IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VII (May-Jun. 2014), PP 102-105
www.iosrjournals.org

Anonymizied Approach to Preserve Privacy of Published Data Through Record Elimination

Sakshi Agrawal
(Computer Science and Engineering, RTMNU University, India)

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 (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. 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 microdatatable for research and statistical analysis, without disclosing sensitive information at the individual level[3].

In today's global network of organizational connections, there is a growingdemand to disseminate and share this information due to various academic, commercial land other benefits. As the records of data frequently include sensitive information thatcould violate the privacy of the corresponding individuals, it is necessary to preprocessthe data prior to its publication in order to limit the disclosure of sensitive data[2]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. 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.

Government and private sectors are publishing micro data to facilitate pure research. Individuals'privacy should be safeguarded. Published data contains sensitive values of record owners. Typically, such information is stored in table format(T). Adversaries (attackers) link more than two dataset and use their background knowledge for deducing the sensitive information.The main objective is to generalize the database before sending it to the any research purpose. It should preserve the privacy of every individual. At the same time information loss should be as low as possible.Certain attributes are linked with external knowledge to identify the individual's records indirectly. Such attributes are calledQuasi Identifiers (QI). Quasi identifiers areassociated with sensitive attribute(S). Such attributes are known as sensitive attributes which should not be discosed.

II. Indentations And Equations

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

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

For instance,if the domainof 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) $IL_{table}(T^*) = \sum_{Vt \, € T^*} IL_{tuple}(t^*)$ `

We calculate the privacy gain by

(3) $ILoss(T) = \sum_{r \, € \, T} ILoss(r)$

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

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

III. Figures And Tables

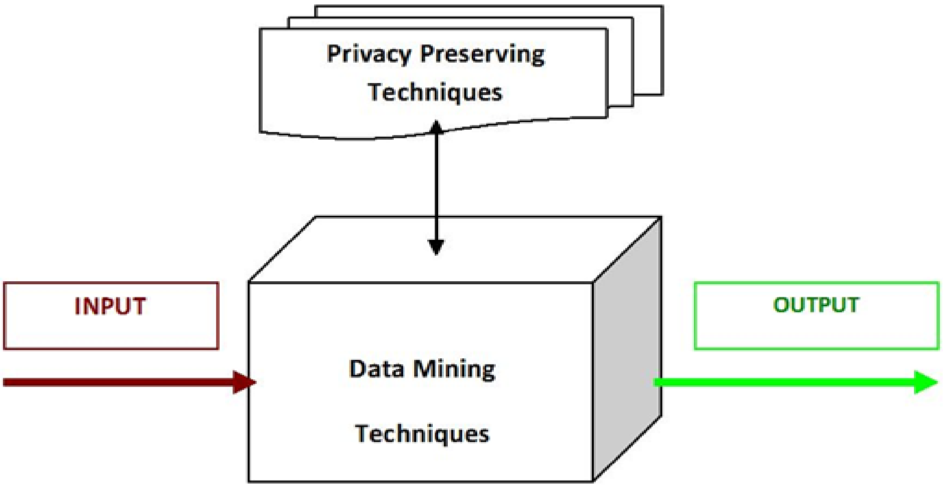
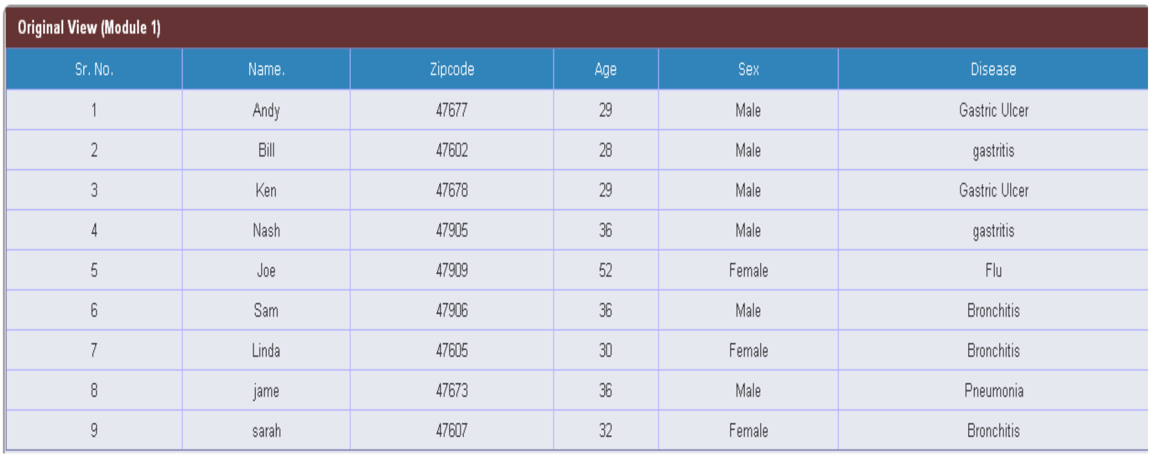


Fig 1: Flow of the process

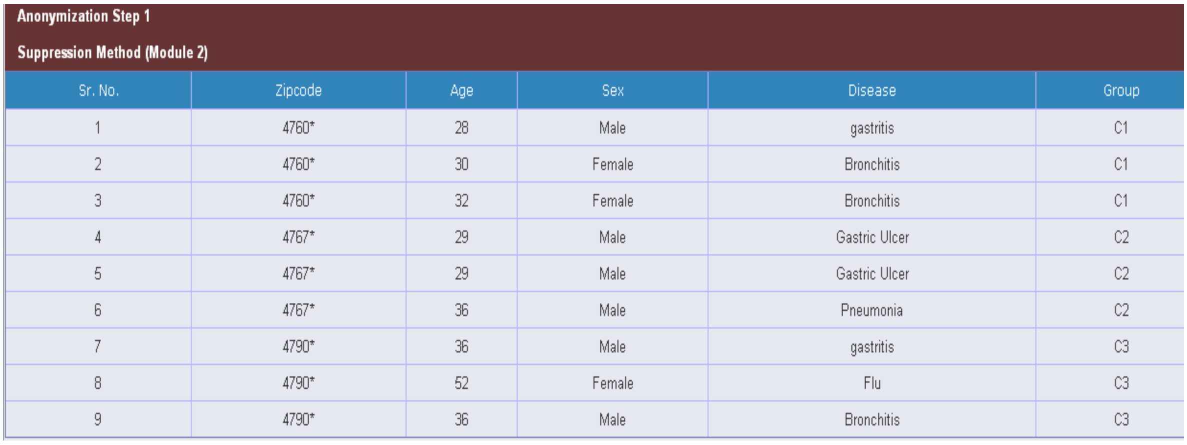
TABLE 1: Original view



Original View (Module 1)

Sr. No.	Name.	Zipcode	Age	Sex	Disease
1	Andy	47677	29	Male	Gastric Ulcer
2	Bill	47602	28	Male	gastritis
3	Ken	47678	29	Male	Gastric Ulcer
4	Nash	47905	36	Male	gastritis
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

TABLE 2: Supression Method

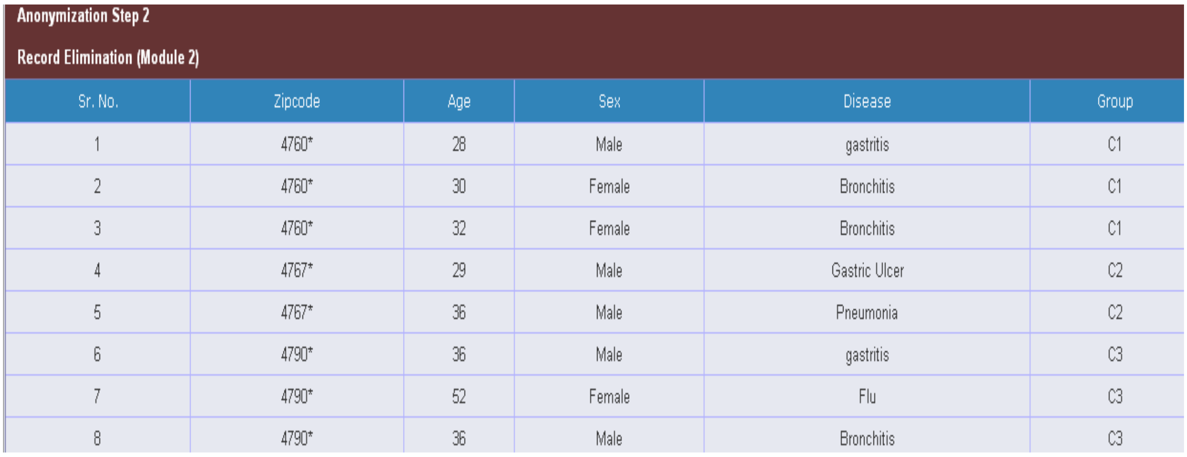


Anonymization Step 1

Suppression Method (Module 2)

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*	29	Male	Gastric Ulcer	C2
6	4767*	36	Male	Pneumonia	C2
7	4790*	36	Male	gastritis	C3
8	4790*	52	Female	Flu	C3
9	4790*	36	Male	Bronchitis	C3

TABLE 3: Record Elimination Method

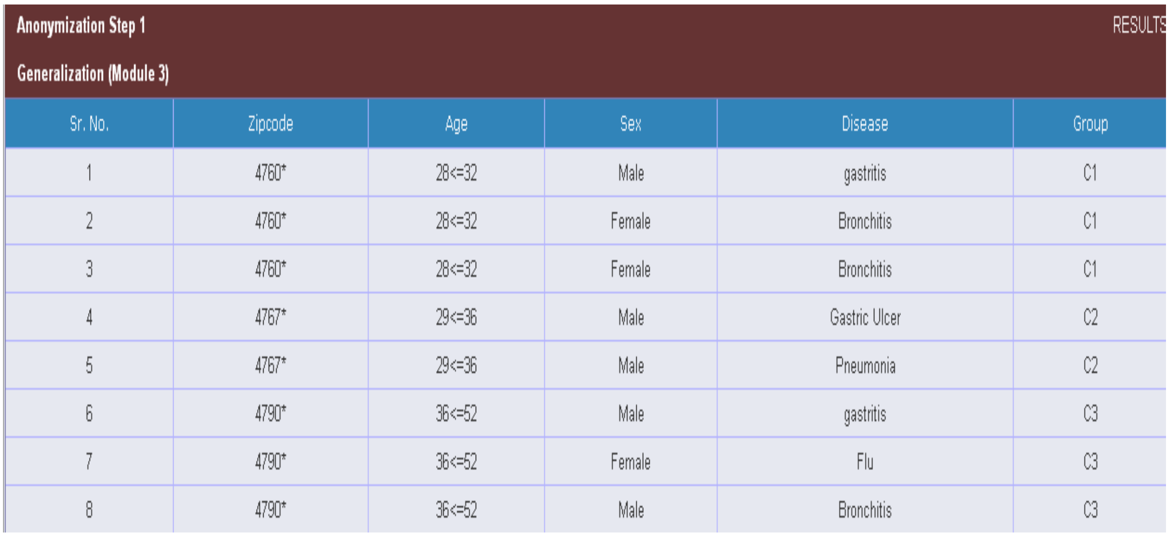


Anonymization Step 2

Record Elimination (Module 2)

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 4: Generalization Method



Anonymization Step 1

RESULTS

Generalization (Module 3)

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. 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 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.

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,MingyangZhang,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