-Cheng Chiu, ChiehYuan Tsai, (2007), "A k Anonymity Clustering method for Effective Data Privacy Preservation", Springer journal on Verlag Berlin Heidelberg, pp.88-99.
- [7] Grigorios Loukides, Aris Gkoulalas - Divanis, and Jianhua Shao, (2012), "Assessing Disclosure Risk and Data Utility Trade-off in Transaction Data Anonymization", International Journal of Software and Informatics, Vol.6, No. 3, pp.359-361
- [8] Ravindra S, Wanjari Prof .Devi, (2013), "Improving the implementation of new approach for Data Privacy Preserving in Data Mining using slicing". International Journal of Modern Engineering Research (IJMER), Vol. 3, Issue. 3.
- [9] L. Sweeney, (2002) "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570.
- [10] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo, (2009), "A Survey on Privacy Preserving Approaches in Data Publishing" in the First International Workshop on Database Technology and Applications
- [11] Mohnish Patel, Prashant Richariya, Anurag Shrivastava, (2013), "A review paper on Privacy-Preserving Data Mining", Review article on Scholars Journal of Engineering and Technology (SJET), pp.359-361
- [12] Agarwa, Srikan R., (2000) "Privacy Preserving Data Mining", In Proc. ACM SIGMOD, conference on management of data (SIGMOD'00), Dallas, TX, pp.439-450.
- [13] Benjamin c.m, Fung, ke wang, rui chen, philips s.yu, (2010), "Privacy Preserving Data Publishing: A Survey of Recent Development" ACM Computing surveys, Vol.42, No.4, pp.523-553
- [14] Byun, J.W, Kamra, A, Bertino, Li, N, (2007), "Efficient k-Anonymization Using clustering Techniques". International Conference on Database Systems for Advanced Applications, pp.188-200.
- [15] N. Li, T. Li, and S. VenkataSubramanian, (2007), "t-closeness: Privacy beyond k-anonymity and L-diversity," Proc. International Conference on Data Engineering (ICDE), pp.106-115.
- [16] Sweeney L, (1996), "Replacing Personally Identifiable Information in Medical Records, the Scrub System". Journal of the American Medical Informatics Association.
- [17] Charu C. Aggarwal, "A General survey of privacy preserving Data Mining Models and Algorithms", IBM, T. J. Watson Research Centre
- [18] Tiancheng Li, Jian Zhang, Ian Molloy, (2012), "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transaction on KDD.
- [19] B.Vani, D.Jayanthi, (2013), "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT.