ISSN:0975-9646

Privacy Preservation of Published Data through Record Elimination by Anonymizied Approach

Sakshi Agrawal

Computer Science And Engineering Department, RTMNU University
Shrinathji Traders,Balaji Plots,,Khamgaon,Dist-Buldhana

Abstract— **Data mining is the process of analyzing data. Data Privacy is collection of data and dissemination of data. Privacy issues arise in different area such as health care, intellectual property, biological data, financial transaction etc. It is very difficult to protect the data when there is transfer of data. Sensitive information must be protected. There are two kinds of major attacks against privacy namely record linkage and attribute linkage attacks. Research have proposed some methods namely k-anonymity, ℓ-diversity, t-closeness for data privacy. K-anonymity method preserves the privacy against record linkage attack alone. It is unable to prevent address attribute linkage attack. ℓ-diversity method overcomes the drawback of k-anonymity method. But it fails to 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 individuals' 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. A proposed method overcomes the drawback of both record linkage attack and attribute linkage attack.**

Keywords— Anonymization , data privacy, data publishing, data mining, privacy preservation

I. Introduction

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Privacy-preserving data publishing (PPDP) aims to publish a microdata table for research and statistical analysis, without disclosing sensitive information at the individual level[3]. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies. Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company.

II. Literature Review

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 modified micro data table T* satisfies (p+, α)-sensitive k-anonymity property if it satisfies k-anonymity, and each QI group has at least p distinct categories of the sensitive attribute and its total weight is at least α. This method significantly reduces the possibility of Similarity Attack and incurs less distortion ratio compared to p-sensitive k-anonymity method.Tamir Tassa [2] proposed an alternative model of k-type anonymity. It is reduce 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. In (α,k)-Anonymity [6] model, a view of the table is said to be an (α, k)-anonymization, if the modification of the table satisfies both k-anonymity and α-deassociation properties with respect to the quasi-identifier. It does not address the identity disclosure attack.

III. Indentations And Equations

We use $IL_{value}(v^*)$ to capture the (amount of) information loss in generalizing v to v^*

$$(1)\ IL_{value}(v^*) = \frac{(\text{the number of values in } v^*) - 1}{\text{the number of values in the domain of A}}$$

For instance, if the domain of Age is [1,60], generalizing age 5 to [1,10] has information loss IL_value([1,10]) = (10- 1)/60.

The total information loss IL $_{table}(T^*)$ of the entire (generalized) relation T^* is given by

(2) $IL_{table}(T^*) = \sum IL_{tuple}(t^*)$
$\forall t \in T^*$ `

We calculate the privacy gain by

(3) $ILoss(T) = \sum ILoss(r)$
$r \in T$

$$PG = avg\{A(QID_j)-As(QID_j)\}$$

Where,$A(QID_j)$ and $As(QID_j)$ denote the anonymity of QID_j before and after specialization.

A. *Figures and Tables*

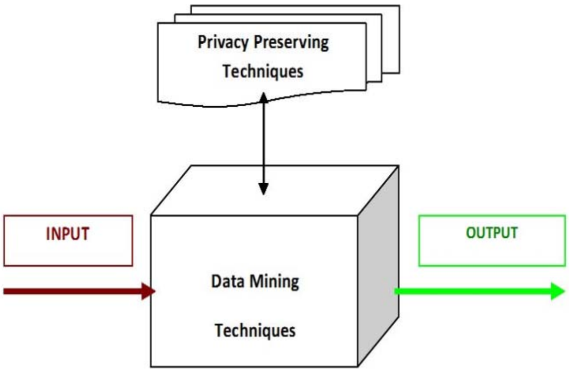

Fig. 1 Flow of the entire data mining process

TABLE I Original View of Module

Sr No.	Name	Zipcode	Age	Sex	Disease
1	Smitha	47677	29	Male	Gastric Ulcer
2	Neeta	47602	28	Male	Gastritis
3	Sunita	47678	29	Male	Gastric Ulcer
4	Amit	47905	36	Male	Gastritis
5	Vaibhav	47909	52	Female	Flu
6	Gaurav	47906	36	Male	Bronchitis
7	Sneha	47605	30	Female	Bronchitis
8	Nimish	47673	36	Male	Pneumonia
9	Pranav	47607	32	Female	Bronchitis

TABLE II Supression view of module

Sr No.	Zipcode	Age	Sex	Disease
1	4760*	28	Male	Gastritis
2	4760*	30	Female	Bronchitis
3	4760*	32	Female	Bronchitis
4	4767*	29	Male	Gastric Ulcer
5	4767*	29	Male	Gastric Ulcer
6	4767*	36	Male	Pneumonia
7	4790*	36	Male	Gastritis
8	4790*	52	Female	Flu
9	4790*	36	Male	Bronchitis

TABLE III Record Elimination view of module

Sr No.	Zipcode	Age	Sex	Disease	Group
1	4760*	28	Male	Gastritis	C1
2	4760*	30	Female	Bronchitis	C1
3	4760*	32	Female	Bronchitis	C1
4	4767*	29	Male	Gastric Ulcer	C2
5	4767*	36	Male	Pneumonia	C2
6	4790*	36	Male	Gastritis	C3
7	4790*	52	Female	Flu	C3
8	4790*	36	Male	Bronchitis	C3

TABLE IV Generalization view of module

Sr No.	Zipcode	Age	Sex	Disease	Group
1	4760*	28<=32	Male	Gastritis	C1
2	4760*	28<=32	Female	Bronchitis	C1
3	4760*	28<=32	Female	Bronchitis	C1
4	4767*	29<=36	Male	Gastric Ulcer	C2
5	4767*	29<=36	Male	Pneumonia	C2
6	4790*	36<=52	Male	Gastritis	C3
7	4790*	36<=52	Female	Flu	C3
8	4790*	36<=52	Male	Bronchitis	C3

IV. CONCLUSIONS

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 challenged. The proposed method attempts at static micro data only which contain numeric quasi identifiers.The Proposed method also attempts to reduce information loss and maximize privacy gain.

ACKNOWLEDGMENT

I would like to thank my professors for their valuable contribution in completing my work.I would also express my deepest appreciation to all my teachers for providing guidance and supervision to my work.

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] TamirTassa,ArnonMazza and Aristides Gionis,"k-Concealment: An Alternative Model of k-Type Anonymity", *transactions on data privacy* 5, 2012, pp189–222

[3] XinJin,Mingyang,Zhang,Nan Zhang and Gautam Das, "Versatile Publishing For Privacy Preservation",2010,*KDD10,ACM*

[4] QiangWang,ZhiweiXu and ShengzhiQu,"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,KEWang,AdaWai-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, JiuyongLi,Ada Fu and Kewang, "(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*