International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(2), pp.228-232
DOI: http://dx.doi.org/10.21172/1.102.36
e-ISSN:2278-621X

AN ADEQUATE AMALGAMATED APPROACH FOR ANONYMIZATION

Deepak Narula[1], Pardeep Kumar[2], Shuchita Upadhyaya[3]

Abstract: Digital era empowered with data sharing for the purpose of research and increasing business prospective. However the collected data some time contains sensitive information that need not to be disclosed publically as if that data is available publically it can be a threat to the privacy. So, data privacy is the most crucial act in data publishing. Various methods for anonymization have been suggested in literature. Out of these methods k-anonymization is one of the fundamental and most popular approach but suffering from the shortcoming of homogeneity and background attack. This paper is an attempt to propose an amalgamated approach with less information loss, discernibility cost and the value of average equivalence class size. Moreover, this also increases the data usability and also supports the privacy of sensitive information. The new proposed method can be practically implemented and works even if the domain set of sensitive information is small.
Keywords: Privacy Preserving Data Publishing (PPDP) , Fuzzy Logic, Discernibility ,Threshold, fuzzify

1. INTRODUCTION

Data collection and data sharing has one of the crucial act in this digital era. Many business organizations for their business and research prospective are collecting data digitally and collected data can be published for the purpose of research. But when the collected data that contains sensitive information is publically available, it may disclose personal information of someone. Sometime, this may attract the attention of attackers who want to extract personal information about an individual. Thus, this remains a challenge to protect the privacy of an individual and keep individual safe and secure. Moreover, the aim of PPDP is to publish the information but by keeping the individual's information secret.

Various techniques are available in literature for the purpose of PPDP, but selection of an appropriate technique is always a matter of concern and challenge for a professional. k- anonymity [1][2] is one of widely discussed approach used for anonymization that is based on the process of generalization and suppression but suffers from the problem of homogeneity and background attack. Moreover, after applying the process of generalization and suppression to anonymize the data always results in information loss [2][3] , value of discernibility[2] [4] and the value of average equivalence class size[2] [5] . A new amalgam method of anonymization have been proposed that does the process of anonymization with less information loss and increases the data utility. Moreover, proposed algorithm also enhances the domain set of sensitive attribute by assigning appropriate weight. Further S-shaped fuzzy is applied on confidential attribute to incur the uncertainties. S-shaped membership function is an efficient approach which is not only used to obtain sanitized data but also gives an illusion as the original one [6][7]. The proposed algorithm uses an amalgamation of shuffling the records, S-Membership function based fuzzy values for sensitive attribute and the process of k-anonymization only for sensitive records.

2. BACKGROUND AND RELATED WORK

Due to increase in the growth of digitalization and huge collection of data, the size of data base is growing exponentially which is an asset for the purpose of research, analysis and for business prospective [6]. But while publishing the data publically the aim of attacker is to obtain personal information of an individual and use this information with some malafide intensions. A variety of techniques have been available in the literature.

2.1 Anonymization:

In literature various method(s) of anonymization exist such as k-anonymization, l-diversity and t-closeness etc. but k-anonymization is one of the fundamental for all types of anonymization and base for all types of approaches [1]. This was the first model for data anonymization and base for the others. The formal definition of k-anonymity for relation is as [9,10]. "A table T is k-anonymous with respect to Quasi-Identifiers $Q_i(Q_1,\ldots\ldots, Q_d)$ if every unique tuple $(q_1,\ldots q_d)$ in the projection of T on $Q_1,\ldots Q_d$ occurs at least k times". For example Table1 represents the original table containing data about school employees where as Table 2 represents the anonymized data with k=3.

[1] Research Scholar, Dept. of Computer Science & Applications, KU, Kurukshetra, Haryana, India
[2] Associate Professor, Dept. of Computer Science & Applications, KU, Kurukshetra, Haryana, India
[3] Professor, Dept. of Computer Science & Applications, KU, Kurukshetra, Haryana, India

Table 1 Records for School Employees

Sno	ID	QID			Sensitive Attribute
	Name	Designation	Age	Pin Code	Salary
1	Ana	TGT	49	132042	42000
2	Ali	PGT	40	132021	58000
3	Joe	PPRT	44	132024	35000
4	Karim	TGT	48	132046	43000
5	Durga	PPRT	45	132045	34000
6	Raghav	PGT	43	132027	55000

Table 2 Anonymized table (k=3) for School Employees

Sno	EQ	QID			Sensitive Attribute
		Designation	Age	Pin Code	Salary
1	A	Teaching	[45-50)	13204$	42000
4		Teaching	[45-50)	13204$	43000
5		Teaching	[45-50)	13204$	34000
2	B	Teaching	[40-45)	13202$	58000
3		Teaching	[40-45)	13202$	35000
6		Teaching	[40-45)	13202$	55000

2.2 Randomization:

Randomization is an ability to anonymize the whole dataset, to preserve certain semantics. In existing privacy preserving techniques, 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 data utility was achieved by adding noise to the data but this also affects the originality of original data.

2.3 Fuzzy Based Approach:

Fuzzy set has dominated the field of knowledge representation and interpretation from a long time. Various fields such as machine intelligence, artificial intelligence, data mining etc. have widely accepted the theory of fuzzy sets in their application areas. This is an aid to represent uncertainty, possibility as well as approximation. If something is fuzzy, it means that it is difficult to define precisely its boundaries. Human reasoning is not only able to make decisions but also to classify the situations based on partial or ambiguous information. Transforming the given set of data points from the crisp set to fuzzy set is known as fuzzification and by applying the process of fuzzification, each data point is associated with certain degree of membership to every fuzzy set. When there are a number of numerical or categorical values, it is extremely unlikely that two individuals have exactly the same set of values for all variables. Hence, in the current work fuzzy S-shaped function is applied to fuzzify the sensitive attribute as S- shaped membership function is an efficient approach which is not only use to obtained the sanitized data but also gives an illusion as the original one [6][7]. Our aim is to reduce the information loss, value of discernibility, value of average equivalence class size and provides an approach to keep the privacy of the data. For this purpose concept of S –shaped fuzzy membership function is used and given by:

$$F(x,a,b)=\begin{cases} 0, & x\le a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a\le x\le\frac{a+b}{2} \\ 1-2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2}\le x\le b \\ 1, & x\ge b \end{cases}$$

Where x is the value of the sensitive attribute, a & b are the minimum and maximum values in the sensitive attribute list.

3. PROBLEM FORMULATION:

In this paper, a new amalgam approach for anonymization of data has been taken which is based on the concept of shuffling the records of original data set, generating the fuzzy values for sensitive attribute and further applying k-anonymization only on high sensitive records.

The new algorithm not only protects the privacy but also reduces the information loss[3] [2] ,Value of discernibility [4] [2] and the value of average equivalence class size[5][2].

4. PROPOSED ALGORITHM

The proposed amalgamated approach is a combination of three different approaches and based on shuffling the records, generation of Fuzzy values for sensitive attribute and anonymizing only high sensitive records using k-anonymization. The main reason behind using shuffling of records is to distinguish the pattern of tuples as compared to original data set. Moreover, fuzzy is applied to sensitive attribute to preserve the privacy of an individual.

1. Input DS_0 || DS_0=Original DataSet
2. Select (A_U,A_Q,A_S) Ɛ DS_0 || {A_U: Unique Attribute, A_Q : Quasi Attribute, AS: Sensitive Attribute }
3. $DS_1 \leftarrow DS_0 - \{A_U\}$ || {DS_1 : New Data set after removing unique attributes}
4. $DS_2 \leftarrow$ Sort(DS_1,As) || {DS2:Obtained dataset after sorting existing dataset on Sensitive attribute}
5. $AS_W \leftarrow$ GenerateW(A_S) || {Generate weight on sensitive attribute}
6. $DS_3 \leftarrow$ SFuzzy(AS_W) || {Obtained dataset DS_3 after applying SFuzzy on generated weight of sensitive attribute.}
7. $DS_H \leftarrow$ Split(DS_3 , T_H) || {Split DS_3 in to DS_H and DS_L where DS_H={X: 0<X<=T_H}, $DS_L \leftarrow$ Split(DS_3 , T_H) || DS_L={ X: X>T_H}}
8. Apply KANONY(DS_H)|| Apply k-anonymity only on High Sensitive dataset
9. $DS_4 \leftarrow$ Combine(DS_H,DS_L) || Combine Low and high Sensitive data
10. Publish data set DS_4

In this algorithm DS_i ∀ i Ɛ n {Different states of dataset}

Figure 1: Algorithm of Proposed Approach

The proposed approach will produces less information loss, value of disernibility and value of average equivalence class size as process of anonymization is applied only on the high sensitive records cut off by threshold value.

Example : Table 3 shows a small dataset containing employee records which is to be published. Dataset is a combination of key attribute, quasi attributes and sensitive attribute. In table 3, Name is considered to be a key attribute as it describes unique identification. So, this is removed from table 3 and a new table 4 is generated containing quasi and sensitive attributes.

Table 3: Original records for School Employees

Name	Age	Sex	Designation	Pincode	Salary
Ansu	49	Male	TGT	132042	60000
Durga	40	Female	PGT	132021	88000
Zen	33	Male	PPRT	132024	35000
Akshu	48	Male	TGT	132046	48000
Gargi	45	Female	PPRT	132045	34000
Pankja	43	Female	PGT	132027	75000
Ana	30	Male	PPRT	132024	35000
Ali	28	Male	TGT	132046	43000
Sahiba	54	Male	PPRT	132045	34000
Kiran	26	Female	PGT	132027	85000

Remove the unique attribute(s) from the table for the process of anonymization

Table 4: Table with removed key attribute for School Employees

Age	Sex	Designation	Pincode	Salary
49	Male	TGT	132042	60000
40	Female	PGT	132021	88000
33	Male	PPRT	132024	35000
48	Male	TGT	132046	48000
45	Female	PPRT	132045	34000
43	Female	PGT	132027	75000
30	Male	PPRT	132024	35000
28	Male	TGT	132046	43000
54	Male	PPRT	132045	34000
26	Female	PGT	132027	85000

Now by applying the proposed approach will accumulate all high sensitive records together by sorting them on high sensitive values followed by generation of weights for sensitive attribute. So, that fuzzy values can be generated for sensitive attribute using S-shaped membership function table 5 and table 6 shows the splitted records on the basis of threshold value. Table 5 shows records with high sensitive values whereas table 6 shows records with less sensitivity.

Table 5: Splitted table containing high sensitive records for School Employees

Age	Sex	Designation	Pincode	Salary
40	Female	PGT	132021	Very High
43	Female	PGT	132027	Very High
26	Female	PGT	132027	Very High
49	Male	TGT	132042	High

Table 6: Splitted table containing Less sensitive records for School Employees

Age	Sex	Designation	Pincode	Salary
45	Female	PPRT	132045	Moderate
54	Male	PPRT	132045	Moderate
33	Male	PPRT	132024	Low
48	Male	TGT	132046	Low
30	Male	PPRT	132024	Low
28	Male	TGT	132046	Low

Now as to apply process of anonymization on table 4 containing high sensitive records. Table 7 shows an anonymized table using k-anonymization, whereas table 8 shows finally merged anonymized table for the complete dataset which is ready for publishing.

Table 7: Annonymized Splitted table containing high sensitive records for employees

Age	Sex	Qualification	Designation	Salary
[30:50)	Male	TGT	132042	High
[30:50)	Female	PGT	132021	Very High
[30:50)	Female	PGT	132027	Very High
[20:70]	Female	PGT	132027	Very High

Table 8: Merged File For publishing

Age	Sex	Qualification	Designation	Salary
[30:50)	Male	TGT	132042	High
[30:50)	Female	PGT	132021	Very High
[30:50)	Female	PGT	132027	Very High
[20:70]	Female	PGT	132027	Very High
45	Female	PPRT	132045	Moderate
54	Male	PPRT	132045	Moderate
33	Male	PPRT	132024	Low
48	Male	TGT	132046	Low
30	Male	PPRT	132024	Low
28	Male	TGT	132046	Low

Finally table 8 is generated after applying amalgamated proposed approach that reflects a combination of shuffling of records, Fuzzy values for sensitive attributes and the process of anonymization only on high sensitive records based on the threshold value. In future, work will be carried out in the direction of verifying and comparing the proposed amalgamated approach with existing anonymization algorithm using various data utility metrics.

5. CONCLUSION AND FUTURE WORK

A new amalgamated approach for preserving privacy with less information loss and minimum overhead has been proposed. The proposed approach implies shuffling of records to restrict the attacker to identify the pattern of data, generation of fuzzy values for sensitive attribute using S-shaped membership function and to incur uncertainty and further k-anonymization method to procure anonymization only on high sensitive data. In future, verification of the proposed approach has been done experimentally.

6. REFERENCES

[1]. J. Soria-Comas , J. Domingo-Ferrer, D. Sanchez, and S. Martınez. Improving the Utility of Differentially Private Data Releases via k-Anonymity. In Proceedings of the 12th IEEE International Conference on Trust , Security and Privacy in Computing and Communications, TRUSTCOM '13, pages 372–379, 2013.

[2]. Vanessa Ayala Rivera, Patrick McDonagh, " A Systematic Comparison and Evaluation of k-anonymization algorithms for practitioners", Transactions on data privacy Volume 7: 337-378,2014

[3]. Nergiz, M. E. and Clifton, C. "Thoughts on k-Anonymization", Data and Knowledge Engineering, 63(3):pp622–645, 2007.

[4]. Bayardo, R. J. and Agrawal, R., "Data Privacy Through Optimal k-Anonymization", In Proceedings of the 21st International Conference on Data Engineering, ICDE '05, pages 217–228, 2005.

[5]. Kristen LeFevre, David J. DeWitt. Mondrian Multidimensional K-Anonymity, In proceeding of 22nd International Conference on Data Engineering, ICDE'06, page 25,2006.

[6]. Thanveer Jahan, Dr. G. Narsimha, Dr. C.V. Guru Rao," A Hybrid Data Perturbation Approach To Preserve Privacy", International Journal of Scientific and Engineering Research, Volume 6,Issue6, June 2015.

[7]. G Manikandan , N Sairam , V Harish, Noka Saikumar, Survey on the use of Fuzzy Membership Functions to Ensure Data Privacy",Research Journal of Pharmaceutical ,Biological and Chemical Sciences, 7(3):pp 344-348, May-June, 2016, ISSN:0975-8585.

[8]. J. Gantz and D. Reinsel. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Technical report, IDC, sponsored by EMC, December, 2012.

[9]. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5):571{588, 2002}.

[10]. P. Samarati. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Engineering, 13(6), 2001.

[11]. Manjusha S. Mirashe , Kapil N. Hande , Survey on Efficient Technique for Annonymized Microdata Preservation, International Journal of Emerging Trends in Engineering and Development , Issue 5, Vol.2 (Feb.-Mar. 2015), ISSN 2249-6149