A New Method for Preserving Privacy in Data Publishing

R.Mahesh[1] and Dr.T.Meyyappan[2]

[1]Department of Computer Science and Engineering, Alagappa University, Karaikudi
aummahesh@gmail.com
[2]Department of Computer Science and Engineering, Alagappa University, Karaikudi
meylotus@yahoo.com

ABSTRACT

Protection of individuals' privacy is a vital activity in data publishing. Government and public sector websites publish enormous amount of data for sharing the data among their departments and also to public for research. Sensitive information of individuals, whose data are published must be protected. Privacy is challenged through two kinds of attack namely attribute disclosure and identity disclosure. Early Research contributions were made in this direction and new methods namely k-anonymity, ℓ-diversity, t-closeness are evolved. K-anonymity method preserves the privacy against identity disclosure attack alone. It fails to address attribute disclosure attack. ℓ-diversity method overcomes the drawback of k-anonymity method. But it fails to address identity disclosure attack and attribute disclosure attack in some exceptional cases. t-closeness method is good at attribute disclosure attack. but not identity disclosure attack. Also, t-closeness method is more complex than other methods. In this paper, the authors propose a new method to preserve the privacy of individuals' sensitive data from attribute and identity disclosure attacks. In the proposed method, privacy preservation is achieved through generalization of quasi identifier by setting range values.The proposed method is implemented and tested with various data sets. The proposed method is found to preserve the privacy of published data against attribute and identity disclosure attacks.

Keywords

Data Privacy, generalization, anonymization, suppression, privacy preservation, data publishing.

1. INTRODUCTION

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 stored in table format (T). Adversaries (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(QI). Quasi identifiers are associated with sensitive attribute(S). Such attributes are known as sensitive attributes which should not be disclosed. Information leakage occurs by coordination of quasi identifiers and external knowledge. There are two types of disclosure namely attribute disclosure and identity disclosure. Anonymization techniques [4] are used to convert the micro data table T to T*. Generalizations, suppression, data

Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, CRYPSIS, ICAIT, ICDIP, ITCSE, CS & IT 07,
pp. 261–266, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2423

swapping are common operations of anonymity. In this paper, a new anonymization based method is proposed to preserve the privacy against identity and attribute disclosure attacks.

2. EXISTING METHODS

Existing methods find solution for privacy problem to some extent. *k*-anonymity[14] 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, it is weaker in case of identity disclosure attack. *t*-closesness method[8] is good at attribute disclosure attack. It is computationally complex in achieving the privacy. Moreover, it fails to protect the privacy against attribute disclosure attack.Sweeny introduced *k*-anonymity [14] as a property that each record is indistinguishable with atleast *k*-1 records. In this method, privacy cannot be achieved if sensitive value has same value in equivalence class. This method fails to preserve the privacy against background knowledge and homogeneity attacks. The modified micro data table T* satisfies (p, α)-sensitive *k*-anonymity[6] property if it satisfies *k*-anonymity, and each QI-group has at least p distinct sensitive attribute values with its total weight at least α . p-sensitive *k*-anonymity[10] is insufficient to prevent Similarity Attack. In Enhanced P sensitive *k*-anonymity model[6], 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.

ℓ-diversity method [9] overcomes the drawbacks of *k*- anonymity. In this method, an equivalence class is said to have ℓ-diversity if there are at least ℓ "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. This method fails to preserve the privacy against skewness and similarity attacks.In (α,*k*)-Anonymity [5] 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. In *t*-closeness method[8], 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.Tamir Tassa [1] 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. Versatile publishing method[3] preserves the privacy by splitting the anonymized table T* by framing the privacy rules. Privacy can be breached by applying the conditional probability in published table.

Qian Wang[2]proposed the model to make up the shortage of k-anonymity in protection of attribute disclosure. It can prevent attribute disclosure by controlling average leakage probability and probability difference of sensitive attribute value.

3. PROPOSED METHOD

Existing Anonymization techniques preserve the individual privacy against either identity or attribute disclosure attack, but not both. The method proposed in this paper attempts to overcome the disadvantages of existing anonymization techniques. It adopts generalization operation of anonymization technique and preserves the privacy in a new way.

A table T of published data contains quasi identifiers Q_i(i=1,2…n) and sensitive attribute S. Table T is first suppressed on selected quasi identifier Qi and then generalized by following a new procedure proposed in this paper.

The proposed new method described below performs generalization operation and creates the anonymized table T. Suppression technique [12] [13] is applied over selected Quasi identifier Q_i having more frequency. After the suppression process, the tuples in the table T are arranged in n groups $G_1,G_2,G_3 \ldots G_n$ ordered by suppressed value of Quasi identifier attribute B_i (i=1, 2, … m). Among the Quasi identifiers Q_i, one with more distinct values is selected. In each group G_i, the next nearest integer value L_i, and next largest integer value M_i are found. Attribute values of quasi identifier Q_i in group G_i is rewritten as a range value Li <= M_i. This process is repeated until all the Qi values in each group G_i are suppressed.

An unanonymized database table T can be generalized on quasi identifier to maintain privacy of a sensitive attribute S. A new method given below performs generalization operation that converts the table T to T*. Dataset in a table T with k number of tuples, n number of Quasi identifiers and sensitive attribute S are chosen.

Input: Table T with k tuples containing Quasi identifier Q and Sensitive attribute S.

Step 1:

Arrange the tuples in the table T into n groups $G_1,G_2,G_3 \ldots G_n$ by value s_i of Quasi Identifier B_i. where i = 1, 2, … k

Step3:

Repeat steps 4 to 6 varying j from 1 to k

Step 4:

Let $L_j = q_j$.
Find the next nearest integer L_j less than q_j in group G_i where i=1,2,3,…n and if found, Let Lj = q_j

Step5:

Let $M_j = q_j$.
Find the next nearest integer M_j greater than q_i in group G_i where i=1,2,3,…n and if found, Let $L_j = q_j$

Step6:

If L_j and M_j are found in the same group G_m, the generalization condition is set as
set $q_j = L_j <= M_j$

Output:

Generalized table T^*

3.1 Computational Procedure for anonymizing table T*

```
Function Anonymize(T)
        Array q,w;
        Int, nextmin,nextmax,i;
        String sv1,sv2;
        T=funcgroup(T[B])
        T=funcsort(T[Q],G);
        q=T[Q];
        w=T[S];
        While (u=0 to G.count)
          While(i=0 to T.rowcount in u )
                Nextmin=if(Findnextmin(q[i],T[Q]))
                If(Nextmin==0)
                        Nextmin=q[i];
                end if
                Next Max=FindNextmax(q[i].T[Q])
```

```
                q[i]=Nextmin+"<="+Nextmax
        End while
End While
        T*=Arrange(T[Q],q)
End function
```

Sub procedure funcgroup() finds number of groups in the given data set. Sub procedure funcsort() sorts the quasi identifier values within each group.

4. RESULTS AND DISCUSSION

Proposed method is applied over the published data in table T as shown below. There are three Quasi identifiers Zipcode, Age, Salary and one sensitive attribute Disease present in original table I. Quasi identifier attribute age has unique values.

Table: 1 Original Table

Sno	Zipcode	Age	Salary	Disease
1	47677	29	3000	Gastric ulcer
2	47602	22	4000	Gastritis
3	47678	27	5000	stomach cancer
4	47905	43	6000	Gastritis
5	47909	52	11000	Flu
6	47906	47	8000	Bronchitis
7	47605	30	7000	Bronchitis
8	47673	36	9000	Pneumonia
9	47607	32	10000	stomach cancer

As a first step, suppression is applied over Zipcode attribute to transform the dataset into Table1

Table: 2 Transformed from Table 1

Sno	Zipcode	Age	Salary	Disease	
1	4760*	22	4000	Gastritis	- G_1
2	4760*	30	7000	Bronchitis	
3	4760*	32	10000	stomach cancer	
4	4767*	27	5000	Stomach cancer	- G_2
5	4767*	29	3000	Gastric ulcer	
6	4767*	36	9000	Pneumonia	
7	4790*	43	6000	Gastritis	- G_3
8	4790*	47	8000	Bronchitis	
9	4790*	52	11000	Flu	

The proposed method is applied over Table II to get the data set transformed to T*

Table: T* Anonymized table

Sno	Zipcode	Age	Salary	Disease	
1	4760*	22<=30	4000	Gastritis	- G_1
2	4760*	22<=32	7000	Bronchitis	
3	4760*	30<=32	10000	Stomach cancer	
4	4767*	27<=29	5000	Stomach Cancer	- G_2
5	4767*	27<=36	3000	Gastric Ulcer	
6	4767*	29<=36	9000	Pneumonia	
7	4790*	43<=47	6000	Gastritis	- G_3
8	4790*	43<=52	8000	Bronchitis	
9	4790*	47<=52	11000	Flu	

It is observed that the proposed method preserves the privacy and reduced the information loss, as shown in table T*.

5. 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 ciphertext, privacy methods protect the privacy of owners whose data are published on web pages. New method proposed in this paper is implemented with MatLab coding and tested with various data sets. 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] Tamir Tassa,ArnonMazza and Aristides Gionis,(2012), "k-*Concealment: An Alternative Model of k-Type Anonymity*", *TRANSACTIONS ON DATA PRIVACY* 5, pp189–222

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

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

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

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

[6] Xiaoxun Sun, Hua Wang, Jiuyong Li and Traian Marius Truta,(2008) "Enhanced P-Sensitive K-Anonymity Models for privacy Preserving Data Publishing", *Transactions On Data Privacy*, pp53-66

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

[8] Ninghui Li, Tiancheng Li, Suresh Vengakatasubramaniam,(2007) "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity*", International Conference on Data Engineering*, pp106-115

[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,(2006) "ℓ-diversity: Privacy beyond k-anonymity*", In Proc. 22nd Intnl international Conference on data engineering. (ICDE)*, pp24

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

[11] X. Xiao and Y. Tao,(2006) "Personalized privacy preservation", *In Proceedings of ACM Conference on Management of Data (SIGMOD'06")*, pp229-240

[12] B.C.M. Fung, Ke Wang and P.S.Yu ,(2005) ," Top-down specialization for information and privacy preservation", *In Proc. of the 21st IEEE international Conference on data engineering(ICDE)*,pp205-216

[13] L. Sweeney,(2002) "An Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*,pp571-588

[14] L. Sweeney,(2002) "k-anonymity: a model for protecting privacy *International Journal on Uncertainty,Fuzziness and Knowledge-based Systems*, pp557-570.

Authors

Mahesh.R M.C.A, Currently Ph.D Research Scholar, Department of Computer Science and Engineering, Alagappa University,Karaikudi, Tamilnadu,India, His research work is in Data mining and publishing, web mining

Dr. T. Meyyappan M.Sc., M.Phil., M.B.A., Ph.D., currently, Professor, Department of Computer Science and Engineering, Alagappa University, Karaikudi, TamilNadu, India. He has obtained his Ph.D. in Computer Science and Engineering in January 2011 and published a number of research papers in National, International journals and conferences. He has been honored with Best Citizens of India Award 2012 by International Publishing House, New Delhi. He has developed Software packages for Examination, Admission Processing and official Website of Alagappa University. His research areas include Operational Research, Digital Image Processing, Fault Tolerant computing, Network security and Data Mining.