ISSN:0975-9646

Privacy Preservation of Published Data Using Anonymization Technique

Isha K.Gayki[1], Arvind S.Kapse[2]

[1]ME (CSE) Scholar, [2]Assitantant Professor
Department of CSE, P R Patil College of Engg. & Tech.,
Amravati-444602, India

***Abstract*—Data Privacy is collection of data and dissemination of data. How we analyze the data is called as Data mining? Need of data privacy arise in different area such as health care, intellectual property, biological data, financial transaction etc. It is very difficult to preserve privacy of the data when there is transfer of data. Sensitive information must be protected. Mainly there are two kinds of major attacks against privacy namely record linkage and attribute linkage attacks. Some methods are proposed namely k-anonymity, ℓ-diversity, t-closeness for data privacy. K-anonymity method preserves the privacy against record linkage attack alone but fails to prevent attribute linkage attack. ℓ-diversity method overcomes the drawback of k-anonymity method but again does not prevent identity disclosure attack and attribute disclosure attack. t-closeness method preserves the privacy against attribute linkage attack but not identity disclosure attack. A proposed method used to preserve the privacy of person sensitive data from record and attribute linkage attacks. In the proposed method, privacy preservation is achieved through generalization by setting range values and through record elimination of duplicate data. A proposed method overcomes the drawback of both record linkage attack and attribute linkage attack.**

***Keywords*— Data privacy, Data mining, Privacy preservation, anonymization**

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful data. Government and private sectors are publishing micro data to facilitate pure research. Attackers links more than two dataset and use their background knowledge for deducing the sensitive information Certain attributes are linked with external knowledge to identify the individual's records indirectly. Such attributes are called Quasi Identifiers attributes. Quasi identifiers are associated with sensitive attribute. Such attributes are known as sensitive attributes which should not be disclosed. Information leakage occurs by combination of quasi identifiers and external knowledge. There are two types of attack namely attribute attack and identity attack.

Microdata contain information about a person, a household or an organization. Agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data is stored in a table, and each record corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories. (1) Attributes that clearly identify individuals. These are known as *explicit identifiers* and include *Social Security Number*, *Address*, and *Name*, and so on. (2) Attributes whose values when taken together can potentially identify an individual. These are known as *quasi-identifiers*, and may include, e.g., *Zip-code*, *Birthdate*, and *Gender*. (3) Attributes that are considered sensitive, such as *Disease* and *Salary*. When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, means the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure.

II. RELATED WORK

Generally when people talk of privacy, they say .keep information about me from being available to others. It is this intrusion, or use of personal data in a way that negatively impacts someone's life, that causes concern. As long as data is not misused, most people do not feel their privacy has been violated. The problem is that once information is released, it may be impossible to prevent misuse. Existing techniques find solution for privacy problem to some extent. k-anonymity[7] can prevent the identity disclosure attack but not attribute disclosure attack. Another method, ℓ-diversity[9] method preserves the privacy against attribute disclosure attack. But, not identity disclosure attack. t-closeness method[9] is good at attribute disclosure attack. It is computationally complex. but, it fail to protect the privacy against attribute disclosure attack

P sensitive k-anonymity model[7], the micro data table T* satisfies (p+, α)-sensitive k-anonymity property if it satisfies k-anonymity.

Tamir Tassa [2] proposed an alternative model of k-type anonymity. It reduces the information loss than k-anonymity and obtained anonymized table by less generalization. It preserves the privacy against identity disclosure alone. Qian Wang[4] proposed the model k-anonymity in protection of attribute disclosure. It can prevent attribute disclosure by controlling average leakage probability and probability difference of sensitive attribute value Mahesh, Meyyappan[11] proposed a new method to anonymize the dataset by setting range values in Quasi identifiers. If the Quasi identifier consists of same attribute values in any class.

In t-closeness method[9], 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. It preserves the privacy against homogeneity and background knowledge attacks.

III. PROPOSED METHOD

The given method provides new anonymization technique comprising record elimination and generalization

A. Definition

The Let T(A1,A2,A3⋯An) be a microdata table. A1,A2,A3⋯An are set of attribute A in table T and tuple $t \subseteq T$. Each tuple is represent the information of individual. There are two types of attributes such as Quasi Identifier Attributes and Sensitive attributes. Let Q be quasi identifier attributes, $Q \subseteq A$ and sensitive attributes $S \subseteq A$. Quasi Identifiers could be known to attacker. attacker find the individual sensitive information through Quasi Identifiers. Hence Sensitive attributes should be protected. Let QT be Quasi Identifier of table T where

$$\{ Q_1....Q_j \} \subseteq \{A_1 ....A_n \}$$

TABLE I

ORIGINAL DATASET

Sno	Name	Zipcode	Age	Sex	Diseases
1	Andy	47677	29	Male	Gastric Ulcer
2	Bill	47602	28	Male	Gastric
3	Ken	47678	29	Male	Gastric Ulcer
4	Nash	47905	36	Male	Gastric
5	Joe	47909	52	Female	Flu
6	Sam	47906	36	Male	Bronchitis
7	Linda	47605	30	Female	Bronchitis
8	Jame	47673	36	Male	Pneumonia
9	Sarah	47607	32	Female	Bronchitis
10	Raj	47609	24	Male	Flu

B. Suppression

Each class should contain atleast a record which has unique sensitive attribute value and unique quasi identifier value. Each class may contain variable number of records. Suppression technique[13] is applied over selected Quasi identifier QT{zipcode} until the stated precondition is satisfied. Table shows each class with atleast one unique quasi identifier value QT{age, disease}in k records. Record elimination operation checks for identical attribute values in each class

TABLE II

Sno	Zipcode	Age	Sex	Diseases	Group
1	4760*	28	Male	Gastric	C1
2	4760*	30	Female	Bronchitis	C1
3	4760*	32	Female	Bronchitis	C1
4	4797*	29	Male	Gastric Ulcer	C2
5	4797*	29	Male	Gastric Ulcer	C2
6	4790*	36	Male	Pneumonia	C2
7	4790*	36	Male	Gastric	C3
8	4790*	52	Female	Flu	C3
9	4790*	36	Male	Bronchitis	C3
10	4790*	24	Male	Flu	C3

C. Record Elimination

Record elimination operation checks for identical attribute values in each class it will check for the duplicate attribute in same cluster of class ,if duplicate value is present in next cluster it will not eliminate that. In same cluster it will keep first one as it is and remove the next one within same tuple.

TABLE III

Sno	Zipcode	Age	Sex	Diseases	Group
1	4760*	28	Male	Gastric	C1
2	4760*	30	Female	Bronchitis	C1
3	4760*	32	Female	Bronchitis	C1
4	4797*	29	Male	Gastric Ulcer	C2
5	4790*	36	Male	Pneumonia	C2
6	4790*	36	Male	Gastric	C3
7	4790*	52	Female	Flu	C3
8	4790*	36	Male	Bronchitis	C3
9	4790*	24	Male	Flu	C3

D. Generalization

Generalization process follows the elimination process, In each class Ci, the next minimum integer value Li, and next largest integer value Mi are found for each attribute value. Attribute values of quasi identifier Qi in class Ci are rewritten as a range value Li <= Mi. This process is repeated until all theQi values in each class Ci are suppressed.

TABLE IV

Sno	Zipcode	Age	Sex	Diseases	Group
1	4760*	28<=32	Male	Gastric	C1
2	4760*	28<=32	Female	Bronchitis	C1
3	4760*	28<=32	Female	Bronchitis	C1
4	4797*	29<=36	Male	Gastric Ulcer	C2
5	4790*	29<=36	Male	Pneumonia	C2
6	4790*	24<=52	Male	Gastric	C3
7	4790*	24<=52	Female	Flu	C3
8	4790*	24<=52	Male	Bronchitis	C3
9	4790*	24<=52	Male	Flu	C3

IV. Result and discussion

Proposed method is applied over the published data in table The performance of the proposed method is evaluated in terms of information loss .The proposed method and existing method namely k-anonymity (k=3),experimented with the same data set and their performance were compared in terms of information loss The following formulae are used to measure information loss ILoss[9]

Let v be value in the domain of attribute A,we use ILvalue (V*) to capture the amount of information loss in generating v to v* which is partition in the corresponding general domain of A

$$\text{ILvalue}(v*) = \frac{(\text{The number of values in } v*)}{(\text{The no of values in domain A})}$$

$$\text{Total information loss ILtable}(T*) = \sum_{\forall t* \in T*} \text{ILtuple}(t*)$$

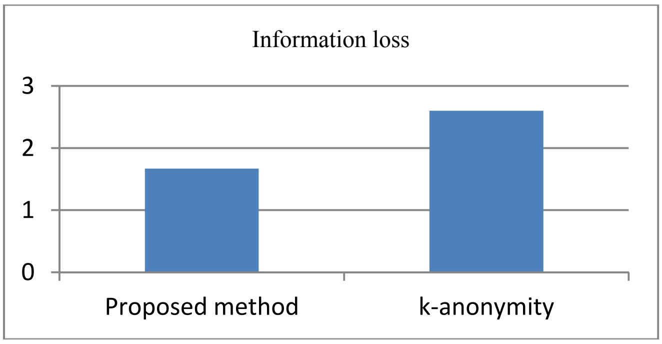

Fig 1.Comparision for information loss

TABLE V

Methods	Information loss
k-Anonymity	2.6
Proposed method	1.67

V. Conclusion

In this information age, data published in web pages are growing enormously every year. While utilizing the data for research purpose, privacy of the individuals whose data are published should not be vulnerable to adversary attacks. In contrast to cryptographic methods which transform the plain text to cipher text, privacy methods protect the privacy of owners whose data are published on web pages. The proposed method preserves the privacy of published data against attribute and identity disclosure attacks. The proposed method is developed only for quasi identifiers with numeric values. Further research is in progress to include non-numeric quasi identifiers as well.

References

[1] Mahesh R, Meyyappan T, “Anonymization technique through record elimination to Preserve Privacy of Published data”, International workshop on pattern recognization, Informatics and mobile engineering, proceedings,,978-1-4673-5845-3,2013

[2] Tamir Tassa,Arnon Mazza and Aristides Gionis,“k-Concealment: An Alternative Model of k-Type Anonymity”, TRANSACTIONS ON DATA PRIVACY 5,2012, pp189–222

[3] Xin Jin,Mingyang Zhang,Nan Zhang and Gautam Das, “Versatile Publishing For Privacy Preservation”,2010,KDD10,ACM

[4] Qiang Wang,Zhiwei Xu and Shengzhi Qu,“An Enhanced K-Anonymity Model against Homogeneity Attack”, Journal of software,2011, Vol. 6, No.10, October 2011;1945-1952

[5] Benjamin C.M.Fung,KE Wang,Ada Wai-Chee Fu and Philip S. Yu, Introduction to Privacy-Preserving Data Publishing Concepts and techniques,ISBN:978-1-4200-9148-9,2010

[6] Raymond Wong, Jiuyong Li,Ada Fu and Ke wang, “(α,k)-anonymous data publishing”, Journal Intelligent Information System, 2009,pp209- 234.

[7] Xiaoxun Sun, Hua Wang, Jiuyong Li and Traian Marius Truta, “Enhanced P-Sensitive K-Anonymity Models for privacy Preserving Data Publishing”, Transactions On Data Privacy, 2008,pp53-66

[8] B.C.M. Fung, Ke Wang and P.S.Yu, “Anonymizing classification data for privacy preservation”, IEEE Transactions on Knowledge and Data Engineering(TKDE), 2007,pp711-725

[9] Ninghui Li, Tiancheng Li, Suresh Vengakatasubramaniam,“t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity”, International Conference on Data Engineering, 2007, pp106-115

[10] X. Xiao and Y. Tao,“Personalized privacy preservation”, In Proceedings of ACM Conference on Management of Data (SIGMOD’06”),2006,pp229-240

[11] Mahesh R, Meyyappan T, “A New Method for Preserving Privacy in Data Publishing”,International workshop on cryptography and Information Security, CS&IT proceedings,2012,pp 261-266

[12 Neha V. Mogre, Girish Agarwal, Pragati Patil:“A Review On Data Anonymization Technique For Data Publishing” *Proc. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181*

[13] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, “Anonymizing Transaction Databases for Publication,” *Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008*

[14] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and „-Diversity,” *Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), pp. 106-115, 2007.*

[15] T. M. Truta and V. Bindu,“Privacy Protection: “p-sensitive k-anonymity property”, International Workshop of Privacy Data Management (PDM2006), In Conjunction with 22th International Conference of Data Engineering (ICDE),2006,pp94

[16] X. Xiao and Y. Tao,“Personalized privacy preservation”, In Proceedingsof ACM Conference on Management of Data (SIGMOD’06”),2006,pp229-240

[17] L. Sweeney,“An Achieving k-Anonymity Privacy Protection Using Generalization and Suppression”, International Journal of Uncertainty, Fuzziness and Knowledge-Based System,2002,pp571-588